

Computational biology methods for identifying leaderless secretory proteins in *Arabidopsis thaliana* and other plant species

Andrew Gregory Lonsdale

ORCID 0000-0002-0292-2880

Doctor of Philosophy

November 2019

School of BioSciences
THE UNIVERSITY OF MELBOURNE

Submitted in total fulfilment of the requirements
of the degree of Doctor of Philosophy.

Abstract

LEADERLESS secretory proteins (LSPs) are proteins that are secreted yet lack the classical canonical signal peptide sequences and, therefore, by definition are undergoing unconventional protein secretion (UPS). We cannot necessarily assume all such proteins found in plant secretomes are LSPs due to the high possibility of cellular contamination arising from cellular disruption, leading to a mix of LSPs and contaminant proteins that are both characterised by the absence of signal peptides. The aims of this Thesis are to use computational biology methods to identify LSPs.

In **Chapter 1** the current knowledge of secretory pathways (LSPs and UPS) in plants, motifs for secretion, the difficulties of isolating the plant secretome (cell wall/apoplastic fluids) without contamination, and the lack of appropriate bioinformatics tools for plants to distinguish them from LSPs is reviewed.

Chapter 2 evaluates a commonly used computational tool, SecretomeP which was trained on mammalian data and is the most widely used prediction tool for LSPs, including in plant studies. Exploring whether this tool is applicable to plants required using conventionally secreted proteins as a proxy, and evaluations were made on SecretomePs premise that conventional and unconventional secretory proteins will share properties. By removing the signal peptide from sequence data, the research shows a bias in scores due to the signal peptide, and only a marginally higher true positive rate compared to false positives. The use of the tool on further plant studies was not recommended and suggested previous inferences of plant LSP status based solely on SecretomeP predictions needed to be re-evaluated. This work was published as "Better Than Nothing? Limitations of the Prediction Tool SecretomeP in the Search for Leaderless Secretory Proteins (LSPs) in Plants" (Lonsdale et al., 2016)

Chapter 3 details the creation of a putative LSP database for *Arabidopsis thaliana* by taking the entire proteome and applying a workflow of collating annotation, literature observation and experimental relationships between proteins. This results in a framework to identify proteins without signal peptides that have been observed (unclassified set) or unobserved (non-secreted set) and compare them to conventionally secreted proteins that have been observed (secreted set) or unobserved (theoretically secreted sets). Protein-protein interactions, GO term and PFAM distributions that are similar between the secreted and unclassified protein sets are used to create confidence lists of putative LSPs.

In **Chapter 4** new prediction tools were created. Firstly, a SecretomeP-like tool is created using the properties of the observed secreted set, with signal peptides removed. Lessons from **Chapter 2** on SecretomeP allowed the bias to be minimized. Secondly, tools were created based on the candidate list from the putative LSP database. Each tool is based on a Random Forest (RF) using protein-derived features and trained on subsets of the LSP database. Consensus predictions between them on new data was used to identify further LSP candidates. Additional reporting tools provide a convenient way to map data from the LSPDB onto new sequences, and the results of these tools on LSPs from plants and other organisms shown.

In **Chapter 5** additional putative candidates for LSPs in *Arabidopsis* and other species were identified by applying the tools and databases to additional data excluded from the original database created in **Chapter 3**, including recent cell wall proteomes, secretory pathway experiments and analysis of extracellular vesicles. In conclusion, **Chapter 6** discusses the limitations of the computational methods developed, and suggestions for improvement to predict LSPs in plants for further experimental investigation and confirmation of location.

Declaration

This is to certify that

1. the thesis comprises only my original work towards the PhD,
2. due acknowledgement has been made in the text to all other material used,
3. the thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Andrew Gregory Lonsdale, November 2019

Acknowledgements

First and foremost, I must thank my wife Mina. Life doesn't stop for a PhD, and our wedding, first home and baby have all made life with a Thesis an eventful time. I will forever be grateful for your support, sacrifice and endurance in helping me to get this finished, and the opportunities that the future now holds for us are all thanks to you.

My supervisors, Tony Bacic, Monika Doblin and Melissa Davis. Thank you for guiding my research, your perspectives on balancing the worlds of computational biology and the plant cell wall, and especially your patience. I have learnt a tremendous amount about how to do research and be a researcher from, and I consider myself lucky to have had you as supervisors. I appreciate your support especially over the final stretch as I have balanced working life and the Thesis, and grateful for the feedback, at times under very short timeframes, during this final push. I think that after the haze of submission has subsided I will be truly able to appreciate all that you have done, and hope to repay your dedication in the future.

To my advisory committee members, Ed Newbigin and Ben Rubinstein, thank you for your guidance at key points in the candidature that helped steer the research in the right direction.

To all members of the ARC COE in Plant Cell Walls, thank you for the great work that was achieved. Being involved in such a large cross-disciplinary endeavour widened my scientific horizons and has left an indelible mark on how I can see what is possible when a dedicated effort is focused on one part of biology. The centre meetings especially were a highlight of collaboration and scientific discussion, and I hope to be involved in something similar again in my career.

To the Melbourne node in particular, thank you for welcoming another bioinformati-

cian into your group and thanks all for lab meetings, coffee runs and the friendship over the years: Kim Johnson, Kris Ford, Andrea Watt, Allison van de Meene, Kelsey Picard, Edgar Liu, John Humphries, Roshan Cheetamun, Edwin Lampugnani and everyone else in the COE, and the Plant Cell Biology Research Centre. To the mezzanine gang, Yin Ying Ho, Joan Narciso, Eva Reda, Lachlan Tegart, Kelsey Picard and Melissa Bain, thank you for your friendship and support in both science and life, and I'll cherish those times up their, despite the lack of temperature control! Mel, your encouragement and help through the last stages of the PhD has been vital in keeping me going, and I truly appreciate it.

Beyond the home of my candidature, I have many people to thank in the wider bioinformatics community. Firstly to Andrew Cassin for your mentorship in the early years, and for showing me the bioinformatic ropes of the COE, as well as for inviting me to join the weekly Proteomics lunch where I have Gabriel Keeble-Gagnre, Nick Williamson, Dave Perkins, Patrick James and everyone else to thank for all their friendship.

To the Davis lab, thank you for making me a not so lonely bioinformatician and welcoming me as an adjacent member of your lab group. Your perspectives and suggestions were very helpful, and I'm thankful for always welcoming me to lab meetings and events: Joseph Cursons, Soroor Zadeh, Greg Bass, Dharmesh Bhuva, Holly Whitfield and Sepideh Foroutan.

Andrew Lonie for coordinating the MSc(Bioinformatics) and accepting me into the course all those years ago that started this long strange journey to a PhD. The original MSc(Bioinformatics) intake especially Peter Diakumis, David Clarke, Tom Ma and Ehtesham Mofiz. I don't think we knew what we were in for and I'm glad to have shared that new path with you. Thanks also Phil Batterham, for looking after this group when we and the course were all new to the university.

To the members of COMBINE, firstly those driving the group when I first heard about it - Geoff Macintyre, Ben Goudey, Karin Klotzbuecher, Kian Ho - thanks for the encouragement and for being the custodians of such an important part of the local and international bioinformatics community. As COMBINE strove for world domination, I am grateful for the trust and support of everyone in the overlapping COMBINE, Bioinformatics Graduate Student Association and Journal Club crews that shaped who I am as a

member of these groups: Harriet Dashnow, Jane Hawkey, Tim Rice, Scott Ritchie, Sabrina Rodrigues, Tlio Campos, Marek Cmero, Zoe Dyson, Luke Zappia, Jocelyn Penington, Thomas Coudrat and many many more.

As the community grew from strength to strength, the opportunity to have the students form an integral part of the newly established Australian Bioinformatics And Computational Biology Society (ABACBS) was appreciated, and I am thankful for the respect the executive committee showed me balancing being a member and a PhD student: Melissa Davis, Ashley Waardenberg, Tony Papnefuss and David Lovell.

Harriet Dashnow gets her own paragraph for being involved in so many of these groups (COMBINE, ABABCS, Journal Club, BGSA) and for being a friend and confidant throughout. Some of the best work I've done is when we work together and our papers (proper and side-projects) have been highlights and I look forward to many more throughout our careers. For side-projects, I have a few ideas!

Thanks also to Harriet for reminding me that I had always wanted to work with Alicia Oshlack, and encouragement to apply for a part-time job led to us working properly together and an amazing experience working in Bioinformatics and Kidney groups at the Murdoch Childrens Research Institute. Thanks, especially to Melissa Little and the Kidney lab, and in particular Alicia Oshlack and Bioinformatics group who as a whole have helped me to balance working life and finishing this Thesis, and I couldn't have made it without your support in both: Belinda Phipson, Jovana Maksimovic, Nadia Davidson, Simon Sadedin, Katrina Bell, Breon Schmidt, Christoffer Flensburg, Luke Zappia, Marek Cmero, Andreas Halman and Harriet Dashnow.

To my family, I could not have done this without your support, and look forward to being able to spend more time with you all. To my parents Michael and Maureen, siblings Kate, Elizabeth and Stephen, and extended family of Dave, Jess, Eamon, Alice, Tim, Anne and all my nieces, thank you all for your support and understanding. Finally, I would like to thank my wife, again, and our daughter Nora for being a delight when the family needed it the most. It's been a team effort, and I am looking forward to our little team being able to work on something else!

Preface

Figures adapted from publications were:

- **Figure 1.1** Guerra-Guimarães et al. (2016)
- **Figure 1.2** Höfte and Voxeur (2017)
- **Figure 1.3** Rabouille (2016)
- **Figure 1.4** Davis et al. (2016)
- **Figure 1.5** Albenne et al. (2013)
- **Figure 3.2** Uhlén et al. (2015).

Chapter 2 consists of the publication in its entirety:

- **Lonsdale A, Davis MJ, Doblin MS and Bacic A (2016).**

Better Than Nothing? Limitations of the Prediction Tool SecretomeP in the Search for Leaderless Secretory Proteins (LSPs) in Plants. *Front. Plant Sci.* 7:1451. doi: 10.3389/fpls.2016.01451

For Mina.

Contents

List of Figures	xix
List of Tables	xxiii
List of Terms	xxvii
1 The Leaderless Secretory Protein Problem in Plants	1
1.1 Introduction	1
1.2 The plant apoplast	2
1.3 Conventional secretory pathway	7
1.4 Unconventional secretory pathway	8
1.5 Unconventional protein secretion (UPS) in plants	13
1.6 The plant LSP problem	19
1.7 Filtering for plant LSPs with SecretomeP	22
1.8 The common feature hypothesis	24
1.9 Validating putative leaderless secretory proteins in plants	27
1.10 Thesis scope	29
2 Better Than Nothing	31
2.1 Introduction	31
2.2 Better than Nothing: Limitations of the Prediction Tool SecretomeP in the Search for Leaderless Secretory Proteins (LSPs) in Plants	31
2.3 Supplementary Material to Lonsdale et al. (2016)	45
2.4 Discussion	56
2.5 Conclusion	57
3 The Unusual Suspects	59
3.1 Introduction	59
3.1.1 Leaderless Secretory Protein Database (LSPDB)	60
3.2 Layers of the LSPDB	63
3.2.1 Features layer	65
3.2.1.1 Secretory proteins and secretory isoforms	67
3.2.1.2 PFAM and GO terms	69
3.2.2 Observations layer	70
3.2.3 Classifications layer	72

3.2.4	Interactions layer	74
3.2.5	Integrating the layers	74
3.3	Results	77
3.3.1	Frequency	77
3.3.1.1	Secretory Isoforms	83
3.3.2	Classifications	83
3.3.2.1	PFAM domains	84
3.3.2.2	GO terms	89
3.3.3	Networks	93
3.3.3.1	Interactions with secretory proteins	93
3.3.3.2	Networks of classified proteins	100
3.3.3.3	Co-occurrence within the PPI network	108
3.3.3.4	Homodimers	111
3.4	Discussion	111
3.4.1	Comparing the evidence types	112
3.4.2	Confidence criteria for GO terms and PFAM domains	113
3.4.3	Confidence criteria for network methods	115
3.4.4	Metrics for prioritising candidates	117
3.4.5	Properties of LSP candidates	118
3.4.5.1	High confidence	120
3.4.5.2	Medium confidence	123
3.4.5.3	Low confidence	134
3.4.6	Using the LSPDB	136
3.4.6.1	Direct database availability	137
3.4.6.2	Web interface	138
3.5	Conclusions	145
4	Follow The Leaderless	147
4.1	Introduction	147
4.2	Methods	149
4.2.1	Available data	150
4.2.2	Feature extraction	151
4.2.3	LSPpred	152
4.2.4	SPLpred	153
4.2.5	LSPblast	153
4.2.6	LSPfam	156
4.2.7	Training random forest models on unbalanced data	157
4.2.8	Evaluation	159
4.2.9	Applying trained models to new data	160
4.3	Results	160
4.3.1	Performance and relevant features of LSPpred	160
4.3.2	Performance and relevant features of preferred SPLpred model	162
4.3.3	Results from reporter tools	166
4.4	Discussion	170
4.4.1	SPLpred3 and LSPred2 are the preferred models	170

4.4.2	Comparison of results between LSPpred, SPLpred, LSPfam and LSPblast	172
4.4.3	Proposed ensemble workflow for secreted proteins in plants	173
4.4.4	Predictor feature importances	175
4.4.5	Testing on plant LSPs	182
4.4.6	Testing on LSPs from other organisms	186
4.4.6.1	Human and disease related LSPs	187
4.4.6.2	Fungi	187
4.4.6.3	Bacteria	188
4.5	Conclusion	191
5	The Secret Secretome	195
5.1	Introduction	195
5.2	LSP workflow on the rice proteome	197
5.3	LSPDB as a dynamic database	206
5.4	Enhancing the LSPDB	215
5.4.1	Co-expression	216
5.4.2	Post translational modifications (PTMs)	217
5.4.3	Abundance	221
5.4.4	Effect of new protein data on the LSPDB	226
5.5	LSP case studies	230
5.5.1	Prediction tools on the remaining UNCLASSIFIED proteins	230
5.5.2	Secretory isoforms	231
5.5.3	Secretory vesicles	233
5.5.4	Extracellular vesicles (EVs)	236
5.6	LSP candidates	238
5.7	Conclusion	241
6	Conclusion	243
6.1	Thesis conclusion	243
6.2	Limitations	244
6.3	Future work	246
	References	255
	Appendices	279
A	Appendix	281
B	Appendix	287
B.1	Secretory Isoform Observations	287
B.2	LSPDB low confidence list	290
B.3	LSPDB orange - ribosomal	314
B.4	LSPDB orange - heat shock	315
B.5	Homodimers	315
B.6	Common GO and PFAM	327

C Appendix	335
D Appendix	359
E Appendix	381

List of Figures

1.1	Diagram of the plant extracellular space/apoplast	3
1.2	Molecular structure of monosaccharides (lower panel) and polysaccharides common in plant cell walls, adapted from Höfte and Voxeur (2017)	5
1.3	Summary of unconventional protein secretion pathways in eukaryotes . .	11
1.4	Summary of conventional and unconventional protein secretion (UPS) pathways in plants	14
1.5	Methods for performing a plant cell wall/apoplastic proteome	20
1.6	Bioinformatics approaches to LSPs	25
2.1	Estimated accuracy for SecretomeP as published for mammalian and bacterial proteins (Figure 1)	34
2.2	Modifications to proteins for testing (Figure 2)	36
2.3	Kernel density estimation and correlation plots for scores after modification to WallProt database proteins (Figure 3)	38
2.4	Kernel density estimation for scores after modification to ASURE database proteins (Figure 4)	39
2.5	Correlation plots for scores after modification to ASURE database proteins (Figure 5)	40
2.6	Receiver operating characteristic (ROC) for <i>Arabidopsis</i> proteins after modifications (Figure 6)	41
2.7	Kernel density estimation and correlation plots for scores after modification to WallProt by species (Supplementary Figure S1)	52
2.8	Scores for WallProtDB sequences with modifications involving random shuffling (Supplementary Figure S2)	53
2.9	ROC curve for WallProtDB vs unmodified proteins (Supplementary Figure S3)	54
2.10	ROC curve for ASURE extracellular vs modified and unmodified proteins (Supplementary Figure S4)	55
2.11	ROC curve for WallProtDB vs unmodified nuclear, cytosolic, plastid, mitochondrion and plastid proteins (Supplementary Figure S5)	55
3.1	Overview of the LSPDB layered structure	61
3.2	Proportions of secretory protein types in the human proteome	68
3.3	Database design overview of the LSPDB	76

3.4	Classification categories from the LSPDB v1.0 database	84
3.5	Venn diagram of PFAM domains in SEC, NONSEC, UNCLASSIFIED classes	85
3.6	Venn diagram of PFAM domains in SEC+SPT, NONSEC, UNCLASSIFIED classes	86
3.7	Venn diagram of GO terms in SEC, NONSEC, UNCLASSIFIED classes	91
3.8	Venn diagram of GO terms in SEC+SPT, NONSEC, UNCLASSIFIED classes	91
3.9	Scatter plot of SEC vs NONSEC interactions	94
3.10	Scatter plot of SEC vs NONSEC interactions for each LSPDB class	94
3.11	Scatter plot of SEC+SPT vs NONSEC interactions for each LSPDB class	95
3.12	Scatter plot of SEC vs SPT interactions	96
3.13	Matrix of scatter plots for PPIs class interactions	102
3.14	Network of LSPDB proteins with 3 SEC interactions	103
3.15	BINGO analysis of Figure 3.14 proteins, with over-represented molecular function terms	104
3.16	Network of NONSEC proteins with 3 SEC interactions	105
3.17	Network of proteins with 10 or more SEC interactions	107
3.18	Network of NONSEC and UNCLASSIFIED proteins with 4 or more SEC interactions	107
3.19	Network of co-occurring PPI interactions	109
3.20	Upset plot of proteins using combined scoring metrics	119
3.21	Web interface to the LSPDB	138
3.22	Example of the LSPDB summary interface, using SAMS1 and SAMS2	140
4.1	Positive, negative and query data for A) LSPpred and B) SPLpred	154
4.2	Interface for LSPblast	155
4.3	ROC curve for LSPpred2	162
4.4	Score distribution curve for LSPpred2	163
4.5	ROC curve for SPLpred1	164
4.6	Score distribution curve for SPLpred1	165
4.7	ROC curve for SPLpred3	167
4.8	Score distribution curve for SPLpred3	168
4.9	Circos visualisation from LSPblast	171
4.10	Recommended workflow for identifying secretory proteins including LSPs	176
4.11	Top 20 features for LSPpred2	178
4.12	Top 20 features for SPLpred3	179
4.13	Prediction results for four plant LSPs	184
4.14	Circos plot for four plant LSPs	186
4.15	ROC curve of SPLpred and LSPpred applied to bacterial LSPs	190
5.1	Venn diagram of results applying the LSP workflow to rice	199
5.2	Overlap of LSP predictions from LSPpred, SPLpred and ApoplastP	204
5.3	Updated workflow for identifying secretory proteins including LSPs	207
5.4	Upset plot of potential LSPs from additional studies	208
5.5	Clusters of 3 or more proteins that co-occur and are co-expressed using both RNA-seq and microarray data	218
5.6	Venn diagram of PTMs in each class of the LSPDB	220

5.7	Boxplots of the abundance of LSPDB proteins	222
5.8	KDE distribution of LSPDB classes	223
5.9	KDE distribution of abundance in <i>Arabidopsis</i> , <i>Oryza sativa</i> , <i>Zea mays</i> and <i>Brachypodium distachyon</i>	225
5.10	Upset plot of the overlaps between evidence types	227
5.11	Upset plot of potential LSPs in secretory vesicles	235
A.1	SP and LSP counts in plant secretomes	282
C.1	Top 20 features by importance for feature selected LSPpred version 1	336
C.2	Top 20 features by importance for feature selected SPLpred version 1	337
C.3	Top 20 features by importance for feature selected SPLpred version 2	338
C.4	Top 20 features by importance for feature selected SPLpred version 4	339
C.5	ROC curve for LSPpred1	340
C.6	Score distribution for LSPpred1	341
C.7	ROC curve for SPLpred2	342
C.8	Score distribution for SPLpred2	343
C.9	ROC curve for SPLpred4	344
C.10	Score distribution for SPLpred4	345
C.11	Confusion matrix for LSPpred version 1 at 0.5 threshold	346
C.12	Confusion matrix for LSPpred version 1 at 0.645 threshold	347
C.13	Confusion matrix for LSPpred version 2 at 0.5 threshold	348
C.14	Confusion matrix for LSPpred version 2 at 0.68 threshold	349
C.15	Confusion matrix for SPLpred version 1 at 0.5 threshold	350
C.16	Confusion matrix for SPLpred version 1 at 0.61 threshold	351
C.17	Confusion matrix for SPLpred version 2 at 0.5 threshold	352
C.18	Confusion matrix for SPLpred version 2 at 0.70 threshold	353
C.19	Confusion matrix for SPLpred version 3 at 0.5 threshold	354
C.20	Confusion matrix for SPLpred version 3 at 0.70 threshold	355
C.21	Confusion matrix for SPLpred version 4 at 0.5 threshold	356
C.22	Confusion matrix for SPLpred version 4 at 0.66 threshold	357
C.23	Prediction scores from bacterial programs on LSPs	358
D.1	Network of co-occurring and co-expressed proteins	380

List of Tables

1.1	Unconventionally secreted proteins in plants	17
1.2	Plant studies using SecretomeP	23
2.1	Details of non classical secretion prediction programs (April 2016) (Table 1)	33
2.2	Protein features used in SecretomeP (Table 2)	34
2.3	Mean SecretomeP scores for all data (Supplementary Table S1)	46
2.4	Spearman correlation scores between original and modified sequences (Supplementary Table S2)	48
2.5	Thresholds, TPR and FPR rates calculated from ROC curve (Supplementary Table S3)	49
2.6	Annotated putative LSPs from Cheng at al. (2009) with scores (Supplementary Table S4)	50
3.1	Summary of literature included in LSPDB with counts of observed proteins (gene level)	71
3.2	Frequent observations (10 or more) in the LSPDB, all of which contain secretory features	78
3.3	Frequent observations (5 or more) of UNCLASSIFIED proteins in the LSPDB, regardless of previous annotations or established non-secretory roles. . .	81
3.4	PFAM domains occurring exclusively in the SEC, SPT and UNCLASSIFIED classified proteins	88
3.5	PFAM domains occurring in the majority in the SEC classified proteins . .	89
3.6	GO terms occurring exclusively in the SEC, SPT and UNCLASSIFIED classified proteins	89
3.7	GO terms occurring in the majority in the SEC classified proteins	92
3.8	Proteins from SPT, NONSEC and UNCLASSIFIED with more than 5 SEC and 5 SPT interactions	97
3.9	UNCLASSIFIED proteins identified in **Figure 3.14**	100
3.10	NONSEC proteins identified in Figure 3.19	105
3.11	UNCLASSIFIED proteins identified in Figure 3.22	110
3.13	Putative LSPs from the high (green) category	120
3.14	Putative LSPs from the medium (orange) category	124
3.15	LSPDB results from UNCLASSIFIED interactions with SAMS1 and SAMS2	143
4.1	Summary of LSP tools created	150

4.2	Summary of LSPpred and SPLpred balanced accuracy at the default (0.5) and FPR controlled threshold.	161
4.3	Consensus LSPfam predictions from low confidence LSPs	170
4.4	Consensus LSPpred and SPLpred predictions from low confidence LSPs	174
5.1	LSPpred and SPLpred results for putative LSPs from Shinano et al. (2011) as selected by Ding et al. (2012) that are also manually curated in UniProt	203
5.2	Counts of the repetitions and transitions of proteins classes in the LSPDB after the addition of 1715 proteins from excluded studies. Other includes secretory isoforms or multi-TM containing proteins.	210
5.3	Distribution of UNCLASSIFIED (repeat and new observations) and SEC (new observations) proteins over the studies. No proteins in these categories appeared in all four excluded studies.	210
5.4	Prediction results for repeat unclassified proteins from excluded studies	213
5.5	Prediction results for new unclassified proteins from excluded studies	214
5.6	New SEC proteins from held out studies	215
5.7	Application of PTM profiles on LSPDB classes	221
5.8	UNCLASSIFIED proteins with a PaxDB abundance above 2500ppm	223
5.9	Consensus LSPpred and SPLpred predictions from UNCLASSIFIED LSPs	232
A.1	Updated citation counts for Table 2.1 for November 2019	281
A.2	List of putative LSPs from Ding et al. (2012) with duplicate entries removed	283
B.1	Summary of combinations of LSPDB proteins with secretory isoform features, stratified by observation	287
B.2	Putative LSPs from the low (red) category	291
B.3	All homodimer UNCLASSIFIED candidates	316
B.4	Most common GO terms in the SEC class of the LSPDB	327
B.5	Most common PFAM domains in the SEC class of the LSPDB	328
B.6	Most common GO terms in the SPT class of the LSPDB	329
B.7	Most common PFAM domains in the SPT class of the LSPDB	329
B.8	Most common GO terms in the UNCLASSIFIED class of the LSPDB	330
B.9	PFAM domains most occurring in the UNCLASSIFIED group	331
B.10	Most common GO terms in the NONSEC class of the LSPDB	332
B.11	Most common PFAM domains in the NONSEC class of the LSPDB	333
D.1	Consensus proteins between LSPpred and SPLpred in results the of applying LSP workflow, excluding those with SP or TM, to the UniProt curated subset of rice proteins.	359
D.2	LSPblast results on proteins predicted solely by ApoplastP	361
D.3	Significantly over-represented PTMs for each LSPDB classification, for PTMs occurring in more than 10 proteins. PTMs were labelled according to the UniProt controlled vocabulary.	362
D.4	UNCLASSIFIED proteins with post-translational modifications identified in SEC proteins	364

D.5	LSPDB proteins pairs that co-occur in the same experiment, and co-express within the top 100 mutual rank, ...	367
D.6	LSPDB proteins that co-occur in the same experiment, and co-express within the top 100 mutual rank	372
E.1	Revised LSP candidate list	382

List of Terms

BRF Balanced Random Forest.

CFH common-feature hypothesis.

CSP Conventionally/Classically Secreted Protein.

ER Endoplasmic reticulum.

GA Golgi apparatus.

GH glycosyl hydrolase.

GO gene ontology.

GT glycosyltransferase.

LSP Leaderless Secretory Protein.

LSPDB Leaderless Secretory Protein Database.

MS Mass Spectrometry.

MVB Multi-vesicular bodies.

NONSEC protein without SP and other secretory features not observed in LSPDB literature.

PCW plant cell wall.

PFAM protein families.

PPI protein-protein interactions.

PTM post translational modification.

RF Random Forest.

SEC SP containing protein observed in the LSPDB literature.

SP_THEORY see SPT.

SPT SP containing protein but not observed in LSPDB literature.

TGN *trans*-Golgi network.

UNCLASSIFIED protein without SP and other secretory features observed in LSPDB literature.

UPS Unconventional protein secretion.

Chapter 1

The Leaderless Secretory Protein Problem in Plants

1.1 Introduction

PLANTS are essential to human existence on Earth as they produce the oxygen we breathe, are major sources of our food and dietary fibre, are the raw ingredients for many industrial and domestic products including wood, textiles and paper, as feedstock for animals, and increasingly a source of biofuels. At the cellular level, a characteristic feature of plants is the plant cell wall that surrounds each cell, a polysaccharide-rich extracellular matrix that surrounds the protoplast and can be considered the plant “skeleton” (Doblin et al., 2010). Plant cell walls (PCWs) provide structural support, form a physical barrier against pathogens, act as a reservoir of signalling and adhesion molecules, and are critical determinants of plant form and function (Keegstra, 2010; van de Meene et al., 2016). The properties of the cell wall are important for these *in planta* functions, as well as contributing to the functional and quality characteristics of plant-based bioproducts.

The cell walls of plants fall into two broad categories; the primary and secondary cell wall. The primary wall is laid down first during cell growth, with material secreted into the apoplast through fusion of vesicles with the plasma membrane (PM). The secondary wall is deposited once growth has ceased and differentiation begins. The secondary wall has a different composition, including lignin in some cases, and typically these walls have structural functions, for example, xylem elements. Understanding the process involved in cell wall assembly and their regulation is vital to our ability to manipulate the wall to

enhance growth and development as well as optimize their properties for biotechnological applications.

The cell wall is a dynamic assembly that changes throughout the processes of plant growth and development and in response to its environment. During cell expansion, its structure must be flexible to allow loosening of existing bonds between molecules and incorporation of new material. Cell wall changes also occur in response to a/biotic or mechanical stress. Since every cell is surrounded by a cell wall, any organism desiring access to its internal contents must penetrate the wall which has the capacity to sense the pathogen and trigger a cellular response. Pathogens may secrete hydrolytic enzymes to ‘soften’ the wall to increase accessibility and towards which the plant responds. To combat this attack, plants can reshape the cell wall in order to decrease the chance of infection (Bellincampi et al., 2014). Abiotic stresses such as soil salinity will also elicit changes to PCWs in addition to mechanical stresses such as wind (Novaković et al., 2018).

How can the cell wall do all these things, be strong and flexible, grow when needed, respond to pathogen challenges and unfavourable environments? A comprehensive understanding of the plant response at the physiological level requires deep knowledge of molecular processes at the cellular level, including the cell wall. This knowledge is vital in our goal to grow and conserve plants in diverse environments to meet the food, fibre and fuel needs of a growing human population, particularly in an age of increasingly uncertain climatic conditions.

1.2 The plant apoplast

The cell wall surrounds the cell on the external surface of the PM, forming a continuous space between adjacent cells. Collectively, the cell wall (primary and secondary layers), the middle lamella arising from the initial cell wall, the phragmoplast, marking the outer layer of the wall, and the fluid contained within them (apoplastic fluid) is termed the extracellular space (ECS) or simply the apoplast (Guerra-Guimarães et al., 2016) (**Figure 1.1**). Communication via the transport of materials between cells of an organ or tissue,

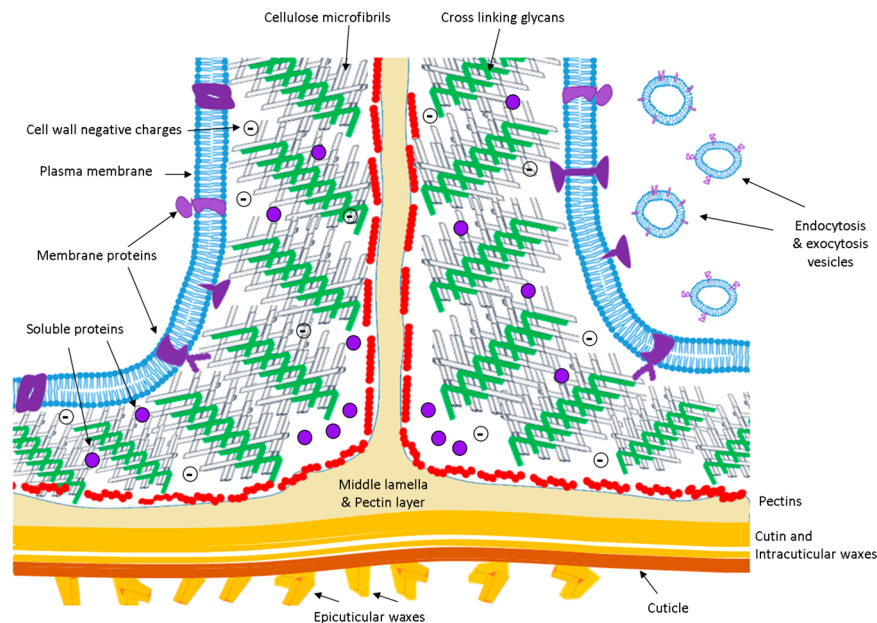


Figure 1.1: Schematic diagram outlining the major components and structure of a non-grass angiosperm cell wall. Adapted from Figure 1 of Guerra-Guimarães et al. (2016).

and within the entire plant, is mediated through the cell wall. This is through both the plasmodesmata channels that link cells through their cell walls (symplastic transport) and via the secretion of cellular contents to the cell surface at the PM/apoplast (apoplastic transport). PCWs act as the external sensor, perceiving their environment through receptors within the wall or at the PM which in turn signal internally into the cell and then effect a molecular response at multiple (transcriptional and post-transcriptional) levels.

The cell wall is a matrix made predominantly of two classes of macromolecules; proteins and polysaccharides. The polysaccharides make up ~90% of the dry weight of the growing primary cell wall (Doblin et al., 2010). The remaining 10% is protein which acts to build, reshape and repair the cell wall. There are many cell types in plants, approx. 40, in the various plant organs including stomata, root, root tip, branches, buds and leaves. Amongst these plant cell types there is variation in their cell wall composition and architecture (van de Meene et al., 2016). Often cell walls are categorised as either a growing primary or the thicker secondary wall, though all plant cells will vary in their wall composition and can be placed somewhere between these two states (Keegstra, 2010). There can also be exceptions to the thin primary wall, thick secondary wall paradigm. Sec-

ondary cell walls also include a significant proportion (up to 30%) of lignin. Important small molecules such as ions, $\text{Ca}^{\text{R}+}$ in particular can be present in the cell wall (Demarty et al., 1984).

The major cell wall polysaccharides and their monosaccharides components are shown in **Figure 1.2**. Polysaccharides are chains of carbon, hydrogen and oxygen molecules. Polysaccharides are either unbranched or branched chains of monosaccharides. Polysaccharides can be composed of a single type of monosaccharide or mixtures. The most common polysaccharides found in plant walls include cellulose (β -D-(1,4)- linked glucose units) that forms the fibrillar phase of the wall and matrix phase polysaccharides that include non-cellulosic polysaccharides ('hemicelluloses') and pectins. Some common hemicelluloses in angiosperms include xyloglucans (mostly in dicots, chains β -D-(1,4)- linked glucose with side branches of xylose, galactose and fucose), mixed linkage glucan (commelinid monocots only, chains of β -D-(1,3;1,4)- linked glucose), heteroxy-lans (comprising β -D-(1,4) linked xylose chains with either arabinose or glucuronic acid side branches), heteromannans (chains of β -D-(1,4)- linked mannose and or glucose with galactose as side chains), and two types of pectins, homogalacturonan (chains of α -D-(1,4)- linked galacturonic acid) and rhamnogalacturonan I (RG I, chains of alternating α -D-(1,4)- galacturonic acid and α -L-(1,2)- rhamnose). The third type of pectin called RG II (a backbone of chains of α -D-(1,4)-linked homogalacturonan with side chains) is an essential but minor component of the wall.

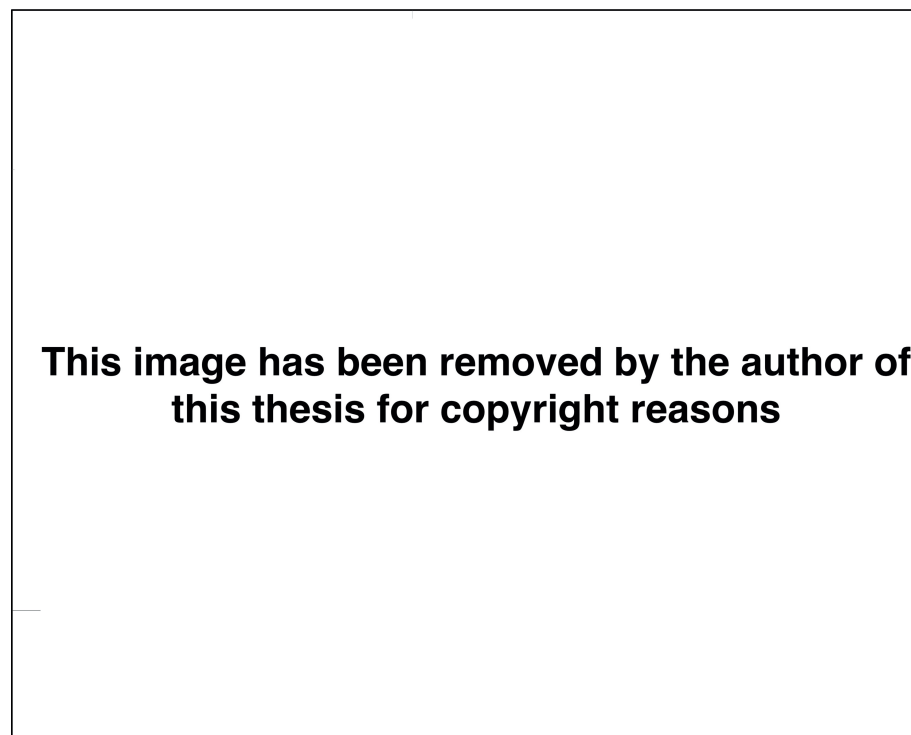


Figure 1.2: Molecular structure of monosaccharides (lower panel) and polysaccharides common in plant cell walls, adapted from Höfte and Voxeur (2017)

There is a large variability between cell types, and between plant species. A useful historical classification has been Type I walls, typical of dicots, gymnosperms and non-commelinid monocots, composed of a higher proportion of pectins and xyloglucans than Type II walls within the Poales and related commelinid monocots which contain mixed linkage glucan and higher levels of heteroxylans van de Meene et al. (2016). In reality, there is a continuum between the wall types. The strength and structure of the wall comes from the presence of these polysaccharides and proteins and the covalent and non-covalent links between them which gives rise to their physicochemical properties.

Proteins in the apoplast have various functions. Since the cell wall is built as a cell grows, or changes due to (a)biotic stress, and the cell wall is polysaccharides rich, it is unsurprising that a large proportion of these proteins are those that modify and/or cross-link polysaccharides (Jamet et al., 2006; Guerra-Guimarães et al., 2016). These include proteins known as Carbohydrate Active enZymes (CAZY), a classification for the enzymes that act on glycosidic bonds, such glycosyl hydrolases (GH) that cleave glycosidic bonds. Estimates of the number of cell wall proteins place proteins acting on polysaccharides as accounting for around 29.5% of the known cell wall proteins (Jamet et al., 2006). This includes GHs, as well as non-enzymatic protein such as expansions that loosen the cell wall and extensins that cross-link the cell wall.

Structural proteins account for a small proportion of known cell wall proteins (approx 1.5%, (Jamet et al., 2006)) yet form an important interconnected network with the polysaccharides. These include the diverse range of hydroxyproline-rich glycoproteins (HRGPs), which are critical to the wall despite the small proportion. These glycoproteins are combinations of a backbone of protein, with added carbohydrates through a process of post-translational modifications (PTM), in this case the PTM is glycosylation. Proteins with domains for interacting with other cell wall components, such as non-enzymatic lectin proteins that can interact with carbohydrates, make up approximately 10% of known cell wall proteins (Jamet et al., 2006).

Other proteins are involved in sensing and signalling responses to the environment that are not strictly cell wall proteins, including some that are classified PM-associated, yet can be considered related to cell wall functions. This can include transmembrane

(TM) containing proteins embedded in the PM, such as receptor-like kinases (RLKs) that include wall-associated kinases (WAKs), or glycosylphosphatidylinositol (GPI) anchored proteins, which are tethered to the CW-side of the PM.

Pathogens and plants first interact in the apoplast (Sperschneider et al., 2018), and the plant response includes proteins as well as other compounds. This can include proteases (estimated 10.0% of known apoplastic proteins (Jamet et al., 2006)) to breakdown the pathogenic proteins invading the apoplast, and other proteins to inhibit the bacterial or fungal organisms activity. Abiotic stress (e.g. salt, cold, heat, drought) can also illicit protein responses in a variety of sub-cellular locations. In the apoplast, these include cell wall modifying and signalling proteins, as well as specific protein classes related to stress such as molecular chaperones (e.g. heat shock proteins (HSPs)), and oxidoreductases to response to toxic levels of reactive oxygen species (ROS). Up to 10% of apoplast proteins have no known function (Jamet et al., 2006). When used in this Thesis, proteins will in general refer to those from the cell wall or the apoplastic fluid, i.e. proteins found anywhere in the apoplast, under all of these native and stress-related conditions.

1.3 Conventional secretory pathway

How do these (glyco)proteins and polysaccharides get to the PM/apoplast? Polysaccharides are typically synthesized in either the Golgi or at the PM. In the Golgi case, glycosyltransferase (GT) proteins act within the Golgi to build up the polysaccharide in a stepwise fashion. The assembled molecules are then transported via vesicles through the trans-Golgi network (TGN) and secreted into the apoplast after fusion with the PM. Synthesis of some wall polysaccharides occurs directly at the PM via integral membrane complexes. For example, for cellulose this occurs by complexes of cellulose synthase (CESA) proteins, which form a rosette of between 18 and 24 proteins (Oehme et al., 2015). CESA adds glucose from UDP-Glc from the cytosolic space onto a growing glucan chain, which then passes through a pore into the cell surface to form a larger microfibril.

These PM bound proteins also arrive via the TGN and secretory vesicles. Unlike polysaccharides that are secondary gene products, these and all other apoplastic proteins

are synthesised by transcribing DNA in the nucleus to mRNA, and translating mRNA to proteins. Secretory proteins include a leading N-terminal of such proteins that contains a signal peptide; a short chain of amino acids with a defined hydrophobic motif (Armenteros et al., 2019). This leader sequence is recognised and facilitates the co-translation of the protein into the endoplasmic reticulum (ER). This starts the classical secretory pathway for proteins, and at this point the SP is cleaved. Single TM domains can also act as signal anchors for translocation but are not cleaved, and multiple TM containing domains also enter the ER membrane.

From the ER, proteins are transported to the Golgi via coat protein complex II (COPII) vesicles. Retention signals for the ER can be recognised in certain proteins, and returned to the ER via a COPI vesicle. Some PTM processes begin in the ER, and continue in the Golgi, such as for ER for N-linked glycosylation and HRGPs. The Golgi is where many other PTMs occur, including glycosylation, and from the Golgi proteins are trafficked to their destination. This final destination is determined in the cell through either sorting signals to either another organelle, or other protein properties. Nikolovski et al. (2012) found differing properties of the TM domain of single-pass ER, Golgi, and PM bound proteins. Without any such other signal, the default pathway will be to the cell surface (PM/cell wall). The return of components from the PM to the TGN, endocytosis, also occurs. The conventional/classical pathway including ER, Golgi and TGN in plants is shown in **Figure 1.4**, however an understanding of alternative methods for protein (and some polysaccharide) secretion that deviate from standard ER, Golgi and TGN mechanisms is emerging.

1.4 Unconventional secretory pathway

Almost by definition, any kind of deviation from the conventional secretory pathway is unconventional, and it is a term that covers a number of possible mechanisms for secretory trafficking. In plants it has been the presence of extracellular proteins without known secretory features such as SP or TM that has sparked interest in alternative routes to the extracellular space. The deviations from conventional protein secretion that has

given these routes their various names; non-classical, non-conventional, unconventional, signal-less and leaderless. Leaderless in this case refers to proteins without a SP, called leaderless secretory proteins (LSPs), and the broader process they undergo is unconventional protein secretion (UPS). These exceptions have led to proteins with the SP or TM that are secreted to be known as classical or conventionally secreted proteins (CSPs). Another term that could encompass UPS is moonlighting protein (MP), as defined in MoonProt database (Chen et al. 2018) and MPFit prediction tool (Khan and Kihara 2016; Khan et al. 2017) which are proteins with distinct functions in the same protein, possibly in multiple locations. The broad definition of any deviations from the norm allows for other forms of UPS have been shown that still involve a SP or TM sequence, such as a pathway that still enters the ER via a classical mechanism, but bypasses the Golgi or trans Golgi network (TGN). In this work, LSP and UPS will be the preferred nomenclature.

Eukaryotic UPS is best understood in mammals and yeast (Ding et al., 2014). There are also examples from *Drosophila melanogaster* and *Leishmania major* (Dimou and Nickel, 2018). The proposed UPS pathways derived from these studies and as reviewed by Rabouille (2016) are shown in **Figure 1.3**, synthesising the types of UPS into general categories for eukaryotes though notably plant specific routes were not taken into account. The defined categories are; type I pore-mediated across the PM; type II ABC transporter-based secretion (which is poorly understood (Rabouille, 2016) and therefore not shown on **Figure 1.3**); type III autophagy or organelle related; and type IV Golgi bypass (not necessarily LSPs). Recent work on another UPS pathway is the release of vesicles to the cell surface, membrane bound extracellular vesicles that can form in multivesicular bodies (MVBs) and are released from the cell after fusion with the PM. Though similar to Type III, Nickel and Rabouille (2018) suggest it should be a distinct type under that naming system. Another UPS review related to human diseases by Kim et al. (2018a) merged the Type I-IV system, noting that some UPS proteins are characteristic of more than one type (also suggested by Popa et al. (2018)) and/or are not well understood. Under this naming system, there is leaderless non-vesicular UPS (e.g. combining Type I and Type II from Rabouille (2016)), leaderless vesicular UPS, and Golgi-bypass route (Type IV). Research in UPS is rapidly changing (Dimou and Nickel, 2018), so perhaps this overlapping

nomenclature demonstrate this changing landscape of understanding. An advantage of the less specific pathway definitions is that can generalize to cover UPS routes that may be distinct to only certain organisms such as via specific organelles. However as the number of confirmed plant UPS proteins is limited, discussed further in Section 1.5, it was not included in these generalisations for eukaryotic UPS and as such these broad definitions do not capture explain the variety of proposed plant UPS mechanisms.

**This image has
been removed by the
author of this thesis for
copyright reasons**

Figure 1.3: Summary of generalised unconventional protein secretion (UPS) pathways in eukaryotes, modified from Rabouille (2016). The conventional (CPS) pathway, and unconventional (Type I, III and IV UPS) pathways are shown.

Despite the lack of clarity on the pathways, known UPS proteins in non-plant eukaryotes are noted in several reviews (Dimou and Nickel, 2018; Rabouille, 2016; Kim et al., 2018a). Selected examples, mostly in mammals and yeast, for each of the four generalised eukaryotic pathways are of note. Stress conditions are thought to largely contribute to UPS and in mammalian systems stress such as inflammation leads to leaderless secretion of interleukin 1 (IL-1 β), transglutaminase 2 (TG2) and thioredoxin (Rabouille, 2016) through pores (Type I). There are also constitutively expressed LSPs. Interestingly, two fibroblast growth factors, FGF1 and FGF2 are two human LSPs where FGF1 is stress response LSP compared to the constitutively secreted FGF2, yet both are expected to be Type I (Prudovsky et al., 2013). Adding an SP to FGF2 lead to biological inactivity, indicating that processing in the ER-Golgi disrupted its function (Wegehingel et al., 2008). FGF2 was one of the first UPS proteins found, and (Dimou and Nickel, 2018). In human immunodeficiency virus (HIV), transactivator of transcription (TAT) is also proposed to constitutively make pores for unconventional secretion (Dimou and Nickel, 2018). A reason proposed for such constitutive UPS is to avoid modifications that would occur in the Golgi, such as glycosylation.

Omitted from **Figure 1.3** is Type II secretion, mediated by ATP-binding cassette transporters (ABC transporter) as it is not well understood (Rabouille, 2016). One example from yeast is Mating type MAT α (Michaelis and Barrowman, 2012). The yeast Acb1 protein is a characteristic LSP of the organelle based LSPs (Type III) pathway. Heat shock proteins (HSPs) in mammals have been proposed to mediate this pathway, translocating LSPs inside organelles in order for secretion to occur when the membrane bound organelle fuses with the PM. Another recent type III LSP protein identified in yeast is superoxide dismutase 1 (SOD1) (Dimou and Nickel, 2018).

A critical distinction though, regardless of the nomenclature for Type IV, is that Golgi-bypass pathway does not consist of LSPs. UPS via Golgi-bypass is due to TM containing proteins trafficking from the ER to the PM directly (Kim et al., 2018a; Rabouille, 2016). Known examples of the ER to PM route use a variety and combinations of molecules to facilitate this UPS (e.g. GRASP, COPII, SEC16), indicating no single molecular mechanism between them, and this pathway has been associated with disease states including kidney

disease, cancer, and Cystic fibrosis Kim et al. (2018a).

In bacteria, the concept of UPS as any deviation from the conventional secretory pathway is generally well established. Earlier work defined CSP in bacteria as those that either follow the conventional Sec (an N-terminal signal peptide) or Tat (twin-arginine translocation) secretion pathways (Bendtsen et al., 2005), however there are now at least 12 categories of secretion in bacteria (reviewed in (Green and Mecsas, 2016)). Since these do not involve membrane bound organelles, they are most similar to Type I UPS mechanisms. Recent work, reconsidered what should be considered non-classical in light of the additional known secretion systems Zhang et al. (2019), and a new prediction tool for UPS in Gram-positive bacteria, PeNGaRoo, was developed. Although some information about UPS in mammals and yeast is known, in most cases the molecular mechanisms are still not well understood (Nickel and Rabouille, 2018). In plants, even less is known.

1.5 Unconventional protein secretion (UPS) in plants

Despite the focus on mammalian and yeast UPS, there is growing evidence and interest in plant LSPs and plant UPS (Davis et al., 2016). The existence of unconventional protein secretion (UPS) in plants is an ongoing and rapidly changing field of research (Ding et al., 2012; Davis et al., 2016). In plants UPS pathways have similarly been defined by Ding et al. (2014), though confusingly this refers to types that differ from general eukaryotic types Rabouille (2016). An updated annotation of plant UPS pathway types is shown in **Figure 1.4**, where type 1 refer to LSPs that are suggested could be secreted via translocation across the PM as in other eukaryotes. Direct evidence for this mechanism in plants has not been established, rather this pathway consists of LSPs that are not thought to be secreted by organelles. Four further subtypes of plant UPS are similar to the single Type III classification of organelle bound UPS; type 2 which is vacuole-mediated UPS in response to pathogens; type 3 via multivesicular bodies (MVBs); type 4 via the exocyst positive organelle (EXPO) organelle.

Notably, however, these five pathways may represent both differing actions yet also different protein specificity. For example, the direct translocation across the PM in type 1

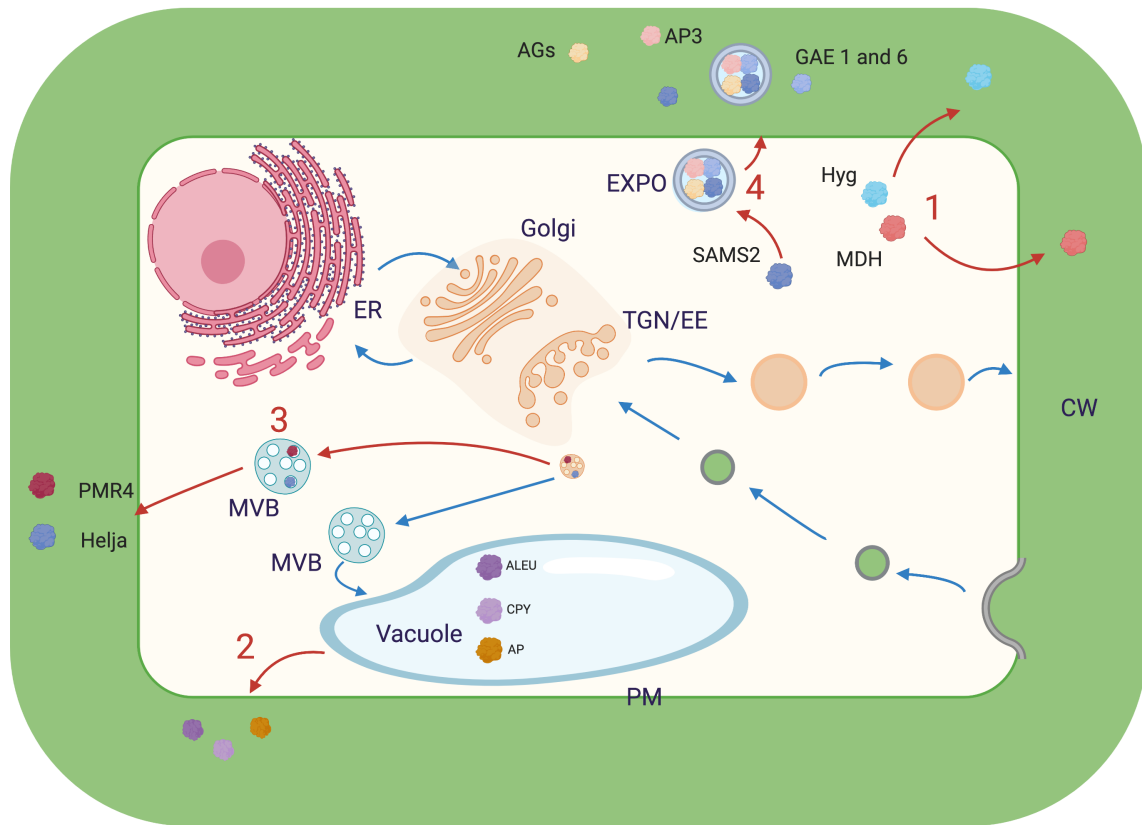


Figure 1.4: Summary of conventional and unconventional protein secretion (UPS) pathways in plants, adapted from Davis et al. (2016) and created using BioRender (<http://biorender.com>). The conventional/classical secretory pathway is shown with blue arrows, including transitions between the ER, Golgi, TGN and early endosome (EE), and secretory vesicles to and from the PM/CW in the processes of exocytosis and endocytosis, respectively. Unconventional secretory pathways, including associated proteins as listed in **Table 1.1**, are shown with red arrows and numbered: 1) LSPs secreted without organelles, directly across the PM including Hyg and MDH; 2) Vacuole mediated UPS, where multivesicular bodies (MVBs) and their proteins such as ALEU, CPY and AP are transported to the vacuole conventionally, before being secreted unconventionally via the vacuole; 3) MVBs containing proteins such as Helja and PMR4 fusing with the PM; 4) Proteins such as SAMS2, AGs, AP3 and GAE1/6 are contained in the EXPO organelle which subsequently fuses to the PM/CW.

secretion is thought to require a specific sequence motif present in the protein structure, whereas UPS via intermediate organelles that form in the cytosol could also passively include cytosolic proteins, and this passive cargo secreted without such specificity.

Given the overlap between the general eukaryotic type systems and these plant specific ones, there is room for confusion in scope and terminology, and it would be useful to have a common set of terms to describe them. For example, Golgi bypass is mentioned in both Ding et al. (2014) and Davis et al. (2016) to describe Type I proteins. The Golgi-bypass route shown for humans (Kim et al., 2018a) is different, as it is used to describe multiple TM domain containing proteins, that after entry to the ER, are directly transported to the PM, and therefore ‘bypasses’ the Golgi. For plants, Golgi-bypass and Golgi-independent are used interchangeably, yet Golgi-independent should be preferred as type 1 refers to proteins without SP, and so both ER and Golgi are bypassed. This clarification is also required because a direct ER to PM pathway is not yet known for plants. In a comprehensive review of ER-mediated UPS in plants Bellucci et al. (2017) explored the possibility of ER directed proteins that then bypass the conventional Golgi-TGN pathway. They refine the definitions of UPS to include any deviation from the Golgi pathway, though note there are no conclusive examples of direct ER to apoplast trafficking in plants.

UPS in plants is thought to be maintained as both an alternative pathway for constitutive secretion as well as for responding to stress. The pathways in **Figure 1.4** are derived from the limited number of plant UPS proteins (**Table 1.1**), each with varying degrees of evidence. Mannitol dehydrogenase (MDH) in celery is amongst the few plant proteins generally accepted to undergo UPS (Cheng et al., 2009c; Davis et al., 2016), as it was found to be secreted after the application of salicylic acid, which simulated a response to pathogens. This response continued even when brefeldin A was applied, which inhibits COPII, indicating the secretion route did not involve the Golgi. It is one of two type 1 LSPs in **Figure 1.4**, however it is the only true plant candidate. The reference to Hyg in **Figure 1.4** is not strictly a plant protein, but actually the *Escherichia coli* hygromycin B phosphotransferase resistance gene (HYG^R), which inactivates Hygromycin B. Despite this it has been included in plant UPS reviews (Ding et al., 2014; Davis et al., 2016) and fea-

tured in the diagrams summarising plant UPS in each (e.g. **Figure 1.4** reproduced here). UPS has been demonstrated for HYG in transgenic *Arabidopsis* and *Nicotiana benthamiana* plants, using electron microscopy and the extracellular localisation of a HYG-GFP fusion construct (lacking a SP) to infer leaderless secretion (Zhang et al., 2011a; Zhang and Li, 2016). This example of UPS was shown to involve the synaptotagmin 2 (SYT2), which localised to the Golgi yet HYG^R is not secreted in a SYT2 knockout mutant (Zhang et al., 2011a). However this inclusion of HYG from *E. coli* in the plant LSP literature is questionable, and similarly, whether to consider it as an example of plant LSPs for this Thesis needs consideration. Since it offers a unique example over the LSPs from other systems, as there is evidence of UPS when it is expressed in plants, we will use it to contrast how the methods developed in this Thesis that apply to native plant LSPs compare to other species LSPs, and note will be made of its origins when doing so, particularly in **Chapter 4** when we compare the outcomes of new tools on plant LSPs.

Table 1.1: Unconventionally secreted proteins in plants either directly (1), vacuole-PM route (2), multi-vesicular bodies (MVBs) (3) or via EXPO (4)

Gene	Organism	UPS route	Secretory Features	LSP
MDH	Celery	Direct	-	Yes
Hyg	E. coli	Direct	-	Yes
Aleurain (ALEU, At5g60360)	Arabidopsis	Vacuole	SP	No
Carboxypeptidase Y (CPY, At3G10410)	Arabidopsis	Vacuole	SP	No
Aspartyl protease (AP, At1g49050)	Arabidopsis	Vacuole	TM	No
Helja	Sunflower	MVB	-	Yes
Glucan synthase like 5 (PMR4/GSL5, AT4G03550)	Arabidopsis	MVB	Multiple TM	No
Arabinogalactan glycosyltransferase (AG, At1G32930)	Arabidopsis	EXPO	SP	No
UDP-glucuronate epimerase 1 (GAE1, At4g30440)	Arabidopsis	EXPO	TM	No
UDP-glucuronate epimerase 6 (GAE6, At3g23820)	Arabidopsis	EXPO	TM	No
Apyrase 3 (AP3, At1G14240).	Arabidopsis	EXPO	TM	No
S-adenosylmethionine synthetase (SAMS2, At4G01850)	Arabidopsis	EXPO	-	Yes

Three known UPS proteins in *Arabidopsis* that take the type 2 vacuole-PM route were found in the extracellular fluid after infection by bacteria (Hatsugai et al., 2009). Each of aleurain (ALEU, AT5G60360), carboxypeptidase Y (CPY, AT3G10410) and aspartyl protease (AP, AT1G49050) contain conventional secretory SP or TM features (**Table 1.1**). The fusion of the vacuole to the PM is thought to be a programmed cell death response in response to pathogens. This mix of CSP and LSP cargo in UPS is also present in the MVB and EXPO pathways. CSPs in these instances represent the form of UPS where expected classical routes, or secretory locations such as the ER or GA, are circumvented. MVBs are involved in conventional secretory pathway from the TGN to the vacuole in type 2, yet are also shared as a direct form of unconventional secretion in type 3. The MVB pathway is implicated in the secretion of the only well characterised plant LSP, Helja (Pinedo et al., 2012a, 2015; Del Rio et al., 2018) from the sunflower (*Helianthus annuus*). A MVB mediated CSP is glucan synthase like 5 (PMR4/GSL5, AT4G03550) which has multi-TM proteins and is involved in callose synthesis, particularly in response to pathogens (Meyer et al., 2009).

Of the 5 candidates from the EXPO pathway, four contain secretory features. These were identified either from Arabinogalactan glycosyltransferase (AG, AT1G32930, Poulsen et al. (2014)), UDP-glucuronate epimerase 1 and 6 (GAE1 AT4G30440, GAE6 AT3G23820,), and Apyrase 3 (AP3, AT1G14240). They are implicated in the EXPO to PM pathway of UPS by the colocalisation with the EXPO marker EXO70E2 (Poulsen et al., 2014, 2015) when transiently expressed in *Nicotiana benthamiana*. S-adenosylmethionine synthetase (SAMS2, AT4G01850) however is an LSP, is also shown to co-locate with EXPO using a SAMS2-GFP construct in *Arabidopsis* (Wang et al., 2010). The extracellular functions of the EXPO mediated proteins is not yet clear. This relatively recent interest in plant UPS has identified just these handfuls of plant UPS proteins, however for LSPs in particular the true number in plants is unknown. As noted, hygromycin phosphotransferase (HYG^R) from *E. coli* could be considered a pseudo-plant LSPs. Only three of the plant proteins discussed here fall into the LSP category, using different UPS mechanisms, and from three different plant species: mannitol dehydrogenase (MDH) in celery, Helja in sunflower and S-adenosylmethionine synthetase (SAMS2) in *Arabidopsis*.

1.6 The plant LSP problem

This scarcity is due in part to the way the LSPs are experimentally observed and classified in plants, and leads to uncertainty about the true number of plant LSPs. Up to 50% of proteins found in plant cell wall/apoplast secretomes lack a SP (Agrawal et al., 2010; Krause et al., 2013) yet the possibility that these are all LSPs is confounded by the possibility that they could be cellular contamination. Estimates of the size of classical secretomes are around 18% of the *Arabidopsis* proteome (Alexandersson et al., 2013), so if all these observed proteins were indeed LSPs than approximately 36% of the proteome would be secreted. If the uncertainty around whether such proteins are truly LSPs is resolved, then the true size and contents of plant secretomes will be better understood, and improve our understanding of the apoplast.

Resolving this issue is complicated by the nature of the plant cell wall, open as it is between adjacent cells and both in contact with apoplastic fluid. This means that to purify and correctly identify wall-related proteins, several steps are required. Other cellular compartments of the secretory pathway, such as the Golgi, can be enriched using differential centrifugation methods for further analysis although some level of contamination is expected. The cell wall however is particularly difficult to purify (Canut et al., 2017) as it is not enclosed in a membrane. Isolating it in order to inspect only the contents of the cell wall is complicated by this ambiguity. These difficulties mean that various techniques to isolate the cell wall and apoplastic fluid have been devised **Figure 1.5**. They are broadly defined as destructive or nondestructive (Albenne et al., 2013). Vacuum infiltration can be used to extract the apoplastic fluid, and the contents then run through mass spectrometry (MS) to identify proteins present. This requires some processing that can lead to other plant sub-cellular components leaking into the mixture due to membrane rupture, but largely leaves the cell wall intact.

Destructive measures instead grind up the plant material, which makes access the plant cell wall material easier but also leads to an increase in contamination with other proteins from different sub-cellular locations. Both methods have advantages and disadvantages; common to them though is that contamination from other parts of the plant cell, in particular the cytosol, can be introduced. If only conventional secretion was con-

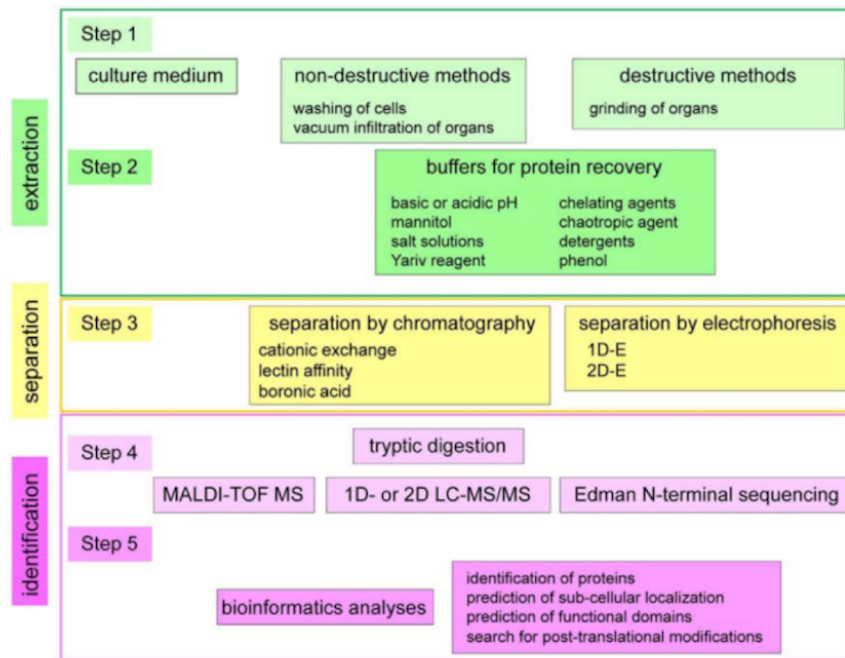


Figure 1.5: Variety of methods for performing a plant cell wall proteome as defined by Albenne et al. (2013), split into the extraction of apoplast proteins, the separation techniques and the identification stages

sidered this is irrelevant; SPs found in the canonical sequences of proteins identify them as secretory proteins, and those without an SP must therefore be assigned as cytosolic contamination.

Estimates of cytosolic contamination can be done using a variety of techniques using markers for well characterised cytosolic proteins. These can include taking these markers and comparing their enzyme activity, fluorescence or using antibodies to contrast their prevalence in secretomes Alexandersson et al. (2013). Although this is a useful approach, it only indicates the extent to which contamination could be present, and cannot rule in or out individual proteins as being contamination. Another is to consider the proportion of apoplastic proteins found. For example, Ge et al. (2011) propose that if greater than 50% of proteins have a classical SP then contamination is low.

The existence of proteins that undergo UPS are typically inferred by the presence of proteins in the secretory space in circumstances where they are not expected. The use of pharmacological intervention has been used to disrupt the secretory pathway, and the

unexpected secretion of proteins in these circumstances has been used as evidence for UPS. For example in plants, brefeldin A (BFA), which disrupts ER-Golgi transport, and wortmannin which interferes with the trafficking of vesicles (Foissner et al., 2016) have been used. Functional differences in the presence of these is used to assign dependency on parts of the secretory pathway. Notably however the specificity of the interference function of BFA is not thought to be relevant to all *Arabidopsis* tissues, considered ineffective in root tissue yet effective in non-root (Zhang et al., 2011a), and suggested to be effective in tobacco leaves (Kaneko et al., 1996). Continued secretion in the presence of BFA is therefore often used yet an inconclusive measure that unconventional secretion is occurring (Krause et al., 2013).

Depending on the experimental methods used, the number of proteins lacking SP varies greatly. As summarised by Albenne et al. (2013), this can range from no potential LSPs to 50% to up to around 85% in one case (**Appendix Figure A.1**). Thus from a proteome there is often a division between the proteins with SPs that are considered true, and those without that are then to be considered as either contamination or LSP candidates. This concept is shown in **Figure 1.6A**. The strategies used in experiments to deal with this concept varies greatly, and falls into three broad categories in published experiments which we will term: accept, reject and filter.

The accept strategy (Strategy I, **Figure 1.6B**) is the most optimistic, allowing all proteins found in a plant cell secretory proteome to be considered to be secreted, the CSP as expected and assuming proteins without SP are LSPs. No contamination is considered. It is the simplest strategy, requires no bioinformatics, though is likely to lead to false positives when there are a large number of contaminants, as well as no discernible way to prioritise the proteins for validation.

Reject (Strategy II, **Figure 1.6C**) is the most common strategy that follows the conventional protein theory and rejects proteins without an SP which are then taken as contamination. This strategy will reduce false positives for secretory proteins, since no incorrect LSPs are identified but at the risk of increasing false negatives, i.e. losing any genuine LSPs as they do not have an SP. The inference of SP is typically done with a bioinformatics program such as SignalP (Armenteros et al., 2019). In practice this has been the

strategy used in many plant experiments by default, i.e. the absence of consideration for LSPs would fall under this category. This also applies when the presence of LSPs is noted yet not considered due to the lack of reference data or computational tools. Proteins inferred to be intracellular contamination are often left off the main paper, and included in the supplementary material.

The third strategy is to filter (Strategy III, **Figure 1.6D**), and identify both CSP and LSP proteins. Firstly, the results are analysed for the classical SP, and included if they satisfy this criterion. The remainder are then filtered to determine if these remaining proteins are LSPs or contamination. This filtering stage will use a bioinformatics program, typically returning predictions with an estimated confidence score that a protein is an LSP. The accuracy of the prediction tool influences what balance is found between the discovery of genuine LSPs with the possibility of that contamination may also be erroneously identified.

1.7 Filtering for plant LSPs with SecretomeP

The most often used program for such a filter is SecretomeP (Bendtsen et al., 2004). This tool is the focus of the publication included in **Chapter 2** due to this popularity, relative to others UPS prediction tools as detailed in **Table 2.1**. The ranking of the popularity of UPS predictions programs in this table remain unchanged since publications (from April 2016 to November 2019), with the increasing numbers of citations for each of the tools shown in **Appendix A.1**. As the most used and usable tool SecretomeP is the primary computational method for plant LSPs in the published literature. Given the need to distinguish the true number of plant LSPs from potential contaminants, it is vital that the accuracy of any tool used in the filtering strategy be well characterised. For SecretomeP there is an important aspect to prompt thorough analysis given its use on plant proteins; SecretomeP was trained on human proteins.

Caveats about its applicability to plants were included in several publications (e.g. Agrawal et al. (2010); Albenne et al. (2013); Cheng et al. (2009a)), yet it has been applied to a wide variety of plant species (**Table 1.2**). The continued use of SecretomeP lacks an under-

standing of just how accurate is it on plants, and the other tools for UPS prediction are similarly not trained on plant data. Do the properties captured in the prediction model transfer from humans to plants? What settings of the program should be used? Is it better than other programs at detecting plant LSPs?

Table 1.2: Plant studies using SecretomeP

Species	References
<i>Beta vulgaris</i>	Ceballos-Laita et al. (2015)
<i>Gossypium hirsutum</i>	Zhang et al. (2014)
<i>Coffea arabica</i>	Guerra-Guimarães et al. (2014)
<i>Brassica juncea</i>	Sehrawat and Deswal (2014)
<i>Physcomitrella patens</i>	Lehtonen et al. (2014)
<i>Lupinus albus</i>	Scarafoni et al. (2013)
<i>Vitis vinifera</i>	Delaunois et al. (2013)
<i>Oryza sativa</i>	Wang et al. (2012b); Song et al. (2011); Zhou et al. (2011); Cho et al. (2009)
<i>Solanum tuberosum</i>	Fernández et al. (2012)
<i>Helianthus annuus</i>	Pinedo et al. (2012b)
<i>Arabidopsis thaliana</i>	Ge et al. (2011); Davis et al. (2009); Cheng et al. (2009a); Jamet et al. (2008)
<i>Cicer arietinum</i>	Gupta et al. (2011); Bhushan et al. (2011)
<i>Linum grandiflorum</i>	Ushijima et al. (2011)
<i>Populus deltoides</i>	Pechanova et al. (2010); Dafoe and Constabel (2009)
<i>Zea mays</i>	Laohavisit et al. (2009); Zhu et al. (2006); Gray-Mitsumune and Matton (2005)
<i>Medicago sativa</i>	Kusumawati et al. (2008)
<i>Nicotiana</i> spp.	Juarez-Diaz et al. (2006)

Part of the difficulty in answering these questions, and a possible explanation for the inertia the current reject and filter strategies have, is the difficulty in assessing the accuracy of these programs. A typical approach is to have a truth set, that us a ‘gold standard’ of known plant LSPs. The higher bar, often biochemical evidence, recommended for confirmation of LSPs means that there are few generally accepted plant LSPs. As will be

discussed further in **Chapter 4** when new tools are developed, the results of tool on just three example LSPs is too few gauge its accuracy. Using SecretomeP to identify likely candidates as LSPs, or taking LSPs from the literature that used SecretomeP in identifying them, does not provide an unbiased set of data for evaluating the accuracy of SecretomeP on plants. This leaves an impasse in critically analysing the current approaches to plant LSP identification. Fortunately, the theory on which SecretomeP is based provides a silver lining in which this can be analysed further.

1.8 The common feature hypothesis

When initially published in 2004, SecretomeP lacked a gold standard method with which it could be compared to. There were 13 known human LSPs at the time. In order to build a prediction tool, many more positive examples were required. The authors overcome this obstacle by making the following claim: regardless of the mechanisms of secretion (classical or leaderless), proteins that are secreted will have common properties that lead them to be secreted. Therefore, classically secreted proteins, with the SP ignored, should contain similar properties to LSPs. The concept is illustrated in **Figure 1.6E**, with the unified class of CSPs and LSPs treated as positive data (red).

They therefore trained a tool using known human CSPs, with the SP portion of the protein removed. A set of known nuclear and cytosolic proteins was used as a negative set. The prediction model used was a neural network, a machine learning algorithm often used for classification. This involves extracting features from the positive and negative data sets, and training a neural network to give a probability score between 0 and 1 that data is from the positive class. Internal validation results in a specificity of 40% and controls for a false-positive rate (FPR) of 5%. The 13 LSP known proteins were applied after training, and they found it predicted 10/13. Under the “Accept/Reject/Filter” framework, this approach seeks to conflate the identification of both SP and LSPs, separate from the contamination. It then applies a simple mathematical operation as a second filter. Given a list of proteins that exhibit these common features for all secreted proteins, a further filtering for SPs can divide the secretory proteins into CSP or LSP (**Figure**

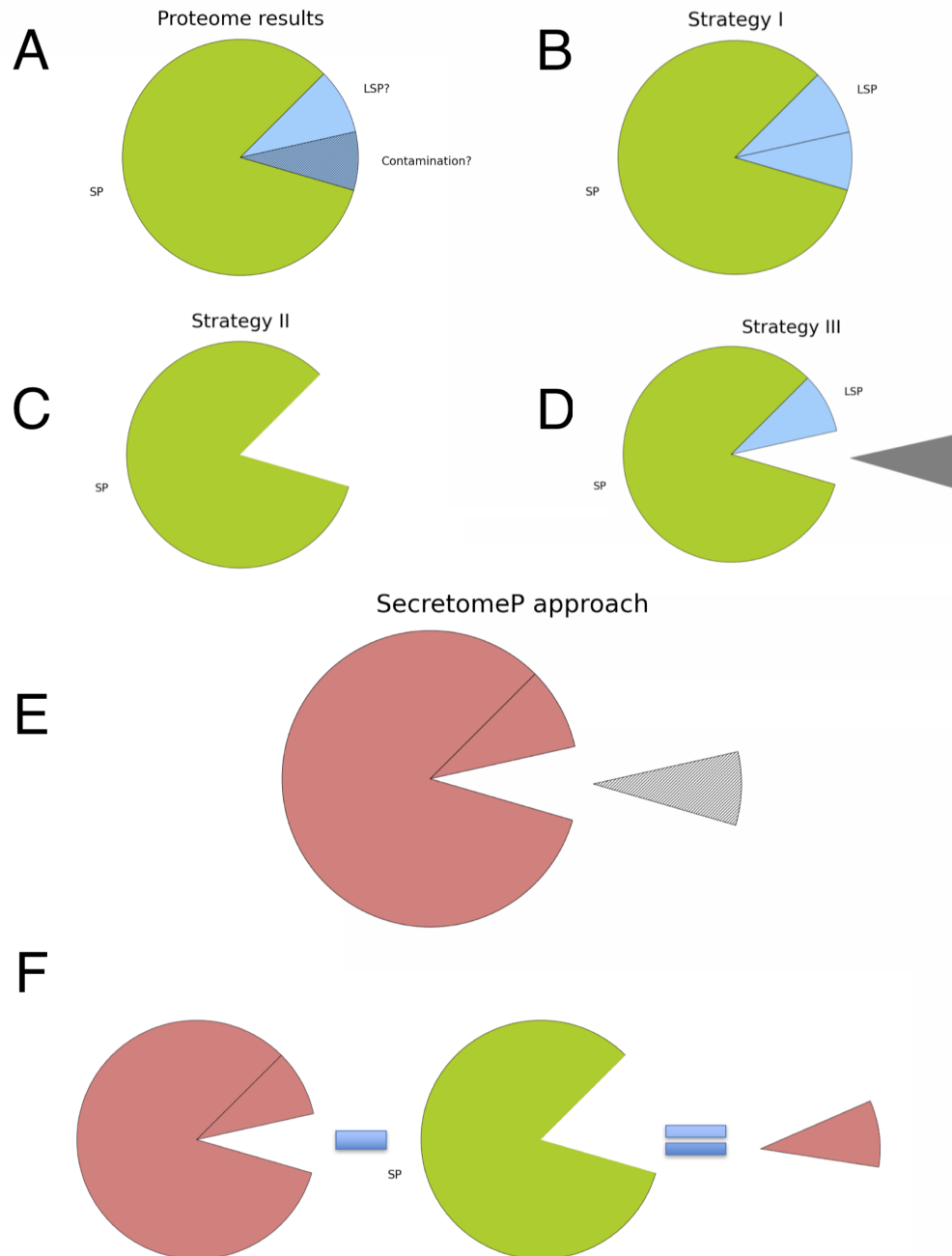


Figure 1.6: Summary of bioinformatics approaches to identify LSPs. (A) Classification of proteins within a cell wall proteome as either having a SP or being possible LSPs or contaminants. (B) Strategy I where all proteins are accepted as secreted, (C) Strategy II where only SP-containing proteins are accepted, (D) Strategy III where SP-containing proteins plus predicted LSPs are accepted, (E) Common feature hypothesis used in SecretomeP, where all secretory proteins (CSP and LSP, red shading) are assumed to share common features different to contaminants (grey shading) (F) LSPs are identified with the SecretomeP tool by subtracting CSPs (green shading) from all secretory proteins (red shading), leaving predicted LSPs.

1.6F). Any proteins that are not identified in this approach are therefore contamination. In a follow up paper, the common feature hypothesis (CFH) was also applied to bacteria (Bendtsen et al., 2005). A similar approach was used, and notably different features were selected as the most predictive for Gram-positive and Gram-negative bacteria. More experimentally verified examples of proteins were available for bacteria, and the accuracy of these was in excess of 80%.

Though not considered due their limited use, the hypothesis behind the alternative programs in **Table 2.1** have some degree of overlap with SecretomeP. Although SecretomeP is the focus, the citation count was also not the only criteria when considering which prediction tools to investigate for LSP identification in plants. Difficulties accessing or installing the programs detailed at the time in **Table 2.1** have also subsequently noted by others (Nielsen et al., 2018), with several tools are either unavailable or produce errors when accessing their website. The throughput, either through web servers or limited number of downloadable tools also varied, which impacted the speed at which large numbers of proteins can be analysed. The hypothesis behind several of them also meant the lack of a gold standard for plants limited the ability to evaluate them, even if they had been available. SPRED (Kandaswamy et al., 2010) also uses the common feature hypothesis with SP removed secretory proteins as positive data, however the website in the publication is not currently available. Other tools such as Sec-GO Huang (2012) GO term annotation on data to predict secretion, yet is also unavailable. SRTpred Garg and Raghava (2008) trained on unmodified CSPs, with features selected to avoid reliance on the SP, and can still be accessed. The SecretP iterations (Yu et al., 2010a,b, 2013) attempts to train on CSPs, LSPs, and non-secreted proteins (NSPs) using annotations for extracellular location without correspond annotations SP to infer the LSP data set, yet none are functioning. The bacteria based NClassG+ Restrepo-Montoya et al. (2011) is also offline. Analysis into the applicability of these tools in plants is limited by these issues, and the overwhelming use of SecretomeP in general and on plants led to focused analysis on it.

Given SecretomeP has been used on plants extensively, what is the expected accuracy? The Cheng et al. (2009a) study of the secretome of *Arabidopsis* suspension cells in response to salicylic acid found that over 60% of the proteins secreted were predicted leaderless

by SecretomeP. This usage was appropriately balanced noting the program missed several CSPs and that the predictions tools are limited by training data, and should not be considered exact. One aspect to the 60% estimate however this used a threshold of any output score of over 0.5 (i.e. greater than 50% probability the protein is secreted). This differs from the 0.6 threshold established during the creation of the mammalian version of SecretomeP (Bendtsen et al., 2004). If the 0.6 cutoff is used, the purported 60% of proteins in the study predicted to be LSPs drops down to 37.5%. Which cutoff should be used? This use also does not take into account false positives. The percentage estimate is only judging what extent of proteins found were predicted by the program, assuming they are all positive. Putting aside that this may not be true given the possibility of contamination, a lack of an estimate of false positives hampers being able to use the tool on plants confidently. Prior to the work in this Thesis, a compelling gap in the literature was just how appropriate SecretomeP was for plants. This question is addressed in **Chapter 2**, where the common-feature hypothesis is explored in more detail.

1.9 Validating putative leaderless secretory proteins in plants

In addition to the few well generally established LSPs in plants (**Table 1.1**), a review by Ding et al. (2012) lists 76 proteins putatively thought to be plants LSPs from selected studies (**Appendix Table A.2**). Albenne et al. (2013) stated that repeated appearances of a protein is not enough to infer true LSPs, instead requiring biochemical evidence. Robinson et al. (2015) concede on this point that immunological evidence is required, yet counter that proteins with low abundance at distinct developmental points make these difficult to identify in this manner, meaning that such confirmation can be extremely challenging. The frequency with which LSPs appear however may be worth investigating. One notable disadvantage to the published review is the occurrence of repeated proteins, e.g. glutathione transferase (AT1G78380) is listed 3 separate times in the putative LSP list. These three come from different experiments and do not warrant any discussion in the review despite this recurrence. After further duplicates are removed from the list includes 65 putative LSPs. There are fewer LSPs than listed, and these duplicates also

illustrate that listing LSPs does not easily allow for repeated observations of proteins to be interrogated, for example to query if the same LSPs appeared each time under similar experimental conditions. Another aspect to the Ding et al. (2012) list is that it includes sources that themselves identified LSPs using SecretomeP (Cheng et al., 2009a) and prior to investigating its accuracy, it is unclear if these should be the selected candidates for likely LSPs in such studies.

Even with a list of candidates confirming the sub-cellular location of a protein is a low throughput and expensive task. The two most common approaches are to either raise an antibody to the protein of interest and confirm subcellular location via electron microscopy, for example, or use a protein tag such as green fluorescent protein (GFP) to follow the pathway of secretion of the fusion protein using confocal laser scanning microscopy. This is time consuming but also requires candidates to be identified first in order to follow up with validation work. Converging on a strict set of criteria for verifying plant LSPs is made difficult by the fact that the methods for validation also have limitations. Fluorescent tags may affect the expression/secretion of proteins (Tanz et al., 2013), and therefore the inference of whether they are actively secreted. Microscopy-based analysis can confirm sub-cellular location yet does not add insight into mechanism. The possibility of active and passive pathways discussed previously necessitates prior knowledge of the pathway and mechanisms for secretion. Even then, knowledge of the context of these pathways (e.g. either constitutive or in response to stress) would be needed for the ultimate validation through biochemical experiments designed to show if LSP candidates are secreted. Given that each potential pathway(s) and mechanism(s) could require a set of different experiments (each with limitations), achieving a complete set of overarching criteria for validation is difficult at this stage of plant LSP knowledge. For the purposes of this Thesis, driven by a dearth of true positive data, we settled for a less stringent set of criteria of either microscopy- or fluorescent- based approaches to establish the few examples of plant LSPs. Without a reliable method to identify candidate proteins such as a LSP prediction tool or a curated database of LSPs then there are no guidelines as to which proteins within a cell wall proteome should be followed up.

1.10 Thesis scope

Although they represent $< 10\%$ of the dry weight of the plant cell wall, knowledge of the proteins (and proteoglycans) within the extracellular space is vital to understanding the function of the plant cell wall. The existence of UPS and LSPs, combined with the difficulty in isolating complete, uncontaminated plant secretomes leads to the situation where the true number of LSPs is unknown, and therefore our knowledge of plant cell wall proteins is incomplete. There are many putative LSP candidates, particularly in *Arabidopsis* but few are verified LSPs.

Prediction programs have tried to close this gap, however, work to date with plant LSPs has lacked critical bioinformatics analysis. In part, the difficulty of working without a sufficient number of positive examples has led to the adoption of human-trained programs, and by default, acceptance of the hypothesis behind them. A requirement for biochemical evidence to verify true secretion has led to the exclusion of putative LSPs from established databases, disallowing the possibility for cumulative insights. The call for experimental evidence is rightly made for plant LSPs, but should not be made to the exclusion of a parallel bioinformatics approach. This Thesis will apply computational methods to identify the most likely LSPs in plants.

In this Chapter, the ways in which LSP analysis is currently undertaken in plants has been discussed, and several critical issues identified. The lack of a thorough investigation of SecretomeP's applicability to plants is investigated in **Chapter 2**, and answers the question of whether it is accurate for plants. The issues of using SecretomeP as a filtering tool, and the need to revisit possible LSPs across many experiments, is addressed in **Chapter 3**, with the construction of the Leaderless Secretory Protein Database (LSPDB). The outputs of this database, including a curated list of high-confidence putative LSPs are then used in **Chapter 4** to develop new predictive tools for plant LSPs. These tools were then applied to case studies in other plant species, as described in **Chapter 5**, and used to prioritise LSP candidates for validation. The conclusion in **Chapter 6** discusses the insights gained from the new tools, their limitations and future directions for LSP research.

Chapter 2

Better Than Nothing

2.1 Introduction

EXISTING approaches to the LSP problem in plants are explored through the publication “Better Than Nothing? Limitations of the Prediction Tool SecretomeP in the Search for Leaderless Secretory Proteins (LSPs) in Plants” (Lonsdale et al., 2016) that is the basis for this chapter. As discussed in the previous chapter, SecretomeP is by far the most common tool used to identify LSPs in plants, even though it is trained on mammalian datasets. This chapter explores whether this tool and the underlying hypothesis of shared properties between leaderless and conventionally secreted mammalian proteins should be applied to plant proteins. This chapter consists of the publication (Section 2.2) and supplementary material (Section 2.3) verbatim, followed by an updated Discussion (Section 2.4) and Conclusion (Section 2.5).

2.2 Better than Nothing: Limitations of the Prediction Tool SecretomeP in the Search for Leaderless Secretory Proteins (LSPs) in Plants



Better Than Nothing? Limitations of the Prediction Tool SecretomeP in the Search for Leaderless Secretory Proteins (LSPs) in Plants

Andrew Lonsdale¹, Melissa J. Davis^{2,3}, Monika S. Doblin¹ and Antony Bacic^{1*}

¹ ARC Centre of Excellence in Plant Cell Walls, School of BioSciences, The University of Melbourne, Parkville, VIC, Australia,

² The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, Australia, ³ Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Parkville, VIC, Australia

OPEN ACCESS

Edited by:

Nicolas L. Taylor,
University of Western Australia,
Australia

Reviewed by:

Ján A. Miernyk,
University of Missouri, USA
Xiangjia Min,
Youngstown State University, USA

*Correspondence:

Antony Bacic
abacic@unimelb.edu.au

Specialty section:

This article was submitted to
Plant Proteomics,
a section of the journal
Frontiers in Plant Science

Received: 26 July 2016

Accepted: 12 September 2016

Published: 27 September 2016

Citation:

Lonsdale A, Davis MJ, Doblin MS
and Bacic A (2016) Better Than
Nothing? Limitations of the Prediction
Tool SecretomeP in the Search
for Leaderless Secretory Proteins
(LSPs) in Plants.
Front. Plant Sci. 7:1451.
doi: 10.3389/fpls.2016.01451

In proteomic analyses of the plant secretome, the presence of putative leaderless secretory proteins (LSPs) is difficult to confirm due to the possibility of contamination from other sub-cellular compartments. In the absence of a plant-specific tool for predicting LSPs, the mammalian-trained SecretomeP has been applied to plant proteins in multiple studies to identify the most likely LSPs. This study investigates the effectiveness of using SecretomeP on plant proteins, identifies its limitations and provides a benchmark for its use. In the absence of experimentally verified LSPs we exploit the common-feature hypothesis behind SecretomeP and use known classically secreted proteins (CSPs) of plants as a proxy to evaluate its accuracy. We show that, contrary to the common-feature hypothesis, plant CSPs are a poor proxy for evaluating LSP detection due to variation in the SecretomeP prediction scores when the signal peptide (SP) is modified. Removing the SP region from CSPs and comparing the predictive performance against non-secretory proteins indicates that commonly used threshold scores of 0.5 and 0.6 result in false-positive rates in excess of 0.3 when applied to plants proteins. Setting the false-positive rate to 0.05, consistent with the original mammalian performance of SecretomeP, yields only a marginally higher true positive rate compared to false positives. Therefore the use of SecretomeP on plant proteins is not recommended. This study investigates the trade-offs of using SecretomeP on plant proteins and provides insights into predictive features for future development of plant-specific common-feature tools.

Keywords: protein localisation prediction, secretome, unconventional protein secretion, leaderless secretory protein, plant cell wall, SecretomeP

INTRODUCTION

The plant cell wall, a complex extracellular matrix of carbohydrate and some protein, is one of the defining features of plant cells. Protein accounts for up to 10% of the dry weight of the primary cell wall (Doblin et al., 2010) yet despite being a minor component, cell wall proteins (CWPs) play important roles in plants. CWPs can function in the plant cell's normal growth and developmental processes, as well as in response to biotic and abiotic stresses. They can act to modify other wall

components thereby altering the structure and composition of the wall in response to internal and external stimuli. Separating individual plant cells within either a tissue or organ is an open continuous compartment called the apoplast, which also contains (glyco)proteins and, together with the cell wall, collectively refers to the extracellular space. The collection of proteins found in the extracellular space is called the secretome.

Proteins at the cell surface (plasma membrane/apoplast) are typically glycosylated and usually trafficked via the highly conserved eukaryotic process of classical protein secretion involving vectorial transport through the ER and Golgi apparatus. These classically secreted proteins (CSPs) are typically targeted to the ER by an N-terminal signal peptide (SP) where it is subsequently cleaved. However, not all proteins destined for the cell surface have a SP in their sequence and those lacking this feature are referred to as leaderless secretory proteins (LSPs; Krause et al., 2013). These types of proteins are secreted via a route called unconventional (i.e., non-classical) protein secretion (UPS). UPS is a term that covers the atypical cases of LSPs but also some secreted proteins containing a SP that are non-classically secreted, either via intermediate organelles or vesicular bodies (Ding et al., 2014; van de Meene et al., 2016).

In plant proteomic studies, it has been estimated that over half the detected secreted proteins lack a classical SP (Krause et al., 2013). However, due to the possibility of contamination either from already lysed cells (as occurs in cell suspension cultures; Miernyk et al., 2016) or from cells undergoing programmed cell death (*in planta*) or from other sub-cellular compartments during the protein extraction procedure, not all proteins lacking a SP in a secretome experiment are expected to be true LSPs (Rose and Lee, 2010; Albenne et al., 2013). Determining which plant proteins are candidate LSPs that have undergone UPS and which have been introduced as contamination is therefore a challenge (Rose and Lee, 2010; Albenne et al., 2013; Krause et al., 2013). Given the high likelihood of contamination during sub-cellular fractionation procedures, a separate and independent assessment is needed in order to conclude that such proteins are indeed located in the ECS and are legitimate components of the plant secretome. There are three broad assessment strategies that can be implemented for an *in silico* approach: (1) accept all proteins

found, a strategy likely to have a high false-positive rate; (2) include only proteins with SPs and reject others, a strategy likely to discard most real examples of LSPs; or (3) balance true and false positives by including proteins with SPs and filter others based on a prediction of a protein being a LSP. Ultimately, there is a need to adopt experimental approaches, for example immunolocalisation to verify the *in silico* predictions, irrespective of their veracity.

SignalP (Petersen et al., 2011) is often used to identify SPs, yet bioinformatics tools to predict LSPs have only been developed for mammals and bacteria (summarized in Table 1). None were trained on plant proteins and the lack of plant-specific tools for predicting LSPs using the third strategy above is a recognized problem for filtering approaches (Agrawal et al., 2010; Ding et al., 2012; Albenne et al., 2013). However, this has not prevented the research community from using tools such as SecretomeP (Bendtsen et al., 2004, 2005) as a filter in proteomics experiments attempting to characterize the plant secretome. As the oldest and most highly cited tool, it is perhaps unsurprising that the plant proteomics community has adopted it when using a filtering strategy. It has been applied to studies in various plant species, for example *Arabidopsis thaliana* (e.g., Jamet et al., 2008; Ge et al., 2011), *Oryza sativa* (rice) (e.g., Song et al., 2011; Wang et al., 2012), *Helianthus annuus* (sunflower) (e.g., Pinedo et al., 2012) and the moss *Physcomitrella patens* (e.g., Lehtonen et al., 2014).

SecretomeP was trained on mammalian and bacterial proteins and the lack of plant training data is of concern when applying it to plants, as was noted by Agrawal et al. (2010) when they recommend the inclusion of SecretomeP in a plant proteomic analysis workflow. Albenne et al. (2013) clearly point out the deficiencies of this approach given the program was not designed for plants, and although some studies include caveats on its suitability, the hypothesis behind SecretomeP is often overlooked. This hypothesis assumes that secreted proteins will share common properties, regardless of their mechanism of secretion. This approach was taken to overcome the lack of known mammalian LSPs available for use as a positive training dataset, and enabled the authors to train their method using CSPs. This also allowed for the few known LSPs to be used after training to test the accuracy of the tool. The authors of SecretomeP

TABLE 1 | Details of non-classically secreted protein (CSP) prediction programs, as of April 2016.

Prediction tool	Reference	Target ¹	Method ²	Number of predictions ³	Available ⁴	Citations ⁵
SecretomeP (v1)	Bendtsen et al., 2004	M	W, D	500 (W), – (D)	Yes	642
SecretomeP (v2)	Bendtsen et al., 2005	B	W	100	Yes	342
SecretP (v1)	Yu et al., 2010b	M	W	1	Error	17
SecretP (v2)	Yu et al., 2010a	M, B	W	1	Error	76
SecretP (v2.1)	Yu et al., 2013	G–	W	1	Error	3
SPRED	Kandaswamy et al., 2010	M	D	–	Yes	15
SRTpred	Garg and Raghava, 2008	M	W	1	Yes	35
Sec-GO	Huang, 2012	M, B	W	?	No	7
NClassG+	Restrepo-Montoya et al., 2011	G+	W	?	No	6

¹Target: (M)ammalian, (B)acteria, (G+) Gram-positive bacteria, (G–) Gram-negative bacteria. ²Method: (W)ebserver, (D)ownload. ³Number of sequences that can be predicted in a single run, – indicating no limit. ⁴Available refers to web server or program availability at the published location. Error indicates where the program is available but not operational. ⁵Citation counts from Google Scholar (<http://scholar.google.com>) as of April 2016.

propose that common properties would be captured by removing the SP from the sequences of CSPs to generate training sequences. Proteins sharing these common features, but lacking a SP are then predicted to be LSPs.

Using the mammalian version of SecretomeP as a tool for LSPs in plants assumes that any common features are also shared between mammalian and plant-secreted proteins. The software programs used to capture protein features in the mammalian versions of SecretomeP are listed in **Table 2**. It also implicitly assumes that the reported threshold and accuracy metrics of the mammalian version of SecretomeP will apply to plants. The threshold value used to generate positive predictions is of particular importance: SecretomeP outputs scores in a range from 0 to 1 to indicate increasing confidence that a protein is secreted. The trade-off between true and false positives for any given threshold is essential in applying the tool to experimental output. The authors of SecretomeP recommended using a threshold of 0.6 when using the method on mammalian proteins, giving a true-positive rate (TPR) of 0.40 and false-positive rate (FPR) of

TABLE 2 | Protein features used in SecretomeP and the programs it utilizes internally to calculate them.

Protein feature	Dependency	Reference
Number of atoms	—	—
Number of positively charged residues	—	—
Low-complexity regions	SEG	Wootton and Federhen, 1996
Sub-cellular localisation	PSORT II	Nakai and Horton, 1999
Transmembrane helices	TMHMM 2.0c	Krogh et al., 2001
Pro-peptide prediction	ProP 1.0c	Duckert et al., 2004

— indicates feature calculated internally without any external dependency.

0.05 (**Figure 1**). This score was derived from cross-validated subsets of the modified CSP training data. When applied to the 13 human LSPs at the time, 10 of these were observed to be predicted at this threshold (Bendtsen et al., 2004).

The equivalent TPR and FPR when applied to plants has not previously been studied, although some accuracy values have

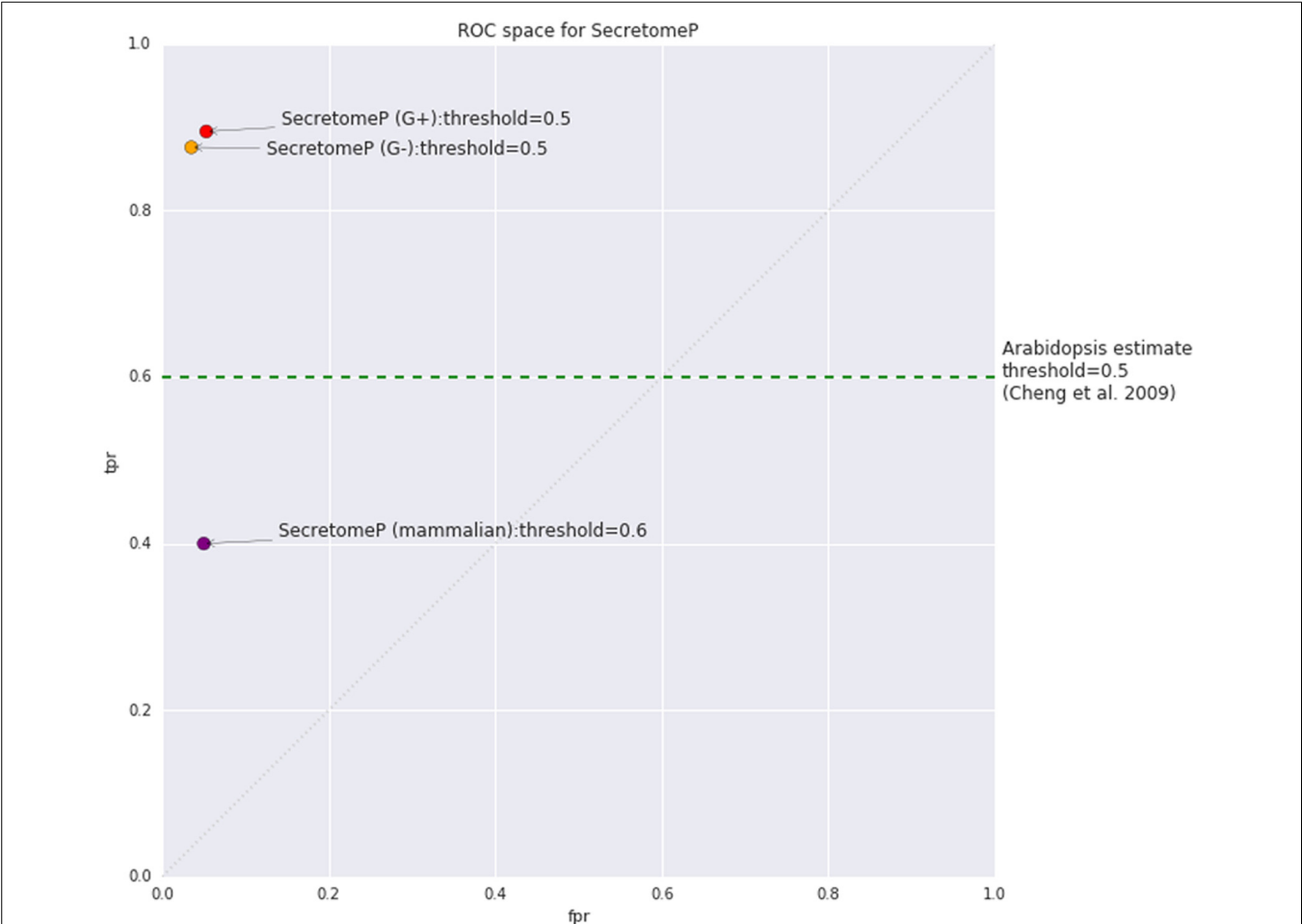


FIGURE 1 | Estimated true positive rate (TPR) and false positive rate (FPR) for SecretomeP as published for mammalian proteins (v1) based on internal cross-validation and bacterial proteins (v2) based on performance on classically secreted proteins (CSPs) from the SignalP 3.0 dataset. The estimated TPR of SecretomeP on plants as stated by Cheng et al. (2009) was 0.6, with no FPR given. The random line diagonal represents equal TPR and FPR, equivalent to random selection of classes.

been reported based on observation from proteomic studies. In an investigation of plant defense responses simulated by the application of salicylic acid, Cheng et al. (2009) stated that 60% of *Arabidopsis* LSPs were predicted to be secreted using a threshold of 0.5, based on the assumption that all the leaderless proteins they found were genuine. This value has also been reported in a more recent review (Agrawal et al., 2010) and both papers include the caveat that the program outputs should be used as indicative only due to the inconsistencies they found with predictions on SP-containing proteins. Even assuming no contaminants, the value only indicates how many true positives are correctly predicted. A TPR of 0.6 seems quite good, so its appearance in the literature could contribute to the continued use of this non-plant based tool on plant proteins. Interestingly, accepting at face value 0.6 as a TPR for SecretomeP on plant proteins this would indicate that the algorithm performs better on plant proteins than mammalian proteins (Figure 1). Whether this score should be extrapolated to be a measure of predictive performance on *Arabidopsis* under other conditions or on other plant species, however, is questionable. Even if all the proteins found are assumed to be true positives and extrapolated to other plant species and/or conditions, then it is only an estimate of the TPR and the FPR is unknown.

Establishing an optimal threshold and determining the performance metrics at that threshold is essential if the results of SecretomeP on plant proteins are to be useful. Ideally, evaluation would require positive (LSPs) and negative (non-secretory proteins) data to investigate the TPR and FPR for a given prediction threshold. However, since the issue is whether SecretomeP can accurately predict if a given plant protein is a LSP, the evidence that such a protein is unconventionally secreted needs to be obtained by means other than the output of the program under evaluation. There are a number of experimental approaches to determine the sub-cellular location of a protein, some relying on cell disruption and others on *in situ* microscopic approaches, assuming the relevant probes (e.g., antibodies) are available. The need for molecular approaches to validate putative LSPs is acknowledged in the literature (Albenne et al., 2013; Robinson et al., 2015; van de Meene et al., 2016) yet practically, in order to perform such a step, candidate proteins from a secretome will often need to be identified first. Since there are very few known plant LSPs and only one protein with experimental (biochemical) validation (Pinedo et al., 2012), a 'gold standard' to evaluate against does not currently exist. Although there are putative lists of LSPs [e.g., Ding et al. (2012)] the uncertainty surrounding their prediction means that their use as a training set would be rendered invalid.

Accepting the validity of the common-feature hypothesis used in SecretomeP provides a proxy positive data set, namely CSPs of the plant cell wall and apoplast. If this hypothesis is correct, we can estimate SecretomeP's accuracy on plant LSPs by evaluating its performance on these known secreted proteins. This study investigates the trade-offs of using such an indirect tool on plant proteins and provides insights into predictive features for future development of plant-specific common-feature tools. If SecretomeP is to be used on plant secreted proteins, then its use

needs to occur with knowledge of the true and false positive rates involved when the method is applied to plants.

MATERIALS AND METHODS

Data Sources

Arabidopsis proteins were the focus of this study due to the availability of large databases of proteins with curated sub-cellular location. Protein sequences were obtained from WallProtDB (San-Clemente and Jamet, 2015) and the *Arabidopsis thaliana* sub-proteome reference (ASURE) (Hooper et al., 2014) databases. After excluding proteins less than 40 amino acids in length due to the SecretomeP cutoff, a total of 1983 WallProtDB and 975 ASURE proteins were included in analyses. WallProtDB proteins were used as positive CSP data. Various sub-sets of WallProtDB were also used as positive data to ensure conclusions from WallProtDB are not biased by its protein composition. Five sub-sets were made by restricting to maximum 30% sequence identity (278), *Arabidopsis* only (522), excluding *Arabidopsis* (1461), rice only (208) and *Brachypodium distachyon* only (358).

The ASURE database contains a smaller set of proteins representing each sub-cellular compartment in the broader SUBA database (Tanz et al., 2013) and as such is a mixture of proteins that are examples of positive, negative, and neutral for this study. The sub-set of CSPs labeled as Extracellular (42) was used as positive and sub-sets of non-secretory proteins located to the nucleus and/or cytosol (352) used as a negative. Sub-sets of other subcellular locations were also used as comparisons such as non-secretory proteins of the plastid (103) (Supplementary Table S1).

Evaluation Data and Protein Sequence Modifications

The use of CSPs as a proxy for SecretomeP prediction of LSPs could be influenced by the components of CSPs rather than by any shared protein features. Removing the SP, in a similar way to the training of SecretomeP, removes this possible bias yet introduces a departure from actual protein data that could also influence the results. To ensure that the evaluation is based on common features rather than a feature unique to CSP such as the SP, the proxy data were modified in various ways to investigate the effect of SP changes on SecretomeP prediction scores.

Under the common-feature hypothesis, modifications to only the SP of CSPs should not influence the prediction score of SecretomeP. Five modifications were made under this assumption (Figure 2): (1) The SP was removed (SP Remove) but because this results in a shorter sequence, sequences were also modified to alter the SP without either changing the length of the original sequence or its amino acid composition. These modifications were (2) removing the SP from the N-terminus and placing it at the C-terminus (SP C-term) and (3) random shuffling the amino acids of the SP at the N-terminus (SP Random). Further modifications that (4) reversed (Reverse) or (5) shuffled (Random) the entire sequence were made to compare the effect on the prediction score when the presumed common features of plant CSPs and LSPs were purposely disrupted. Both these latter

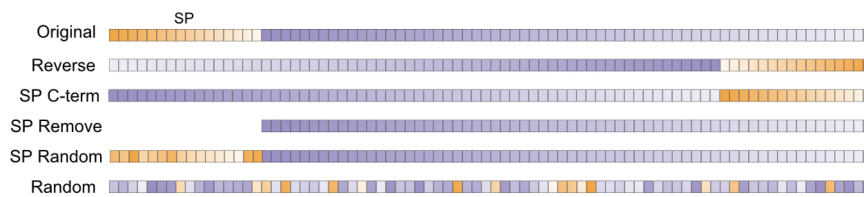


FIGURE 2 | After the SP region (orange) was identified using SignalP 4.1 for positive data or a fixed length of 30 amino acids for negative, five modifications were made to each original sequence in a dataset. Reverse inverted the amino acid order of the entire sequence; SP C-term placed the SP at the C-terminus, leaving the sequence length unchanged; SP Remove excluded the SP region and shortened the sequence length. SP Random and Random involved random shuffling of the SP region and the entire coding sequence, respectively. The sequences were then submitted to SecretomeP. Orange, SP region; Purple, mature protein sequence.

modifications were expected to negatively impact SecretomeP prediction scores.

Details on how these modifications were performed via Python scripts is available online (see Prediction and modification scripts). Briefly, the amino acids for each sequence were treated as strings of text and for the reverse dataset the amino acid sequence was simply reversed. For others, SignalP 4.1 (Petersen et al., 2011) was used to predict the SP of CSPs. The predicted SP was then either automatically removed or moved to the C-terminus to generate the SP Remove and SP C-term datasets, respectively. Random shuffling (with 500 bootstrap replicates) of either the amino acids of the SP (SP Random) or the entire sequence (Random) was done using the Python random library on the respective section of the sequence and a mean prediction score from all bootstraps calculated. When comparisons to negative data were required, i.e., non-secreted proteins where no SP was predicted, the first 30 amino acids of each sequence were modified in the same manner as the SP of CSPs. For each sequence in a test set, this resulted in 1003 additional sequences.

Prediction and Modification Scripts

The downloadable version of SecretomeP v1 was used for all predictions. The large number of permutations required for the shuffling of sequences meant that parallel processing of the prediction results was required. The downloadable version was therefore executed on multiple virtual machines using Docker containers¹ on the NeCTAR cloud service². The features used by SecretomeP rely on several other programs, hence SEG (Wootton and Federhen, 1996), PSORT II (Nakai and Horton, 1999), TMHMM 2.0c (Krogh et al., 2001) and ProP 1.0c (Duckert et al., 2004) was also installed on the same machines, as well as the optional dependency of the older version 3 of SignalP.

For the generation of modified sequence data, sequences from the data sources were modified using the steps described above (see Evaluation Data and Protein Sequence Modifications) on virtual machines with SignalP version 4 installed. The modified sequences were then split into smaller subsets for processing, and SecretomeP containers executed on these in parallel. Each set of results was collated together for analysis.

¹<http://docker.com>

²<http://nectar.org.au>

The files describing the steps for the creation of these containers, as well as the scripts to modify sequences, are available at: https://github.com/lonsbio/lsp_modification_analysis.

Accuracy Analysis

To quantify if changes in SecretomeP output scores were significant, a Student's *T*-test between the original scores and each protein modification was calculated. Since the modified sequences are dependent on the original sequences, a paired *T*-test was used. Due to these multiple tests, an adjusted *p*-value of ≤ 0.01 was used as the significance threshold to reject the null hypothesis that a modification did not influence the SecretomeP results.

Receiver operating characteristic (ROC) curves were plotted to determine TPR and FPR for SecretomeP on plant proteins. ROC curves are useful to evaluate and visualize classifiers (Fawcett, 2006). Given a binary classifier and both positive and negative data, ROC curves are built by gradually reducing the threshold for classification and plotting each point on ROC space (as in Figure 1). ROC curves were plotted for pairs of positive and negative datasets (as defined in Supplementary Table S1) with the 'scikit-learn' package in Python. Each ROC plot represents one data set with multiple curves, with each curve representing either the original scores or one of the protein modifications. Given a random selection of a positive and negative protein from their respective datasets, the area under-the-curve (AUC) is the probability the positive protein will be classified higher than the negative protein (Fawcett, 2006).

RESULTS

Mean SecretomeP Prediction Scores for Plant Classically Secreted Proteins (CSPs) Are Higher Than for Non-secreted Proteins

Various sets of proteins obtained from either WallProtDB or ASURE databases were subjected to SecretomeP prediction and the mean prediction score calculated (Supplementary Table S1). The 1983 unmodified WallProtDB proteins (Original) that were analyzed had an average prediction score of 0.707. Subsets of WallProtDB proteins, selected by protein redundancy or species,

had prediction scores ranging from 0.644 to 0.746, indicating the average scores exceeded previously used thresholds of 0.5 and 0.6. For the subsets of ASURE-derived proteins, scores were lower in every case, including the sub-cellular locations with no secretory pathway involvement such as the nucleus and cytosol (0.494 and 0.446 individually, 0.479 combined). The highest score of 0.582 was obtained for the Extracellular sub-set that contains only secreted proteins. Since these scores are averaged over uneven datasets they are not conclusive, however, they do indicate that the average scores for secreted proteins are higher than for non-secreted proteins and hence that SecretomeP appears to distinguish between plant secreted and non-secreted proteins.

SecretomeP Prediction Scores Are Influenced by the Presence of a Signal Peptide

Additional protein datasets were created through modifications to each original protein to explore the possible influence of the SP, a protein feature unique to CSPs (see Materials and Methods, Evaluation Data and Protein Sequence Modifications), and SecretomeP was applied to each modified protein dataset. It was expected that modifications to the SP only (SP Remove, SP C-term and SP Random) should not impact SecretomeP output scores whereas modifications to the entire sequence (Reverse/Random) should as protein features common to both CSPs and LSPs would be disrupted. For the full WallProtDB dataset, the mean SecretomeP output scores were lower for each modification from the Original (0.707) to: 0.522 (Reverse), 0.568 (SP Remove), 0.541 (SP C-term), 0.603 (Random), and 0.648 (SP Random). For significance, rather than compare the difference between mean scores of the entire set, the changes in scores were tested via a paired Student's *T*-test ($p \leq 0.01$) between original scores and the modifications (Supplementary Table S1) due to the dependence between the original and modified protein. Every modification to WallProtDB was significant, indicating that the modifications reduced the confidence with which the method assigned "secreted" status.

In contrast, the equivalent modifications to the non-secreted proteins using the nucleus and/or cytosol dataset as the exemplar, produced no substantial change in the SecretomeP scores with means of 0.479 (Original), 0.483 (Reverse), 0.486 (SP Remove), 0.460 (SP C-term), 0.545 (Random), and 0.482 (SP Random) (Supplementary Table S1). None of the changes were significant, with the exception of some randomized sequences, which also resulted in higher average prediction scores for these non-secreted proteins. Modifications affecting the N-terminal region (where a SP is located in CSPs) significantly changed the prediction score for CSPs only, suggesting that a portion of the high predictive score for these proteins is attributable to properties of the SP, rather than any general common-feature in the mature protein.

Since the mean score of predictions is a summary statistic and susceptible to outliers, the distribution of scores was plotted to visualize the effect of protein modifications. A kernel density

estimation (KDE) plot was used for all original proteins in each dataset, with the distribution of each modification overlaid. The curve represents the smoothed Gaussian distributions of scores. The rug plot marks along the x-axis indicating the original scores for each protein. As suggested by the significant lowering of the SecretomeP output scores, a change in the score distribution was seen in KDE plots for each protein modification compared to the original WallProtDB dataset (Figure 3A). Protein modifications that were both expected and not expected to alter scores did so. Sub-sets of WallProtDB, based on either plant species or maximum sequence identity threshold, exhibit the same shifts showing that this observation is not due to the composition of WallProtDB (Supplementary Figure S1).

The KDE plot of the modified WallProtDB dataset is distinct from the original sequences. The SP Remove, Reverse and SP C-term modifications all appear to shift the distribution toward the left, i.e., a higher density of lower scores (Figure 3A). Random shuffling of the SP region has the least change compared to other modifications. Given that it is the least disruptive modification to the sequence (altering the order of amino acids in only a small region of the protein that over 500 bootstraps could often resemble the original sequence), and yet it still results in a significant drop in scores, is further evidence that high prediction scores of SecretomeP are reliant on the SP. The density plot of the completely shuffled sequences (Random) were narrower with a high density of scores between 0.5 and 0.6 in all datasets. As a modification expected to alter the prediction score, this illustrates that disruption of the entire sequence does result in a change in scores, although as each score is an average of 500 bootstraps the exact range of these scores may not be informative. A more detailed investigation of the effect of SP sequence randomization and full sequence randomization on scores shows that the average value from bootstraps can hide some interesting variations (Supplementary Figure S2) but ultimately support the conclusion that the SP region can have a strong effect on SecretomeP output scores. For the ASURE dataset, the KDE plots of the secretory extracellular protein sub-set (Figure 4A) exhibit a similar pattern with a shift to lower scores. The mean scores did not quite reach a *p*-value to reject the common-feature hypothesis (Supplementary Table S1), likely due to the smaller sample size, however, these data suggest the same reliance of output scores on the SP as seen in the full WallProtDB dataset.

This effect of a shift when the SP region is altered is not seen for the non-secretory sub-sets of ASURE (nucleus, cytosol, nucleus and/or cytosol, plastid) proteins (Supplementary Table S1). Apart from the Random modification, each modification has a similar distribution to the original scores (Figures 4C–E) even though the equivalent to the SP region of the sequence was altered in the same way. The ASURE sub-sets defined as neutral also showed the same change in Random modification scores (Figures 4B,F). This contrast between the positive and negative datasets, and together with the significance tests and changes in mean scores demonstrate the influence of the SP on SecretomeP output scores and diminishes the likelihood that LSPs lacking this feature will be accurately predicted as "secreted."

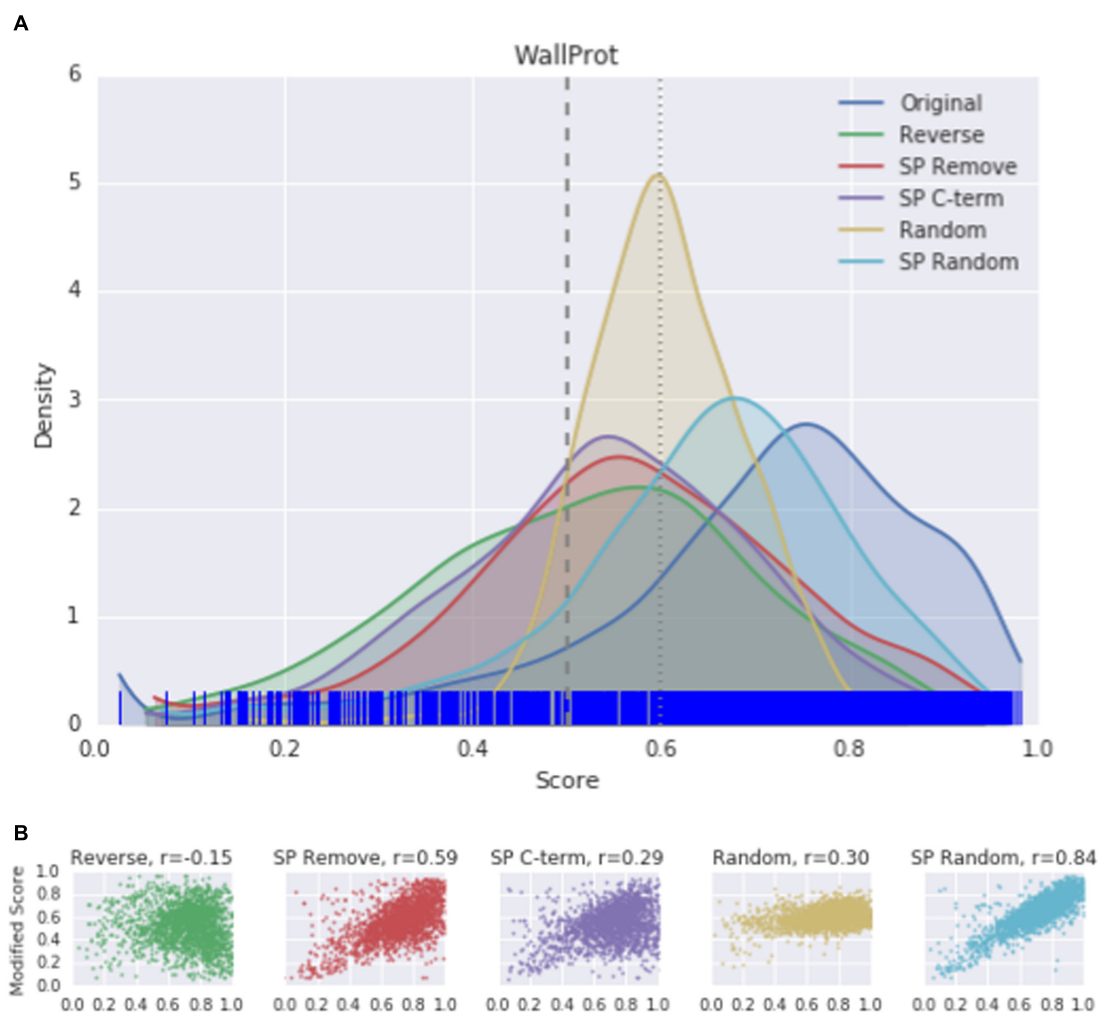


FIGURE 3 | (A) The kernel density estimation (KDE) distributions of SecretomeP scores (v1.0) for original and modified sequences of the entire WallProtDB dataset. (Individual KDE distribution plots for sub-sets of WallProtDB are provided in Supplementary Figure S1). Rug plot marks along the x-axis indicate position of original scores. Vertical dashed lines at 0.5 and 0.6 represent the most common cutoff scores used in plant studies. **(B)** Correlation plots between original scores (x-axis) and modified scores (y-axis). Spearman correlation values are indicated for each modification.

Correlation between Modified and Original Sequences Differs between Secreted and Non-Secreted Proteins

To explore whether high-scoring proteins remain high-scoring after sequence modification, the Spearman correlation coefficient was calculated between the SecretomeP scores obtained from the unmodified data and the scores following sequence modification. The correlation between original score and each modification of the WallProtDB dataset is shown in **Figure 3B**. Shuffling the SP (SP Random) was least divergent at $\rho = 0.84$. The most divergent change was reversing the sequence, which showed a weak negative correlation. The deviation from the diagonal shows that some modifications largely dropped scores (bottom-right of correlation plot), although some improved the prediction score from low to high (top-left of correlation plot). Of note is the relatively high correlation for the SP Remove modification.

The sub-sets of WallProtDB overall show consistent results with similar levels of correlation indicating conclusions drawn from these correlations holds across the plant species included in WallProtDB (Supplementary Table S2).

Correlations for the ASURE subsets are also informative (**Figure 5**). The relative positive correlation value for each modification is maintained from the WallProtDB analysis, though mostly higher values were obtained (compare **Figure 5** with **Figure 3B**, Supplementary Table S2). The exception was the Reverse modifications, which showed a small positive correlation, particularly for nucleus, cytosol and the combined sets, and a negative correlation for the WallProtDB and ASURE (Extracellular) sub-sets. This modification does not rely on a substitute SP region, and the difference in scores between positive and negative data suggests that the weak negative correlation found in WallProtDB and ASURE (Extracellular) is a feature of SP-containing proteins.

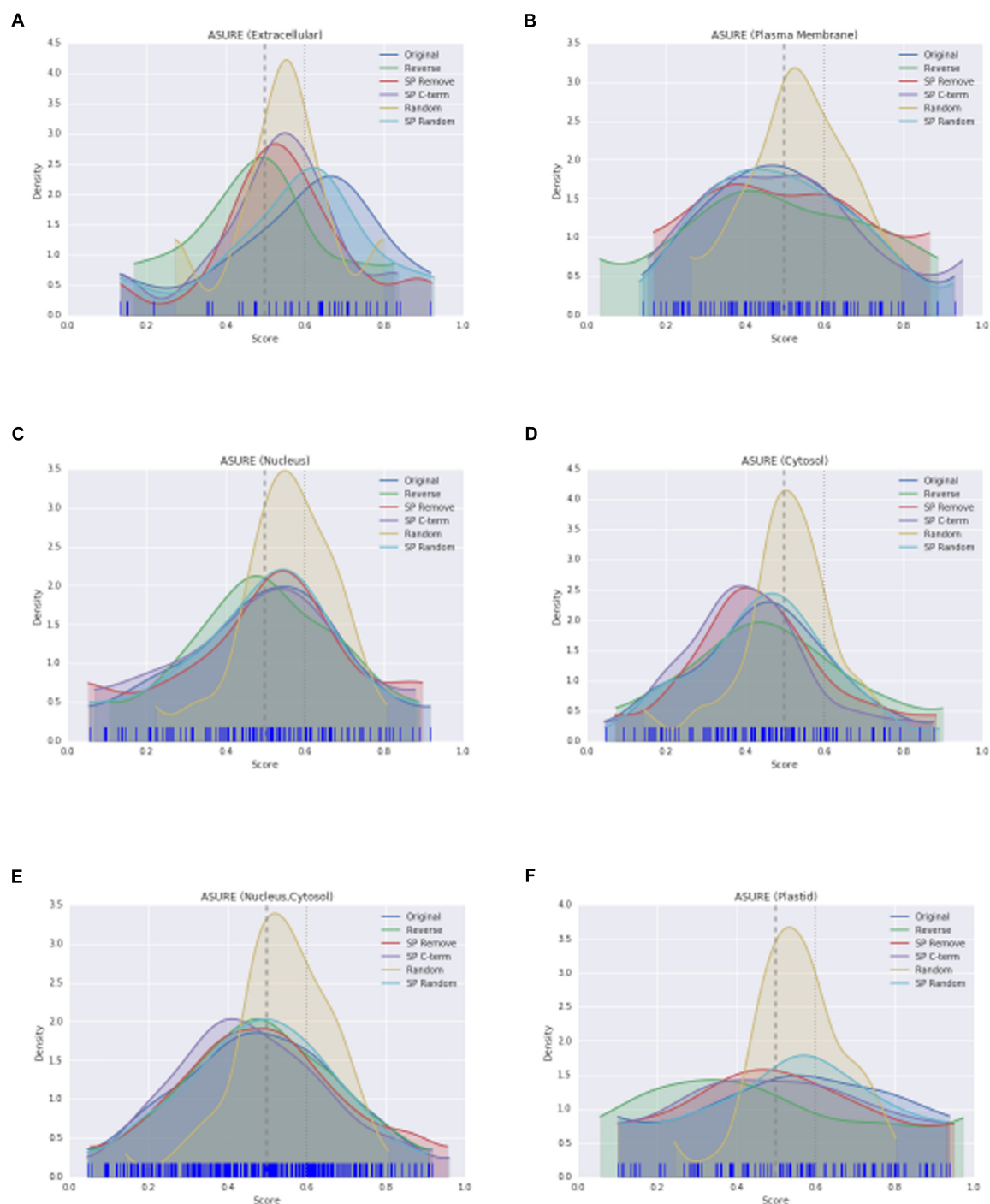
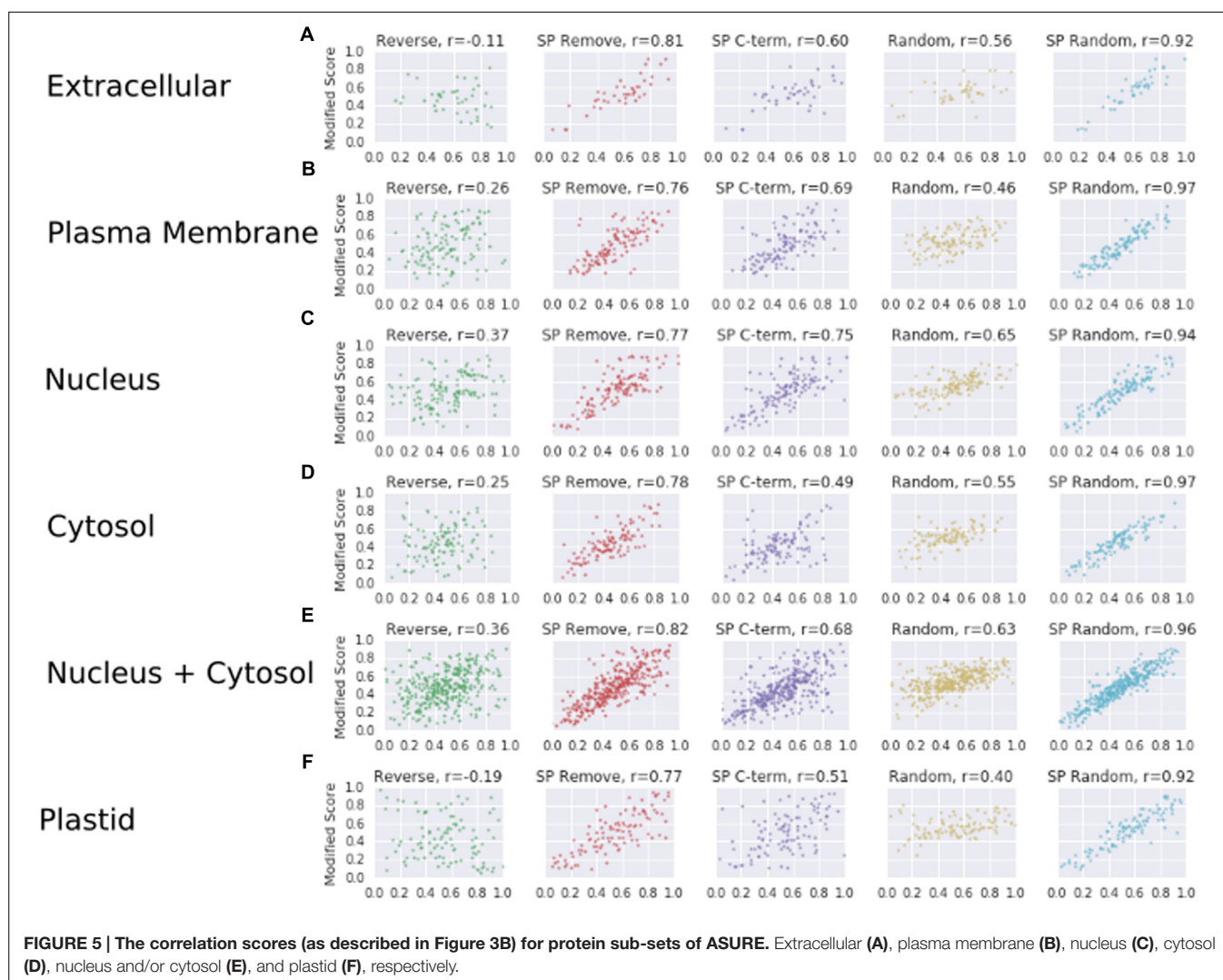


FIGURE 4 | The KDE distributions of SecretomeP scores (as described in Figure 3A) for protein sub-sets of ASURE. Extracellular (A), plasma membrane (B), nucleus (C), cytosol (D), nucleus and/or cytosol (E), and plastid (F), respectively.

Taking into account the combination of different SecretomeP output score means, shifted distributions and correlation patterns across WallProtDB and ASURE sub-cellular organelle protein sub-sets, it is reasonable to conclude that: (1) the scores of plant CSPs from SecretomeP are influenced by the presence of a SP;

(2) unmodified SP sequences are a poor proxy for LSPs; and (3) the SP Remove modification has the highest correlation with original scores (except for those involving any random shuffling) across both positive and negative data and largely produces a lower score than the original sequence, thereby making it the



most suitable proxy to use when comparing the performance of SecretomeP.

SecretomeP Performs Marginally Better Than Chance on Plant Proteins

WallProtDB contains many similar sequences. Therefore, to be able to compare against the *Arabidopsis* negative data in an unbiased manner, we used the *Arabidopsis* WallProtDB proteins and the ASURE (nucleus and/or cytosol) sub-set to create ROC curves comparing modified positive to modified negative datasets (Figure 6). The SP Remove modification is used to infer the TPR and FPR scores, although the results from other modifications are also shown.

The path of the SP Remove ROC curve shows the TPR and FPR are similar and near to the rate equivalent of “luck/chance.” The stark difference between the ROC curve based on the original sequences (dark blue) and protein sequences where the SP has been removed (SP Remove; red) confirms that the ability of SecretomeP to distinguish between secretory and non-secretory plant proteins is largely influenced by the presence of the SP in

CSPs. Therefore, the curve for the SP Remove sequences is the more accurate for evaluating the use of SecretomeP to predict LSPs.

Three points of interest on the SP Remove curve are annotated with dotted lines in Figure 6. Each corresponds with the nearest discrete points on the ROC curve corresponding to where values of either the FPR or cut-off threshold are fixed, the exact values of which are shown for all modifications in Supplementary Table S3. Firstly, the threshold and subsequent TPR are identified when the FPR is set to be ≤ 0.05 (Supplementary Table S3A). For SP Remove this results in a threshold of 0.860 which leads to a TPR of 0.029 for a FPR of 0.048, indicating more false predictions than true.

Secondly, we note the TPR and FPR for thresholds around 0.5–0.6, which are the most commonly used thresholds for plant proteins (Supplementary Table S3B). These two points have higher TPR at the cost of higher FPR: the threshold near 0.6 would improve the TPR to 0.341, however, at the cost of 0.264 FPR. Further reduction to a 0.5 cut-off results in a TPR of 0.598 with a 0.469 FPR (Figure 6).

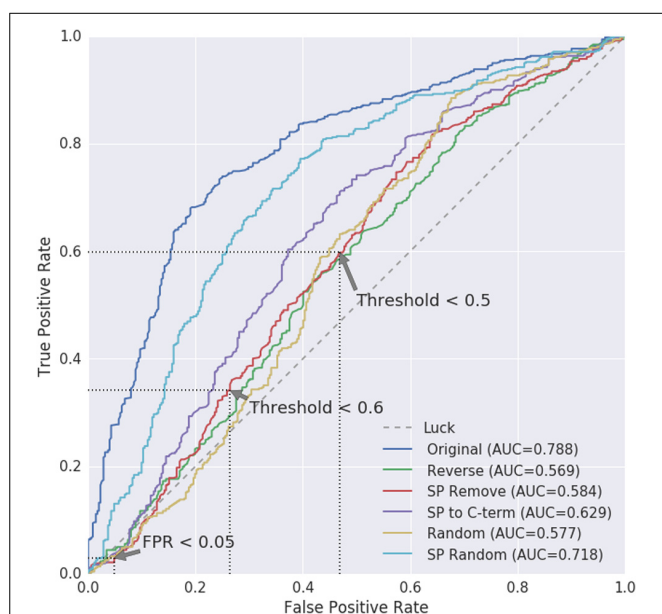


FIGURE 6 | Receiver operating characteristic (ROC) curves for positive WallProtDB *Arabidopsis* proteins vs. the negative class of ASURE proteins from the nucleus and/or cytosol. Area Under Curve (AUC) values are shown for each protein modification. The AUC for random (pure chance/luck) is 0.5. As the threshold for classification is reduced, the true positive (TPR) and false positive (FPR) rates are mapped to the axes, giving visual insight into how the classifier balances true and false positives and which thresholds might be considered most appropriate. The preferred path of an ROC is toward the upper left hand corner signifying a high true positive rate and low false positive rate. Three points on the SP Remove curve are annotated corresponding to when either the FPR is <0.05 or threshold is set to 0.5 or 0.6. The exact values are listed in Supplementary Table S3.

If the maximum area under the curve (AUC) of 0.788 and upper-left curve trajectory of the Original scores represented a realistic evaluation, then performance on plant proteins would be fair. However, the shift of the curve back toward the diagonal as soon as the SP is removed shows how dependent this predictive power is on the SP, as the AUC decreases to 0.584, and performance is poor at higher thresholds. There is no reasonable trade-off between error rates; regardless of where the threshold is drawn, the true positive and false positive rates are roughly equivalent, indicating prediction is effectively equal to random chance.

DISCUSSION

Common-Features between CSPs and LSPs in Plants Are Not Captured in SecretomeP

It is difficult to assess the accuracy of a prediction tool without both positive and negative examples of the prediction target of interest. In order to evaluate how well an *in silico* method predicts LSPs, the best kind of positive data would be experimentally confirmed LSPs, but unfortunately there are very few known plant LSPs and only one protein with biochemical validation

(Pinedo et al., 2012). If we accept the common feature hypothesis that underlies SecretomeP prediction, then CSPs are the next best option. For LSPs in plant studies, SecretomeP has previously been estimated to have a 0.6 TPR by assuming all proteins found in a study are positive examples, without considering negative examples. Therefore in this current study, we attempted to overcome this difficulty of assessment by virtue of the common-feature hypothesis behind SecretomeP and the expectation that performance for CSPs is informative if the hypothesis is true. Since modifications intended to disturb the SP of a sequence have a significant effect on the prediction output of positive data and not on negative data (Supplementary Table S1), we concluded that modified sequences were required to evaluate predictions based on these common features. This evaluation revealed that performance, when the FPR is limited to 0.05, resulted in true positive predictions that were not better than chance alone. Previous suggestions that SecretomeP is not well suited to plants (Agrawal et al., 2010; Ding et al., 2012; Albenne et al., 2013) were confirmed, and the estimated accuracy is so low that its use should be avoided for plant proteins. Furthermore, previous use of SecretomeP to predict putative LSPs on plant proteins must be re-evaluated in light of our findings.

Effects of Hydrophobic Regions

WallProtDB includes some proteins from the plasma membrane involved in cell wall metabolism, such as cellulose synthase (CESA). Under the common-feature hypothesis, it is reasonable to include these in the positive dataset given the adjacency to the extracellular space, since plasma membrane-associated proteins should have some features suitable to or related to the extracellular space and would be expected to traffic through the ER/Golgi compartments. We do not consider their inclusion in WallProtDB to alter the conclusions, since the ROC results for the balanced ASURE tests are consistent with poor performance on other data.

Transmembrane domains were also excluded from the mammalian training set of SecretomeP, although the output of the TMHMM predictor was included as a candidate feature and found to be amongst the most predictive when the features were selected. Because of the 'black box' nature of SecretomeP's internal workings, the exact interpretation of TMHMM outputs in determining mammalian LSPs is unknown. As noted when published, TMHMM misclassifies about 20% of SPs as helices in eukaryotes (Krogh et al., 2001). The original training set featured no SPs or transmembrane helices and so from the influence of SPs on plant CSP scores, we suggest that the presence of an amino acid sequence with similar hydrophobic properties is taken as a positive predictor of secretion. This reliance on TMHMM to identify such regions could be one of the causes for the difference in prediction scores between CSPs and truncated versions without the SP region. The inclusion of other general tools, such as PSORT II may also detect these features of CSPs and boost the prediction scores.

Choice of Threshold

Our conclusions are drawn directly from the available data in WallProtDB and ASURE, and the selection of which sub-sets

of both datasets to use for ROC analysis. As such the choice of data determines the precise TPR, FPR and threshold results. Modifications were made to both positive and negative data to create ROC curves (Figure 6), though comparisons between modified positive data to unmodified negative data produce similar results (Supplementary Figure S3). The *Arabidopsis* sub-set of WallProtDB and ASURE (nucleus and/or cytosol) were selected due to their size and clear exclusion from the secretory pathway and extracellular destination. Broadly, however, the results hold even if a more balanced yet smaller set, such as using ASURE for both positive and negative data is selected or additional sub-cellular locations are included as negative data (Supplementary Figures S4 and S5). The trend of the ROC curve of partial sequences toward random continues, and although the AUC is higher, there are only 42 positive proteins in the ASURE (Extracellular) sub-set as comparison, and thus it lacks power. Although the results are specific to *Arabidopsis*, the use of a general mammalian tool on plant proteins pre-supposes that the features are shared between all mammals and plants. Poor performance on *Arabidopsis* does not rule out better performance on other plant proteins, but it would require that the conserved features are poor on *Arabidopsis* only which is unlikely given the high degree of conservation of proteins across species.

The previous thresholds used for SecretomeP in the plant literature vary. Albenne et al. (2013) indicate a small number of *Arabidopsis* proteins exceed the threshold, but don't equivocally state what threshold they used. Jamet et al. (2008) use 0.6, yet Cheng et al. (2009) use 0.5 in their finding of 0.6 TPR based on 37 LSPs they assumed were genuine. This TPR drops to 0.3 when using a more stringent 0.6 threshold (Supplementary Table S4). The output of the web server version for SecretomeP³ previously indicated 0.5 should be used as a cutoff (accessed August 2015) which matches the more recent bacterial version. Currently (accessed April 2016) the server indicates both 0.5 and 0.6 thresholds for bacterial and mammalian proteins, respectively, in line with the published results. This change could contribute to the multiple values used, however, since SecretomeP is not explicitly listed as a tool for plant proteins, which threshold score to use to match the expected sensitivity and specificity it provides when applied to plant proteins is lacking. The pervasiveness of a precise value in the literature, however, indicates that it may be useful to use the thresholds in Supplementary Table S3 to recommend a different value with a known FPR, which could be preferable to the status quo.

The original cross-validation estimate of SecretomeP sensitivity (TPR) was 0.4 when FPR is 0.05 and it was acknowledged that this would not classify all LSP proteins without using a much lower threshold, thereby incurring a high cost of false positives (Bendtsen et al., 2004). From Supplementary Table S3, the upper bound of unmodified sequences would indicate that the TPR around 0.276 occurs for a threshold of 0.794, and that estimates of the effect of the SP will only reduce this. Given our results showing the reliance on the SP when using CSPs as proxy, and based on the performance of SecretomeP on modified CSPs with the SP Removed, controlling

the FPR to match the mammalian version at below 0.05 requires the threshold to be 0.86 and the subsequent TPR is 0.029 (Supplementary Table S3A). These values represent the values at which we estimate SecretomeP can predict based on conserved features with some degree of certainty and therefore the use of SecretomeP on plants does not appear justified. Although it would represent an improvement over current usage and match the FPR of the original program, excluding proteins on this basis would mean losing the vast majority of true positives from any proteome; the expectation would be that more false-positives would still be identified than true secreted proteins.

Upper and Lower Limits on Performance

The discrepancy between scores for original and truncated CSPs are evidence against common features of plant proteins being recognized. The theory behind SecretomeP is that the mechanism of secretion should not be captured, and that other shared aspects of proteins that make it suitable for secretion are what predictions are based on.

There remains the possibility, however, that shared features in plants between CSPs and LSPs could also resemble the SP mechanism (e.g., the similar hydrophobic properties between SPs and GPI anchor regions) and that performance on true LSPs may be similar to the original sequence results seen in this study, if such features are highly predictive. Under this scenario, the ROC curve for the original sequence could represent an approximate upper bound on the accuracy, and the SP Remove the lower. This would still not make SecretomeP suitable for plant proteins, as the predictions would not be consistent with its original hypothesis based on common-features between CSPs and LSPs, but rather on similarities between the SP mechanism and LSP specific features. This scenario would also not account for reduced scores for CSPs, as these features should exist in addition to the SP region if they are truly conserved.

This illustrates one difficulty with the mammalian trained common-feature hypothesis being transferred to plants. This work shows that in plants the secretion mechanism is influential to the prediction score, but masks the true performance value, which is likely to be somewhere between the Original and SP Removed sequence curves (Figure 6). Future work using the common-feature hypothesis in plants should take these types of protein regions into account, namely if sequences are modified for training a tool, then during development they must be checked to ensure both modified and unmodified sequences have a similar accuracy so that common-features are captured rather than secretion mechanisms.

New Prediction Tools and Approaches Are Required

SecretomeP was not trained on plant proteins. There is no claim made in either the original publication or in the currently available web service that SecretomeP has any applicability to plants. Although version 1.0 could be considered 'eukaryotic' by inference from the text available on the archived website⁴, the

³<http://www.cbs.dtu.dk/services/SecretomeP/>

⁴<http://www.cbs.dtu.dk/services/SecretomeP-1.0/>

abstract clearly states the method is for mammalian secretory proteins. The current version offers modes of operation for mammals and bacteria only. Applying SecretomeP to plant protein data implies a chain of assumptions that the features shared between mammalian CSPs and LSPs are conserved and also shared with plant CSPs and LSPs. Our results show that these assumptions do not hold, and that observations of the accuracy in the literature are not supported. While it is tempting to reduce a large dataset into something manageable for further analysis the current study has shown that the predictive power of SecretomeP on plant proteins is such that true-positives will be found at a rate close to the false-positive rate, with marginal improvement over random selection of classes.

Instead of SecretomeP being ‘better than nothing,’ we propose that it is actually ‘worse than nothing’ so it is better to do nothing than to use it on plant proteins. Independent experimental lines of evidence (biochemical/cell biological) are required confirm possible LSPs, but in plants, SecretomeP should not be used to identify putative LSPs. Depending on the consequences of either false positives or false negatives in a proteomic study, strategies to accept all proteins found, or reject those without a SP, are advised. A filtering strategy should not be used until a plant-specific tool is developed.

REFERENCES

- Agrawal, G. K., Jwa, N.-S., Lebrun, M.-H., Job, D., and Rakwal, R. (2010). Plant secretome: unlocking secrets of the secreted proteins. *Proteomics* 10, 799–827. doi: 10.1002/pmic.200900514
- Albenne, C., Canut, H., and Jamet, E. (2013). Plant cell wall proteomics: the leadership of *Arabidopsis thaliana*. *Front. Plant Sci.* 4:111. doi: 10.3389/fpls.2013.00111
- Bendtsen, J. D., Jensen, L. J., Blom, N., Von Heijne, G., and Brunak, S. (2004). Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng. Des. Sel.* 17, 349–356. doi: 10.1093/protein/gzh037
- Bendtsen, J. D., Kiemer, L., Fausbøll, A., and Brunak, S. (2005). Non-classical protein secretion in bacteria. *BMC Microbiol.* 5:58. doi: 10.1186/1471-2180-5-58
- Cheng, F.-Y., Blackburn, K., Lin, Y.-M., Goshe, M. B., and Williamson, J. D. (2009). Absolute protein quantification by LC/MS(E) for global analysis of salicylic acid-induced plant protein secretion responses. *J. Proteome Res.* 8, 82–93. doi: 10.1021/pr800649s
- Ding, Y., Robinson, D. G., and Jiang, L. (2014). Unconventional protein secretion (UPS) pathways in plants. *Curr. Opin. Cell Biol.* 29, 107–115. doi: 10.1016/j.cceb.2014.05.008
- Ding, Y., Wang, J., Wang, J., Stierhof, Y.-D., Robinson, D. G., and Jiang, L. (2012). Unconventional protein secretion. *Trends Plant Sci.* 17, 606–615. doi: 10.1016/j.tplants.2012.06.004
- Doblin, M. S., Pettolino, F., and Bacic, A. (2010). Evans review: plant cell walls: the skeleton of the plant world. *Funct. Plant Biol.* 37, 357–381. doi: 10.1071/FP09279
- Duckert, P., Brunak, S. R., and Blom, N. (2004). Prediction of proprotein convertase cleavage sites. *Protein Eng. Des. Sel.* 17, 107–112. doi: 10.1093/protein/gzh013
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognit. Lett.* 27, 861–874. doi: 10.1016/j.patrec.2005.10.010
- Garg, A., and Raghava, G. P. S. (2008). A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search. *In Silico Biol.* 8, 129–140.
- Ge, W., Song, Y., Zhang, C., Zhang, Y., Burlingame, A. L., and Guo, Y. (2011). Proteomic analyses of apoplastic proteins from germinating

AUTHOR CONTRIBUTIONS

AL devised and performed the analyses, in consultation with MoD, MeD, and AB. All Authors contributed to data interpretation and writing of the manuscript.

FUNDING

This work was supported by a grant to the ARC Centre of Excellence in Plant Cell Walls (MoD, AB; CE1101007)). The ARC Centre of Excellence in Plant Cell Walls also acknowledges the support of funds from the Australian Government’s National Collaborative Research Infrastructure Program (NCRIS) administered through Bioplatforms Australia (BPA) Ltd. Computation was performed on the Nectar Research Cloud, a collaborative Australian research platform supported by NCRIS.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fpls.2016.01451>

- Arabidopsis thaliana* pollen. *Biochim. Biophys. Acta* 1814, 1964–1973. doi: 10.1016/j.bbapap.2011.07.013
- Hooper, C. M., Tanz, S. K., Castleden, I. R., Vacher, M. A., Small, I. D., and Millar, A. H. (2014). SUBAcon: a consensus algorithm for unifying the subcellular localization data of the *Arabidopsis* proteome. *Bioinformatics* 30, 3356–3364. doi: 10.1093/bioinformatics/btu550
- Huang, W.-L. (2012). Ranking gene ontology terms for predicting non-classical secretory proteins in eukaryotes and prokaryotes. *J. Theor. Biol.* 312, 105–113. doi: 10.1016/j.jtbi.2012.07.027
- Jamet, E., Albenne, C., Boudart, G., Irshad, M., Canut, H., and Pont-Lezica, R. (2008). Recent advances in plant cell wall proteomics. *Proteomics* 8, 893–908. doi: 10.1002/pmic.200700938
- Kandaswamy, K. K., Pugalethi, G., Hartmann, E., Kalies, K.-U., Möller, S., Suganthan, P. N., et al. (2010). SPRED: a machine learning approach for the identification of classical and non-classical secretory proteins in mammalian genomes. *Biochem. Biophys. Res. Commun.* 391, 1306–1311. doi: 10.1016/j.bbrc.2009.12.019
- Krause, C., Richter, S., Knöll, C., and Jürgens, G. (2013). Plant secretome—From cellular process to biological activity. *Biochim. Biophys. Acta* 1834, 2429–2441. doi: 10.1016/j.bbapap.2013.03.024
- Krogh, A., Larsson, B., Von Heijne, G., and Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580. doi: 10.1006/jmbi.2000.4315
- Lehtonen, M. T., Takikawa, Y., Rönholm, G., Akita, M., Kalkkinen, N., Ahola-Iivari, E., et al. (2014). Protein secretome of moss plants (*Physcomitrella patens*) with emphasis on changes induced by a fungal elicitor. *J. Proteom. Res.* 13, 447–459. doi: 10.1021/pr400827a
- Miernyk, J. A., Jett, A. A., and Johnston, M. L. (2016). Analysis of soybean tissue culture protein dynamics using difference gel electrophoresis. *J. Proteom.* 130, 56–64. doi: 10.1016/j.jprot.2015.08.023
- Nakai, K., and Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* 24, 34–36. doi: 10.1016/S0968-0004(98)01336-X
- Petersen, T. N., Brunak, S., Von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8, 785–786. doi: 10.1038/nmeth.1701

- Pinedo, M., Regente, M., Elizalde, M. Y., Quiroga, I. A., Pagnussat, L., Jorrin-Novo, J., et al. (2012). Extracellular sunflower proteins: evidence on non-classical secretion of a jacalin-related lectin. *Protein Pept. Lett.* 19, 270–276. doi: 10.2174/092986612799363163
- Restrepo-Montoya, D., Pino, C., Niño, L. F., Patarroyo, M. E., and Patarroyo, M. A. (2011). NClassG+: a classifier for non-classically secreted gram-positive bacterial proteins. *BMC Bioinformatics* 12:21. doi: 10.1186/1471-2105-12-21
- Robinson, D. G., Ding, Y., and Jiang, L. (2015). Unconventional protein secretion in plants: a critical assessment. *Protoplasma* 253, 31–43. doi: 10.1007/s00709-015-0887-1
- Rose, J. K. C., and Lee, S. J. (2010). Straying off the highway: trafficking of secreted plant proteins and complexity in the plant cell wall proteome. *Plant Physiol.* 153, 433–436. doi: 10.1104/pp.110.154872
- San-Clemente, H., and Jamet, E. (2015). WallProtDB, a database resource for plant cell wall proteomics. *Plant Methods* 11:2. doi: 10.1186/s13007-015-0045-y
- Song, Y., Zhang, C., Ge, W., Zhang, Y., Burlingame, A. L., and Guo, Y. (2011). Identification of NaCl stress-responsive apoplastic proteins in rice shoot stems by 2D-DIGE. *J. Proteomics* 74, 1045–1067. doi: 10.1016/j.jpro.2011.03.009
- Tanz, S. K., Castleden, I., Hooper, C. M., Vacher, M., Small, I., and Millar, H. A. (2013). SUBA3: a database for integrating experimentation and prediction to define the SUBcellular location of proteins in *Arabidopsis*. *Nucleic Acids Res.* 41, D1185–D1191. doi: 10.1093/nar/gks1151
- van de Meene, A., Bacic, A., and Doblin, M. S. (2016). The plant secretory pathway seen through the lens of the cell wall. *Protoplasma* doi: 10.1007/s00709-016-0952-4 [Epub ahead of print].
- Wang, Y., Wu, J., Lee, D. Y., Kim, Y. J., Kim, Y. C., Park, S.-R., et al. (2012). Comparative secretome analysis of differentially-induced proteins in rice lesion mimic mutant spotted leaf 11 (spl11). *J. Plant Mol. Biol. Omics* 5, 567–575.
- Wootton, J. C., and Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* 266, 554–571. doi: 10.1016/S0076-6879(96)66035-2
- Yu, L., Guo, Y., Li, Y., Li, G., Li, M., Luo, J., et al. (2010a). SecretP: identifying bacterial secreted proteins by fusing new features into Chou's pseudo-amino acid composition. *J. Theor. Biol.* 267, 1–6. doi: 10.1016/j.jtbi.2010.08.001
- Yu, L., Guo, Y., Zhang, Z., Li, Y., Li, M., Li, G., et al. (2010b). SecretP: a new method for predicting mammalian secreted proteins. *Peptides* 31, 574–578. doi: 10.1016/j.peptides.2009.12.026
- Yu, L., Luo, J., Guo, Y., Li, Y., Pu, X., and Li, M. (2013). In silico identification of gram-negative bacterial secreted proteins from primary sequence. *Comput. Biol. Med.* 43, 1177–1181. doi: 10.1016/j.combiomed.2013.06.001

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Lonsdale, Davis, Doblin and Bacic. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

2.3 Supplementary Material to Lonsdale et al. (2016)

Supplementary Material

Better than nothing? Limitations of the prediction tool SecretomeP in the search for leaderless secretory proteins (LSPs) in plants.

Andrew Lonsdale¹, Melissa J Davis², Monika S Doblin¹ and Antony Bacic^{1,*}

* Correspondence: Antony Bacic: abacic@unimelb.edu.au

1 Supplementary Figures and Tables

Supplementary Table S1. Mean SecretomeP output scores for the all positive and negative input data from WallProtDB and ASURE databases. Shaded scores indicate statistically significant changes based on a paired student T-test between modified and unmodified sequences which considers the mean of the differences in scores is due to the dependence between original and modified protein (* $p \leq 0.01$, ** $p \leq 0.001$ and *** $p \leq 0.0001$).

Data	# ¹	Data ²	Original	Reverse	SP Remove	SP c-term	Random	SP Random
WallProt	1983	Positive	0.707	0.522***	0.568***	0.541***	0.603***	0.648***
WallProt (30% maximum sequence identity)	278	Positive	0.644	0.499***	0.542***	0.517***	0.580***	0.598***
WallProt (<i>Arabidopsis</i>)	522	Positive	0.683	0.525***	0.536***	0.534***	0.573***	0.614***
WallProt (excl. <i>Arabidopsis</i>)	1461	Positive	0.716	0.521***	0.579***	0.544***	0.615***	0.660***
WallProt (Rice)	208	Positive	0.746	0.483***	0.591***	0.529***	0.616***	0.688***
WallProt (<i>Brachypodium</i>)	358	Positive	0.728	0.471***	0.556***	0.521***	0.605***	0.665***
ASURE Proteins (all)	975 (73 dual)	Mixed	0.503	0.480**	0.492**	0.477***	0.550***	0.499
ASURE (Nucleus)	130	Negative	0.494	0.503	0.517	0.492	0.565***	0.499

Supplementary Material

ASURE (Cytosol)	112	Negative	0.446	0.452	0.445	0.412	0.507***	0.445
ASURE (Nucleus + Cytosol)	352 (110 dual)	Negative	0.479	0.483	0.486	0.460	0.545***	0.482
ASURE (Mitochondrion)	70	Neutral	0.578	0.520	0.555	0.537	0.584	0.571
ASURE (Peroxisome)	12	Neutral	0.524	0.450	0.464	0.393	0.555	0.506
ASURE (Plastid)	103	Neutral	0.541	0.437*	0.502*	0.503	0.565	0.532
ASURE (Nucleus, Cytosol, Plastid, Mitochondrion, Peroxisome)	555 (18 dual)	Mixed	0.505	0.479	0.497	0.478**	0.553***	0.504
ASURE (Extracellular)	42	Positive	0.582	0.475	0.543	0.527	0.555	0.573
ASURE (Plasma Membrane)	119	Neutral	0.492	0.494	0.490	0.486	0.539*	0.482
ASURE (Extracellular + Plasma Membrane)	177 (16 dual)	Mixed	0.524	0.490	0.500	0.495	0.545	0.511**
ASURE (Golgi)	73	Neutral	0.487	0.446	0.467	0.455	0.523	0.479
ASURE (Vacuole)	50	Neutral	0.506	0.521	0.505	0.503	0.616*	0.509
ASURE (Endoplasmic Reticulum)	32	Neutral	0.483	0.449	0.479	0.443	0.516	0.474
ASURE (Endoplasmic Reticulum + Golgi)	120 (15 dual)	Neutral	0.487	0.451	0.475	0.462	0.529	0.480

1. Number of proteins in data set. Where multiple locations are included in sub-sets, they contain additional dual-located proteins excluded from individual sub-sets.
2. Indicates data (sub)set contain: Positive – CSPs as proxy for LSPs, Negative - non-secreted proteins, Neutral – secretory pathway or other organelles not used as negative data for ROC curves, Mixed – combinations of positive, negative and neutral data

Supplementary Table S2. Spearman correlation scores between original and modified sequences. Negative correlation are shaded.

Data	Reverse	SP Remove	SP C-term	Random	SP Random
WallProt	-0.15	0.59	0.29	0.30	0.84
WallProt (30% maximum sequence identity)	-0.09	0.65	0.44	0.42	0.89
WallProt (<i>Arabidopsis</i>)	-0.13	0.61	0.44	0.34	0.82
WallProt (excl. <i>Arabidopsis</i>)	-0.15	0.58	0.23	0.27	0.84
WallProt (Rice)	-0.14	0.61	0.18	0.21	0.87
WallProt (<i>Brachypodium</i>)	-0.22	0.59	0.27	0.16	0.87
ASURE (Extracellular)	-0.11	0.81	0.60	0.56	0.92
ASURE (Plastid)	-0.19	0.77	0.51	0.40	0.92
ASURE (Nucleus)	0.37	0.77	0.75	0.65	0.94
ASURE (Cytosol)	0.25	0.78	0.49	0.55	0.97
ASURE (Nucleus + Cytosol)	0.36	0.82	0.68	0.63	0.96
ASURE (Plasma Membrane)	0.26	0.76	0.69	0.46	0.97

Supplementary Table S3. Corresponding thresholds, TPR and FPR rates calculated from Figure 6 based on values where the FPR is below 0.05 (a) and using the common thresholds of 0.5 and 0.6 as found in the plant literature (b). In each, the values from the original sequences are in bold, and from the SP Remove modification in italics.

(a)

	TPR	FPR	Threshold
Original	0.276	0.043	0.794
Reverse	0.044	0.045	0.810
<i>SP Remove</i>	<i>0.029</i>	<i>0.048</i>	<i>0.860</i>
SP C-term	0.036	0.045	0.817
Random	0.033	0.043	0.735
SP random	0.130	0.048	0.772

(b)

	TPR	FPR	Threshold
Original	0.858	0.469	0.500
Original	0.753	0.298	0.600
Reverse	0.573	0.457	0.500
Reverse	0.330	0.281	0.598
<i>SP Remove</i>	<i>0.598</i>	<i>0.469</i>	<i>0.500</i>
<i>SP Remove</i>	<i>0.341</i>	<i>0.264</i>	<i>0.600</i>
SP C-term	0.617	0.389	0.500
SP C-term	0.360	0.236	0.600
Random	0.839	0.656	0.500
Random	0.347	0.324	0.600
SP random	0.814	0.486	0.499
SP random	0.634	0.278	0.599

Supplementary Table S4. Putative LSPs adapted from Table 1 in Cheng et al. (2009), with the prediction scores and classification using 0.5 and 0.6 thresholds.

AGI number	protein name and annotation	SecretomeP Score	LSP (0.5)	LSP (0.6)
AT3G16450	Jacalin- related protein (lectin)	0.93	YES	YES
AT1G23410	Ubiquitin extension protein	0.85	YES	YES
AT3G32980	PER32, peroxidase	0.79	YES	YES
AT3G07390	AIR12, extracellular matrix constituent	0.76	YES	YES
AT5G53560	Cytochrome b5 reductase	0.76	YES	YES
AT5G18100	CSD3, Cu/Zn superoxide dismutase	0.74	YES	YES
AT1G08830	CSD1, Cu/Zn superoxide dismutase	0.73	YES	YES
AT4G29680	Phosphodiesterase	0.73	YES	YES
AT3G43810	CAM7, Calmodulin	0.64	YES	YES
AT1G11580	Pectinesterase (methylesterase)	0.60	YES	YES
AT2G36530	LOS2 (low expression of osmotically responsive prot.), phosphopyruvate hydratase	0.60	YES	YES
AT3G16410	jacalin- related protein (lectin)	0.57	YES	NO
AT4G30920	Cytosolic aminopeptidase family	0.57	YES	NO
AT2G15130	Basic secretory protein (BSP) family protein	0.56	YES	NO
AT3G1639	jacalin- related protein (lectin)	0.56	YES	NO
AT3G03910	Glutamate dehydrogenase	0.53	YES	NO
AT3G08900	RGP, α 1,4-glucan-protein synthase	0.51	YES	NO
AT4G30910	Cytosolic aminopeptidase family	0.51	YES	NO

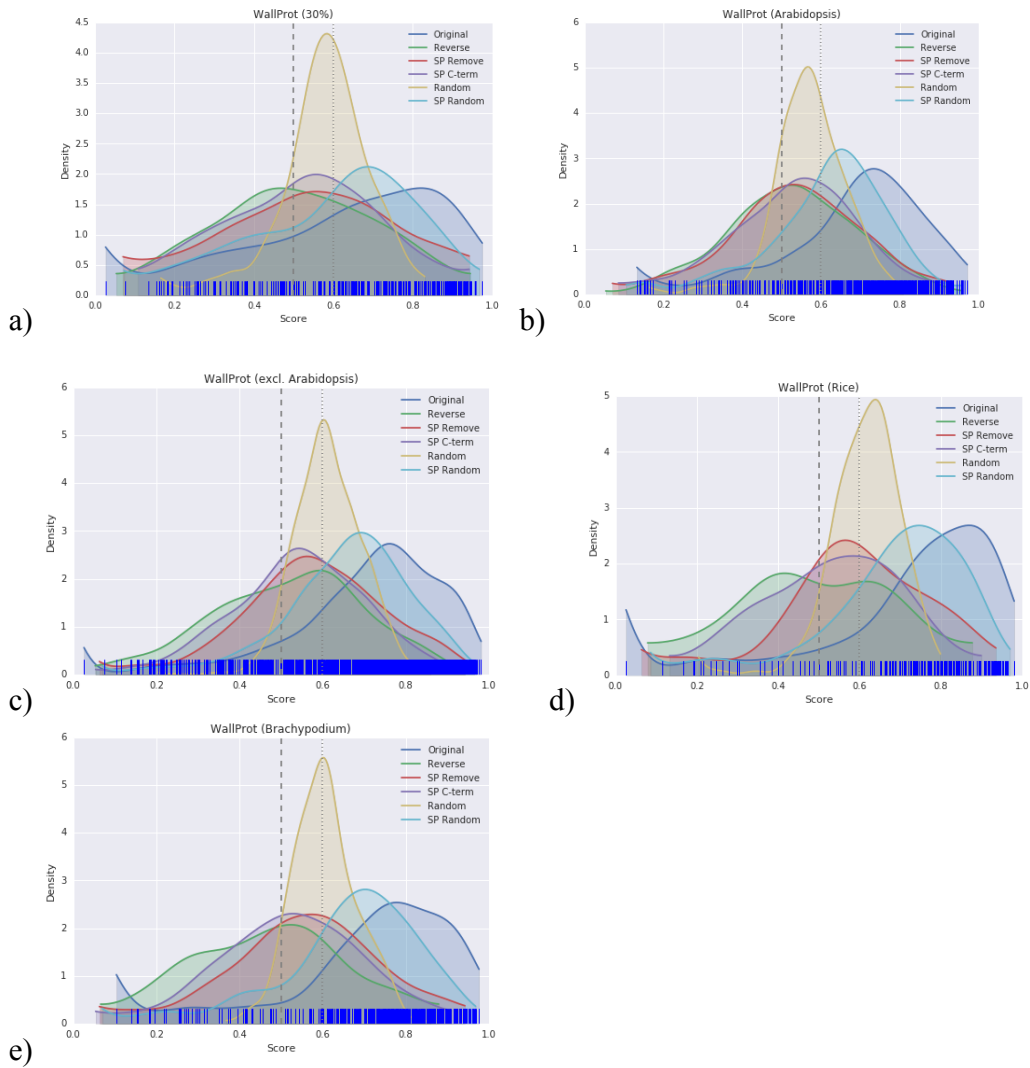
Supplementary Material

AT5G15650	CE±-1,4-glucan-protein synthase	0.51	YES	NO
AT5G16510	Reversibly glycosylated polypeptide	0.51	YES	NO
AT5G18170	GDH1, glutamate dehydrogenase	0.51	YES	NO
AT3G03780	AtMS2, methionine synthase	0.50	NO	NO
AT3G02230	ATRGP1, reversibly glycosylated polypeptide	0.49	NO	NO
AT4G23590	Aminotransferase	0.47	NO	NO
AT5G07440	Glutamate dehydrogenase	0.47	NO	NO
AT5G17920	ATMETS, methionine synthase	0.47	NO	NO
AT2G24200	Leucine aminopeptidase	0.46	NO	NO
AT2G36460	Fructose-bisphosphate aldolase	0.46	NO	NO
AT1G61820	BGLU46, Glycoside hydrolase	0.44	NO	NO
AT1G08110	Lactoylglutathione lyase	0.37	NO	NO
AT1G78380	GST8, glutathione transferase	0.35	NO	NO
AT5G55730	FLA1, fasciclin-like arabinogalactan protein	0.33	NO	NO
AT3G16420	PBP1, assists PYK10 (CE≤-glucosidase complex) activity in pest damaged tissues.	0.29	NO	NO
AT3G16430	Jacalin- related protein (lectin)	0.25	NO	NO
AT1G12080	Similar to vacuolar calcium binding protein	0.23	NO	NO
AT4G25100	FSD1, iron superoxide dismutase	0.20	NO	NO
AT4G20260	DREPP, plasma membrane protein	0.03	NO	NO

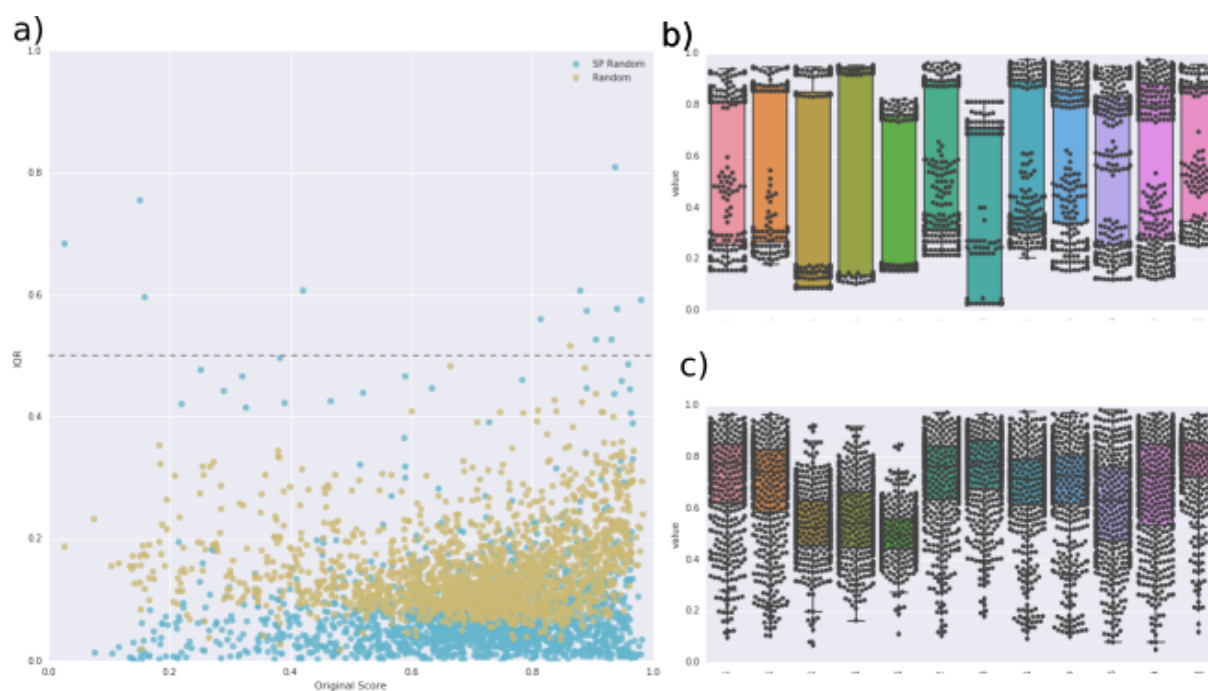
0.5 threshold = 21/37 = 0.567 TPR

0.6 threshold = 11/37 = 0.297 TPR

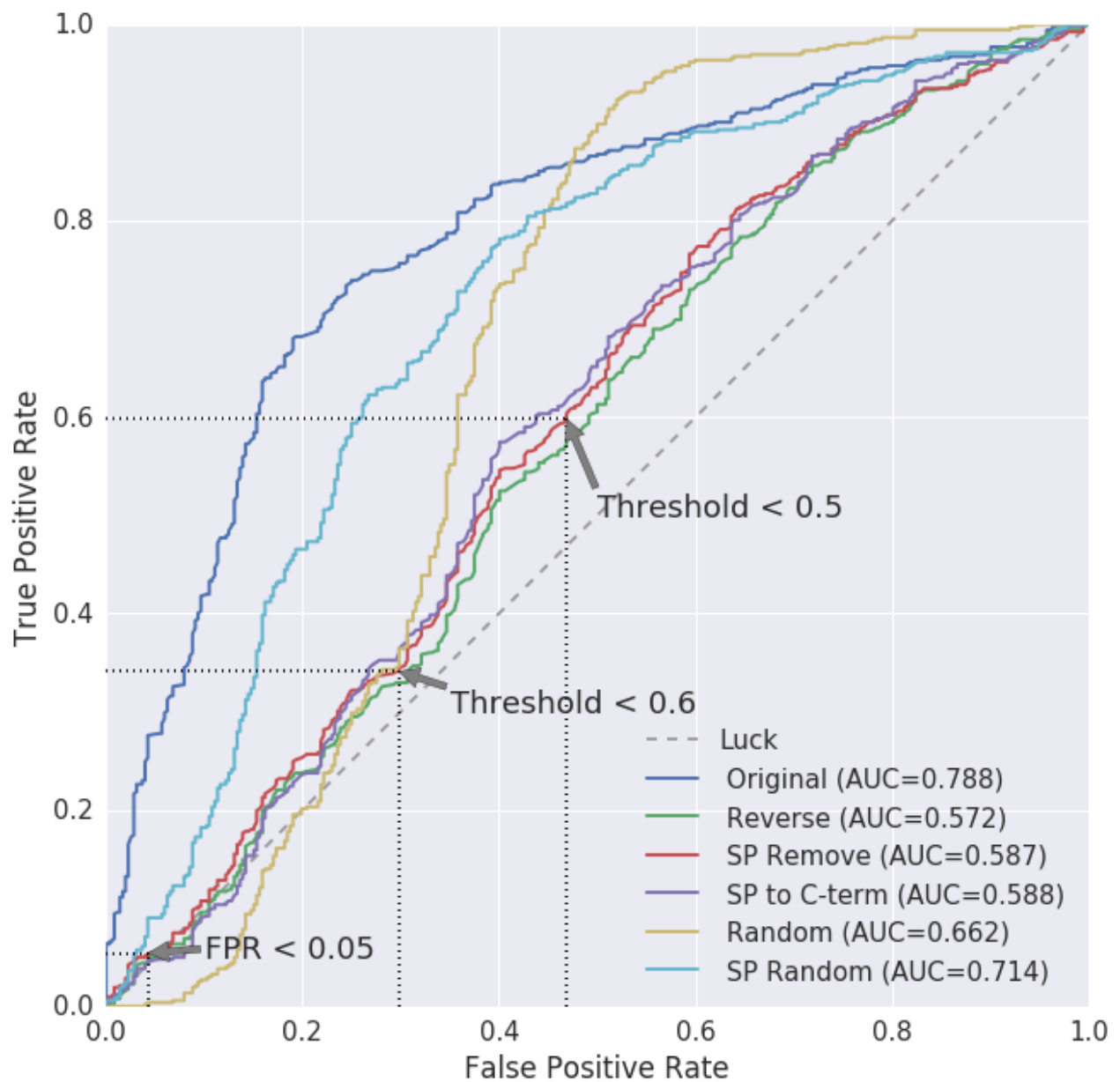
1.2 Supplementary Figures



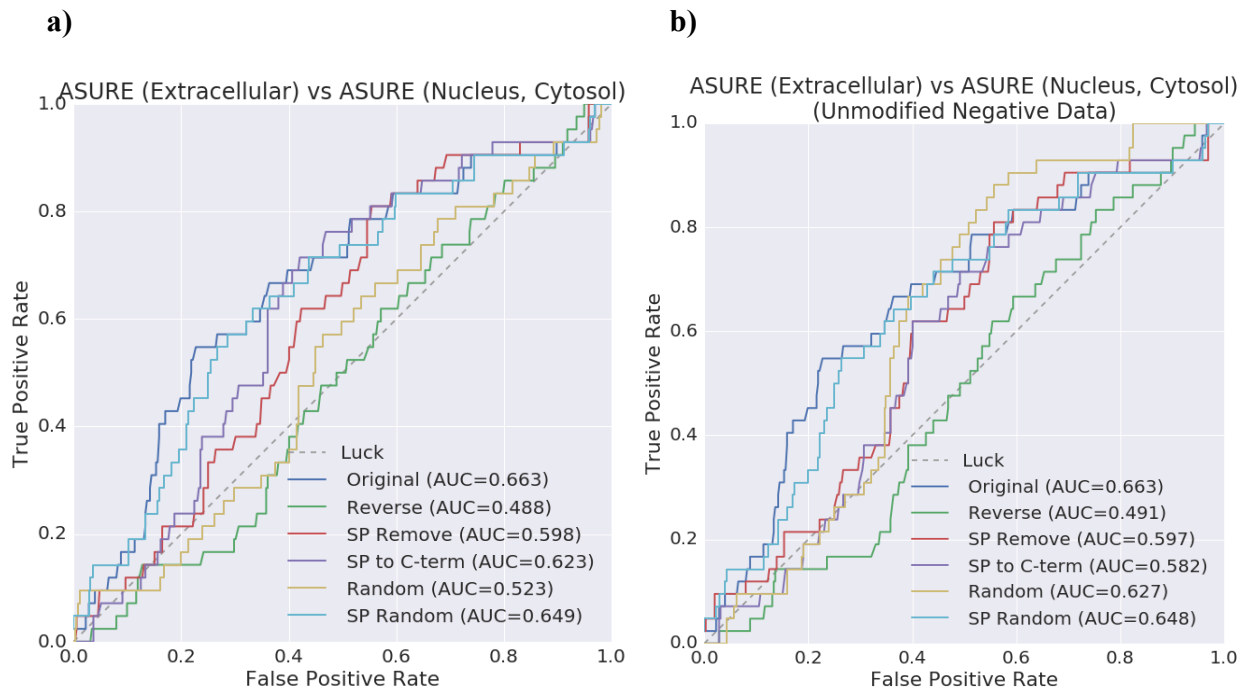
Supplementary Figure S1. The KDE distributions of SecretomeP scores (as described in Figure 3a) for original and modified sequences of subsets of WallProtDB with maximum 30% sequence identity (a), *Arabidopsis* only (b), excluding *Arabidopsis* (c), rice only (d) and *Brachypodium* only (e).



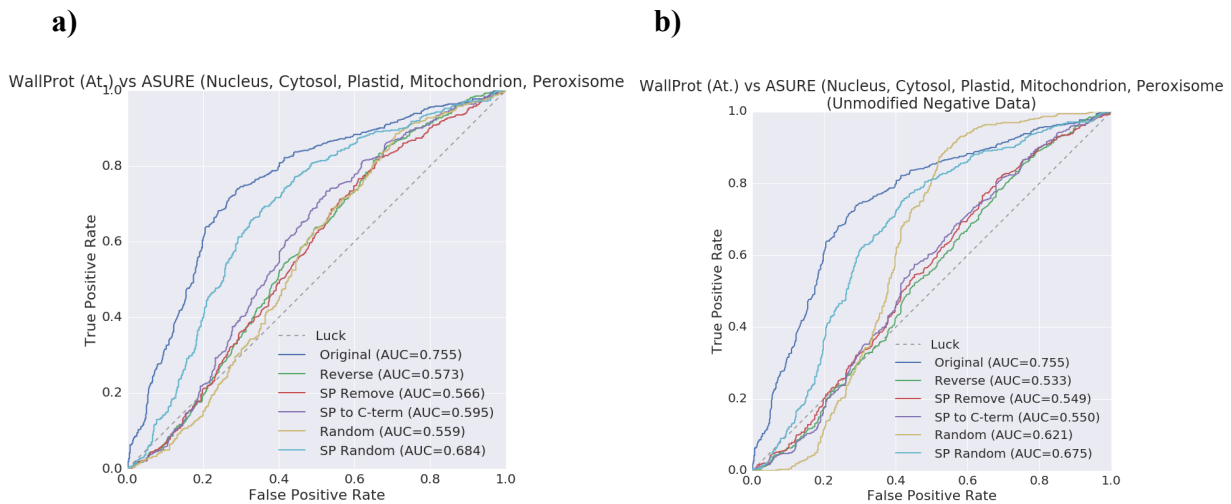
Supplementary Figure S2. WallProtDB sequences with modifications involving random shuffling. The interquartile range (IQR) for each sequence from 500 bootstraps is shown in (a). The 12 proteins with IQR > 0.5 for the SP Random modification are shown in (b) as a boxplot of the individual bootstrap scores and the corresponding Random bootstrap scores for the same 12 proteins (c). In these 12 examples, modifications to the SP region tend to group into either very high or very low scores, and for the completely shuffled equivalent sequence, these extreme scores are possible but not typical.



Supplementary Figure S3: ROC curves (as described in Figure 6) for modified WallProtDB *Arabidopsis* proteins as a positive class versus unmodified ASURE nuclear and/or cytosolic proteins as a negative class. Contrary to Figure 6, for each modification curve only the positive WallProtDB data is modified, and is compared to unmodified (Original) scores of the negative data. Three points on the SP remove curve are annotated corresponding to when either the FPR is <0.05 or threshold is set to 0.5 or 0.6.



Supplementary Figure S4: ROC curves, constructed in the same way as Figure 6 except using modified ASURE (Extracellular) proteins as a positive class. Sub-sets of ASURE with nuclear and/or cytosolic proteins were selected as a negative data either (a) modified or (b) unmodified.



Supplementary Figure S5: Additional ROC curves, with *Arabidopsis* WallProtDB proteins as positive data and an expanded negative data set of ASURE nuclear, cytosolic, plastid, mitochondrion and plastid proteins (a) modified (as per Figure 6) or (b) unmodified (as per Supplementary Figure 3), respectively.

2.4 Discussion

Although not referred to explicitly in this publication, the primary methods comparing the results of SecretomeP between modified and unmodified sequences are akin to metamorphic testing (Xie et al., 2011; Chen et al., 2009a). Metamorphic testing has been used for resolving the lack of either a ‘gold standard’ or ‘oracle’ of known answers for software tools and is proposed as a test case generation method in order to resolve bugs. It has been applied specifically to bioinformatics software (Giannoulatou et al., 2014; Shahri et al., 2019) in such cases where there is an unknown solution, such as a genome assembly where there is no prior knowledge that can be used to ground truth. In this method, metamorphic relationships are established that relate how modifications to program inputs should relate to program outputs. In this way, violation of the metamorphic relationships demonstrates that there is software error. In this research, rather than a bug specifically being resolved, the overall accuracy of the program on proxy data was investigated.

In metamorphic testing, additional data to test a program is created by applying the defined metamorphic relationships to input data. These relationships define a change in the input data that would have (or not) an expected change on output of a program. The changes for LSP prediction using CSP as a proxy are established in this chapter, notably defining a kind of metamorphic relationship between the amino acid sequence and the prediction score. The methods used were less formal than metamorphic testing, since the impact of all the changes was not necessarily known ahead of time. The five relationships that were established were: 1) a reversed polypeptide sequence; 2) the SP placed at the C-terminus; 3) the SP removed; 4) the SP randomised; and 5) the entire sequence randomised. Much of the work in the chapter sought to ensure that the observations of score changes within the SP removed data were symptomatic of an intrinsic bias towards these sequences. The additional metamorphic relationships demonstrated this was the case, including comparing the removal of the SP to iterations of randomised sequences to show that both modifications affected the scores. For future work, such additional relationships may not necessarily be needed, as the basis of the input modification used for training the tool, removing the SP, is the most important relationship to test, and a

critical bias to avoid. In general, it shows that the technique inspired from metamorphic testing is relevant to the LSP problem when modified input data is used as the common feature hypothesis. Reapplying this technique to new tools, in particular to avoid such bias during the creation of a tool, will be discussed in Chapter 4.

2.5 Conclusion

This research, published in 2016, established that SecretomeP performs poorly on plant proteins. The publication was intended to stop its usage by plant researchers and encourage critical evaluation when it is being applied. Comment was made on the prediction thresholds to be used, as a lack of guidance for application to plants led to varying interpretations. Overall, poor performance of the tool is made worse if the SP is removed, indicating reliance on SPs. Finally, very recent published work reinforces this chapter's conclusions. A revaluation of SecretomeP (Nielsen et al., 2018) using the latest UniProt data, split into mammalian and non-mammalian proteins, showed that contrary to the initial estimates, when using a prediction threshold of 0.6, the accuracy is 42.9% for mammalian and 35.6% for non-mammalian proteins, and critically the equivalent FPR is over 20%. Other similar tools, reliant on the same common-feature hypothesis, are similarly not recommended in this revaluation. The work presented in this chapter has cast doubt on inferences made regarding plant LSPs to date using mammalian trained programs. These subsequent results cast further doubt on SecretomeP's applicability to mammalian LSPs, and prompted the subsequent work in this Thesis to consider alternative bioinformatics tools for identifying plant LSPs.

Chapter 3

The Unusual Suspects

3.1 Introduction

What now for LSPs? The absence of a reliable bioinformatics tools for plant LSPs motivated this Chapter. If SecretomeP is not reliable for evaluating a plant cell wall or apoplastic proteome, as demonstrated in Chapter 2, and alternatives are required. Given that many of the putative LSPs were selected using SecretomeP, an unbiased approach is to go “back to the drawing board”, in essence, because any application of Reject or Filter strategies leads to the possibility of true LSPs that have been dismissed as contamination, or unlikely LSPs prioritised.

To identify LSP candidates we consider them without preconceived labels of which proteins are LSPs. We use a data-driven approach, starting with an entire proteome, and define a method to identify LSP candidates by combining protein features with experimental observations. The approach is outlined in **Figure 3.1**, and the method further described in Section 3.2. Critically for this Chapter, we divide the entire TAIR (Berardini et al., 2015)) proteome into “observed proteins”, those identified in secretomes in a defined set of published literature, and “unobserved” for those that have not. The relationship between proteins in these two groups forms the basis of the resource and subsequent analysis. The research in this Chapter provides this resource as the Leaderless Secretory Protein Database (LSPDB).

3.1.1 Leaderless Secretory Protein Database (LSPDB)

The first rule for biological database is to simply establish necessity (Batemen, 2007; Galperin et al., 2017; Helmy et al., 2016): Do we need a new database? To answer this affirmatively, we consider how would such a database be used, and what kinds of queries would be asked. Returning to the central issue with plant LSPs, either contamination or secretion, the purpose of a structured database is to be able to interrogate proteins across multiple studies and assign proteins to either state with some confidence. This can be applied either to new studies or retrospectively to consider which potential LSPs have previously been dismissed as either contamination or misclassified. This collation of data can allow it to be interrogated collectively. For example, which proteins without SPs occur most frequently in plant secretomes? Least frequently? Do protein motifs occur in these proteins more often than expected by chance? Which LSP candidates interact with conventionally secreted proteins? Which do not?

For plant LSPs, there is currently no convenient databases that summarise these data to be able to answer these questions. For example, PlantSecKB (Lum et al., 2014) collates relevant data for plant extracellular proteins and other locations and although the authors explicitly mention LSPs and that SecretomeP is unsuitable, they focus on conventional secretion signals. They recognise that collating protein information relevant to secretion can help identify secretory proteins yet the database lacks any specific mechanisms to investigate LSPs. WallProtDB is a plant cell wall dedicated database (San-Clemente and Jamet, 2015) that collates cell wall proteomes in a convenient interface for reference and research. Critically for LSPs, proteins without either signal peptides or transmembrane helices are excluded from WallProtDB, adhering to the default approach of requiring known features such as a signal peptide for inclusion. This approach does increase the reliability and utility of WallProtDB and makes it a valuable resource for the study of apoplastic/cell wall proteins. Their exclusion of putative LSPs increases the confidence and reliability of the data WallProtDB set. Thus WallProtDB is comprehensive but deliberately conservative with respect to LSPs. Proteins without an SP that are observed are excluded.

Other databases relevant to LSPs include those dedicated to ‘moonlighting’ - proteins known to have dual (or multiple) functions in a cell. MoonProt contains only 5 *Arabidop-*

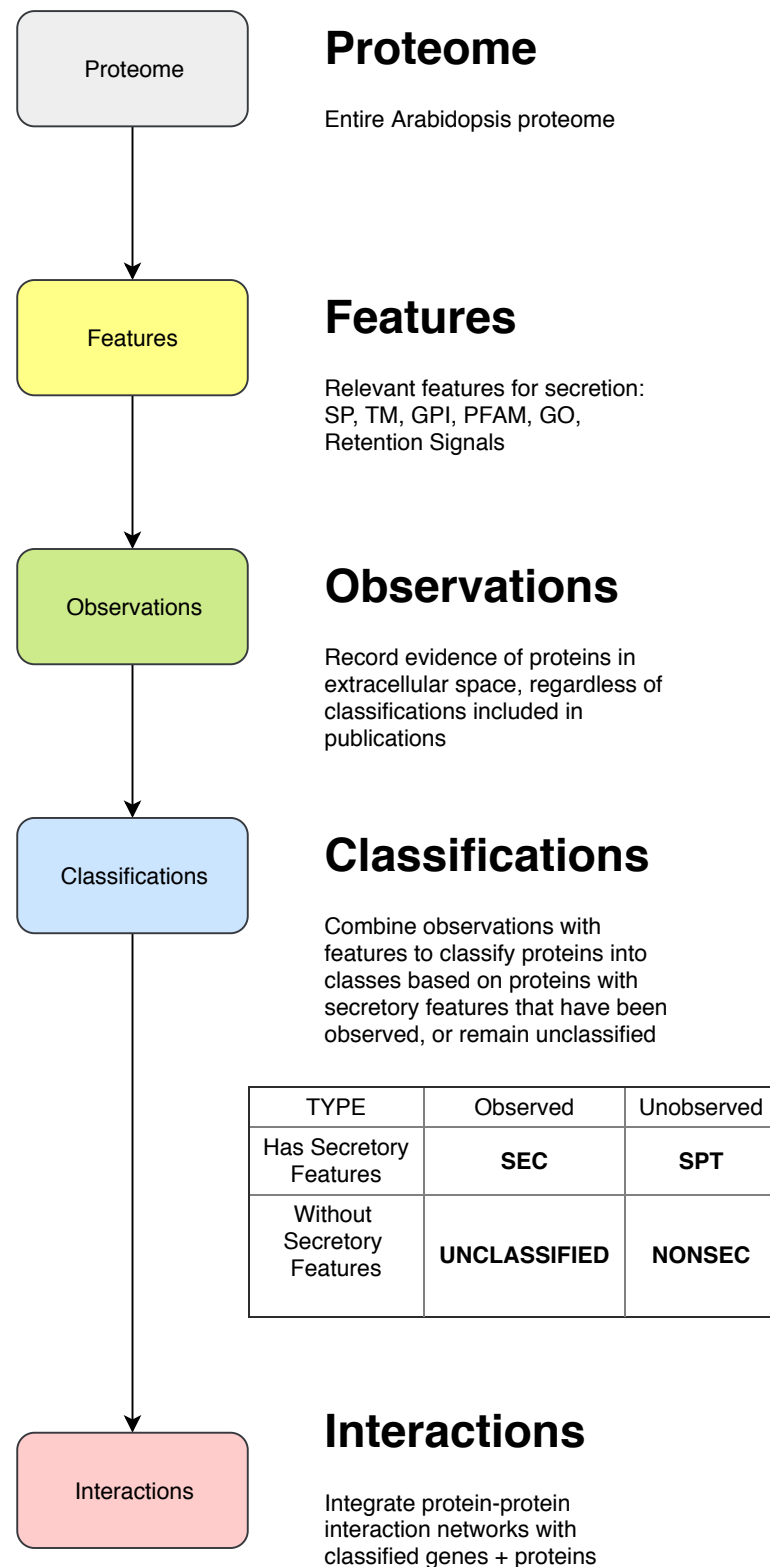


Figure 3.1: Overview of the LSPDB layered approach of taking a proteome and adding secretory features and literature observations to classify proteins as SEC (secretory features and observed), SPT (secretory features and unobserved), NONSEC (no secretory features and unobserved) and UNCLASSIFIED (no secretory features and observed). Interactions between these classes are then calculated.

sis proteins (accessed: July 2018 (Chen et al., 2018)) and MoonDB (Ribeiro et al., 2019) does not contain plant species. Interestingly, MoonDB uses a prediction suite MoonGo to look for ‘moonlighting’ proteins Chapple et al. (2015) including protein-interaction networks for pairs of interacting proteins with dissimilar biological functions, suggesting multiple functions. For LSPs there is some overlap in this methodology, for proteins with known functions in other sub-cellular localisations may be included, however, those without confirmed extracellular functions are not, so is not appropriate for LSP discovery, and does not specifically consider occurrences in the secretome and the contamination/leaderless test. The approach to consider protein-protein interactions and GO terms combinations, however are used in the LSPDB. The LSPDB can be considered an analogue of a moonlighting database, as there are also proteins with unknown functions that may be purely secretory, though many proteins with intercellular functions could be considered moonlighters if established as LSPs.

The LSPDB is designed for dual purposes; to collate relevant information on plant proteins specific to plant secretion for the purposes of exploring potential plant LSPs, and to present these data in an interface to act as a drop-in replacement in the analysis pipeline for plant extracellular proteins. *In lieu* of an accurate prediction tool, the LSPDB is presented as the next best alternative and is driven by the lack of a database to conveniently search for potential LSPs. This user-centric approach is a key component beyond the data captured in the database. Recall the “Accept/Reject/Filter” approach to observed secretory proteins. To filter requires some criteria, whether it be the single output of software tool, or a combination of factors. The LSPDB is intended to be a resource for such a filtering method.

In reviews of plant LSPs, qualitative information has been primarily used to assess plant LSPs (Ding et al., 2012) (**Appendix Table A.2**). By collating plant information the LSPDB allows for some quantitative calculations on the properties and propensity of proteins to appear in plant secretomes. The LSPDB data can then be used as a tool for prioritising the results of cell wall proteomes into a list of putative list for experimental validation.

The LSPDB version referenced in this Thesis is denominated **LSPDB/Plant v1.0**.

The LSPDB is intended to initially promote the research of plant LSPs, but other organisms may be added in the future. Data scarcity, of LSPs in particular, drives the choice of *Arabidopsis thaliana* (*At*) as the initial organism of interest. As a model system there are considerable additional resources for genomics, transcriptomes, proteomics, protein-protein interaction data, and co-expression networks to mine. The publications linked to these are also of importance. Since experimental observations - in various conditions, tissues and developmental stages - is part of the database, well studied organisms are preferred. The LSPDB is initially suited to *Arabidopsis*, but the outputs will not be isolated to “mining” *Arabidopsis* research. Firstly, the methods can be applied to other plant species. Secondly, results establishing a set of high confidence LSPs in *Arabidopsis* could yield homologous proteins in other plant species. Insights into secretion categories can be applied to other plant genomes, and finally it provides a template for other organisms, so that as the secretomes of additional species are added, the same inferences can be made. This initial version (**LSPDB/Plant v1.0**) of the LSPDB sets a template for future iterations for more plant species, and other organisms.

This Chapter details the methods used in classifications derived from the LSPDB in Section 3.2, and the results are presented in Section 3.3. We discuss the implications of these results including applying the LSPDB for the purposes of generating a candidate list of LSPs in Section 3.4. The construction and creation of an online interface for accessing the LSPDB externally is presented in Section 3.5 and finally in Section 3.6 conclude with a summary of the utility of the LSPDB and directions for future development.

3.2 Layers of the LSPDB

The philosophy underpinning the LSPDB is to consider *Arabidopsis* plant proteins without preconceived labels. For each protein, we collate relevant information to facilitate the likelihood of LSPs, and use both protein features and experimental observations as markers for sub-categorisation. Conceptually, the foundation layer to the layered database construction is an entire proteome. From this position, analysis of LSP status from the LSPDB is inferred via layers of evidence and annotation, where each layer can be queried

either in isolation or additional query mechanisms based on combinations of previous layers. The final database schema is therefore a product of these distinct layers.

For a whole organism, initially *Arabidopsis*, all proteins are included in the model. Given the protein-centric approach to LSPs (the P is for protein!) the base layer is actually at the gene level due to the possible ambiguity of mass spectrometry results. Genes may produce multiple proteins due to alternative splicing, yet peptides identified may not cover all disambiguation points in a protein sequence in order to identify which exact protein it is. Reporting of MS results can often include a mix of exact protein matches or gene-level only results, and the lowest common clarity of the gene level is reported. The diversity of gene products is further complicated by the possibility of post-translational modifications (PTMs).

Since the literature reporting does not consistently list PTMs in the observed genes/proteins, the absence of PTMs in results is not indicative of a lack of PTM. Additional issues considered was the large number of PTMs, reliance on predictions, and whether PTMs apply to all isoforms of a protein which could impact the classification. Therefore PTMs were not included in the initial LSPDB structure. Instead, PTMs are an ideal candidate to investigate extending the LSPDB with external data, and the results of joining PTM annotation with the LSPDB classifications are discussed in chapter 5.

In this context the observation of potential LSPs we will add to the LSPDB may only be at the gene level, and so both genes and proteins form the initial database contents, with a link of multiple proteins to each gene. Critically, the amino acid sequence is stored with the protein so that although observations can only be grouped at the gene level, relevant secretory differences between the proteins of that gene can be assessed; for example, if all of the protein products of a gene encode signal peptides, then we can safely consider the proteins at the gene level in the context of the LSP versus contamination issue, whereas proteins from the same gene that differ in their SP composition may be ambiguous in this context and need to be considered separately. We will term these as equivalent classes of secretory genes, and identify non-equivalent genes that we term secretory isoforms, discussed further in Section 3.2.2.1. The gene-protein model is built according the following pseudocode:

- For each gene:
 - Create gene model in LSPDB
 - Get all transcripts of gene
 - For each transcript:
 - Create protein model in LSPDB

3.2.1 Features layer

Features relevant to secretion are primarily those with amino acid stretches with known functions related to the secretory pathway: signal peptides (SP), transmembrane domains (TMs) and GPI anchor signal sequences. For well studied proteins these annotations are available in large databases such as UniProt. Evidence obtained experimentally and/or computationally is collated, however this may not be available for all proteins, in particular those with little experimental evidence and few publications. To remedy this uneven annotation, computational predictions for SP (SignalP, Armenteros et al. (2019)), TM (TMHMM, Krogh et al. (2001)) and GPI (BigPi Eisenhaber et al. (1999) and PredGPI Pierleoni et al. (2008)) were used on the entire proteome.

A trade off for the added depth of using predictions for these secretory features is dealing with ambiguity and overlap. Using the two most recent versions (at the time) SignalP, or between alternative GPI prediction programs BigPi and PredGPI, meant that there was not always consensus. The similarity between these features, with similar hydrophobic properties can also lead to ambiguity, e.g. positive predictions for both SP and TM in the same N-terminal region. Similarly, overlapping predictions for GPI anchors and TM near the C-terminus were also observed.

To resolve a lack of consensus, a positive prediction in either SP or GPI predictor is sufficient for the respective annotation. For overlaps with TM domains, the range each of the SP and GPI combined predictions is compared to TM helix predictions. If a TM prediction overlaps with either SP or GPI, then either the SP or GPI is preferenced and the TM removed from the annotation. This weighs SP and GPI more heavily, but essentially collapses ambiguous predictions into SP-like or GPI-like annotations. Multi-membrane proteins are unaffected by those corrections, as the final category will still include TM,

so the effect is to reduce the number of possible spurious TM annotations and avoid features being assigned from the same amino acid sequence. The algorithm is as follows in pseudocode:

```
- for each protein:
  - collect predictions of SP, TM and GPI regions
  - for all positive predictions collect region start and end
  - for each SP prediction:
    - if SignalP v3 and SignalP v4 == positive:
      - if regions overlap:
        - merge regions start and end
        - SP: Yes
    - else if SignalP v3 or SignalP v4 == positive:
      - take region from positive prediction
      - SP: Yes
  - for each GPI prediction:
    - if BigPi and PredGPI == positive:
      - if regions overlap:
        - merge regions start and end
        - GPI: Yes
    - else if BigP or PredGPI == positive:
      - take region from positive prediction
      - GPI: Yes
  - for each TM prediction:
    - if SP and TM regions are overlapping at N-terminus:
      - remove TM
    - if SP and GPI regions are overlapping at C-terminus:
      - remove TM
  - assign each protein according to the remaining annotations:
    - SP (Yes/No)
    - GPI (Yes/No)
```

- TM (count of TM predictions)

After all annotations are corrected, in order to simplify the data model, the count of TM predictions per protein is assigned, where 0 indicates no TM. The effect of this correction reduces the transcripts annotated with TM from 11194 proteins to 8318, resulting in 2876 transcripts/proteins with corrected features.

3.2.1.1 Secretory proteins and secretory isoforms

With annotations corrected to “unique in sequence” feature assignment, we can assign secretory status to each gene in the context of its protein(s). For the “one to many” relationship of gene to protein, we determine if the gene is an equivalent class of secretory genes if all its protein share the same secretory features. Multiple transcripts from the same gene with identical secretory properties are not necessarily all secretory isoforms, e.g. when all contain a SP, or none contain any such features. Only genes with multiple transcripts, where at least one of the transcripts has a different secretory combination, are included as non-equivalent genes that we term secretory isoforms. This complexity in secretory features has been observed in other species, including humans as summarised by (Uhlén et al., 2015) in **Figure 3.2**.

For the LSPDB using conventional properties of SPs, TM helics and GPI-anchor signal sequences, there are (3^3) 27 possible combinations where a protein can have these as either always present (label ‘Yes’), always absent (label ‘No’) or varying (label ‘Iso’ e.g isoform).

Retention signals for the secretion pathway (motifs such as KKXX and KXE/D Gao et al. (2014)) can also vary between sequences. From a secretory point of view, the combination of SP and retention signal in some cases may be of interest, so we can also assign a retention secretory status as Yes/No/Iso as well, leading to (3^4) 81 possible combinations. Although the presence of a retention motif is generally of interest only in the context of classical secretory mechanisms, these are retained in the LSPDB for future investigation of secretory isoforms.

- For each gene:

**This image has
been removed by
the
author of this
thesis for
copyright reasons**

Figure 3.2: Proportions of secretory protein types in the human proteome, Figure 3B from Uhlén et al. (2015). Overlaps represent secretory isoforms.

- Collect SP status for each protein
- Collect TM status for each protein
- Collect GPI status for each protein
- Collect retention motif status for each protein
- If SP status identical:
 - Assign SP: Yes|No
- Else:
 - Assign SP: Iso
- If TM status identical:
 - Assign TM: Yes|No
- Else:
 - Assign TM: Iso
- If GPI status identical:
 - Assign GPI: Yes|No
- Else:
 - Assign GPI: Iso
- If Retention status identical:
 - Assign Retention: Yes|No
- Else:
 - Assign Retention: Iso

Each gene now has a stored (for succinct analysis) value for the secretory isoform status of each of SP, TM, GPI and retention signals.

3.2.1.2 PFAM and GO terms

Every gene is also categorised according to either known function domains or gene ontology terms. PFAM domains are drawn from sequences, whereas GO terms are not necessarily directly related to sequence as they may come from molecular function, biological process or cellular compartment. In this case the collated annotations from a central database are more useful since we do not have tools to computationally assign all of these

features as was the case for SP etc. Using the `biomart` Python package, the Ensembl Plant Genes 35 dataset available from <http://plants.ensembl.org/biomart>, the PFAM and GO terms were added to the LSPDB as follows:

- For each protein:
 - Connect to Biomart and extract PFAM and GO attributes
 - Take unique occurrences of PFAM and GO in protein only
 - if GO term does not exist in DB:
 - create GO term
 - assign GO term to protein
 - if PFAM does not exist in DB:
 - create PFAM term
 - assign PFAM term to protein

Each protein now has GO and PFAM terms associated with it. Due to the complexity of GO and PFAM terms, and the possibility that multiple combinations of each could appear in various proteins, we do not extend the same model for the sequence features as this would, for example, require an isoform status for every PFAM domain and/or GO term.

3.2.2 Observations layer

A critical layer of the LSPDB is built on relevant observations. Proteins that have been experimentally observed in the extracellular space from proteomics studies are entered in order to combine this observation with other layers in the database. There is scope to include observations from the relevant published literature regarding LSPs. As a source, the scope of WallProtDB overlaps with this purpose and serves as a seed for inclusion into the LSPDB. Exclusions include related but necessarily secreted protein experiments such as studies of enriched plasma membrane preparations. Due to the limits of identifying proteins, the gene level of the LSPDB is linked to the Observations layer.

For each paper, every protein reported in that paper is entered into the database as an observation, along with the lab methodology, experimental conditions, and the classifi-

cation as determined by the authors. This classification is only used to compare against LSPDB classifications (see Results). This step explicitly includes proteins dismissed as either contamination or even identified as putative LSPs. In total, 6053 observations of proteins (genes) were recorded for *Arabidopsis*, 2677 of which are unique. Since the purpose of the LSPDB is to identify contamination from LSPs, observations within the same published paper, i.e due to multiple experimental conditions, are each counted as observations of that gene. A summary of the observations per paper is listed in **Table 3.1**.

Table 3.1: Summary of literature included in LSPDB with counts of observed proteins (gene level)

	Observations
Basu et al. (2006)	48
Nguyen-Kim et al. (2016)	1968
Hervé et al. (2016)	1662
Chen et al. (2015)	27
Boudart et al. (2005)	101
Haslam et al. (2003)	10
Trentin et al. (2015)	116
Cheng et al. (2009a)	75
Charmont et al. (2005)	49
Minic et al. (2007)	104
Kwon et al. (2005)	168
Robertson et al. (1997)	16
Tran and Plaxton (2008)	49
Zhang et al. (2011b)	131
Borderies et al. (2003)	153
Irshad et al. (2008)	257
Oh (2005)	78
Bayer et al. (2006)	716
Feiz et al. (2006)	178

	Observations
Borner (2003)	30
Casasoli et al. (2008)	33
Chivasa et al. (2002)	72
Schultz (2004)	12

Notebook methodology is useful in the creation of a database drawn from the literature due to the inconsistent formatting between papers. For our focussed work on the LSPDB for plants, the manageable number of papers and need to record as much data as possible necessitated a manual method. Additionally, since the supplementary methods are often the place for proteins without canonical signals to be included, the unstructured format nature of these sections would lead to difficulties in automatic extraction. Therefore “notebook style” allows for a fine-tuned approach to data curation. A template notebook for accessing the database abstraction layer was created, and then one notebook per paper was created. This gave flexibility with each individual publication, which could have had differences in either formatting and/or annotation that means there is no universal approach to the data import. The manual approach allows for the variety of ways that proteins are reported (i.e gene or protein level, grouped as SP and non-SP containing, or split between main and supplementary materials) to be normalised. Minor points like accession codes for converting from other IDs to a common format such as AT accession style can be done at this manual stage, and data saved into the database in a consistent manner.

3.2.3 Classifications layer

The classification layer is the first layer to combine data from previous layers, namely observations and secretory features. Since secretory isoforms contain some ambiguity, they are not included in the classification layer. Another combination not currently included for direct classification are multi-transmembrane proteins. From a UPS perspective that could be relevant to a Golgi-bypass route (Gee et al., 2018), however for the soluble LSP-

centered classes they are excluded. This avoids protein isoforms with differing numbers of TM which could be considered a kind of secretory isoforms and additional complexity for retention considerations for either Golgi or ER residence of these proteins. These two classes of proteins remain in the database, but their interactions and features were not used to infer LSP status. Inclusion of both in the future could enhance the LSPDB.

For all other genes/proteins, they can be assigned to either secretory or non-secretory categories based on features, and further divided on whether they have been observed/classified experimentally. These two primary categories are labelled SEC (secretory) for proteins with such features that have been observed and NONSEC (non-secretory) for proteins that lack the SP, GPI and TM features that have been not observed in the LSPDB. The remaining two categories invert these first two classes. The UNCLASSIFIED label applies to proteins that have no relevant secretory signals and have been seen experimentally observed (i.e. **Table 3.1**), noting that the way authors classified these proteins in these studies is not included at this level. They are the best candidates for LSPs (or contamination), and further analysis will seek to identify which of these they are most likely to belong. The final classification includes those genes/proteins that possess the required secretory protein features that have not (yet) been observed experimentally. We label these as theoretical secretory proteins (SPT or SP_THEORY). The assignment of proteins to these categories in version 1 of the LSPDB is shown in **Figure 3.1**.

Each of the four classes can be used for analysis, though the UNCLASSIFIED is of particular interest as it represents a superset of possible LSP candidates. The NONSEC class may contain, as yet unobserved LSPs. Applying LSP prediction tools to such a class could assist in the identification of LSPs prior to their experimental observation and hence focus these approaches.

The classifications layer is also a convenient exit point for LSPDB data. The division of proteins into those with expected and unexpected secretory observations can be used to join with external data to extend the use of the LSPDB beyond the information included in the database. Combining classifications with additional external data is explored in chapter 5.

3.2.4 Interactions layer

The interactions layer infers connections between proteins based on protein-protein interaction(s) data, and is applied to the inferred classifications layer. It is suggested that the interactions of proteins can inform their function. For potential LSPs, interactions with known secretory proteins may be evidence for LSP status. We insert the AtPIN (Brandão et al., 2009) Arabidopsis protein-protein interaction network on top of the LSPDB for all proteins, assigning interactions directly. Given the SEC/NONSEC/SPT/UNCLASSIFIED classifications on a subset of the LSPDB, the subset of interactions that occur between classified proteins can be summarised and stored for analysis:

```
- for each protein:
  - get proteins classification
  - get all interactions
  - for each interaction:
    - if with SEC:
      sec_count +=1
    - if with NONSEC:
      nonsec_count +=1
    - if with SPT:
      spt_count +=1
    - if with UNCLASSIFIED:
      unclassified_count +=1
```

3.2.5 Integrating the layers

With each of these layers, the database can be interrogated at either individual layer or as a whole. Each layer corresponds to one or more database classes implemented in the Python Django framework as shown in **Figure 3.3**. This shows that the conceptual overview of the database in **Figure 3.1** is more complex when actually implemented as a database. The proteome foundation (yellow: Gene, Protein), features layer (yellow: Secretory Features, PFAM , GO Term), observations layer (green: Observation), classifi-

cations layer (Sec Type), interactions layer (Interactions, Interaction summary) each relate to the colour-coded database layers in **Figure 3.1**. Directed arrows between each box represent the relationship between classes.

Of note, the information available in the LSPDB is restricted to what is minimally required to categorise into secretory categories based on the hypothesis of features and interactions. The database schema is accessible as a web interface (see Section 3.5) and as Django classes for direct queries. These queries extend the database beyond an informative framework for users to assess their gene lists for LSPs, and collate the data in order to observe and test hypotheses regarding LSPs as a whole. Using Python scripts, these queries can be implemented as simple object orientated queries. For example, to count all the Gene level objects in the LSPDB using Python code:

```
import django
django.setup()
from secfeatures.models import Gene

total_gene_count = Gene.objects.all().count()
```

More complex queries can be chained together. For example, to collect all GO terms that occur in proteins labelled as SEC, the following query can be used:

```
import django
django.setup()
from secfeatures.models import *
from secinteractions.models import *
from secobservations.models import *

sec_goterm = SecType.objects.filter(
    base_classified = "SEC",
    gene__isoform_class = "SP=Yes,TM=No,GPI=No"
).values_list('gene__protein__secretoryfeatures__go__term', flat=True)
```

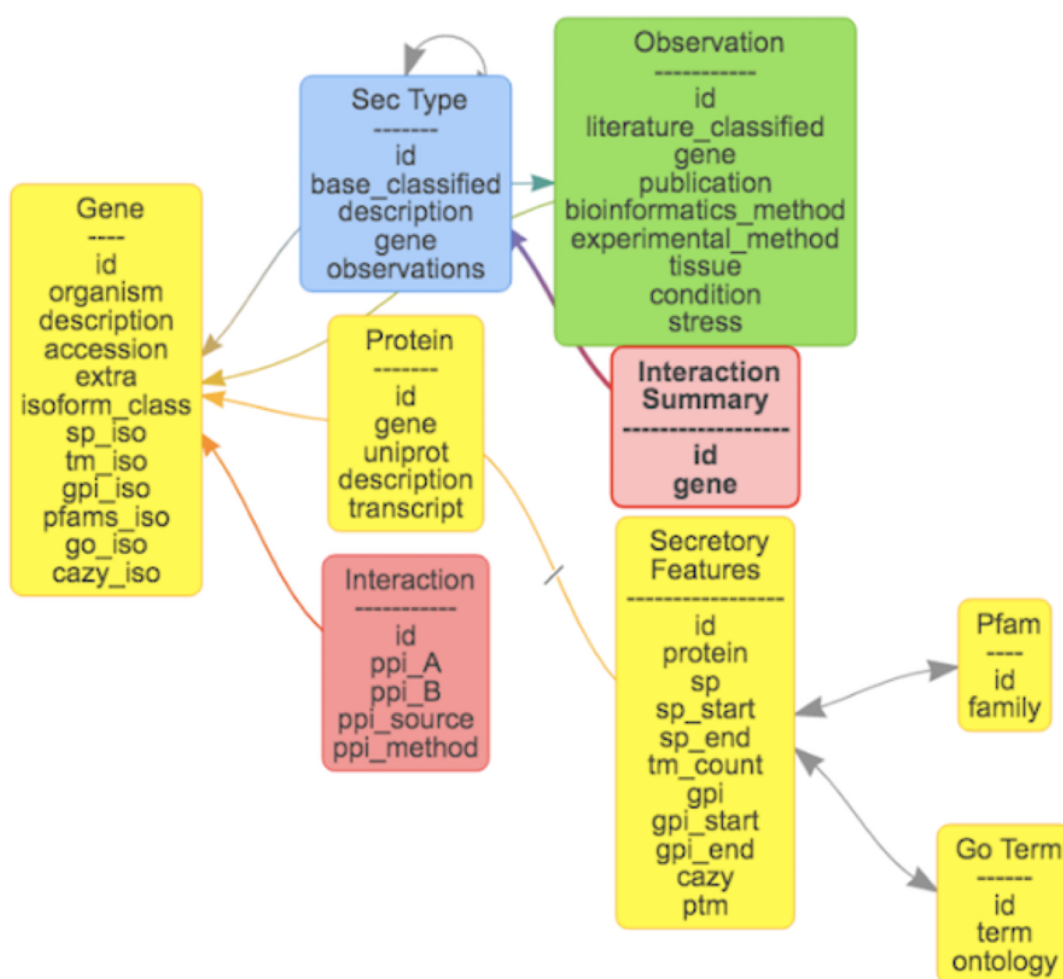


Figure 3.3: Database design overview of the LSPDB, with each colour corresponding to a Python Django database class

This abstract layer can then be used to analyse the LSP landscape in order to ask questions such as: What is the most frequently seen protein in plant secretomes? The least? Which PFAM and GO terms occur most frequently overall, or in the SEC/NONSEC etc classes? The database also allows testing for over-representation of features or motifs in proteins, or the prevalence of protein-protein interactions within or between classes. The results of these and further queries are presented in Section 3.3, utilising the layered approach used when building the LSPDB.

3.3 Results

Results of the LSPDB are divided below into observations from the input based layers (features and observations layers) and those built on the classification structure (classifications and interactions layers). First we consider the most frequent secretory features, PFAM and GO terms based results of the features and observations layers (Section 3.3.1), followed by results derived from the classifications including the distribution of PFAM domains and GO terms amongst the classes (Section 3.3.2), and the interaction network profiles of classified proteins (Section 3.3.3).

3.3.1 Frequency

Collating observations of genes allows the distribution of occurrences to be calculated. These distributions show that most proteins appear once, and the most frequent protein (18 observations) is in the SEC category. Since the bar for CSP appearance is typically once, a frequency based approach seeks to consider the equivalent for LSPs. In terms of the “Accept/Reject/Filter” paradigm, “Accept” can be thought of as 1:1 ratio - one appearance of a protein with a SP is sufficient for it to be accepted. Repeated observations are typically not required. “Reject” dismisses all LSPs so can be considered 1+:0, as “no” number of occurrences will suffice.

Table 3.2: Frequent observations (10 or more) in the LSPDB, all of which contain secretory features

Accession	Description
AT1G68560	alpha-xylosidase 1
AT1G78850	D-mannose binding lectin protein with Apple-like carbohydrate-binding domain
AT3G45970	expansin-like A1
AT1G33590	Leucine-rich repeat (LRR) family protein
AT1G03220	Eukaryotic aspartyl protease family protein
AT5G67360	Subtilase family protein
AT3G08030	Protein of unknown function, DUF642
AT1G78830	Curculin-like (mannose-binding) lectin family protein
AT4G25900	Galactose mutarotase-like superfamily protein
AT1G53070	Legume lectin family protein
AT5G20950	Glycosyl hydrolase family protein
AT3G49120	peroxidase CB
AT2G06850	xyloglucan endotransglucosylase/hydrolase 4
AT1G03230	Eukaryotic aspartyl protease family protein
AT3G54400	Eukaryotic aspartyl protease family protein
AT5G06860	polygalacturonase inhibiting protein 1
AT1G09750	Eukaryotic aspartyl protease family protein
AT3G13790	Glycosyl hydrolases family 32 protein

Accession	Description
AT5G08380	alpha-galactosidase 1
AT5G07030	Eukaryotic aspartyl protease family protein
AT3G14310	pectin methylesterase 3
AT5G11420	Protein of unknown function, DUF642
AT5G55730	FASCICLIN-like arabinogalactan 1
AT1G47128	Granulin repeat cysteine protease family protein
AT1G76160	SKU5 similar 5
AT3G52500	Eukaryotic aspartyl protease family protein
AT5G25460	Protein of unknown function, DUF642
AT3G21770	Peroxidase superfamily protein
AT3G32980	Peroxidase superfamily protein

For secretory proteins in **Table 3.2**, proteins such as glycosyl hydrolases and FASCICLIN-like arabinogalactan-proteins (FLAs) appear as expected. The most observed is a GH, Alpha-xylosidase 1, and is a well characterised apoplastic protein. Using only frequency for the unclassified however yields no additional insight, with a maximum of 10 observations for Cobalamin-independent synthase family protein (ATMS1,AT5G17920). **Table 3.3** shows all UNCLASSIFIED proteins observed more than the median (5) number of observations. Several proteins in the table are well characterised as cytosolic proteins, with others that are further afield such as either mitochondrial or chloroplast located genes. Simply counting the number of times they appear does little to resolve whether they are contaminants, or if UPS pathways are leading to genuine secretion. Therefore this result alone does not resolve the “Is it an LSP or a contaminant?” issue, and as such a threshold for the number of appearances is difficult to justify. For the “Filter” strategy, no clear empirical threshold to identify frequently occurring candidate LSPs and establish a cut-off for the number of appearances required to accept is shown. Therefore, “frequency of occurrence” is a poor parameter to use to identify LSPs, however it can be compared to frequency-independent methods of LSP identification, discussed further in **Chapters 4 and 5**, to compare if those candidates have been repeatedly observed.

Table 3.3: Frequent observations (5 or more) of UNCLASSIFIED proteins in the LSPDB, regardless of previous annotations or established non-secretory roles.

Accession	Description
AT5G17920	Cobalamin-independent synthase family protein
AT1G13440	glyceraldehyde-3-phosphate dehydrogenase C2
AT2G36530	Enolase
AT3G16420	PYK10-binding protein 1
AT3G55440	triosephosphate isomerase
AT4G09320	Nucleoside diphosphate kinase family protein
AT4G39260	cold, circadian rhythm, and RNA binding 1
AT4G34870	rotamase cyclophilin 5
AT5G20080	FAD/NAD(P)-binding oxidoreductase
AT5G42980	thioredoxin 3
AT1G26630	Eukaryotic translation initiation factor 5A-1 (eIF-5A 1) protein
AT1G78380	glutathione S-transferase TAU 19
AT3G04120	glyceraldehyde-3-phosphate dehydrogenase C subunit 1
AT5G07440	glutamate dehydrogenase 2
AT1G79550	phosphoglycerate kinase
AT1G53240	Lactate/malate dehydrogenase family protein
AT2G24200	Cytosol aminopeptidase family protein

Accession	Description
AT4G17260	Lactate/malate dehydrogenase family protein
AT3G16450	Mannose-binding lectin superfamily protein
AT2G16600	rotamase CYP 3
AT1G67090	ribulose biphosphate carboxylase small chain 1A
AT4G01900	GLNB1 homolog
AT1G11840	glyoxalase I homolog
AT1G67280	Glyoxalase/Bleomycin resistance protein/Dioxygenase superfamily protein
AT3G11630	Thioredoxin superfamily protein
AT4G23670	Polyketide cyclase/dehydrase and lipid transport superfamily protein
AT1G08830	copper/zinc superoxide dismutase 1
ATCG00490	ribulose-bisphosphate carboxylases
AT3G02230	reversibly glycosylated polypeptide 1
AT4G20260	plasma-membrane associated cation-binding protein 1
AT4G38740	rotamase CYP 1
AT4G25100	Fe superoxide dismutase 1
AT1G04410	Lactate/malate dehydrogenase family protein
AT3G54660	glutathione reductase
AT2G17420	NADPH-dependent thioredoxin reductase A

3.3.1.1 Secretory Isoforms

For the secretory isoforms assigned to each protein, we can also summarise based on which have been experimentally observed observation. **Appendix Table B.1** includes all combinations of the SP, TM, and GPI features (as well as retention signals) that are either present in all proteins of a gene (annotated “Yes”), absent (annotated “No”), or present only in some proteins (annotated “Iso”). The total observed secretory isoforms represents approximately 7% of all observations, or on average 11.43% when considering the proportion on a per experiment basis. Given these proteins represent a consistent portion of observed experimental results, the frequency warrants consideration when looking at proteins found in the secreted space. From a secretory point of view, they are quite ambiguous when observed - the possibility of secretory isoform could be taken as an obvious explanation that the secreted version has been observed, or it could be the non-secretory version as either contamination or as a LSP. If the experimental observation was derived from MS data, then delving into the mass spectra data can in some cases definitively identify which peptides were observed, and if those peptides resolve to SP regions of the protein, then the ambiguity could be resolved. Given this ambiguity the secretory isoforms are excluded in the classifications.

3.3.2 Classifications

Outcomes of the classification of genes based on observations and secretory properties for analysis of possible LSPs into SEC (647), NONSEC (16850), SPT (2786) and UNCLASSIFIED (1696) is summarized in **Figure 3.4**. The result is that the majority of the proteome is unobserved in the studies included in the LSPDB. From the observed categories (SEC, UNCLASSIFIED) a total of 72% of the observed proteins lack SP, exceeding the 50% estimates (Agrawal et al., 2010; Krause et al., 2013). Interpretations of these data needs to be tempered by the fact that a few studies contributed a large portion of the observations, and that the excluded proteins from this categorisation (secretory isoforms, or proteins with many TM helices which would relate to a Golgi-bypass form of UPS) means that the percentages are potentially skewed. Nevertheless, the UNCLASSIFIED category repre-

	OBSERVED	UNOBSERVED
+ SECRETORY FEATURES	 SEC (647) Secretory	 SPT (2786) Signal Peptide/ secreted in theory
- SECRETORY FEATURES	 UNCLASS (1696) Unclassified, LSP or contamination	 NONSEC (16850) Non-secretory

Figure 3.4: Classification categories from the LSPDB v1.0 database

sents a sizeable number of proteins to consider in the search for LSPs.

3.3.2.1 PFAM domains

The collective features of these four classifications can be used to identify either unique or common aspects of the proteins within the groups. For PFAM domains, we can collect all of the terms observed in proteins in each class and consider if they are unique or shared between the classes. In **Figure 3.5** genes common and unique to each of SEC, NONSEC and UNCLASSIFIED are shown. The intersections of the groups indicate that each class has a different proportion of shared and unique. The majority in 69% (153/3099) of PFAM in NONSEC are unique, whereas the opposite is true for UNCLASSIFIED where 78% (799/1025) are shared. We can also take the hypothetical secretory proteins (SPT) as secreted, and compare a combined SEC and SPT to the other categories (**Figure 3.6**) giving a similar distribution. In particular, domains shared only between SEC and UNCLASSIFIED classes can be used to look for LSPs under the hypothesis that protein function

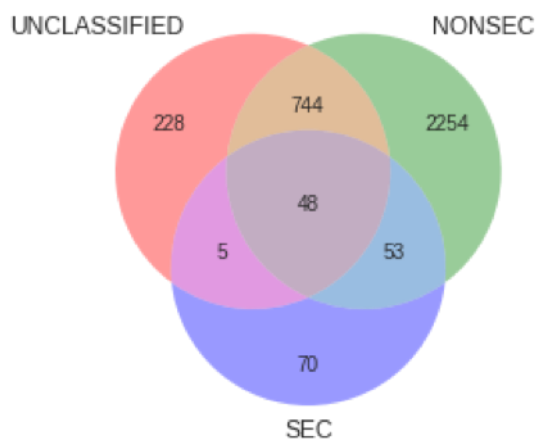


Figure 3.5: Venn diagram of distinct PFAMs in each of three classes (SEC, NONSEC, UNCLASSIFIED) in the LSPDB

relevant to the secretory space is conserved. The overlap between SEC and UNCLASSIFIED is consistent with the extended SEC plus SPT yielding two more PFAM domains for a total of 7 that are predominantly secretory listed in **Table 3.4**. Notably, PF00251 and PF08244 are glycosyl hydrolases that hydrolyse the glycosidic bond of polysaccharides.

Having identified these PFAM domains, we can search the UNCLASSIFIED category to extract LSP candidates. Essentially, this identifies proteins that contain the PFAM domains that have only appeared in experimentally observed proteins, both with and without a SP. From the UNCLASSIFIED category this results in 6 proteins without secretory features that possess these domains and have been observed. One criticism of exclusive criteria is that it is susceptible to domains that are unevenly distributed across the categories, for example a PFAM domain that appears predominantly in SEC and UNCLASSIFIED would be excluded if it also appears in a single NONSEC protein. Since we could expect that not all potential LSPs have been observed, there may be numerous proteins in NONSEC that could skew this approach. We can therefore relax the criteria further to select PFAM domains with the majority of their appearances within a category. Of interest are the domains that are mostly (as opposed to exclusively) secretory. These are listed in **Table 3.5**. A further glycosyl hydrolase domain (PF00232) with cell wall functionality is identified using this approach, though some additional domains either have unknown function or are false-positives.



Figure 3.6: Venn diagram of distinct PFAMs in each of four classes (SEC, SPT, NONSEC, UNCLASSIFIED) in the LSPDB, with SEC and SPT combined

Firstly, a domain of unknown function (DUF) in DUF538 (we return to this in Section 5.5) and secondly a photosystem I reaction centre subunit IV which may be considered a likely false-positive. The photosystem I reaction center subunit IV A protein (PSAE1,AT4G28750) is SEC classified in the LSPDB, with predictions from both SignalP3 and SignalP4 methods. It is however a chloroplast protein, and the chloroplast transit peptide could be the cause of an erroneous SP prediction. The recent DeepLoc and SignalP programs correctly categorise it as plastidic (chloroplast) and with no signal peptide.

This domains approach can similarly be applied to the potential LSPs to infer which observed proteins have these domains. These criteria are also susceptible to bias, as the magnitude of SEC to other categories observations is not taken into account; 2:1 is treated the same as 220:10, but rather than defining an arbitrary threshold they are currently both treat equally.

Another approach to interrogate these categories is to look at the domains that may be unique or predominantly restricted to one in each class, such as the UNCLASSIFIED. These domains may be of interest to identify subsets of potential LSPs or contamination that have a unique signature, however until likely LSPs can be identified in this data whether any such signature relates to LSPs or contamination is unclear.

Of note are also the large number of families shared between NONSEC and UNCLASSIFIED, and other combinations such as SEC and NONSEC **Figures 3.5, 3.6**. These multi-

grouping PFAMs are not directly applicable to identifying secretory properties. Given that proteins can contain more than one domain, and not all are going to be relevant to secretion, these are not used to further categorise proteins.

Table 3.4: PFAM domains occurring exclusively in the SEC, SPT and UNCLASSIFIED classified proteins

PFAM	Symbol	Description
PF00251	Glyco_hydro_32N	Glycosyl hydrolases family 32 N-terminal domain
PF02427	PSI_PsaE	Photosystem I reaction centre subunit IV / PsaE
PF08244	Glyco_hydro_32C	Glycosyl hydrolases family 32 C terminal
PF14368	LTP_2	Probable lipid transfer
PF14543	TAXi_N	Xylanase inhibitor N-terminal
PF16499	Melibiase_2	Alpha galactosidase A
PF16845	SQAPI	Aspartic acid proteinase inhibitor

Table 3.5: PFAM domains occurring in the majority in the SEC classified proteins

PFAM	Symbol	Description
PF00141	peroxidase	Peroxidase
PF00190	Cupin_1	Cupin
PF00232	Glyco_hydro_1	Glycosyl hydrolase family 1
PF01156	IU_nuc_hydro	Inosine-uridine preferring nucleoside hydrolase
PF02427	PSI_PsaE	Photosystem I reaction centre subunit IV / PsaE
PF04398	DUF538	Protein of unknown function, DUF538
PF07676	PD40	WD40-like Beta Propeller Repeat
PF16499	Melibiose_2	Alpha galactosidase A
PF16845	SQAPI	Aspartic acid proteinase inhibitor

3.3.2.2 GO terms

For GO terms we can apply the same principles between the groups of exclusive and majority membership. We consider the overlap and extended overlap with SEC+SPT in **Figure 3.7** and **Figure 3.8**, respectively. This identifies 11 (up to 21 with SPT included) GO terms indicative of secretion listed in **Table 3.6**. As with the PFAM domains, there are mixes of terms that would appear to be canonically involved in cytoplasmic process, yet their additional presence in SP containing proteins identifies them here.

Table 3.6: GO terms occurring exclusively in the SEC, SPT and UNCLASSIFIED classified proteins

GO ID	Description
GO:0004034	aldose 1-epimerase activity
GO:0004128	cytochrome-b5 reductase activity, acting on NAD(P)H
GO:0004557	alpha-galactosidase activity
GO:0006212	uracil catabolic process
GO:0006491	N-glycan processing

GO ID	Description
GO:0009311	oligosaccharide metabolic process
GO:0009805	coumarin biosynthetic process
GO:0009807	lignan biosynthetic process
GO:0010037	response to carbon dioxide
GO:0010306	rhamnogalacturonan II biosynthetic process
GO:0016813	hydrolase activity
GO:0019318	hexose metabolic process
GO:0019483	beta-alanine biosynthetic process
GO:0033499	galactose catabolic process via UDP-galactose
GO:0034722	gamma-glutamyl-peptidase activity
GO:0042344	indole glucosinolate catabolic process
GO:0047622	adenosine nucleosidase activity
GO:0047724	inosine nucleosidase activity
GO:0090506	axillary shoot meristem initiation
GO:1903426	regulation of reactive oxygen species biosynthetic process
GO:2000122	negative regulation of stomatal complex development

Applying these to the UNCLASSIFIED set indicates 20 proteins of interest. We also look at the GO terms in the majority in SEC, using the same reasoning as expanding the criteria for PFAM domains. These are listed in **Table 3.7**, and several terms that could be applicable to secretory proteins, such as protein deglycosylation (GO:0006517) appear. One notable difference with GO terms as opposed to PFAM domains is their origin; rather than being sequence based, they can be annotated from a variety of sources. One term identified as a SEC majority GO term that illustrates this point is the cell wall, GO:0005618; counts of SEC:198, SPT 122, UNCLASSIFIED 34 and NONSEC 197, respectively. The annotation of various proteins to this GO term originates in some cases from the very studies we consider as observation input to the LSPDB, which are correctly cell wall studies, however the nuances around contamination versus secretion are ignored. This places a limit on the usefulness of GO terms, whereas the sequence based PFAM

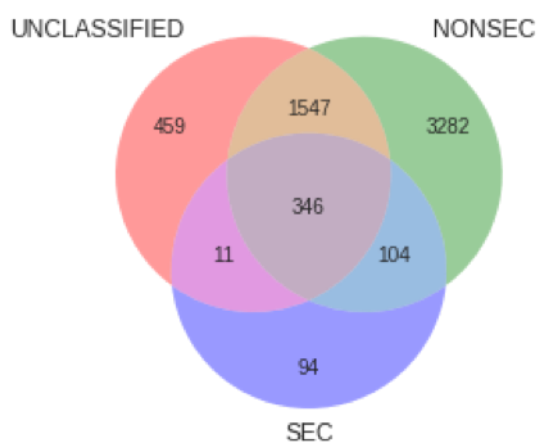


Figure 3.7: Venn diagram of distinct GO terms in each of three classes (SEC, NONSEC, UNCLASSIFIED) in the LSPDB

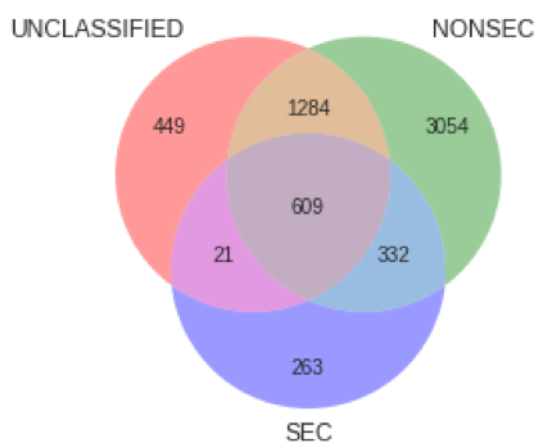


Figure 3.8: Venn diagram of distinct GO terms in each of four classes (SEC, SPT, NON-SEC, UNCLASSIFIED) in the LSPDB, with SEC and SPT combined

domains are not influenced by the associated publications they are observed in.

Table 3.7: GO terms occurring in the majority in the SEC classified proteins

GO ID	Description
GO:0004128	cytochrome-b5 reductase activity, acting on NAD(P)H
GO:0004557	alpha-galactosidase activity
GO:0004601	peroxidase activity
GO:0004869	cysteine-type endopeptidase inhibitor activity
GO:0005618	cell wall
GO:0006517	protein deglycosylation
GO:0006751	glutathione catabolic process
GO:0008061	chitin binding
GO:0009505	plant-type cell wall
GO:0010466	negative regulation of peptidase activity
GO:0016614	oxidoreductase activity, acting on CH-OH group of donors
GO:0016756	glutathione gamma-glutamylcysteinyltransferase activity
GO:0019318	hexose metabolic process
GO:0019762	glucosinolate catabolic process
GO:0030414	peptidase inhibitor activity
GO:0034722	gamma-glutamyl-peptidase activity
GO:0042344	indole glucosinolate catabolic process
GO:0042744	hydrogen peroxide catabolic process
GO:0045735	nutrient reservoir activity
GO:0047622	adenosine nucleosidase activity
GO:0047724	inosine nucleosidase activity
GO:0052692	raffinose alpha-galactosidase activity

3.3.3 Networks

The network layer of the LSPDB takes protein-protein interactions (PPIs) for *Arabidopsis thaliana* from the AtPIN interaction network, with 95,841 interactions combining yeast two-hybrid and other experimental methods. The driving hypothesis is that LSPs could interact with other secretory proteins, as supplementary evidence of a genuine extracellular function, and/or a possible method for secretion through a protein complex.

3.3.3.1 Interactions with secretory proteins

A subset of interactions was further analysed for interactions that occur between the inferred classifications of SEC, SPT, NONSEC and UNCLASSIFIED. For each of the proteins in these classes, the number of interactions with remaining classes was assigned to an interaction summary. Summaries for 11,404 genes of the 21,656 inferred classes were made. This subset of proteins with interactions can be analysed for interaction patterns in each class. Initially we consider the number of interactions to SEC and NONSEC (**Figure 3.9**), also shown split by network class (**Figure 3.10**). We exclude some genes with more than 200 interactions (AT3G62870, AT5G52640, AT3G12280) from visualisation.

The distribution of the interactions shows two clear patterns for the expected classes of SEC and NONSEC; proteins with secretory features interact with other secretory proteins with greater prevalence than non-secretory proteins (at most 33), and non-secretory proteins have a fewer interactions with secretory proteins than with themselves (over 150 in some cases), predominantly 1 or 2 PPI in the SEC proteins. Partly due to the class imbalance the non-secretory role the second pattern is not entirely unexpected, however the fact that fewer SEC proteins do not largely interact with NONSEC proteins is despite this imbalance. On the assumption that these two classes exhibit interaction patterns consistent with the underlying theory that interactions with like proteins occur within classes, we looked further at the other classes.

The two unexpected observed classes, SP_THEORY and UNCLASSIFIED, have similar SEC/NONSEC ratios to their paired classes of SEC and NONSEC, respectively. This suggests that a proportion of the SPT have similar interaction partners to SEC and could



Figure 3.9: Scatter plot of the number of PPIs to SEC and NONSEC proteins for each protein, coloured by class

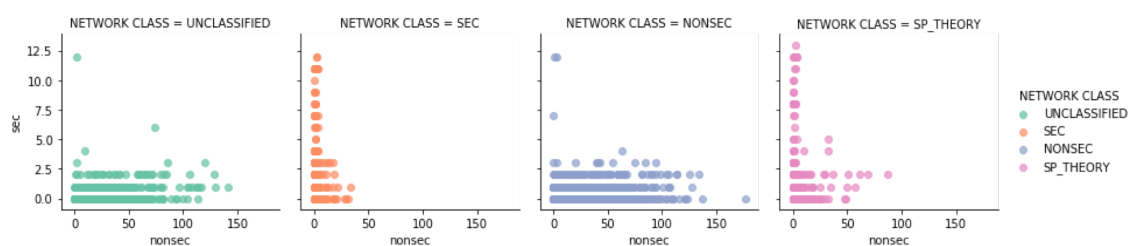


Figure 3.10: Scatter plot of the number of PPIs to SEC and NONSEC proteins for each protein, divided by class

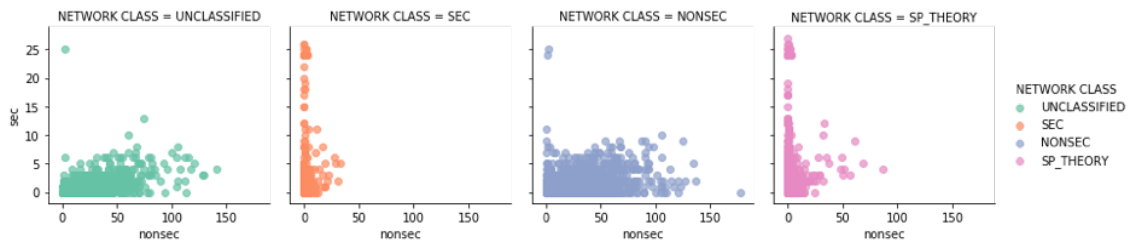


Figure 3.11: Scatter plot of the number of PPIs to SEC+SPT and NONSEC proteins for each protein, coloured by class

be expected to be observed in future studies. The majority of UNCLASSIFIED appear similar to NONSEC in terms of the large number of NONSEC interactions. From a LSP perspective this does not rule them out, however from a network perspective the evidence is minimal in those cases. There are also notable outliers in UNCLASSIFIED with more SEC interactions similar to the SEC profile.

We can also consider a combined metric for SEC+SPT interactions compared to NONSEC in the same manner as PFAM domains and GO term class membership, plotting combined interactions SEC+SPT versus NONSEC (**Figure 3.11**).

In this combined comparison, the patterns are predominantly emphasised, indicating the SPT as SEC assumption is generally held. This assumption of equivalence between SEC and SPT can be explored by comparing the relationship between SEC and SPT interactions (**Figure 3.12**). The upper right quadrant represents proteins with positive interactions for both, with guide lines at 5 interactions in each as an arbitrary line to identify genes with interactions to both.

We can take these upper quadrant proteins and further filter by examining those with few (5 or less) NONSEC interactions and identify SPT (41), NONSEC (2) and UNCLASSIFIED (1) proteins that fit the profile for secretion with a SEC+SPT score of at least 10 interactions (summarised in **Table 3.8**). Many proteins from the same family occur in this list, including beta-glucosidases that act on glucosidic bonds, cysteine proteinases superfamily proteins that have plant defence functions (Balakireva and Zamyatnin, 2018) and the glycoprotein serine carboxypeptidase-like.

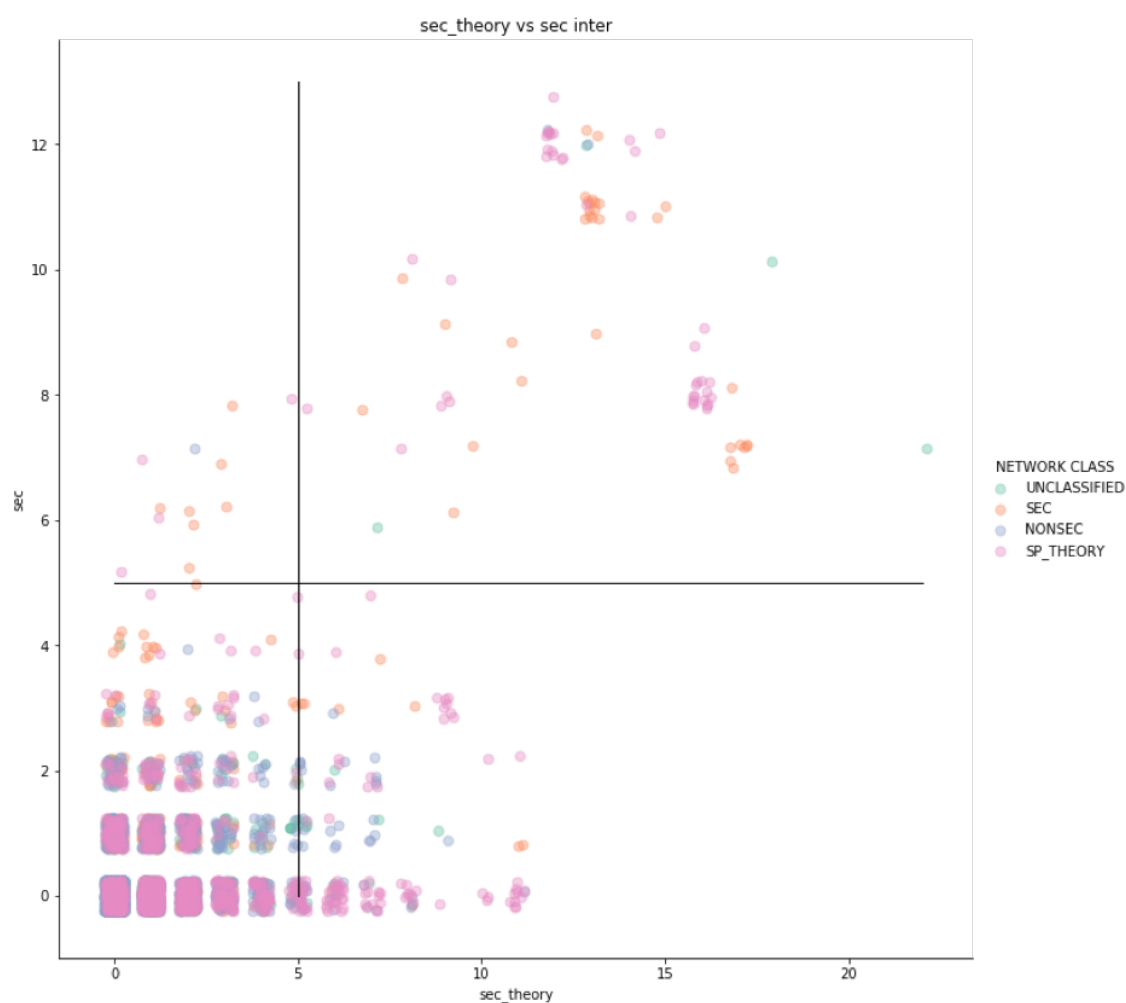


Figure 3.12: Scatter plot of the number of PPIs to SEC and SPT proteins for each protein, coloured by class

Table 3.8: Proteins from SPT, NONSEC and UNCLASSIFIED with more than 5 SEC and 5 SPT interactions

Accession	Description	SPT	SEC	UNCLASSIFIED	NONSEC	CLASSS
AT3G60120	beta glucosidase 27	13	12	0	2	UNCLASSIFIED
AT1G51490	beta glucosidase 36	12	12	1	1	NONSEC
AT5G36890	beta glucosidase 42	13	12	3	3	NONSEC
AT1G11080	serine carboxypeptidase-like 31	9	10	0	0	SP_THEORY
AT1G75940	Glycosyl hydrolase superfamily protein	12	12	1	3	SP_THEORY
AT2G34080	Cysteine proteinases superfamily protein	16	8	0	0	SP_THEORY
AT3G19400	Cysteine proteinases superfamily protein	16	9	0	0	SP_THEORY
AT3G52010	serine carboxypeptidase-like 37	13	11	0	1	SP_THEORY
AT4G11320	Papain family cysteine protease	16	9	0	0	SP_THEORY
AT4G16190	Papain family cysteine protease	16	8	0	1	SP_THEORY
AT4G23520	Cysteine proteinases superfamily protein	16	8	0	0	SP_THEORY
AT5G42260	beta glucosidase 12	12	12	1	2	SP_THEORY
AT5G44640	beta glucosidase 13	12	12	1	2	SP_THEORY
AT5G45890	senescence-associated gene 12	16	8	0	0	SP_THEORY
AT1G73300	serine carboxypeptidase-like 2	9	8	0	0	SP_THEORY
AT5G50260	Cysteine proteinases superfamily protein	16	8	0	2	SP_THEORY
AT3G25420	serine carboxypeptidase-like 21	8	10	0	0	SP_THEORY
AT3G52000	serine carboxypeptidase-like 36	14	11	0	2	SP_THEORY

Accession	Description	SPT	SEC	UNCLASSIFIED	NONSEC	CLASS
AT2G44470	beta glucosidase 29	12	12	1	2	SP_THEORY
AT2G05850	serine carboxypeptidase-like 38	14	12	0	1	SP_THEORY
AT5G48375	thioglucoside glucosidase 3	12	12	1	1	SP_THEORY
AT3G62740	beta glucosidase 7	12	12	1	2	SP_THEORY
AT1G60090	beta glucosidase 4	12	12	1	2	SP_THEORY
AT3G17180	serine carboxypeptidase-like 33	15	12	0	0	SP_THEORY
AT2G27420	Cysteine proteinases superfamily protein	16	8	0	0	SP_THEORY
AT3G12230	serine carboxypeptidase-like 14	9	8	0	0	SP_THEORY
AT5G28510	beta glucosidase 24	12	12	1	2	SP_THEORY
AT2G21430	Papain family cysteine protease	16	8	0	0	SP_THEORY
AT1G29090	Cysteine proteinases superfamily protein	16	8	0	0	SP_THEORY
AT1G06260	Cysteine proteinases superfamily protein	16	8	0	1	SP_THEORY
AT1G73270	serine carboxypeptidase-like 6	9	8	0	0	SP_THEORY
AT1G26560	beta glucosidase 40	12	13	1	2	SP_THEORY
AT3G49340	Cysteine proteinases superfamily protein	16	8	0	0	SP_THEORY
AT3G60140	Glycosyl hydrolase superfamily protein	12	12	1	3	SP_THEORY
AT3G21370	beta glucosidase 19	12	12	1	3	SP_THEORY
AT1G29080	Papain family cysteine protease	16	8	0	0	SP_THEORY
AT1G29110	Cysteine proteinases superfamily protein	16	8	0	0	SP_THEORY

Accession	Description	SPT	SEC	UNCLASSIFIED	NONSEC	CLASS
AT1G61130	serine carboxypeptidase-like 32	14	12	0	0	SP_THEORY
AT3G54940	Papain family cysteine protease	16	8	0	2	SP_THEORY
AT1G73290	serine carboxypeptidase-like 5	8	7	0	0	SP_THEORY
AT4G22100	beta glucosidase 2	12	12	1	2	SP_THEORY
AT3G48350	Cysteine proteinases superfamily protein	16	8	0	1	SP_THEORY
AT3G43960	Cysteine proteinases superfamily protein	16	8	0	1	SP_THEORY
AT5G24550	beta glucosidase 32	12	12	1	2	SP_THEORY

The comparison of interactions with the UNCLASSIFIED proteins is also of interest. Extending the scatter plots to all combinations (**Figure 3.13**), we visualise the relationship between each class.

By visualising all the interaction relationships, the symmetry between SEC and SPT, and NONSEC and UNCLASSIFIED reinforces that there are few UNCLASSIFIED with SEC-like network profiles, and most seem to be NONSEC-like. As with previous comparisons, there are outliers in each comparison, indicating that the exceptions are few and worth further consideration as LSP candidates.

3.3.3.2 Networks of classified proteins

Extending the network interactions we can consider more than the ratios, and take advantage of the network structure. The combined PPI networks with LSPDB data essentially relate objects (nodes: proteins) by relationships (edges: protein-protein interaction), with added annotation of classifications at the node level. Such proteins (11,530) with network classifications form a network with 70,909 edges, of which 1,378 of the nodes are UNCLASSIFIED proteins. These nodes of interest can be selected to create sub-networks, where the degree (number of connections) and the classes of those connections in a local network neighbourhood can be considered.

From the interaction count plots, few NONSEC and UNCLASSIFIED proteins exceed three interactions with SEC proteins. Isolating proteins with 3 or more such interactions in the LSPDB network identifies 180 proteins (including SEC and NONSEC proteins). We can consider the immediate network neighbourhood of these proteins to form a sub-network of 2,092 proteins and can further reduce the network to those interactions between the candidates for secretion, excluding the NONSEC class. This network of 166 proteins (1,079 edges) can be further broken down to the parts containing potential LSPs (123 nodes, 991 edges) (see **Figure 3.14**).

Table 3.9: UNCLASSIFIED proteins in identified in **Figure 3.14**

Accession	Description
AT1G24320	Six-hairpin glycosidases superfamily protein

Accession	Description
AT4G26840	small ubiquitin-like modifier 1
AT1G78300	general regulatory factor 2
AT1G19120	Small nuclear ribonucleoprotein family protein
AT5G42980	thioredoxin 3
AT3G60120	beta glucosidase 27
AT5G52640	heat shock protein 90.1

This network represents an interconnected series of proteins that are both observed and unobserved, with a mix of secretory features. Of note are several highly interconnected sub-clusters of SEC and SPT proteins, with potential LSPs linking the network. SPT proteins were not specifically selected for and appear due to their direct interactions as first neighbours. The 7 unclassified proteins identified in this method are listed in **Table 3.9**. In the search for LSPs, this approach again yields results that includes proteins with well known functions in non-secretory locations. For example, AT5G52640 (a heat shock protein (HSP)) is included in this list, which was previously excluded due to the high number of PPI. It, along with thioredoxin are implicated in other eukaryotic UPS, and it is of interest that both appear in this sub-network.

We can also consider this sub-network and its composition. BINGO (Maere et al., 2005) is a method for identifying network components that are over-represented compared to the entire proteome, and terms with an adjusted p-value after multiple hypothesis testing of less than 0.01 are shown in **Figure 3.15**. Terms include hydrolase activities, acting on glycosidic bonds and di-sulphide bonds indicating extracellular utility. These candidates represent the outliers from **Figure 3.11 UNCLASSIFIED** and give context to what kind of proteins are selected. We can also consider the other outliers, namely in the NONSEC category for proteins as yet unobserved that exhibit this increased number of SEC interactions. Selecting those with 3 or more SEC neighbours, and expanding the network to include all nearest neighbours produces a network of 3 sub-components (see **Figure 3.16**).

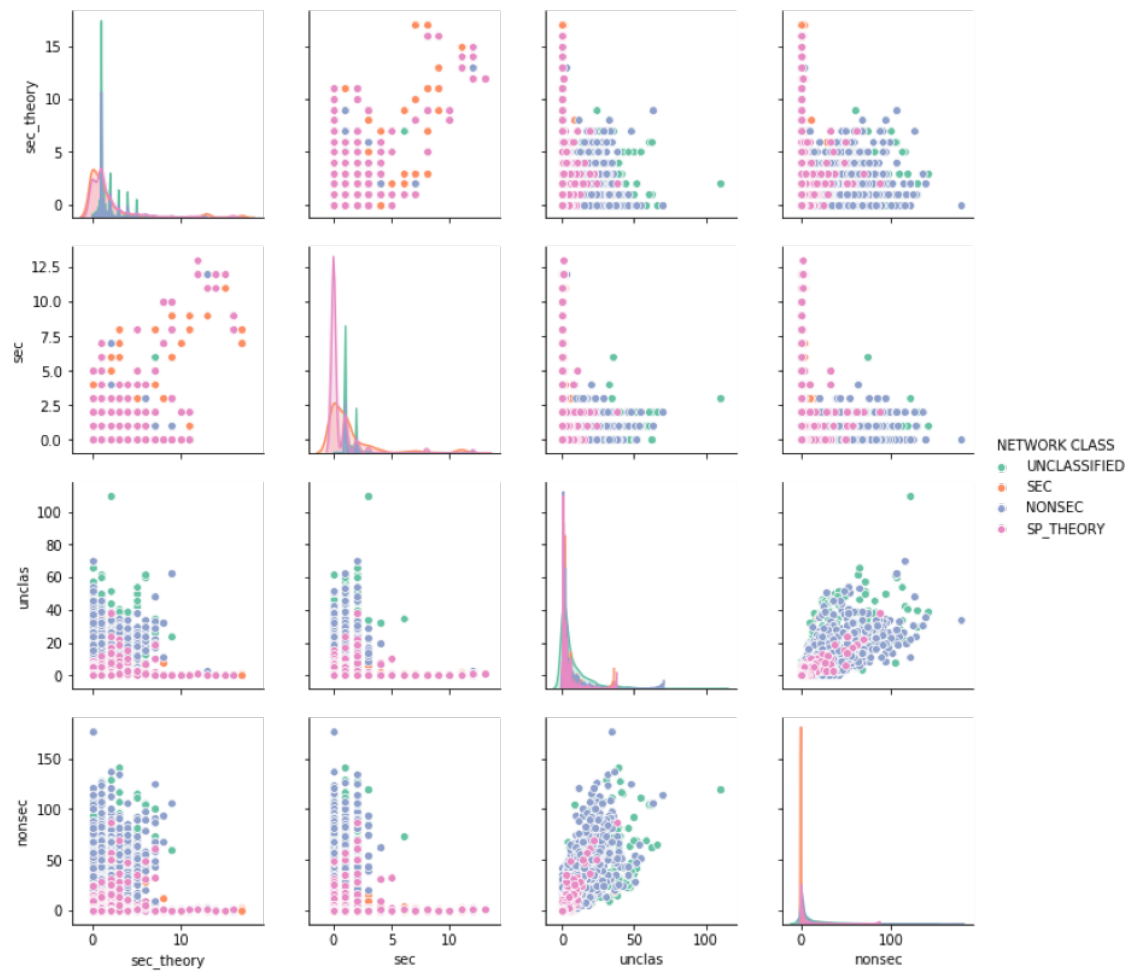


Figure 3.13: Matrix of scatter plots of the number of PPIs to each combination of class of LSPDB proteins, coloured by class

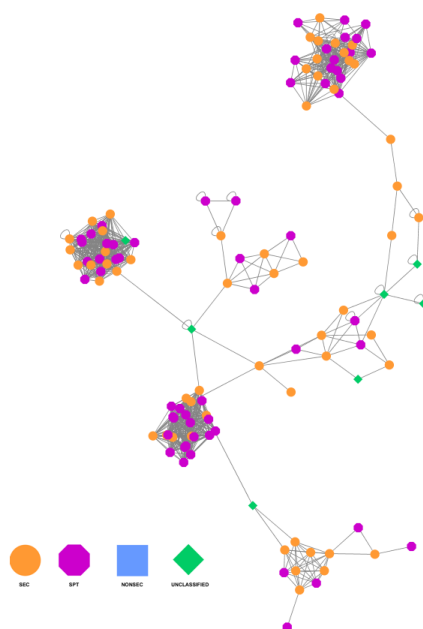


Figure 3.14: Clusters of proteins identified from the LSPDB protein-protein with greater than 3 SEC interactions that contain potential LSPs. UNCLASSIFIED (green diamond); SEC (orange circle); SPT (purple hexagon); NONSEC (blue square).

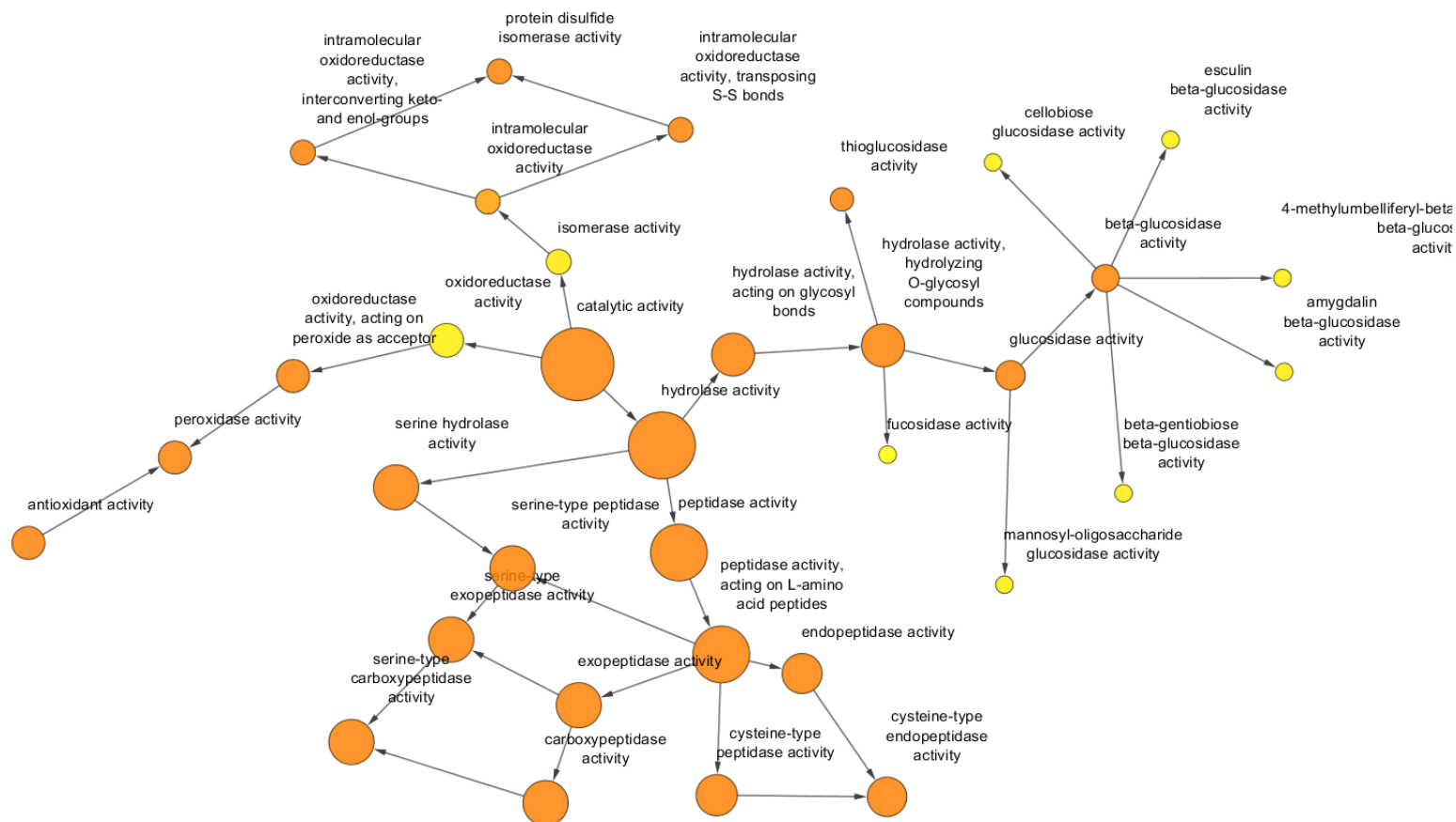


Figure 3.15: BINGO analysis of Figure 3.14 proteins, with over-represented molecular function terms

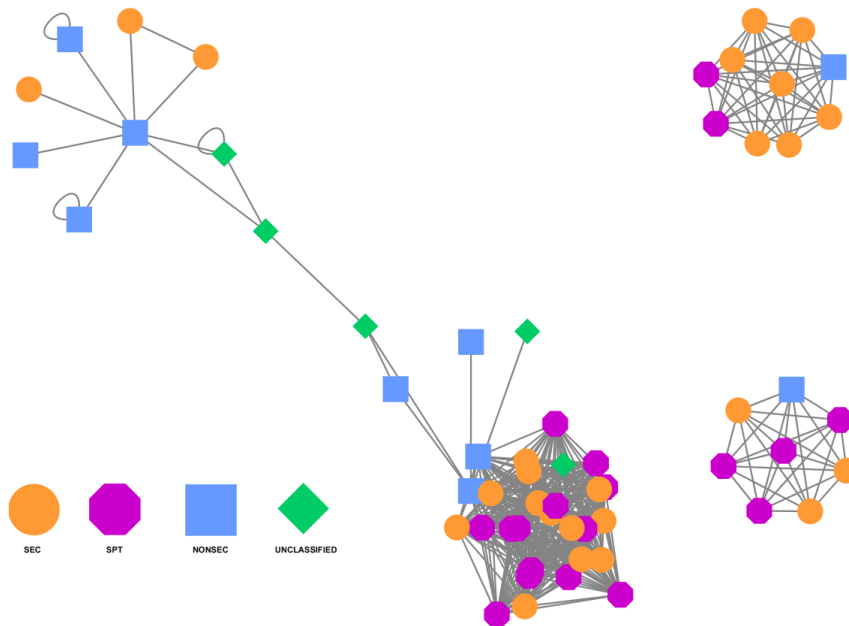


Figure 3.16: NONSEC with 3 or more SEC neighbours from base PPI network. UNCLASSIFIED (green diamond); SEC (orange circle); SPT (purple hexagon); NONSEC (blue square). .

The appearance of any of these proteins in studies that are added to the LSPDB will render them LSP candidates. The local neighbours also represent an advantage of this network approach, as to only look at certain classes loses the nuances, such as either the interactions between the same class or in this case non-secreted protein neighbours. In this network, five further NONSEC proteins are included that do not interact directly with SEC (**Table 3.10**). Comparing these potential LSPs - if they are observed - to the previous results shows similarity in results with two further beta-glucosidases identified. There are also novel candidates that do not appear to have a non-canonical function that remain to be demonstrated, such as the chloroplastic ADC synthase superfamily protein (AT1G74710), though members of this family have implicated in plant pathogen defence (Rossi et al., 2015).

Table 3.10: NONSEC proteins identified in Figure 3.19

Accession	Description
AT5G36890	beta glucosidase 42

Accession	Description
AT1G51490	beta glucosidase 36
AT5G42230	serine carboxypeptidase-like 41
AT4G32320	ascorbate peroxidase 6
AT5G39410	Saccharopine dehydrogenase
AT1G10170	NF-X-like 1
AT1G26910	Ribosomal protein L16p/L10e family protein
AT3G53830	Regulator of chromosome condensation (RCC1) family protein
AT5G40490	RNA-binding (RRM/RBD/RNP motifs) family protein
AT1G74710	ADC synthase superfamily protein

We can also look at the local networks starting from SEC proteins. From the PPI scatter plots, we observe that the upper limit of SEC interactions is around 10, so we can select just these proteins initially, then expand to nearest neighbours (**Figure 3.17**). To avoid hyper-interacting proteins, those with less than 100 neighbours in total are included.

These two distinct clusters include 3 candidates (2 NONSEC, 1 UNCLASSIFIED) that have already been mentioned: AT3G60120 (UNCLASSIFIED), AT5G36890 and AT1G51490 (NONSEC). Again applying BINGO reveals that these clusters are also over-representative for carboxypeptidase and glycosyl bond hydrolysis, beta-glucosidase activities.

The repetition of these candidates reinforces one limitation of approaching networks in this way, namely that multiple subsets of the entire PPI network can appear quite distinct when certain nodes or edges are excluded. However different network queries with overlapping results can reinforce results and view potentially useful targets. We can also combine classes. Considering a broader criteria, informed by the distribution of all interactions, select any NONSEC or UNCLASSIFIED with at least 4 SEC interactions, plus their immediate neighbours (**Figure 3.18**).

Here we observed 4 distinct sub-clusters, including the carboxypeptidase and glycosyl hydrolases, with new clusters that exhibit clear NONSEC or UNCLASSIFIED dominance. In future, this approach beyond single neighbours could be extended more formally. In the previous approaches we consider the 0th order network - direct interactions.

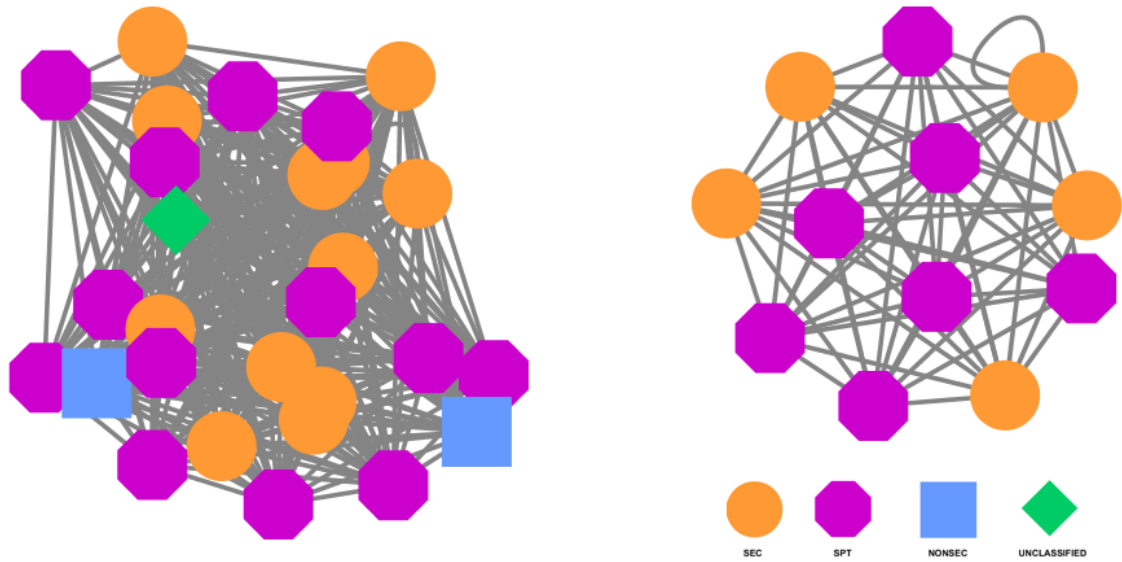


Figure 3.17: Clusters of proteins identified from the LSPDB with 10 or more SEC interactions. UNCLASSIFIED (green diamond); SEC (orange circle); SPT (purple hexagon); NONSEC (blue square).

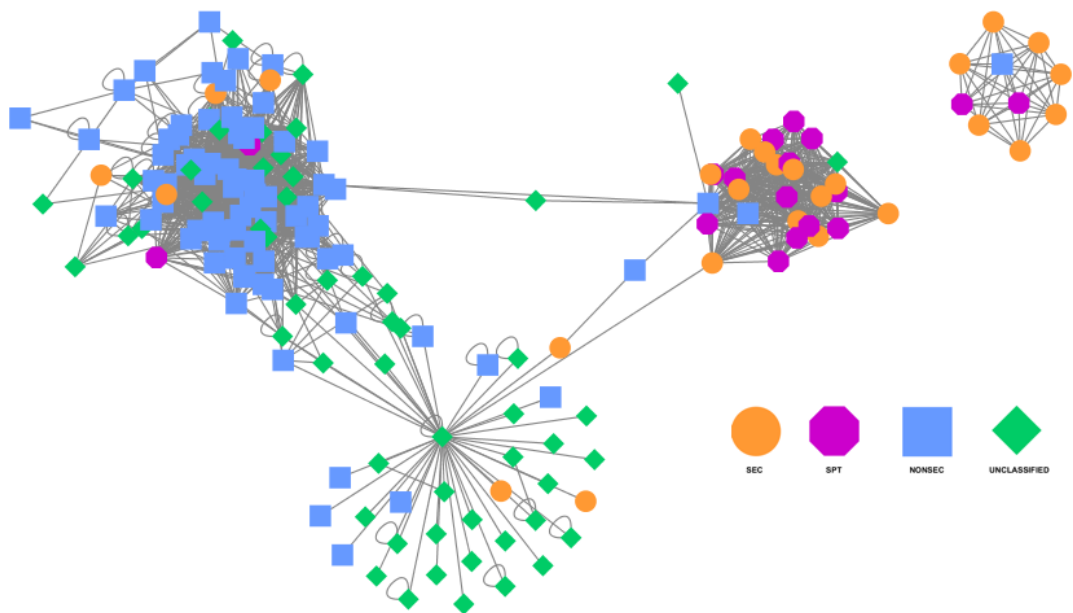


Figure 3.18: Network of proteins identified from the LSPDB that are NONSEC or UNCLASSIFIED with 4 or more immediate SEC interactions, with immediate neighbours shown. UNCLASSIFIED (green diamond); SEC (orange circle); SPT (purple hexagon); NONSEC (blue square).

For proteins with few interactions the above criteria may not yield insight, for example UNCLASSIFIED with only other UNCLASSIFIED interactions. However, the network topology may provide more information, by considering the immediate neighbours (1st order network) of those neighbours. That is, since we can measure how many network connections to each class, for potential LSPs we can ask how many interactions before a SEC neighbour is reached? This is a less strict criteria, and necessarily requires some high interacting proteins to be excluded to prevent the entire network satisfying a condition. In practice we have identified several proteins this way by using the first neighbours, but a more complex approach based on network topology could be used in future.

3.3.3.3 Co-occurrence within the PPI network

One limitation of looking at the PPI network only is that we do not take advantage of the experimental context embedded in the LSPDB, specifically which proteins and therefore which interactions have occurred in the same experiment. Consider UNCLASSIFIED proteins that interact only with other UNCLASSIFIED. In an isolated network, with no other PPI available, it does not meet our criteria for a SEC interaction. However, if these proteins (or clusters of them) occur repeatedly then the proteins could be treated as a complex, and the frequency of their observance re-evaluated as a unit.

Given our network, we can reinforce the relationships between proteins to help with this approach by creating a new network from the classified PPI and considering each experiment :

For each experiment:

- for all proteins in experiment:
 - find all network pairs in the PPI from that experiment:
 - create new network nodes if required
 - create new network edge if required
 - add 1 to weight

We can use these reinforced PPI links to narrow down the “network hairball” into smaller groups. Note that by definition the co-occurrence graph can only include either

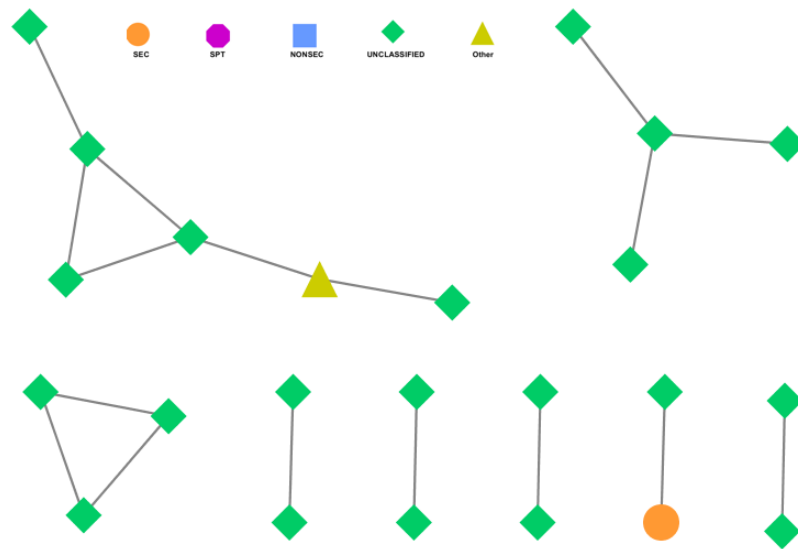


Figure 3.19: Reinforced network with protein-protein interactions, and co-occurrence with most observed edges involving UNCLASSIFIED (green diamond) proteins. SEC (orange circle); SPT (purple hexagon); NONSEC (blue square).

SEC, UNCLASSIFIED, or other observed proteins with other features such as TM domains and secretory isoforms. Each weight to an edge represents a PPI confirmed in a cell wall proteome, 1,431 proteins (4,846 edges) fit this criteria, of which 1,294 (4,311 edges) are either SEC or UNCLASSIFIED.

This reduces the LSPDB network and extends the frequency results. The most highly observed PPI is between two SEC proteins, AT1G09750 and AT5G07030, that are members of the aspartyl protease family proteins. Using this approach, we can rank our network based on the weight of each edge, and take the top 10% of edge weights that include at least one UNCLASSIFIED protein (**Figure 3.19**). This collection of sub-networks includes 8 sub-networks of 23 protein (17 edges) represents the PPI interactions that have passed a filter of occurring in the same experiment (**Table 3.11**). Since this approach uses weights, the number of occurrences is therefore related to frequency of observation, and as with the proteins in **Table 3.3**, defer further analysis in order to compare to frequency independent approaches.

Table 3.11: UNCLASSIFIED proteins identified in Figure 3.22

Accession	Description
AT1G45145	thioredoxin H-type 5
AT4G34870	rotamase cyclophilin 5
AT2G16600	rotamase CYP 3
AT3G52930	Aldolase superfamily protein
AT3G11630	Thioredoxin superfamily protein
AT5G18170	glutamate dehydrogenase 1
AT5G07440	glutamate dehydrogenase 2
AT2G36530	Enolase
AT4G38740	rotamase CYP 1
AT3G20390	endoribonuclease L-PSP family protein
AT2G18020	Ribosomal protein L2 family
AT3G49910	Translation protein SH3-like family protein
AT3G03780	methionine synthase 2
AT5G17920	Cobalamin-independent synthase family protein
AT2G44350	Citrate synthase family protein
AT3G18740	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein
AT1G04410	Lactate/malate dehydrogenase family protein
AT1G47128	Granulin repeat cysteine protease family protein
AT3G04120	glyceraldehyde-3-phosphate dehydrogenase C subunit 1
AT1G47710	Serine protease inhibitor (SERPIN) family protein
AT1G13440	glyceraldehyde-3-phosphate dehydrogenase C2
AT3G57150	homologue of NAP57
AT5G42980	thioredoxin 3

3.3.3.4 Homodimers

PPI interaction data includes proteins that are thought to form homodimers, that is protein with interactions with themselves and have self-edges in the network. These interactions are excluded in the counts and analysis above, i.e a SEC protein with interactions to itself and NONSEC is listed as interacting with NONSEC only. However, if they are included then we can investigate if these occur in different proportions within the groups.

Homodimers were found in the SEC (21), SPT (38), NONSEC (1,118) and UNCLASSIFIED (324) sets. As proportions of the total in each set, 324 from 1,696 total UNCLASSIFIED represents 19.10%, a far larger proportion than for the 3.24%, 6.53% and 1.36% of the SEC, NONSEC and SPT classifications. Chi-squared statistics indicate this is a significant overrepresentation ($p\text{-value} = 1.63 \times 10^{-118}$) and in SPT and SEC under-representation, indicating the homodimers are not distributed evenly between the classifications.

These could be biased by the composition of the interaction data – not all proteins have interaction data and so the proportions may not be truly representative of the total proteome. We recalculated the tests as a proportion of the number of interaction summaries for each protein class. In this case 324/1,378 is still calculated as a significant divergence (Chi-squared $p\text{-value} = 2.60 \times 10^{-22}$) and under-representation for SEC and SPT. In both cases the proportion of homodimers in NONSEC is expected, with no significant difference. The 324 in UNCLASSIFIED were noted however for future analysis, and listed in **Appendix B.5**.

3.4 Discussion

Construction of the LSPDB included an explicit goal to allow for interpretation by the end user, and the results demonstrate several lines of evidence for making LSP decisions. In this section, we compare the results from the use of this hierarchical approach (see **Figure 3.1**), and the implications of the design choices on them in Section 3.4.1. Confidence criteria for each layer in in Section 3.4.2 and Section 3.4.3. A prioritisation scheme for these criteria is outlined in Section 3.4.4, and the properties of subsequent LSP candidates discussed in Section 3.4.5. A case study of user interaction using SAMS2, an EXPO related

LSP, is outlined in 3.4.6.

3.4.1 Comparing the evidence types

The various evidence types in the LSPDB of frequency of protein appearances, GO terms and PFAM domains distribution, and networks of PPI and co-occurrence each provided a different set of results. Though each outcome provides a unique insight into potential LSP candidates, the veracity of each needs to be evaluated critically for bias and assumptions.

The frequency of protein observation was shown in isolation to offer no additional insight into the plausibility of LSP candidates. Repeated observation could be taken equally as evidence of either ongoing UPS or arising from common contamination. The size of proteome results, and bias in the kinds of experiments also needs to be taken into account. Experiments with many conditions provide multiple opportunities for protein appearance, and the influence of a common experimental type in the database could influence which kinds of proteins appear more frequently. At the core of the LSPDB, classification into observations is based on single appearance, so the putative LSPs are in general independent of the number of observations.

GO terms can be based on information from annotations, which often misses the subtlety of whether a protein is an LSP compared to a contamination. That the cell wall GO:0005618 protein is associated with many SEC proteins shows the GO term method works. However its presence as annotation on UNCLASSIFIED proteins is less informative as the annotation is derived from merely being found in the same study as secretory proteins and, hence, it may have classified it as contamination. The completeness of the evidence strategies is also relevant. Network interactions provide strong results for a secretory interaction profile, yet are not exhaustive as many proteins in the classification subsets do not have PPI data. PFAM domains and GO terms are associated with many but not all proteins.

The assumption(s) behind each line of evidence also differ. For example, by looking for either GO or PFAMs shared between SEC and UNCLASSIFIED, commonality is assumed and harks back the common-feature hypothesis. On the other hand, we can examine either PFAM or GO terms unique to the UNCLASSIFIED collection - indicating

possible unique features of LSPs - that are not shared with NONSEC. However, since NONSEC may contain potential LSPs, there is less certainty in this approach. Taking terms that are predominantly in the UNCLASSIFIED could be a future approach, though the distribution of terms between classes would need to be carefully considered, and a more complex approach than the majority used in this work.

Similarly, PPI networks are looking for evidence of interactions with SEC, encouraged by the pattern for SEC-SEC interactions, is another kind of common features approach. They are not mutually exclusive; each is one suggested way to prioritise LSP candidates, so we can proceed with criteria for PFAM, GO and network interactions separately, and collate UNCLASSIFIED proteins that meet the criteria for any of them. One consideration is the class imbalance – NONSEC make up the large majority of proteins in the classes (over 75%), and SEC as a guiding class for common properties, the smallest. Also as noted, uncertainty is inherent in these classes; SPT and UNCLASSIFIED present mixed groups by definition, and NONSEC could contain as either unseen contaminants or LSPs.

This mix of caveats and limitations means that not every protein was visible to every evidence approach. Therefore, a suggested model for the LSPDB is to combine results from each approach when identifying candidates. We can stratify results from each categories based on confidence, and suggest high, medium and low confidence criteria for each and this is discussed in the subsequent sections.

3.4.2 Confidence criteria for GO terms and PFAM domains

Given the similarity of the approaches to PFAM domains and GO term annotation, we used the same logical groups for confidence criteria. As shown in **Figures 3.5** and **3.7**, the highest criteria for PFAM and GO are exclusive items to the SEC & UNCLASSIFIED classes – these can be taken as properties of secretory proteins present in the observed proteins without secretory trafficking features.

To identify medium confidence candidates, we can consider items that in the union of SEC and SPT, with UNCLASSIFIED as shown in **Figures 3.6** and **3.8** Though these proteins may be secreted, they could be retained in either the ER or Golgi, leading to their absence from proteomes thus far. However for the purposes of LSPs, their involvement in

some part of the secretory pathway, and not in the large group of NONSEC proteins, is of interest if they share properties with UNCLASSIFIED proteins, and not with NONSEC.

Finally, we can deal with items present across all classes. Given the uncertainty of putative LSPs as yet unseen, exclusion based solely on presence in the NONSEC class will introduce some false negatives. However, either very common domains or functions may be present across many proteins and so the present SEC class would include false positives. A low priority set can be constructed by taking proteins which exhibit a majority of SEC occurrences. By being observed in the majority of cases in the SEC (which is the smallest class) these protein features are also of interest as they correspond to the proteins containing PFAM domains and GO terms in **Tables 3.5 and 3.7**. In addition to exclusive and majority tests, the most common PFAM domains and GO terms were also calculated. As these will relate to the frequency of observation, as with **Table 3.3** we do not consider them for criteria, but the GO terms and PFAM domains are noted in **Appendix Tables B.4 and B.5** for SEC, **Appendix Tables B.6 and B.7** for SPT, **Appendix Tables B.8 and B.9** for UNCLASSIFIED and **Appendix Tables B.10 and B.11** for NONSEC, respectively. In summary, the confidence criteria, together with the corresponding number of UNCLASSIFIED proteins (in parenthesis) for each class, are listed below:

PFAM requirements

- High: PFAMs exclusive to SEC and UNCLASSIFIED (5 proteins)
- Medium: PFAMs exclusive to SEC, SPT and UNCLASSIFIED (1 protein)
- Low: PFAMs predominantly in SEC, possibly in SPT, must be in UNCLASSIFIED (15 proteins)

GO term requirements

- High: GO exclusive to SEC and UNCLASSIFIED (8 proteins)
- Medium: GO exclusive to SEC, SPT and UNCLASSIFIED (12 proteins)
- Low: GO predominantly in SEC, possibly in SPT, must be in UNCLASSIFIED (186 proteins)

By way of comparison, we can apply these criteria back to the SEC class they are derived from to estimate how well they select the 647 classically secreted proteins. For the PFAM criteria, 190 proteins are recovered (High: 113, Medium: 1, Low: 76). The GO terms recover 301 proteins (High: 10, Medium: 18, Low: 273). Notably the relative order of counts of proteins in High/Medium/Low holds between the SEC and UNCLASSIFIED applications.

3.4.3 Confidence criteria for network methods

For network results, the criteria are less clear. Given the “stickiness” of PPI data (Brettner and Masel, 2012), where protein interaction assays tend to identify interactions *in vitro* that are possible even if not probable *in vivo*, the overall number of interactions means that the exclusive criteria we used for PFAM domains and GO term has no direct analogy. Only two unclassified proteins from the PPI network have exclusive SEC PPI interactions.

This is expected in part due to secretion being a process- there are many non-secretory proteins that may interact along the pathway to secretion. Indeed many of the SEC interact with NONSEC proteins, and is unsuitable to distinguish the unclassified. The low numbers of proteins fitting such strict criteria motives a different approach for network evidence.

Compared to the PFAM and GO terms, where we found a small number of representative domains and terms for the secretory proteins, for PPI interactions with secretory proteins this does not even describe the SEC class of proteins well. Only 57% and 59% of SEC proteins interact with SEC and SPT proteins, respectively. Additionally, 30% and 33% of these interact with some NONSEC (16% overall interact with both SEC and NONSEC). In this context, NONSEC interactions are not a disqualifying criteria for SEC proteins. Contrast with NONSEC: 6% interact with secreted, and 93% interacted with other NONSEC. Also unlike the PFAM domains, we expect interactions either along the secretory pathway or (for dual functioning proteins) interactions based on the non-secretory role. The other network neighbourhood methods (Figures 3.14, 3.16 3.17, 3.18 and 3.19) identify candidates yet do not generalise and are quite selective based on the method at which nodes are selected and then expanded to identify neighbourhoods.

As such, we establish a high-level criteria for network interactions by starting from a position of any SEC interaction for UNCLASSIFIED proteins being of interest. Using the distribution of the known secretory proteins, which can be easily summarised numerically, we can take all 213 UNCLASSIFIED proteins with 1 or more SEC interactions. One pattern of note is the maximum number of connections to NONSEC, which is different between the classes (SEC: 33, SPT: 87, NONSEC: 211, UNCLASSIFIED: 416). By comparing between **Figures 3.10 and 3.11**, we demonstrate that the analogous SEC vs (SEC + SPT) plots show similar distributions. Notably, the SEC classification remains the same with the upper limit on the X axis at 33, so this appears to be a useful profile and is broadly similar to the SEC to SEC+SPT expansion in the previous high and medium splits. Using this cut-off of 33 NONSEC interactions splits our initial set of 213 into high, applying to the 124 UNCLASSIFIED which have SEC interaction and ≤ 33 NONSEC, and the medium set of 89 proteins that have 34 or more NONSEC.

A low confidence class for networks also requires a broad approach to match the GO and PFAM bottom confidence criteria. In the GO and PFAM case, changing the exclusivity requirement for either a GO term or PFAM domain for proteins in the overlap between classes to a simple majority instead identified candidates. Since SEC interactions are already included in the higher criteria, the relaxed restriction for network is to consider proteins that interact with SPT proteins, with caveats on those SPT proteins that they interact with SEC proteins and not NONSEC. This places a limit on the kinds of SPT that are accepted, similar to the filter in **Figure 3.12** to identify the most likely SPTs that are predicted to become SEC. Given this range is using interactions with unobserved proteins, this conservative ruleset identifies 23 UNCLASSIFIED proteins with successful SPT interactions, including 7 that also have SEC interactions and are already in the high/medium sets for 16 unique low network confidence proteins.

Despite the interesting results in the co-occurrence and homodimers analysis, the proteins identified in those are not included in these network criteria. As noted, co-occurrence identifies which pairs (or more) may occur together, but does not infer a cause. The distribution of homodimers across the classes was shown to be uneven, however, whether it could be used to identify leaderless secretion is unclear. There were more homodimers

in UNCLASSIFIED than expected, but the interpretation is unclear. Since most proteins and homodimers come from the NONSEC class, one explanation is that the overrepresentation of these in the UNCLASSIFIED indicates a higher proportion of contaminants in this group. Alternatively, homodimers could represent a signal related to leaderless secretion. Since the NONSEC group could contain proteins that may be LSPs but have not been observed yet, neither explanation is definitive. As noted previously, PPI data can be “sticky” and proteins with many partners are more likely to form homodimers (Brettner and Masel, 2012). Still we note this result for further discussion and exclude these from the high, medium and low criteria below:

- High: All UNCLASSIFIED proteins with SEC interactions matching the profile , >1 sec $\leq$ 33 NONSEC (124)
- Med: All UNCLASSIFIED proteins with SEC interactions matching the profile , >1 sec $>$ 33 NONSEC (89)
- Low: All UNCLASSIFIED proteins with SPT interactions matching the profile of SPT with >1 SEC and 0 NONSEC (16)

3.4.4 Metrics for prioritising candidates

Given the caveats around each of the evidence types, establishing the three confidence categories allow us to take the resulting high/medium/low confidence groups from each and compare which proteins are identified by more than one line of evidence. We can also assign a simple scoring method based those high, medium and low confidence criteria by assigning 3,2 and 1 points, respectively. Additionally, we can assign 0.5 to the homodimer results to note which proteins also appear in this interesting class. Given this simple scoring, we can stratify the resulting overlaps in overall high, medium and low by the following metric:

Class (colour)	Minimal Combination	Score Range
High (green)	High+Low+low or Med+Med+Low or High + Med	≥ 5
Medium (orange)	Low+Low+low or Med+Low or High	3 up to < 5
Low (red)	Low	> 0.5 to < 3

Class (colour)	Minimal Combination	Score Range
None (black)	Homodimer	0.5

Essentially, the top class requires either hits in at least 2 (including one high) or 3 categories (with two medium). The medium and low follow similar patterns. The homodimer class at 0.5 is of interest only, and does not affect the overall category. These ensure that proteins that appear in more than one of the database outputs can be identified. The distribution of these scores is shown in **Figure 3.20**.

The distribution of candidates from each category highlights three points. Firstly, few candidates meet criteria in all three lines of evidence and allow a small subset of higher confidence proteins to be investigated in depth. Secondly, the distribution of homodimers overall is distinct from the other categories, with 117 (36.1%) of the candidates overlapping with one or more of either the PFAM, GO or network methods and often with a low confidence category. This doesn't support the overrepresentation of the homodimers as a leaderless signal. Finally, the total number of primary overlap classes (high, medium, low) of 384 represents 22.6% of the total UNCLASSIFIED proteins indicating the criteria, both individually and combined, cover only a subset of the unclassified proteins and the bar to satisfy them is sufficient to ensure that the method can be used to filter a large number of candidates to a manageable size.

For the purposes of this research, the threshold for these additional sources and inferences need not be high. These LSPs aim to be plausible candidates for secretion, a secondary source to the observation in plant proteomes. We review these below.

3.4.5 Properties of LSP candidates

Applying these criteria to the UNCLASSIFIED proteins in the LSPDB results in 8 high confidence, 149 medium confidence and 227 low confidence candidates listed in **Tables 3.13 and Table 3.14** and **Appendix Table B.2**. These top candidates are examined in the next sections at an individual protein level for high confidence (green), at the group level for medium (orange) , and select observations from the low (red). Experimental

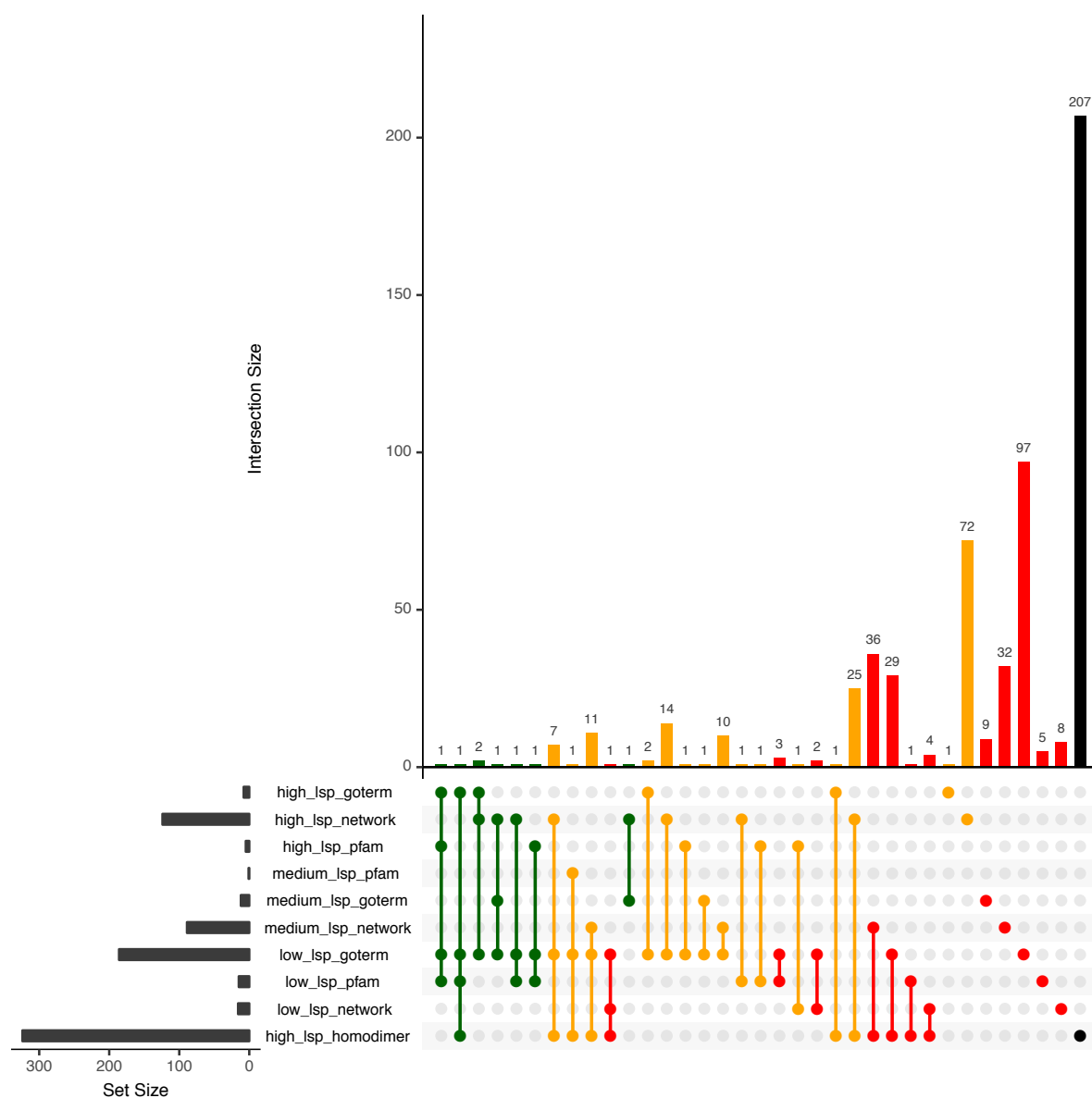


Figure 3.20: Upset plot of high (green), medium (orange), low (red) and homodimer (black) class candidates for LSPs formed by combining the individual confidence criteria from PFAM domains, GO terms and the PPI network.

validation is required for wider assurance in the research community, but in the context of the current state of knowledge of LSPs in plants the list should be considered provisional.

For each candidate LSP, we look at annotation and experimental evidence to support these LSP predictions. In the construction of the LSPDB primary importance has been given to studies of sub-cellular location, i.e. cell wall and apoplastic proteomics studies. Other literature which provides orthogonal evidence for these predictions is also of interest, and the candidates may have appeared in other studies. In particular, cell wall-related studies such as the plasma membrane, plasmodesmata or ER and Golgi may all be relevant to some kind of interaction with the secretory pathway. Prediction program accuracy often vary and are generally not noted, as for many of these proteins, sub-cellular predictions can be obtained for almost any locations depending on the program selected. Given the focus on *Arabidopsis*, annotations in *Arabidopsis*-specific databases such as TAIR, Araport, and SUBA for these proteins are also relevant. Experimental evidence as collated by SUBA includes tagging with fluorescent proteins (FP), MS and PPI data. Since PPI data is already in the LSPDB, only FP and MS experimental evidence (except that used as the basis for these classifications) for protein sub-cellular locations will be noted if available.

3.4.5.1 High confidence

The proteins in **Table 3.13** represent proteins occurring in either three or more of the individual graded confidence classes.

Table 3.13: Putative LSPs from the high (green) category

Accession	Description
AT1G07890	ascorbate peroxidase 1
AT1G24320	Six-hairpin glycosidases superfamily protein
AT2G36310	uridine-ribohydrolase 1
AT3G17940	Galactose mutarotase-like superfamily protein
AT3G26380	Melibiose family protein
AT5G12140	cystatin-1

Accession	Description
AT5G20080	FAD/NAD(P)-binding oxidoreductase
AT5G41520	RNA binding Plectin/S10 domain-containing protein

Ascorbate peroxidase 1 (APX1, AT1G07890) is present in the putative LSP list (Ding et al., 2012) (**Appendix Table A.2**) and inferred by the presence in a cell wall study (Bayer et al., 2006). Canonical function is related to the removal of hydrogen peroxide and is involved in reactive oxygen species (ROS) (Davletova et al., 2005), where it a role was also noted for sensing hydrogen peroxide in concert with heat shock proteins. The presence of AT1G07890 was noted as present in the extracellular space along with other proteins that respond to oxidative stress in an extracellular vesicle purification study (Rutter and Innes, 2017). Interestingly, two the other proteins identified, annexin1 (AT1G35720) and glutathione S-transferase PHI 2 (AT4G02520), are present in LSPDB low confidence list (**Appendix Table B.2**). APX1 has also been noted as putatively apoplastic (Podgórska et al., 2017) in a review of apoplastic ROS. SUBA collates MS/MS localisation to the plastid, PM, cytosol, Golgi, TGN, vacuole. No FP in vivo experimental evidence is available.

Alpha-glucosidase 2 (GCS2, a six-hairpin glycosidases superfamily protein, AT1G24320) has a glycan metabolism function (CAZy: Glycoside Hydrolase Family 63) has few published studies and subsequently little experimental evidence. Interestingly, vascular bundles isolated from switchgrass found a contig nearest to GCS2 in its cell-wall associated proteins (Srivastava et al., 2010). External to the scoring mechanism, AT1G24320 is present in the cluster identified in **Figure 3.14** and **Table 3.9** of interconnected UNCLASSIFIED, SPT and SEC proteins.

Uridine nucleosidase 1 (nucleoside hydrolase 1, NSH1, AT2G36310) has an expected cytoplasmic hydrolysis function (Jung et al., 2011) in the degradation of purine and pyrimidine. Of note however is the function of a homolog, NSH3 (AT5G18860) which is SEC classified due to the presence of a signal peptide. NSH3 has been found to have an apoplastic activity (Jung et al., 2011), and suggested to clean up extracellular ATP after signalling activity ceases. If NSH1 can be leaderlessly secreted then it may perform a similar function.

Aldose 1-epimerase (AT3G17940) has been identified in the cytoplasmic proteome (Ito et al., 2011) via MS. It is a member of Galactose mutarotase-like superfamily. Expression analysis by Ge et al. (2011) found differential expression of AT3G17940 in germinated pollen in the apoplastic proteome.

Alpha-galactosidase (glycoside hydrolase family 27, AT3G26380) represents a false-positive result, as the presence of a canonical signal peptide is unclear via predictions. One consequence of prediction software is results may vary over time as software changes; for SignalP version 3 this protein has prediction scores close to the cutoff, and version 5 of SignalP indicate a SP. In version 4 a low probability of a signal peptide is returned. DeepLoc indicates a probability of 0.9765 that it is secreted, and is actually in WallProtDB as a classically secreted protein. The protein was identified in a cytosolic proteome study (Ito et al., 2011) yet UniProt and AraPort databases indicate extracellular location though do not attribute a signal peptide. It illustrates a potential downside to the methods used for the LSPDB; removing the author inferences when importing data removed bias as to what proteins might be contamination, yet also discards correct interpretations that the automated LSPDB pipeline may inadvertently remove. However, this false positive results shows that classically secreted proteins can be identified from their connections and domains, and other proteins with either ambiguous or missing SP annotations could also be rescued as secretory using the LSPDB, so at least some such errors can be recovered.

Cysteine proteinase inhibitor (AT5G12140) acts in response to abiotic stresses including cold, heat and wounding. It has experimental evidence for location in other parts of the secretory pathway, such as the ER (Nikolovski et al., 2012), vacuole (Jaquinod et al., 2007) and plasma membrane proteomes (Marmagne et al., 2007) despite the lack of secretory features. Database annotations also support evidence in many sub-cellular locations. An in-depth study of this family (Hwang et al., 2010) indicates expression in vascular tissue, and interaction with the jasmonate signalling pathway with JAZ12 (AT5G20900) has been reported (Pauwels et al., 2015). Multiple sub-cellular locations, including the ER, cytosol and plasma-membrane, have been suggested from MS studies.

NADH-cytochrome b5 reductase-like protein (AT5G20080) was thought to be ER-located but more recent localisation work suggests both mitochondrial membrane (Dun-

can et al., 2011) and Golgi (Parsons et al., 2012). It frequently occurs in the LSPDB (**Table 3.3**). Differential abundance due to salt stress has also been shown Jiang et al. (2007). Additional MS data indicates a PM association Mitra et al. (2007).

The RPS10B ribosomal protein (RNA binding Plectin/S10 domain-containing protein, AT5G41520) forms part of the S10e ribosomal subunit in the cytoplasm, and localisation in the nucleus and cytoplasm (Palm et al., 2018). Down regulated expression for this gene was shown after a synthetic plant natriuretic peptide (PNP) treatment (Turek et al., 2014). Of note, PNPs are secreted into the apoplastic space due to both biotic and abiotic stress. Several other ribosomal proteins were identified in the medium confidence set, and the plausibility of the family of ribosomal proteins as LSPs is discussed below. There is however MS data has also shown it to localise to the PM, MVBs, and the TGN (Heard et al., 2015).

3.4.5.2 Medium confidence

The 149 medium confidence proteins (**Table 3.14**) represent those with evidence from at least two categories. Due to the large number we have taken a summary approach to reviewing these as related groups, both formally and informally. Formal selection considers GO term over-representation, with the caveat that the selection of proteins was based in part on GO terms. We highlight several significantly ($p < 0.05$, bonferroni correct with PANTHER Mi and Thomas (2009)) over-represented categories of molecular function in proteins: 1) constituents of the ribosome and RNA binding, 2) unfolded protein binding, protein folding chaperones, misfolded protein binding, and heat shock protein binding. We can also identify other groups informally by inspection of related proteins that occur in the medium list.

Table 3.14: Putative LSPs from the medium (orange) category

Accession	Description
AT1G01100	60S acidic ribosomal protein family
AT1G03330	Small nuclear ribonucleoprotein family protein
AT1G03680	thioredoxin M-type 1
AT1G03860	prohibitin 2
AT1G04480	Ribosomal protein L14p/L23e family protein
AT1G05190	Ribosomal protein L6 family
AT1G07770	ribosomal protein S15A
AT1G08830	copper/zinc superoxide dismutase 1
AT1G12230	Aldolase superfamily protein
AT1G13320	protein phosphatase 2A subunit A3
AT1G14320	Ribosomal protein L16p/L10e family protein
AT1G20020	ferredoxin-NADP(+)-oxidoreductase 2
AT1G21720	proteasome beta subunit C1
AT1G32470	Single hybrid motif superfamily protein
AT1G33040	nascent polypeptide-associated complex subunit alpha-like protein 5
AT1G42970	glyceraldehyde-3-phosphate dehydrogenase B subunit
AT1G47710	Serine protease inhibitor (SERPIN) family protein
AT1G48830	Ribosomal protein S7e family protein

Accession	Description
AT1G55260	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein
AT1G62020	Coatomer, alpha subunit
AT1G62740	stress-inducible protein, putative
AT1G64190	6-phosphogluconate dehydrogenase family protein
AT1G64520	regulatory particle non-ATPase 12A
AT1G66240	homolog of anti-oxidant 1
AT1G66580	senescence associated gene 24
AT1G67430	Ribosomal protein L22p/L17e family protein
AT1G70600	Ribosomal protein L18e/L15 superfamily protein
AT1G70730	Phosphoglucomutase/phosphomannomutase family protein
AT1G71750	Hypoxanthine-guanine phosphoribosyltransferase
AT1G72730	DEA(D/H)-box RNA helicase family protein
AT1G74030	enolase 1
AT1G74050	Ribosomal protein L6 family protein
AT1G75780	tubulin beta-1 chain
AT1G78900	vacuolar ATP synthase subunit A
AT1G79440	aldehyde dehydrogenase 5F1
AT1G80460	Actin-like ATPase superfamily protein
AT2G04030	Chaperone protein htpG family protein

Accession	Description
AT2G09990	Ribosomal protein S5 domain 2-like superfamily protein
AT2G16600	rotamase CYP 3
AT2G17190	ubiquitin family protein
AT2G17200	ubiquitin family protein
AT2G18020	Ribosomal protein L2 family
AT2G20260	photosystem I subunit E-2
AT2G20610	Tyrosine transaminase family protein
AT2G22780	peroxisomal NAD-malate dehydrogenase 1
AT2G27170	Structural maintenance of chromosomes (SMC) family protein
AT2G28720	Histone superfamily protein
AT2G33150	peroxisomal 3-ketoacyl-CoA thiolase 3
AT2G34480	Ribosomal protein L18ae/LX family protein
AT2G37190	Ribosomal protein L11 family protein
AT2G37270	ribosomal protein 5B
AT2G37760	NAD(P)-linked oxidoreductase superfamily protein
AT2G38880	nuclear factor Y, subunit B1
AT2G41970	Protein kinase superfamily protein
AT2G42740	ribosomal protein large subunit 16A
AT2G43160	ENTH/VHS family protein

Accession	Description
AT2G43750	O-acetylserine (thiol) lyase B
AT2G44350	Citrate synthase family protein
AT2G44510	CDK inhibitor P21 binding protein
AT2G45240	methionine aminopeptidase 1A
AT2G45640	SIN3 associated polypeptide P18
AT2G45710	Zinc-binding ribosomal protein family protein
AT3G01590	Galactose mutarotase-like superfamily protein
AT3G05530	regulatory particle triple-A ATPase 5A
AT3G05560	Ribosomal L22e protein family
AT3G07770	HEAT SHOCK PROTEIN 89.1
AT3G09440	Heat shock protein 70 (Hsp 70) family protein
AT3G09630	Ribosomal protein L4/L1 family
AT3G11510	Ribosomal protein S11 family protein
AT3G11630	Thioredoxin superfamily protein
AT3G12580	heat shock protein 70
AT3G13920	eukaryotic translation initiation factor 4A1
AT3G16400	nitrile specifier protein 1
AT3G16420	PYK10-binding protein 1
AT3G17810	pyrimidine 1

Accession	Description
AT3G18740	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein
AT3G18780	actin 2
AT3G23390	Zinc-binding ribosomal protein family protein
AT3G23940	dehydratase family
AT3G25800	protein phosphatase 2A subunit A2
AT3G44110	DNAJ homologue 3
AT3G44320	nitrilase 3
AT3G45780	phototropin 1
AT3G47370	Ribosomal protein S10p/S20e family protein
AT3G48930	Nucleic acid-binding, OB-fold-like protein
AT3G49010	breast basic conserved 1
AT3G52150	RNA-binding (RRM/RBD/RNP motifs) family protein
AT3G57140	sugar-dependent 1-like
AT3G58610	ketol-acid reductoisomerase
AT3G60120	beta glucosidase 27
AT3G60245	Zinc-binding ribosomal protein family protein
AT3G62940	Cysteine proteinases superfamily protein
AT4G02230	Ribosomal protein L19e family protein
AT4G02450	HSP20-like chaperones superfamily protein

Accession	Description
AT4G02930	GTP binding Elongation factor Tu family protein
AT4G10450	Ribosomal protein L6 family
AT4G10480	Nascent polypeptide-associated complex (NAC), alpha subunit family protein
AT4G16563	Eukaryotic aspartyl protease family protein
AT4G17390	Ribosomal protein L23/L15e family protein
AT4G17510	ubiquitin C-terminal hydrolase 3
AT4G17620	glycine-rich protein
AT4G19003	E2F/DP family winged-helix DNA-binding domain
AT4G20440	small nuclear ribonucleoprotein associated protein B
AT4G20980	Eukaryotic translation initiation factor 3 subunit 7 (eIF-3)
AT4G23730	Galactose mutarotase-like superfamily protein
AT4G25550	Cleavage/polyadenylation specificity factor, 25kDa subunit
AT4G26910	Dihydrolipoamide succinyltransferase
AT4G29390	Ribosomal protein S30 family protein
AT4G30530	Class I glutamine amidotransferase-like superfamily protein
AT4G33150	lysine-ketoglutarate reductase/saccharopine dehydrogenase bifunctional enzyme
AT4G34670	Ribosomal protein S3Ae
AT4G35450	ankyrin repeat-containing protein 2
AT4G36130	Ribosomal protein L2 family

Accession	Description
AT4G37910	mitochondrial heat shock protein 70-1
AT4G38740	rotamase CYP 1
AT5G02490	Heat shock protein 70 (Hsp 70) family protein
AT5G02560	histone H2A 12
AT5G03340	ATPase, AAA-type, CDC48 protein
AT5G05460	Glycosyl hydrolase family 85
AT5G06290	2-cysteine peroxiredoxin B
AT5G09590	mitochondrial HSO70 2
AT5G10450	G-box regulating factor 6
AT5G11920	6-&1-fructan exohydrolase
AT5G15090	voltage dependent anion channel 3
AT5G18380	Ribosomal protein S5 domain 2-like superfamily protein
AT5G19090	Heavy metal transport/detoxification superfamily protein
AT5G19550	aspartate aminotransferase 2
AT5G23230	nicotinamidase 2
AT5G37780	calmodulin 1
AT5G37830	oxoprolinase 1
AT5G40770	prohibitin 3
AT5G42220	Ubiquitin-like superfamily protein

Accession	Description
AT5G42980	thioredoxin 3
AT5G44070	phytochelatin synthase 1 (PCS1)
AT5G46430	Ribosomal protein L32e
AT5G46550	DNA-binding bromodomain-containing protein
AT5G48870	Small nuclear ribonucleoprotein family protein
AT5G52650	RNA binding Plectin/S10 domain-containing protein
AT5G56000	HEAT SHOCK PROTEIN 81.4
AT5G56030	heat shock protein 81-2
AT5G56710	Ribosomal protein L31e family protein
AT5G58940	calmodulin-binding receptor-like cytoplasmic kinase 1
AT5G59850	Ribosomal protein S8 family protein
AT5G63120	P-loop containing nucleoside triphosphate hydrolases superfamily protein
AT5G63310	nucleoside diphosphate kinase 2
AT5G64370	beta-ureidopropionase
AT5G65430	general regulatory factor 8
AT5G66190	ferredoxin-NADP(+)-oxidoreductase 1
AT5G67500	voltage dependent anion channel 2

Ribosomal and RNA binding proteins Structural components of the ribosome are over-represented (35; $p\text{-value} = 1.08 \times 10^{-28}$) in the medium set (**Appendix B.3**), and expanding the list to include ribosome-related makes a total of 38 proteins. As noted in the high confidence result for AT5G41520, the inclusion of such a large number of ribosomal proteins in the results is cause for concern and requires careful evaluation. We know that the ribosome complex is bound to either the ER or free in the cytoplasm, with a primary function protein synthesis (Carroll, 2013). Such a large proportion of results can be assessed either positively or negatively; a set of LSP candidates, or false positives and likely contamination when observed in the apoplastic space.

The positive point of view is to reason what non-canonical function ribosomal proteins may have, that would warrant a leaderless secretion. Extra-ribosomal functions in humans include tumourgenesis and immune signalling (Zhou et al., 2015). In *Arabidopsis*, of the just 5 proteins with alternative functions (moonlighting) one is ribosomal, AT1G14320. However, the secondary function of antiviral response in the nucleus (Carvalho et al., 2008), directed by a PTM, differs from the possible secretory role a putative LSP status would imply. Furthermore, the large number of ribosomal proteins suggests that they would either need to have alternative functions, or a role for intact ribosomal unit in the extracellular space. Lacking evidence for these extremes or some combination leaves the negative view.

Related RNA-binding proteins such as the Like-Sm (LSM) proteins (AT1G03330 LSM2, AT1G19120 LSM1A) and similar small nuclear ribonucleoprotein protein appear in the medium and low confidence sets. These Sm-like proteins interact and bind to RNA, localise to the cytoplasm, including process bodies (P-bodies) (Tsuzuki et al. 2017; Perea-Rese et al. 2012). The function of P-bodies is not well understood beyond the accumulation and decapping of RNA, though some plant evidence for P-bodies as a target of fungal effectors has been suggested (Petre et al. 2016), and their involvement in the plant innate immune responses (Maldonado-Bonilla et al. 2014; Roux et al. 2015).

Inspection of the PPI network in detail reveals one ribosomal protein with predicted SP (SEC: AT2G27710). PPI with this SEC interaction occur in the majority of the ribosomal results, and notably are the only PPI with a secretory protein for these candidates. In the

graded scoring system, removal of this signal would shift all ribosomal proteins to the low confidence (red) set. Since ribosomal proteins represents 38 of the 149 proteins, we will refine the medium confidence list further in **Chapter 4** when using these data as the basis for new tools. As with all putative LSPs, FP or antibody-based immuno-microscopy approaches are needed to investigate if the ribosomal proteins truly are secreted via UPS.

Chaperone, folding and heat shock (HSP)-related proteins The related terms of misfolded protein binding (5 proteins; $p\text{-value} = 1.61 \times 10^{-4}$), protein folding chaperone (5 proteins; $p\text{-value} = 9.08 \times 10^{-4}$), HSP binding (7 proteins; $p\text{-value} = 7.98 \times 10^{-6}$) and unfolded protein binding (10 proteins; $p\text{-value} = 5.15 \times 10^{-6}$) are all overrepresented. The union of these results along with additional HSPs from the set yields 12 proteins, including several HSP70 family proteins, HSP 20-like, HSP 80 and rotamase CYP1/3. (**Appendix B.4**)

Roles in humans for extracellular HSPs have been suggested (Krause and Rodrigues-Krause, 2011), including in an innate immune response (Nativel et al., 2019). One proposed mechanism for secretion of HSPs via extracellular vesicles (humans (De Maio and Vazquez, 2013; Reddy et al., 2018); plants Rutter and Innes (2017)) and leaderless secretion are suggested.

Interestingly, the relationship between HSPs and ribosomal release has also been suggested (Merret et al., 2017). In particular, the accumulation of ribosomal RNA during heat shock leads to a rapid release of ribosomal proteins during the restoration of cell activity after heat stress.

Calmodulin-related proteins Calcium binding Calmodulin 1 (CAM1, AT5G37780) and the calmodulin-binding receptor-like cytoplasmic kinase 1 (CRCK1, AT5G58940) both occur in the medium confidence results, and as suggested by their names, interact with each other according to the PPI layer of the LSPDB. Additionally, low confidence results (**Appendix Table B.2**) include three more calmodulin-related proteins (Calmodulin4, Calmodulin7 and Calmodulin-like protein 10). Calmodulin7 is in the Ding et al. (2012) LSP list. Each of these have a cytoplasmic localisation, however, Yu et al. (2016) showed that CAM1 interacts with the receptor protein kinase-like protein (AT2G01210,

ZAR1) at the cytoplasmic side of the plasma membrane, via fluorescence signals joined to both proteins. CAM1 has a possible stress response function.

Prohibitin family proteins Prohibitin 2 (PHB2, AT1G03860) and Prohibitin 3 (PHB3, AT5G40770) are nominally involved in stabilisation in the mitochondria. Marmagne et al. (2004) found PHB3 in an experiment enriched for plasma membrane proteins. PHB2 also has evidence for plasma membrane location Mitra et al. (2007) and also the Golgi (Parsons et al., 2012). Notably a single transmembrane domain is annotated in UniProt for PHB2 and PHB3 that was not identified by the LSPDB pipeline. However, the TM was assigned correctly in PHB4 and the PPI between PHB2, PHB3 and PHB4 rescues these two when computational predictions miss. PHB3 has been suggested to have a stress response function to salicylic acid, also localising to the chloroplast Seguel et al. (2018).

Thioredoxin- and peroxiredoxin- related proteins Several thioredoxin (AT5G42980, AT1G03680, AT3G11630) and 2-cysteine peroxiredoxin (AT5G06290, AT3G11630) proteins occur in the medium confidence categories. Interactions between them in plants have been suggested (Bernal-Bayard et al., 2014), as well as the general involvement in the hydrogen peroxide pathway. Generally, these products of typical metabolism are regulated below toxic levels, however, they also serve a dual signalling function in response to stress (Sevilla et al., 2015). The disulfide bonds of peroxiredoxin are often reduced by thioredoxin. Notably thioredoxin 3 (AT5G42980) occurred in the sub-cluster of SEC and SPT proteins (**Table 3.9**). Several thioredoxins are also found in the low (red) confidence list. Another thioredoxin, AT1G45145, is in the putative LSP list from Ding et al. (2012) and frequently occurs in the PPI as shown (**Table 3.11**).

3.4.5.3 Low confidence

The largest confidence set (227) is also the least confident, consisting of proteins with at most two individual low or a single medium confidence assignments from PFAM, GO and/or network, respectively. At this stage of the work, we defer in depth analysis of

them individually due to this lower bar of evidence, though related proteins in the high and medium list have already been mentioned reinforcing that these confidence levels are a construct of this analysis, and related proteins could be viewed collectively if they appear at multiple tiers. The full low confidence list is displayed in **Appendix Table B.2**. This low confidence list can also contain proteins selected from only the low confidence of PFAM, GO term or network methods. Without complementary scores from another approach, bias or false-positives driven by these categories can therefore be the only reason for inclusions. For the GO terms in particular, the low confidence criteria selects proteins that have GO terms predominantly assigned to SEC proteins that are also in UNCLASSIFIED. Many UNCLASSIFIED had been labelled as the general term cell wall (GO:0005618) or the more specific plant cell wall (GO:0009505), which does not reveal any further information for clarifying if they are genuine LSPs. These annotations could be assigned due to the experiments they were observed in, where their contamination status is unresolved, and the terms erroneously assigned. Alternatively, these could be the result of curation from additional experiments that confirm a genuine cell wall or apoplastic presence. Resolving this would require knowledge of the provenance of each annotation, not captured in the LSPDB data model at present.

This uncertainty is shown in the inclusions of S-adenosylmethionine synthase 2 (SAMS2), the sole *Arabidopsis* LSP, in the low confidence list. The UPS pathway SAMS2 is proposed to undertake is via EXPO, and it was shown to co-locate with an EXPO fluorescence-based marker (Wang et al., 2010). In that study, they noted it's appearance in the Bayer et al. (2006). This observation led to initial UNCLASSIFIED status in the LSPDB, and accessing the LSPDB reveals no others observations from the literature in scope, confirming that it remains the only proteomic apoplastic observation of this LSP. Its LSPDB score of 1 is due to its inclusion in the low confidence GO term, notably due to the cell wall term (GO:0005618). Inspecting UniProt entry (P17562) for SAM2 also illustrates the circular nature of the GO term assignment in the context of LSPs. Cellular component GO terms are each assigned their source, and though cell wall is assigned to this under the "inferred by direct assay" (IDA) evidence category, the assay in this instance is the same Bayer et al. (2006) study the LSPDB uses as source data. This is an example of er-

roneous assignment in the LSP context, yet SAMS2 is also annotated by IDA with the extracellular exosome (GO:0070062) by Wang et al. (2010). In this second example the term is correctly assigned and is of relevance to the LSP problem, despite the identical IDA method. Future strategies to resolve this issue could include a modification of the GO term assignment in the LSPDB, adding the provenance of each term, and then removing any terms that are drawn from literature that was used as a source of observations in the LSPDB. In the interim, due to these issues the low confidence list is used in both Chapter 4 and Chapter 5 as an independent set of proteins to apply other tools to, rather than as a source of training data.

3.4.6 Using the LSPDB

In addition to these confidence lists, the LSPDB is intended to be a resource for users - initially the author of this Thesis but also the broad research community (that require user friendly access interfaces) to enable the rapid “identification” of candidate LSPs so as to minimize the number of proteins needing to be subjected to the laborious but essential experimental validation. Despite the complexity of the LSPDB confidence criteria that identified SAMS2, as it remains an accepted LSP it is used in this section to illustrate how the LSPDB can generally be applied to proteins of interest. This is also motivated by the observation that another related protein, S-adenosylmethionine synthase 1 (SAMS1; AT1G02500) is also in the low confidence list (albeit for the same reason as SAMS2) and can be queried via the LSPDB.

With additional users in mind, the framework that notably includes the SEC, NON-SEC, SPT, UNCLASSIFIED classificationns is made available in a number of ways. In this section access to the LSPDB database is discussed; directly either via the SQL database or the abstraction layer provided by Django (Section 3.5.1), or via a web interface to the LSPDB currently in development (Section 3.5.2). Each of these modes of access suits different user needs. For LSP research, access to the raw data allows for complementary research, such as the PTM and abundance extensions discussed in Chapter 5. For plant cell wall researchers, making data available in order to confirm potential LSPs from contamination is the guiding principle.

One aspect of the LSPDB, explored further in Chapter 5, is that new protein observations will change classifications, recalling that additional observations could move genes from the category of protein that have not been seen, SP_THEORY and NONSEC, into SEC and UNCLASSIFIED, respectively. These can cascade further changes as PPI, PFAM and GO terms associated with each class are recalculated, and can alter the list of potential LSPs. This fluidity strikes to the core concept of the LSPDB – to use additional relevant experiments to converge on consensus classifications for proteins to be putative LSPs. Simple lists of putative LSPs fail to easily capture potential insights, such as the repeated appearances of putative LSPs. Use of the LSPDB in future will allow for a nuanced selection of potential LSPs driven by the researchers own knowledge and the experimental context.

3.4.6.1 Direct database availability

Django (v 2.0) is a Python library for Web applications. It was chosen for this research for two particular reasons: an abstraction layer for database queries which simplifies and separates the research questions from the data storage, and with built-in user management and web interface support for the end goal of providing an accessible resource for LSPs to the research community.

The LSPDB Django framework is maintained in an underlying Structured Query Language (SQL) format. As such, the database contents can be manually accessed from the `sql` file directly, in this case a development version as SQLite. Prior to publication of the LSPDB, the database contents including the Django source code and the `lspdb.sqlite` database file will be available via a private institutional Figshare collection (<https://melbourne.figshare.com/s/0f775540d7dd65f7f4e8>). Access to abstract layer was demonstrated in Section 3.2, and templates that can be used to perform custom analysis will be added to the online repositories at: <https://github.com/LSPtools>.

LSPDB/Plant:
Leaderless Secretory Protein Database
for Plants

Accessions:

Arabisidopsis accessions (AT format) **Submit**

Recently added

No protein are available.

Uncurated

The uncurated layer of LSPDB contains all proteins found in various experiments, including classical and putative leaderless secreted proteins, and intracellular contamination. Observation in a proteome is the only requirement.

In LSPDB/Plant there are currently observations in total of **46320** proteins from various plant species.

For *Arabidopsis thaliana* there are observations of unique proteins. Of these, there are observations of proteins that do not exhibit classical secretory properties.

Sources

The WallProtDB database is used to seed LSPDB with classically secreted proteins.

Literature: of *Arabidopsis thaliana* publications have been reviewed for possible LSPs initially classified as contamination.

Curated

Curation involves identifying putative LSPs based on association with CSPs, biologically relevant features and statistical methods.

Protein-Protein Interactions

PPI networks are used identify **forward** matches (where putative LSPs interact with known CSPs) and **reverse** matches (where known CSPs are used to find putative LSPs).

There are **X** possible LSPs using **forward** matches to protein-protein interaction networks for *Arabidopsis thaliana*.

There are **X** possible LSPs using **reverse** matches to protein-protein interaction networks for *Arabidopsis thaliana*.

Plant Cell Walls
ARC Centre of Excellence
© 2016

Figure 3.21: Web interface to the LSPDB

3.4.6.2 Web interface

The usability of the LSPDB for researchers without Python or SQL database experience is determined by the interface; finding potential LSPs needs to be an easy process, as easy as the current (ineffective) techniques of accepting, rejecting or filtering based on a single program. A web interface to the LSPDB is provided to do this. Users with varying levels of bioinformatics expertise may access the LSPDB website, and its usability and ability to improve the knowledge of LSPs will be critical. Design of the LSPDB interface balances accessibility to the underlying data with an interface intended to improve the study of LSPs, initially in plants but extendable to other species.

Prior to publication, the interface (**Figure 3.21**) is available at:

url: <http://dev.lspdb.org>

user: reviewer

password: LSPDB_review_2019

This landing page summarises the contents of the LSPDB, and accepts input of *Ara-bidopsis* accessions for further analysis. With simplicity the aim, the current interface supports two main views: individual and summary. Either of these allows users to investigate potential LSP proteins, however the summary view considers the relationships between each all of the genes submitted. The landing page leads to the summary view, even if one gene is submitted. Each gene on the summary page then links to the individual view.

For each accession in the database, an analysis summary page is available for that gene including all of the information in the group search page, with additional links to external databases including SUBA. Further detail provided in the gene summary page includes an image of the gene interaction network, and list of PFAM and GO terms associated with that gene.

The summary interface is designed for researchers with a list of gene/proteins to be considered as either cellular contamination or as leaderless. For each gene in the list, the relevant data in the LSPDB is returned at two levels - gene and protein.

For each gene accession, static information including the presence or absence of retention signals, the observations in the literature, the total protein-protein interaction links, the base classification, and secretory isoform status. Then for each protein model of this gene, the status of SP, TM count, GPI status and individual retention motifs are also noted.

Results

Results are:

Gene	Classifications	Description	Observations	Conditions	Stress	Tissue	Local CoExpression (RNA-seq, Microarray)	Local Network (SEC, NONSEC, UNCLASSIFIED)	Total Network (SEC, NONSEC, UNCLASSIFIED)
AT1G02500	UNCLASSIFIED	S-adenosylmethionine synthase [Source:UniProtKB/TrEMBL;Acc:A0A178WES1]	3 (tmp) Basu et al (2006) Plant Soil 286: 357-376, (tmp) Bayer et al. (2006) Proteomics 6: 301-311., (tmp) H Nguyen-Kim et al. (2016) Proteomics, 16:491-503,	18 day-old seedlings, 3 days, Col0 ProteoMiner,	nan, nan, nan,	roots, cell suspension culture, roots,	AT4G01850, , AT4G01850, ,	([], [], ['AT4G01850'])	(0, 21, 11)
AT1G02500.1									
AT1G02500.2									
AT4G01850	UNCLASSIFIED	S-adenosylmethionine synthase [Source:UniProtKB/TrEMBL;Acc:A0A178V3V9]	1 (tmp) Bayer et al. (2006) Proteomics 6: 301-311.,	3 days,	nan,	cell suspension culture,	AT1G02500, , AT1G02500, ,	([], [], ['AT1G02500'])	(0, 15, 6)
AT4G01850.1									
AT4G01850.2									

Figure 3.22: Example of the LSPDB summary interface, using SAMS1 and SAMS2

An assumption in the design of the group interface page is that the list comes from an experiment (or are otherwise related). In addition to the total protein-protein interaction links (number of SEC/NONSEC/UNCLASSIFIED/SPT proteins that the protein interacts with), local networks links are calculated. This takes all PPI interactions for each protein, and only includes those that pair with others in the input list. This represents proteins that interact and have been observed in this same experiment, and is a subset of the total number of PPI for each protein.

These provide a local network context for a researcher to add to the genome-wide LSPDB results, and without subscribing to all of the LSP scoring criteria in Section 3.4, the interaction of two proteins that have been found in the same experiment may be a useful criteria for a researcher to use. For example, a researcher may choose to investigate a potential LSP with SEC PPI interactions if the interacting proteins were also found in the same proteome.

In the analysis of the LSPDB, secretory isoforms were excluded since it is unclear which of the proteins has actually been experimentally demonstrated to be secreted. For group search it is also of interest if these have been observed, and they are reported to guide the researcher to consider if they warrant further scrutiny. The output of the summary search page can be exported to CSV for formatting or further analysis.

As noted, the low confidence list also contained SAMS1. Querying the LSPDB information for both SAMS1 and SAMS2 demonstrates the summary view, shown in **Figure 3.22**. Both are UNCLASSIFIED and SAMS2 has indeed only been observed in the study by Bayer et al. (2006) which was referred to by Wang et al. (2010) in the original EXPO work. The LSPDB highlights that SAMS1 was also observed, and that this experiment was from cell suspension culture after 3 days. SAMS1 has also been observed alone in Basu et al. (2006) and Nguyen-Kim et al. (2016) from a different root tissue proteomics, 10 years apart. None of these observations were under stress growth conditions. The local co-expression and network data, applied to the input list which in this case is just SAMS1 and SAMS2, also reveals they co-express and have a protein interaction. The overall LSPDB network indicates no PPI with secretory proteins from either SAMS1 or SAMS2, but 10 and 5 interactions, respectively, with other UNCLASSIFIED proteins apart from

each other.

Common to both are two additional SAMS family proteins; methionine adenosyl-transferase 3 (MAT3, AT2G36880) and S-adenosylmethionine synthase 4 (MET4, AT3G17390), leading to 13 additional proteins in this local SAMS1, SAMS2 network. We can collate this list, and then resubmit to the LSPDB (**Table 3.15**)

An initial analysis to such results could be as follows, and one way of using the LSPDB. Taking the most observed of this set, phosphoglycerate kinase (AT1G79550) , we can search the UPS literature for previous evidence. Surprisingly, this leads to a review of UPS work in yeast (Nombela et al., 2006). Collating LSPs from *Candida albicans* and *Saccharomyces cerevisiae*, not only is phosphoglycerate kinase present but also S-adenosylmethionine synthase (SAM2). There is also HSPs, thioredoxins and other proteins similar to those identified in this chapter.

This represents a new set of proteins to explore and experiment with as potential LSPs, as well as a set of possible cargo for the EXPO UPS route. It also raises questions about commonality of UPS between plants and fungi. As the best *Arabidopsis* candidate LSP, this demonstrates how the LSPDB can be used to explore potential LSPs using others as bait, and return data of interest in the LSP context that can then be iteratively reapplied to the LSPDB, or as a catalyst for future experimental validation.

Table 3.15: LSPDB results from UNCLASSIFIED interactions with SAMS1 and SAMS2

gene	Description	Count	Publication(s)
AT1G79550	Phosphoglycerate kinase	7	Kwon et al. (2005), Bayer et al. (2006), Tran and Plaxton (2008) (multiple), H Nguyen-Kim et al. (2016) (multiple), Herve et al. (2016)
AT5G26360	T-complex protein 1 subunit gamma	3	Bayer et al. (2006) , H Nguyen-Kim et al. (2016), Herve et al. (2016)
AT3G62870	60S ribosomal protein L7a-2	3	Bayer et al. (2006) , H Nguyen-Kim et al. (2016) (multiple)
AT1G02500	S-adenosylmethionine synthase	3	Basu et al (2006), Bayer et al. (2006) , H Nguyen-Kim et al. (2016)
AT2G36880	S-adenosylmethionine synthase	3	Borderies et al. (2003), Bayer et al. (2006) , H Nguyen-Kim et al. (2016)
AT3G17390	S-adenosylmethionine synthase	3	Bayer et al. (2006) , H Nguyen-Kim et al. (2016), Herve et al. (2016)
AT5G19990	26S proteasome regulatory subunit 8 homolog A	2	Bayer et al. (2006) , H Nguyen-Kim et al. (2016)
AT3G05560	60S ribosomal protein L22-2	2	Bayer et al. (2006) , H Nguyen-Kim et al. (2016)

gene	Description	Count	Publication(s)
AT2G18230	PPa2	2	H Nguyen-Kim et al. (2016), Herve et al. (2016)
AT5G63980	SAL1 phosphatase	2	H Nguyen-Kim et al. (2016) (multiple)
AT3G28730	FACT complex subunit SSRP1	1	Herve et al. (2016)
AT4G01850	S-adenosylmethionine synthase	1	Bayer et al. (2006)
AT2G45300	3-phosphoshikimate 1- carboxyvinyltransferase, chloroplastic	1	H Nguyen-Kim et al. (2016)
AT5G53480	Importin subunit beta- 1	1	Bayer et al. (2006)
AT3G44300	NIT2	1	H Nguyen-Kim et al. (2016)

3.5 Conclusions

Collating the published literature on plant LSPs serves two purposes. Firstly, an online resource for LSPs in plants is a dynamic alternative to the static reviews (e.g Ding et al 2012) and book chapters (Davis et al 2016) on this topic. With the changes in definitions and new discoveries, the fact that plant LSPs studies are still in their infancy should not exclude, and in fact should expedite, the need for shared resources. By deliberately keeping the barrier to entry to the LSPDB low, both technically and with respect to curation standards, the LSPDB should rapidly accumulate experimental information regarding putative plant LSPs and contaminants.

Several classes of candidates for LSP, including HSP thioredoxin and calmodulin families of proteins have been identified. HSPs and thioredoxin proteins have other eukaryotic UPS evidence. Ribosomal protein results have been discussed and on balance considered less likely candidates. For each candidate, the PPI, GO term or PFAM associated data are available and the putative assignment for LSP is transparent, contrasting with the 'black box' of prediction tools. These can give users of the database a theoretical basis for further investigation. For example, the PFAM domains related to glycosyl hydrolases (PF00232, PF00251, PF08244) appear in both PFAM based lists, and can be used to identify UNCLASSIFIED proteins with these domains.

Through the design and implementation of the LSPDB, several limitations and areas for future work have been noted. The majority of the LSPDB analysis relied on the distinct subsets of SEC/NONSEC/SPT/UNCLASSIFIED proteins. Further use of the secretory isoforms, and multiple TM-containing proteins in the network could enrich the analysis. The reliance on computational predictions in the annotation pipeline could also be relieved by using available annotations, noting though that many of these annotations are also computationally predicted.

Secretory isoforms are identified from comparisons of the protein sequences. In an experimental context, future work could entail working this out at the time of experiment by interrogating the mass spectrometry outputs determine if SP peptides are present, allowing the appearance of secretory isoforms to be resolved into the secretory version when SP are found. Secretory isoforms could also be considered in RNA-seq experi-

ments, if there is transcriptional evidence for an SP isoform. In reviewing candidates that appear in literature outside the scope of the custom LSPDB publication layer (e.g. either in Golgi or plasma membrane proteomes), a better integration with databases that cross reference proteins with all publications they appear in could allow these insights to be automated. Such a framework would also allow the database to be extended to other plant species more rapidly.

Despite these limitations and areas for improvement, the LSPDB v1.0 does provide a new approach to considering LSP candidates. The goal of the LSPDB is to exist as long as LSPs are overlooked, incorrectly labelled as contamination, or filtered with inappropriate software. The gradual shift away from a default CSP point of view will allow for LSP annotation to gradually become included in larger consortium (e.g. UniProt) as more accurate bioinformatics tools suggest putative LSPs, which lead to experimental validation, better tools and more candidates. When methods and theory of the plant cell secretory system advance that such knowledge is absorbed into the major databases, there will be less of a need for the LSPDB. Until then, it provides a framework for prioritising candidates for experimental validation, aiding in the search for more plant LSPs.

Chapter 4

Follow The Leaderless

4.1 Introduction

BASED on work in **Chapters 2 and 3**, our confidence in the identification of LSPs in plants has actually diminished. In Chapter 2 we established that the assumptions of the SecretomeP approach, using the common feature hypothesis, are not upheld when considering plant secreted proteins, and thus plant proteins that have either been excluded or included for LSP analysis based on its predictions need to be re-evaluated. This includes the putative list of plant LSPs (Ding et al., 2012) (**Appendix A.2**), collated from apoplastic proteome studies that including those that used SecretomeP to identify LSPs, and other candidates without biochemical evidence. This re-evaluation led to the creation of the Leaderless Secretory Protein Database (LSPDB) in **Chapter 3**, with *A. thaliana* as the base organism. As a standalone alternate tool, the LSPDB can assist researchers to identify candidate LSPs for experimental follow-up, with the caveats that it is limited to *Arabidopsis* and each version of the database is tied to a static release based on the available experimental observations included at the time.

The database does not directly identify LSPs *per se*. Instead of predicting LSPs, the database provides a more transparent approach than reliance on a tool not trained on plant data, and the information required for the user to justify LSP candidates is returned, rather than a context-free prediction value. This information about proteins of interest includes their secretory features, context within the proteome, and the number of other observations, but does not capture information in a manner that can easily be applied to other plants. It has actually reduced the size of the putative LSP list (Ding et al., 2012)

list by identifying several likely false positives. These are proteins with signal peptides and transmembrane domains which are predicted to enter the ER-Golgi pathway, and proteins with secretory isoforms that contain signal peptides and/or transmembrane domains, adding uncertainty as to which isoform may have been experimentally observed.

In order to identify LSP candidates using information from the LSPDB, the evidence available in the LSPDB must move beyond being only reportable, rather, it should rely on some form of confidence criteria for the selection of candidates. We used a data-driven approach, based on the GO terms, PFAM domains and PPI networks to identify aspects of each that applied to conventional secretory proteins, and applied a scoring mechanism to prioritise potential LSPs that either shared these aspects or interacted with other secretory proteins (**Figure 3.20**). This led to three lists of putative LSPs of high (**Table 3.13**), medium (**Table 3.14**) and low (**Table B.2**) confidence. These lists offer a transparent way identify putative LSPs in *Arabidopsis* for experimental validation but is still a static list of proteins, based on retrospective analysis.

An advantage of a prediction tool is that it can be easily applied to new data. The work in **Chapter 2** concluded that there is currently no appropriate tool for finding plant LSPs. To replace a tool like SecretomeP, the LSPDB or putative LSP lists alone will not suffice as they do not encapsulate a model of how to identify LSPs, and cannot be directly applied to other plants. By taking observations from the LSPDB and using those as input data for a new tool, we can extrapolate beyond the *Arabidopsis* data it contains. These new tools will improve on the status quo by offering researchers a method for identifying putative LSPs in all plant species based on plant-related evidence.

One conclusion from **Chapter 2** however was the risks involved in using a common-feature hypothesis, relying solely on prediction scores and in general the consequences of applying a tool not intended for plants on plant data. These lessons should be applied to new tools, even if trained on plant data. The information in the LSPDB and putative LSPs identified are transparent, allowing users to either decide or inspect in *Arabidopsis* what the features are that lead to LSP status. In contrast, machine learning (ML) tools based on some input would be “black box” - once trained on input sequences, the underlying model and reasons for certain predicted outputs is often unclear to the end user. A new

ML-based method built on *Arabidopsis* data and then applied to other species could suffer the same lack of clarity on predictions, however ML methods can learn aspects of the data otherwise unknown that relate features of interest in the input data to the classes being predicted. Therefore in order to retain this openness and also leverage the advantages of ML to learn previously unknown relationships, we create two kinds of tools that can be applied to new data: “black box” where input data is used to create an ML model that yields predictions about LSPs, and “white box” tools that take protein properties from the LSPDB and reports on their presence.

Both kinds allow for insight from the LSPDB to be transferred to software that can be automated and applied to new plant proteins in order to identify LSP candidates in a high throughput manner. In Section 4.2 we will detail the methods used to select input data, choose theoretical models and implement new tools. The results of creating these new tools are shown in Section 4.3. The trade-offs between the models, a recommend a software workflow for LSPs, and the biological insights these tools offer regarding LSPs are discussed in Section 4.4 when applied to both plants and other organisms. We conclude with recommendations for their use, limitations and suggestions for future work in Section 4.5. It should be noted that further discussion relating to the identified proteins in terms of their protein function, and an expanded confidence occurs in **Chapter 5**.

4.2 Methods

We built four tools based on the trade-offs between “black box” and “white box” approaches, given the available data discussed in Section 4.2.1. Described in Section 4.2.2 is a predictor based on the LSPs derived from the LSPDB (LSPpred), in Section 4.2.3 a predictor based on revisiting the common feature hypothesis of CSPs and LSPs in plants (SPLpred), and in Section 4.2.4 a reporter based on secretory PFAMs present in proteins (LSPfam). An alternative approach is also considered in Section 4.2.5 by simply considering sequence similarity via BLAST (Altschul et al., 1997). A BLAST database including class annotation from the LSPDB can be used to establish if queries of interest bear resemblance to putative LSPs (LSPblast). Together, each of these four approaches provides

Name	Description	Type
LSPpred	train tool on LSP proteins from LSPDB	Black box
SPLpred	train tool using SecretomeP-like method	Black box
LSPfam	identify proteins with LSP dominant pfams	White box
LSPblast	identify proteins similar to LSPs from database	White box

Table 4.1: Summary of LSP tools created

complementary methods to identifying plant LSPs, and gives researchers a choice on the assumptions of LSPs and the evidence for them. A summary of these tools and their type is shown in **Table 4.1**.

4.2.1 Available data

The output of the LSPDB applicable to developing these new tools can be divided into two categories: either annotated properties of proteins or protein amino acid sequence. Properties of proteins are the PFAM and GO terms identified in **Figures 3.7** and **3.9** as shared between SEC and UNCLASSIFIED proteins. Proteins from the LSPDB are both the primary classes (SEC, UNCLASSIFIED, SPT and NONSEC) as well as subsets of UNCLASSIFIED. As described in Section 3.5, 384 proteins from UNCLASSIFIED were selected on this basis and divided into subgroups with differing confidence: 8 (high - green), 149 (medium - orange), 227 (low - red) , with the labels introduced in **Chapter 3** in parentheses. An additional 207 (homodimers - black) were identified as proteins of interest. These were identified due to the high number of proteins expected to self-dimerise in the set of UNCLASSIFIED proteins, relative to the other classes.

For each data set, the relevant proteins from the LSPDB gene classes were extracted in FASTA format. Given the possible one to many relationships between genes and proteins, initially all proteins from a classified gene were selected. For the ML-based methods (SPLpred, LSPpred), a second version of the data was created to avoid a bias and over-reliance on similar sequences (either from the same gene, or gene family). Using CD-hit (version 4.6.8) (Li and Godzik, 2006), protein sequences with up to 40% maximum identity in each class were selected . This is similar to the SecretomeP division into non-similar sets with approximately ~26% identity between them for cross validation. A

threshold of 40% was selected in order to balance basing the predictions on independent data with the limited original data size. The sequence similarity measure ensures the data is dissimilar, and prevents accuracy from one particular motif overly influencing the data. However this may reduce the power to detect secretory signals. This resulted in overall sizes of each class; SEC 322, UNCLASSIFIED 1178, SPT 1439 and NONSEC 10523 proteins. The remaining putative LSPs includes 8 high, 93 medium and 353 low confidence LSPs.

4.2.2 Feature extraction

Taking the amino acid sequence and extracting features requires a conversion into numerical values best suited for ML. In the same manner that the results from **Chapter 2** informed the selection of input data, it also informed the way we represented features. Specifically, the possibility of using intermediate program outputs as numerical feature converters (TMHMM, Seg, etc) to introduce unknown biases is purposefully avoided by preferring protein features derived directly from sequences.

The ProFET (Ofer and Linial, 2015) suite of tools was selected to derive these features, in particular due to its goal of deriving a large number of features without relying on additional input sources. An initial set of 203 categories with a total of 1170 features were derived for the inputs. These categories include properties of sequence length, weight, isoelectric point, grand average of hydropathy (GRAVY, calculated as sum of amino acid hydropathy values divided by length), amino acid compositions, composition based on compressed amino alphabets that replace related amino acids with a single value, possible sites of PTMs, autocorrelation, as well as scaled and transformed features. Features are scaled by ProFET (see **Appendix C** for full list). As the features truly associated with LSP status are unknown for the input data, starting with a large number of possible features allows the training stage to determine the most relevant features, and give insight into what aspects of the training data discriminate LSPs. The most relevant ones can then be investigated in detail and are elaborated upon in Section 4.5. Though this does give some view of inside the ‘black box’, these features that are most predictive may offer some insight during the tool development, and are not typically available to a user of the

tool.

4.2.3 LSPpred

LSPpred is based on the putative list of LSPs from the LSPDB for a unique-feature hypothesis. By using those proteins as input data and a ML tool based on the features of those proteins, the tool predicts proteins that share properties with the candidate list. Although some of these curated proteins were selected on shared features, they were also chosen based on PPI, and any features derived from the sequences will be used to predict to LSPs.

For LSPpred, positive data is taken from the high and medium LSPDB results. As discussed in **Sections 3.4.1** and **3.4.5.3**, each of the evidence methods in the LSPDB has the possibility of false-positives, and the low confidence list can include proteins from only one of these sources. Higher confidence scores temper this uncertainty by requiring evidence from multiple LSPDB sources, and so for a tool based on these confidence lists, we exclude the low confidence list. The plausibility of the ribosomal LSPs in the medium confidence lists was also discussed in Section 3.4.5.2. Given the large proportion (38 of 149 in the unreduced list), we removed proteins annotated as ribosomal-related from the LSPpred input.

For the negative class, we consider two possibilities (**Figure 4.1A**). Firstly, we can take the NONSEC proteins, intending for the classifier to distinguish between either LSPs or contamination. This will achieve the goal for an LSP predictor, however, in a similar fashion to **Chapter 2**, we consider what the behaviour of the classifier will be on SP-containing proteins and what kinds of features are used as a basis for prediction. In this version, whether features are also present in SEC proteins is not known. In order to compare, a second version of LSPpred can be created with an expanded negative data set to include the SEC and SPT protein classes. This version by design classifies CSPs as negative data, and the results are indicative of proteins with features only present in the putative LSP list. For LSPpred, the remainder of the unclassified proteins form an independent set of proteins of interest that it can be applied to.

4.2.4 SPLpred

SPLpred takes a SecretomeP-like approach. As concluded in **Chapter 2**, looking for common features between CSP and LSPs is not necessarily shown to be ineffective for plant LSPs, only that the features captured for human CSP do not inform plant ones. Although Nielsen et al. (2018) have also raised doubt as to the effectiveness of SecretomeP, our results focused on the influence of the modified data and possible bias due to the use of TMHMM as a feature. SPLpred revisits the common-feature hypothesis with the benefit of these results.

With an awareness of the potential for bias using modified SPs, we consider four alternatives for selecting training data to use in the SecretomeP-like approach. Each of the input data choices are shown in **Figure 4.1B**. In each case the LSPDB subsets of secretory and non-secretory proteins are used. The first is use secretory proteins with SP removed as positive data, with non-secretory proteins as negative. The second approach is to treat both positive and negative proteins sequences equally by removing the average length of the SP from the negative data (similar to the SPremove method used in **Chapter 2**). The third approach similarly matches the positive and negative data, by removing the SP from the positive class and replacing with a leading methionine in order to ensure that the absence of the typical initiating methionine does not drive the predictions between classes. Finally, we consider countering the potential for bias when using modified data by placing modified data within both positive and negative classes, with SP-removed version SEC as positive, and both intact SEC and either version of NONSEC as negative. This approach aims to ensure that the absence/presence of the SP is not relevant to the classification.

4.2.5 LSPblast

LSPblast was built by creating a conventional BLAST database with the LSPDB classifications appended to the protein labels. For each of the SEC, NONSEC, UNCLASSIFIED and SPT data sets, the FASTA formats were extracted for all protein sequences, and the class name appended to the sequence header. These annotated sequences are

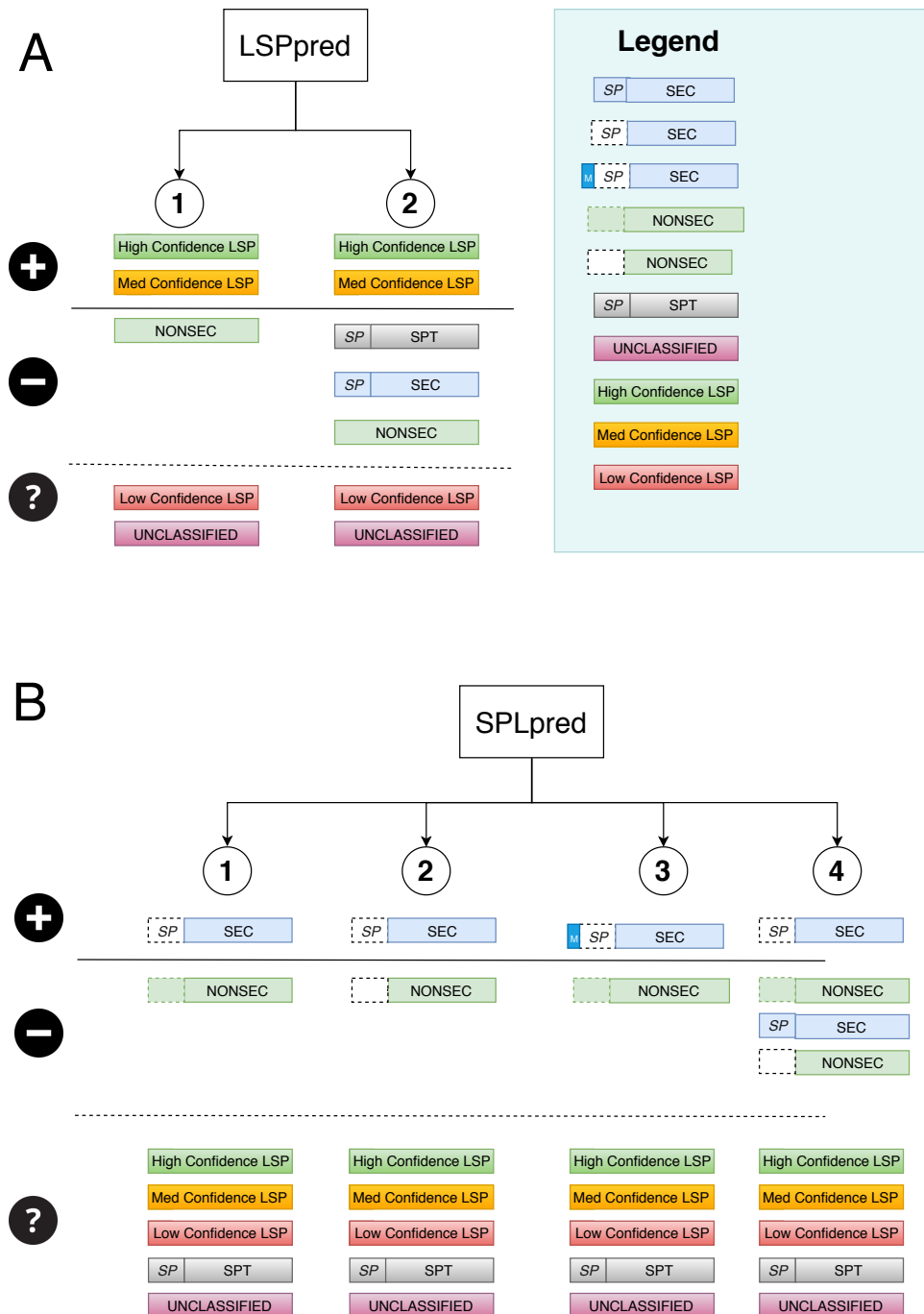


Figure 4.1: LSPpred approaches to training data, with high and medium confidence LSPs as positive compared to alternative negative data sets; A.1 NONSEC classified proteins as negative, or A.2 the combination of NONSEC, SPT and SEC as negative. Low confidence and the remaining UNCLASSIFIED are proteins of interest to apply LSPpred to. B) SPLpred approaches use different positive and negative combinations of B.1 SEC with SP removed vs NONSEC protein, B.2 SEC with SP removed vs NONSEC with average length of SP removed, B.3 SEC with SP removed and initiating methionine restored vs NONSEC and B.4 SEC with SP removed vs combination of SEC with SP retained, NONSEC and NONSEC with average length of SP removed. All high, medium and low confidence LSPs, the remaining UNCLASSIFIED and SPT proteins are of interest to applying LSPpred.

SequenceServer 1.1.0.beta12 [Help & Support](#)

Paste query sequence(s) or drag file containing query sequence(s) in FASTA format here ...

Protein databases [\[Select all\]](#)

- ☒ NONSEC
- ☒ SEC
- ☒ SPT
- ☒ UNCLASSIFIED
- ☐ HOMODIMER
- ☒ HIGH
- ☒ MEDIUM
- ☒ LOW

Advanced parameters: ? [BLAST](#)

Figure 4.2: Example LSPblast interface with available sub-databases correlating to classes or confidence lists from the LSPDB

then processed by *makeblast* and presented via a web interface (**Figure 4.2**) as `http://blast.lspdb.org` using the convenient SequenceServer (Priyam et al., 2019) package. BLAST queries return proteins that are most similar to the input query along with the LSPDB classification, so proteins that have similarity to putative secretory proteins can easily be identified.

There are 6 sub-databases, with each corresponding to a class or confidence lists (NONSEC, SEC, SPT, UNCLASSIFIED, high, medium, low) from the LSPDB. Each sub-database includes all protein isoforms for each class. The SequenceServer interface allows for protein FASTA sequences to be pasted (or a file “dragged” into the interface), and then which databases to BLAST against. BLAST results are displayed, and FASTA files of hits and the conventional BLAST reports are available for download.

The choice of databases is dependent on the the question; selecting only the confidence list or unclassified proteins will identify which query proteins have identified with *Arabidopsis* proteins that observed, but lacks context relative to other proteins. Selecting only some of the databases may obscure that although proteins with some similarity have

been observed in *Arabidopsis*, others with more similarity have not.

A competitive search would select all sub-databases, and compare sequence similarity of putative LSPs relative to SEC, SPT or NONSEC. As a whitebox tool, LSPblast currently allows the user to use these results as desired to relate query proteins to LSP candidates. The relative ratio of LSP candidates to NONSEC, and position in the BLAST hits, may be informative. For example, the total proportion of confidence lists LSPs in the hits, or the relative rank of the best hit. Another strict criteria could take only the top hit. As simple measure could be a percentage approach, considering the number of observed (UNCLASSIFIED, including the confidence lists) compared to the total number of BLAST hits. We will demonstrate this competitive use of LSPblast in Section 4.4.5.

Prior to publication, the interface (**Figure 4.2**) is available at:

url: `http://blast.lspdb.org`

user: `reviewer`

password: `LSPDB_review_2019`

4.2.6 LSPfam

LSPfam is built on top of PfamScan (version 1.6), a perl tool that takes a set of HMM and identifies their presence in FASTA files of sequences. Implemented as a script, the LSPfam algorithm in pseudocode is as follows:

Given an `input_file`, `pfam_database` and `secretory_pfam` files:

- if `input_file.hits` does not exist:

Run `pfam_scan.pl` on all input proteins against `pfam_database` and
save to `input_file.hits`

- For each line of `input_file.hits`:

- if line contains a PFAM is in `secretory_pfam`:

keep line

Collate number of `secretory_pfam` hits per protein, and report True if `> 0`

This has the effect of only keeping lines from `Pfam_scan` that contain the PFAMs of interest. `Pfam_scan` is run with the default settings, with the addition of the `-clan_overlap`

parameter since the presence of a secretory PFAM is the primary goal, and whether there is overlap between PFAM domains is less relevant. The algorithm implemented as a shell script will be available at <https://github.com/LSPtools>. The data sources for this script are the entire PFAM database (available at <https://pfam.xfam.org>) and exclusive and majority secretory Pfams as identified in **Tables 3.4** and **3.5**).

4.2.7 Training random forest models on unbalanced data

Though the positive and negative data were defined separately for the LSPpred and SPLpred approaches, they share the common goal to classify LSPs, so a LSP a classification ML model (i.e. not a regression) is most suitable. Both also use the same large number of initial protein feature representations (ProFET) without prior knowledge of which features are biologically relevant. Both also include the NONSEC data as negative, which is the majority of proteins in the LSPDB and ensures that the data is unbalanced; there are many more negative examples than positive. Though they are independent tools, selecting the same ML approach for both assist with both the creation and testing of the ML tools, and ensures the selected model address the data and feature aspects they have in common.

There are many ML models for classification that have been applied to bioinformatics including support vector machines, neural networks, and decision trees (Larrañaga et al., 2006). Decision trees consist of cascading nodes of tests that lead to classifications, with each node based on the input features. A useful property of decision trees for LSPs is that these features can be measured by their Gini importance (or simply importance) to the tree. These importances for LSP prediction could be used to infer biological meaning and suggestions for future experimental work. Random Forests (RFs) are an ensemble method that uses multiple decision trees of randomly selected features to build the “forest”, and a vote amongst them used to classify (Breiman, 2001). The Gini importance can similarly be averaged for each feature in each of the individual trees.

Another key consideration is the source of input data. For LSPpred in particular, the LSP data is selected from the LSPDB, and far from the ‘gold standard’ desired. Therefore, a general purpose ML model such as RFs is appropriate to explore predicting in

this space, over more complicated methods. With a goal of predicting LSPs along with gaining insight into what features might indicate biological function, RFs were chosen.

The RFs were created using the Python *scikitlearn* implementation (v0.19, Pedregosa et al. (2011)). In general however RF generalise poorly when trained on imbalanced data. To overcome this, there are several techniques to balance out the data. These include under sampling the majority class, oversampling the minority class, as well as more complex weighting or data generation methods such as SMOTE (Chawla et al., 2002). To account for the unbalanced nature of the positive and negative data whilst keeping the feature insight from RF, the imbalanced learning (Lemaître et al., 2017) extension was used which provides a *BalancedRandomForestClassifier* (BRF) extension to RF. The input data for each version of SPLpred/LSPpred was split into training and test sets (75% train, 25% test), with stratification to ensure the relative proportion of positives and negatives was maintained between the training and test data.

The BRF set was then trained using 5-fold cross-validation to estimate the accuracy. Aiming for the convention of a false positive rate of 0.05, from the cross-validated ROC curve we estimate a prediction threshold by taking the mean of the 5 thresholds that do not exceed 0.05 FPR across the folds. We can also estimate the TPR at this threshold by taking the mean TPR at the same thresholds. Cross validation also identified the top most important features, and for each predictor these were ranked and the upper quartile selected. The entire data set was then trained using the RBF model on the reduced set of features. The cross-validated threshold is then used to calculate the accuracy of the complete model on the held out test data by making a positive prediction for any value exceeding it.

Given the class imbalance with fewer positive data points in each version of the predictors, the balanced accuracy, defined as the average of the accuracy of positive and negative data was used. Using this metric to compare the versions ensures that performance on scarce positive data is equally weighted with being able to correctly identify true negatives. To estimate bias in the modified predictors of SPLpred, each trained model was then applied to two versions of the SPT data set, with and without the SP removed. The kernel density estimation (KDE) plot of both prediction distributions was used, and the

mean difference in prediction scores for each protein with and without SP calculated. A similar test was performed using any modified or unmodified versions of the training data; for example, if SP removed data was used as training input, then scores on full length versions of the sequences were used as a comparison. Since these data has been seen by the model, any deviation is an indication of bias due to input modifications.

4.2.8 Evaluation

We treat each version of LSPpred and SPLpred as an independent prediction method: trained on its own modified or selected data, and cross-validated to select the most relevant features and estimate the accuracy and recommended prediction thresholds. The methods used for LSPpred and SPLpred differ only in the choices of input data and modifications to them, if any. LSPpred included training data of putative LSP proteins from the LSPDB, whereas SPLpred used only the SEC and NONSEC data. These reflect the different underlying hypotheses (unique or common features) and inform whether either UNCLASSIFIED or SEC proteins from the LPSDB form the positive class, and what the tools are trained to compare against as negative data. These independent predictors can then be compared to identify which version is best suited for the purpose of predicting LSPs. For LSPpred the accuracy and relevant features are compared between versions. For SPLpred, we compare the accuracy and performance on the SPT data (which is unseen to the SPLpred method, and contains SPs) to estimate the magnitude of modification bias inherent in SecretomeP in this SecretomeP-like method. For both predictors, the putative low confidence LSPs and remainder of the UNCLASSIFIED set are proteins of interest to which the chosen version of each predictor can be applied, as **Figure 4**. Results on these data is not evidence of accuracy, but useful for identifying common and unique proteins predicted by LSPpred and SPLpred. Accuracy can tentatively be estimated using the four (including HYG) generally accepted plant LSPs (see Section 1.5).

4.2.9 Applying trained models to new data

After training, the selected models are written to disk using the *joblib* library. Python scripts for LSPpred and SPLpred take FASTA input, convert it to numerical features using ProFET, and then apply the prediction model that has been stored in the binary *joblib* file to each sequence. If the model probability exceeds the determined threshold, a True prediction is returned. The tool can be invoked using the Python command line:

```
python LSPpred.py input.fasta
python SPLpred.py input.fasta
```

The source code and instructions for the tools will appear in online repositories available at: <https://github.com/LSPtools>

4.3 Results

The comparisons between the alternative versions of the “black box” tools, and application of all the tools on the test data is shown in this section. For each tool accuracy as estimated by cross-validation in a ROC curve, a confusion matrix applied to the randomly selected 25% test data, and an estimate of the bias between modified and unmodified data is shown. The test data is stratified to include relative proportions of positive and negative data to the training data, giving 25 and 53 positive test cases for LSPpred and SPLpred respectively. For each the balanced accuracy is noted, where the accuracy on both positive and negative data equally weighted, and in some cases the predictors can achieve high balanced accuracy despite a low number of true-positives, as these matched by a low number of false-positives.

4.3.1 Performance and relevant features of LSPpred

LSPpred version 1 (negative data drawn from the NONSEC class) had an internal cross fold AUROC of 0.82 ± 0.05 (**Appendix Figure C.5**). The mean threshold to achieve a FPR of 0.05 was 0.701, which achieved a TPR of 0.3675. On the unseen test data it achieved a

Name	Threshold	Balanced Accuracy
LSPpred1	0.5	0.71
LSPpred1	0.65	0.60
LSPpred2	0.5	0.65
LSPpred2	0.68	0.63
SPLpred1	0.5	0.94
SPLpred1	0.61	0.95
SPLpred2	0.5	0.81
SPLpred2	0.70	0.72
SPLpred3	0.5	0.82
SPLpred3	0.70	0.76
SPLpred4	0.5	0.84
SPLpred4	0.66	0.83

Table 4.2: Summary of LSPpred and SPLpred balanced accuracy at the default (0.5) and FPR controlled threshold.

balanced accuracy of 0.71, including predicting 16/25 LSPs using a default threshold of 0.5 (**Appendix Figure C.11**). Applying the more conservative threshold of 0.645 reduced the balanced accuracy to 0.6 with far fewer FPs, and 6/25 in the test set (**Appendix Figure C.12**).

The second version utilising a larger negative data set had a similar mean AUROC of 0.84 ± 0.05 across the 5-fold cross validation, with a slightly lower threshold of 0.68 to achieve a FPR of 0.05 (equivalent TPR 0.355) (**Figure 4.3**). The balanced accuracy was 0.65 (14/25 LSPs predicted) (**Appendix Figure C.13**) and dropped to 0.63 (8/25 LSPs predicted) when using the empirical threshold (**Appendix Figure C.14**). Notably the balanced accuracy remains similar despite almost halving the true positives, as using the higher threshold decreased the false positive count by 627.

Considering the potential for bias in SP-containing proteins, therefore by design neither version of LSPpred uses modified proteins. Calculating the difference in prediction scores of modified and unmodified data results in a mean difference of 0.03 to 0.05 for all sets (**Appendix Figure C.6, Figure 4.4**), which will affect the prediction outcome only in edge cases where the prediction score is within ± 0.05 of the threshold, and could cross the boundary from positive to negative prediction (or vice-versa).

Notably for LSPpred2, each of the SP, SPT and NONSC proteins were included in the training data. Overall these differences are not major and suggests there is no systematic

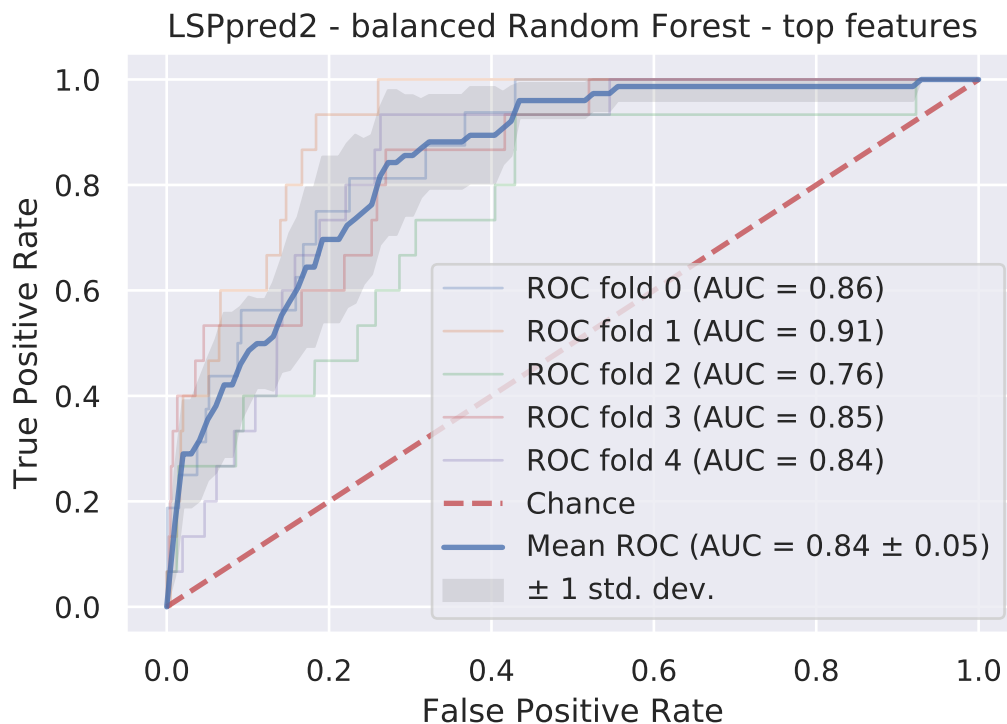


Figure 4.3: ROC curve for LSPpred version 2

SP-related bias in LSPpred1 or LSPpred2. Given the lack of modification in training data, these results serve as a baseline guide for differences in predictors; that is, without any modification, there is still some observed differences that could arise from some protein features occurring in the N-terminal region of the protein, and these differences can be compared to predictors that do specifically include modified data.

4.3.2 Performance and relevant features of preferred SPLpred model

In the first version of SPLpred, modelled directly on the approach by SecretomeP, the cross-validated AUROC is 0.99 ± 0.01 (**Figure 4.5**). This highly accurate model has a true positive rate of 0.946 when controlling FPR to be less than 0.05, using 0.61 as the prediction threshold. The balanced accuracy of 0.94 (weighted 41 / 43 LSPs predicted to 139/2037 false positive) (**Appendix Figure C.15**) on the test data actually increases to 0.95 (40/43 correct LSPs) when using the prediction threshold as there are fewer false

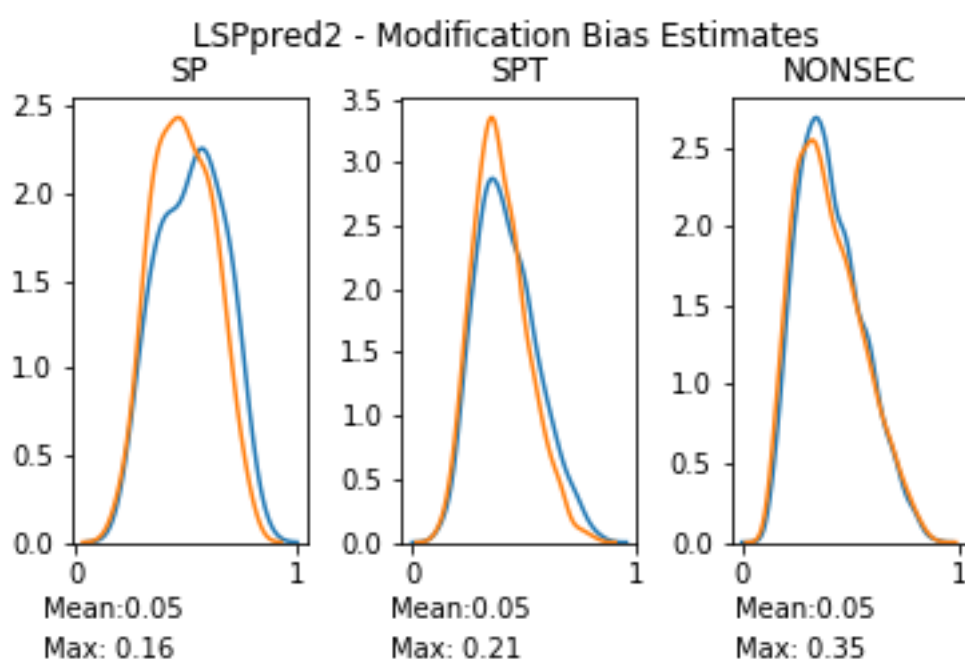


Figure 4.4: Score distribution for LSPpred 2 with (blue) and without (orange) modifications on SEC, SPT and NONSEC proteins

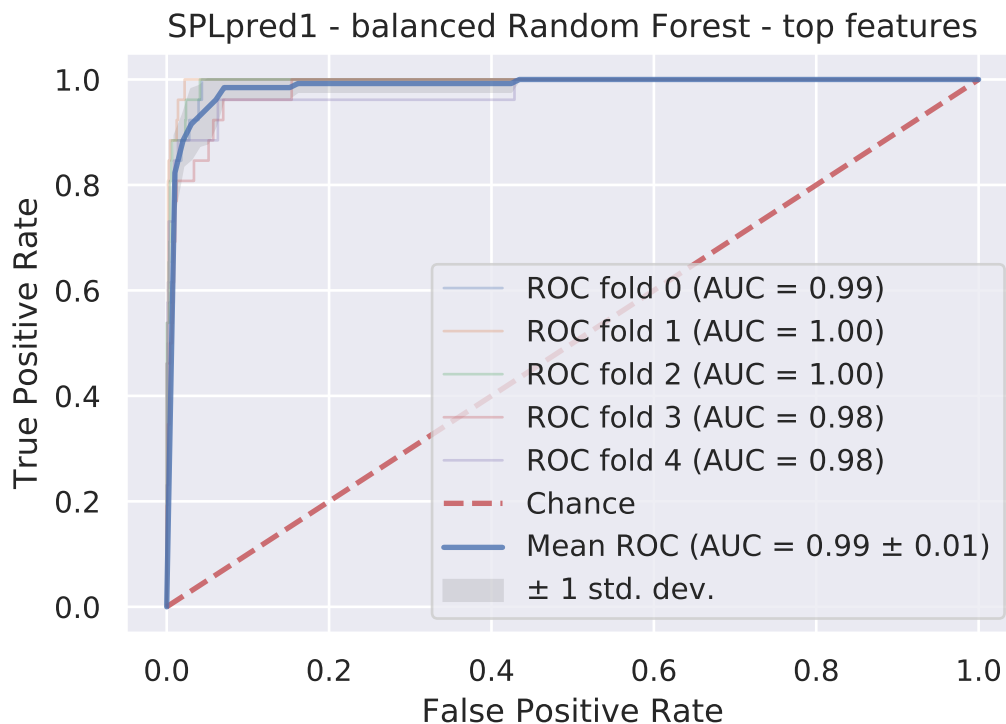


Figure 4.5: ROC curve for SPLpred version 1

positives (**Appendix Figure C.16**).

Bias estimation (**Figure 4.6**) reveals that the highly accurate model has noticeably different distribution of prediction scores between modified and unmodified data. Reminiscent of the work presented in **Chapter 2**, the accuracy on SPT proteins has a noticeable shift to lower scores for some proteins. Furthermore the accuracy on the negative training data (which was unmodified) shifts to higher scores when modified. Since this modification will resemble the cleaved SP training data, the suspiciously high accuracy of SPLpred1 seems to be driven by the modifications, and the absence of the SP (and therefore the initiating methionine) is recognised as a feature of secretory proteins. The mean score difference for all sets is also between 3 and 10 times greater than the LSPpred baselines, and so the modifications in SPLpred1 systematically bias the results.

SPLpred2 was designed to overcome bias by modifying both the positive and negative data, removing the SP from the positive and a sequence of average SP length from the start of negative data sequences. By doing so, the overall accuracy by AUROC is

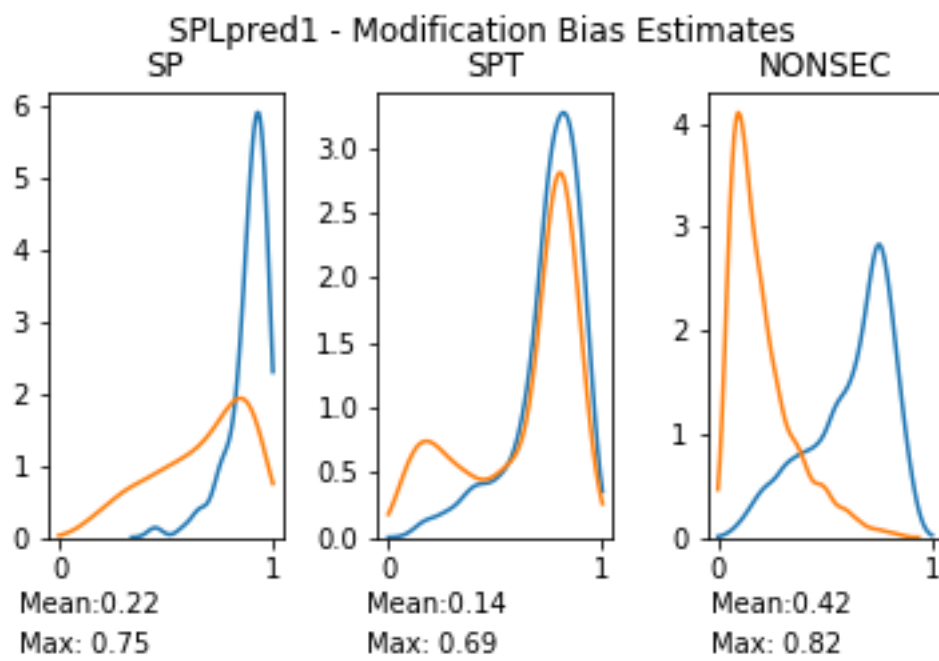


Figure 4.6: Score distribution for SPLpred1 with (blue) and without (orange) modifications on SEC, SPT and NONSEC proteins

0.88 ± 0.06 (**Appendix Figure C.7**), using a threshold of 0.70 to achieve a TPR of 0.62 when FPR is 0.05. When applied to test data, the balanced accuracy of 0.81 (35/43 LSPs) (**Appendix Figure C.17**) drops to 0.72 when using this higher threshold (21/43 LSPs) (**Appendix Figure C.18**). Importantly, the bias shift reduced with the mean difference between scores 0.1 (SP), 0.05(SPT) and 0.06 (NONSEC) (**Appendix Figure C.8**) is lower compared to SPLpred1.

Tempering the modifications by changing both positive and negative data does reduce the bias, at the cost of modifying more data. LSPpred3, which restores the leading methionine to the positive data only has an AUROC of 0.90 ± 0.05 (**Figure 4.7**) is comparable with LSPpred2, and similar threshold of 0.70 achieves a mean TPR of 0.62 controlled for 0.05 FPR (**Appendix Figure C.19**). The test data results have a balanced accuracy of 0.82 (35/43) and 0.76 (24/43) when using threshold (**Appendix Figure C.20**), which are also broadly similar. The bias distribution is also similar, and on average the difference in score is lower than for SPLpred2, with 0.07 (SP), 0.04(SPT) and 0.04 (NONSEC) (**Figure 4.8**).

The final alternative, SPLpred4, approaches balancing the modifications by adding replicates of data to the negative class. The inclusion of modified and unmodified data in the negative class achieves a similar AUROC of 0.89 ± 0.04 (**Appendix Figure C.9**) and with a threshold 0.67 a 0.58 TPR is achieved. The held out test data gave a balanced accuracy of 0.84 (37/43 LSPs) (**Appendix Figure C.21**) and reduced slightly to 0.83 with the threshold applied (31/43 LSPs) (**Appendix Figure C.22**). The bias however increases, with average difference between modified and unmodified scores 0.15 (SP), 0.06(SPT) and 0.1 (NONSEC) (**Appendix Figure C.10**).

4.3.3 Results from reporter tools

By comparison, the “white box” classification results are more transparent, though are not as amenable to accuracy metrics as TPR, FPR etc do not directly apply as they do not make predictions. Nonetheless we demonstrate their use on the low confidence LSP data set (red) of 353 proteins. LSPblast operates as a conventional BLAST database. To gauge it’s utility, we submit the 353 low confidence proteins as a query against the high

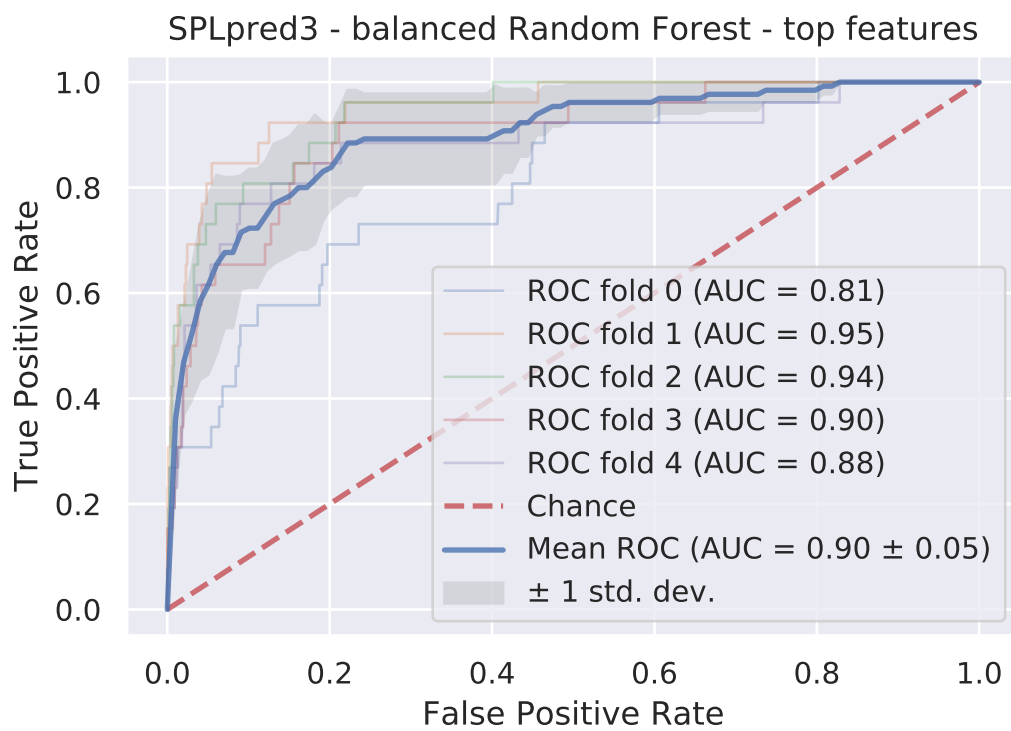


Figure 4.7: ROC curve for SPLpred version 3

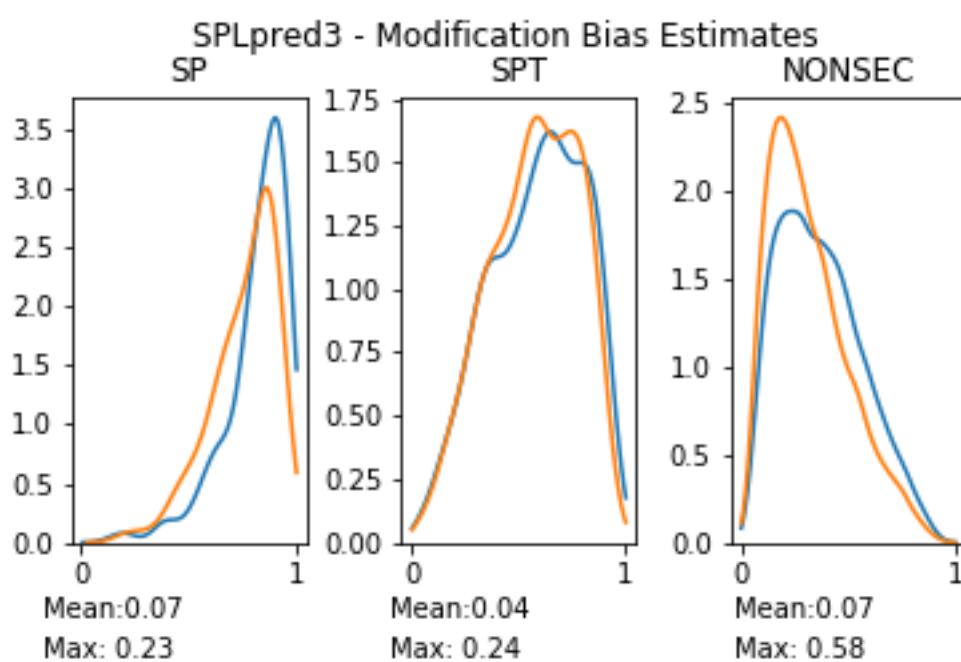


Figure 4.8: Score distribution for SPLpred3 with (blue) and without (orange) modifications on SEC, SPT and NONSEC proteins

(green) and medium (orange) confidence LSPs to demonstrate searching for sequence similarity between proteins of interest and proteins putatively assigned as LSPs. Only 75 of the 353 proteins had hits (e-value cutoff 1×10^{-5}) to the high or medium confidence set. For example, the low confidence ascorbate peroxidase 4 (TL29, AT4G09010) unsurprisingly shares sequence similarity with ascorbate peroxidase 1 (AT1G07890) from the high confidence set. The LSPblast interface recovers this relationship. A Circos visualisation of the queries can also be downloaded from the database as shown in **Figure 4.9** for AT4G09010, demonstrating that several isoforms of ascorbate peroxidase 1 have similar sequence regions to ascorbate peroxidase 4.

LSPfam is similarly hard to quantify in terms of accuracy. One measure is to consider how SEC proteins can be identified by secretory PFAM alone. Applied to LSPfam, 159 proteins of a total 850 SEC proteins in the LSPDB are identified as having a secretory PFAM. Accounting for the inclusion of isoforms of a single gene in this total of 850, the recall of LSPfam is low, and not sufficient to identify all secretory proteins. We would not expect all LSPs to share PFAMs, so a similarly low rate for LSPs could be assumed in the scenario that LSPfam performs the same on CSPs and LSPs alike. Applied to the 353 low confidence proteins (**Table 4.3**), 6 proteins occur that include the secretory PFAMs (see **Table 3.5**). The overlap between LSPfam and LSPblast is small, with only two ascorbate peroxidase proteins common to both from the low confidence set, indicating they can be complementary approaches.

Accession	Description	LSPblast predicted
AT1G07750	RmlC-like cupins superfamily protein	No
AT1G09310	Protein of unknown function, DUF538	No
AT5G01610	Protein of unknown function, DUF538	No
AT4G08390	stromal ascorbate peroxidase (SAPX)	Yes
AT4G09010	ascorbate peroxidase 4 (TL29)	Yes
AT4G14570	acylaminoacyl-peptidase-related	No

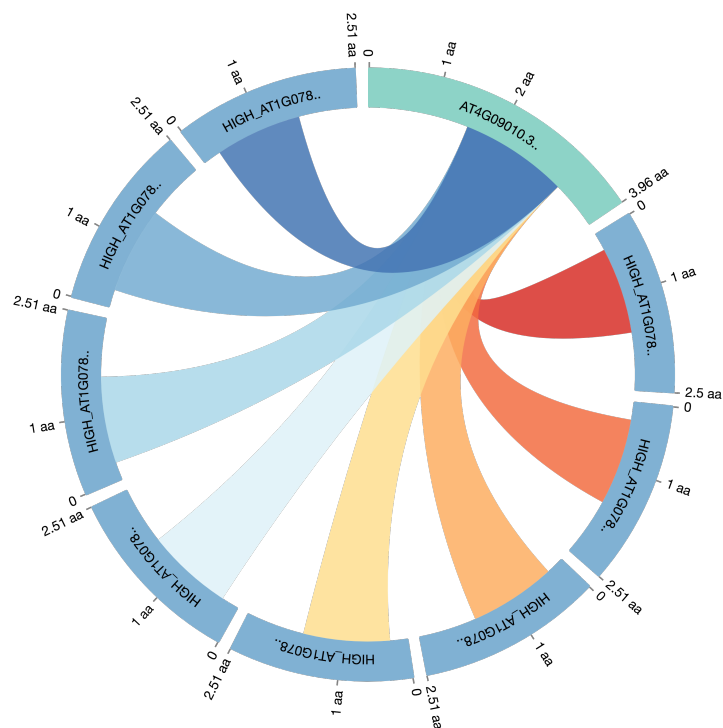
Table 4.3: Consensus LSPfam predictions from low confidence LSPs, with those with sequence similarity to high or medium confidence lists as identified by LSPblast shown

4.4 Discussion

4.4.1 SPLpred3 and LSPred2 are the preferred models

For LSPpred and SPLpred, we now select the preferred model to use for LSP prediction. The LSPpred versions are highly correlated (Pearsons $r = 0.878$), however, LSPpred2 in training has observed a variety of data (i.e. inclusive of SP and SPT proteins) and is therefore trained to differentiate between LSPs and both NONSEC and SP containing proteins. Given this added complexity without loss of accuracy, it is the favoured of the two since the unknown inbuilt features of the model were trained in the context of proteins with SP as negative data. The aim is that the tool therefore not recognise common aspects of CSPs as a signal for LSPs, which is a more appropriate assumption for the SPLpred models.

SPLpred is designed for SP protein-specific features using modified SP data, and the lessons of **Chapter2** motivate using the the bias estimates to separate the predictors. As shown, SPLpred1 appears extremely accurate (a warning sign in itself) yet highly susceptible to bias from the modification to the input data. It was deemed unreliable due to this SP bias, which demonstrated the data modifications of removing amino acids from the N-terminus were recognised within the model. SPLpred4 took another unconventional approach, doubling the training data with modified and unmodified versions of positive and negative data. Though the accuracy is reasonably high, the bias estimates are the second highest after version 1. Between SPLpred2 and SPLpred3, which removed all or most of the SP from the positive data only, accuracy is similar and the bias estimation separates them. SPLpred3 was selected as the preferred model. The bias differences are the



lowest of the set, and comparable with those LSPpred2, which as a tool is not reliant on modifications and gives a good estimate of the inherit difference in protein composition the modifications can make to affect overall score.

These two preferred models will now be referred to as LSPpred and SPLpred respectively. For each, the threshold (LSPpred: 0.68, SPLpred: 0.7) estimated to control the FPR to around 0.05 is used for a positive prediction. Any prediction score either equal to or exceeding the threshold is taken as a positive prediction.

4.4.2 Comparison of results between LSPpred, SPLpred, LSPfam and LSPblast

As with the “white box” tools, their application to the 353 low confidence or unclassified proteins from the LSPDB is of interest. The different models also allow us to apply one tool to the training data of the other, namely apply SPLpred to the high and medium confidence data used to train LSPpred (see **Figure 4.X**). Almost half (3/8) high confidence proteins are predicted, whereas just over 10% (11/93) medium confidence LSPs are predicted by SPLpred. The two tools are not expected to be mutually exclusive or in complete agreement due to the differences in the underlying model. Features specific to LSP proteins and common secretory features would be expected to exist in some LSPs, particularly when there are specific mechanisms for non-classical secretion. Applied to the low confidence set, 132 proteins are predicted by either SPLpred or LSPpred to be LSPs. Of these, only 21 are predicted by both (**Table 4.4**). This subset of the low confidence proteins is discussed below in terms of common predictions with the other tools. Further discussion in **Chapter 5** relates to these proteins in terms of the protein function, and an expanded confidence set. Notably, four of the proteins listed in **Table 4.4** only have positive predictions in either SPLpred and LSP pred if the lower bar of 0.5 confidence is used, rather than the thresholds established to maintain FPR under 0.05, and demonstrate that the threshold used can effect interpretations of these tools. All had one prediction above the higher threshold. As discussed extensively **Chapter 2**, the guiding thresholds are important to use if false-positives are to be minimised, and even though these four have a second tool predicting them, there is no guarantee that the positive pre-

diction overrides the negative, and a conservative approach would be to exclude these.

This overlap, however, does give a measure of increased confidence in the subset, as they are LSPs selected using multiple different theoretical models. This overlap between the black box models shares one prediction with LSPfam and six with LSPblast. In total, 7 proteins are identified in 3/4 for the models, and none in all. Thirty one (31) more proteins were predicted by either LSPpred or SPLpred (not both) and LSPblast, and two by both LSPblast and LSPfam only. Given the relatively unknown status of the low confidence set, consensus on all predictions or at least one positive prediction of all proteins is not expected. However, the small common subsets between some of the models shows that no one tool is in agreement with all the others. LSPpred and SPLpred have an estimated TPR of approximately 60% to control for false positives (of which there are likely some), and so around 40% of LSPs are typically expected to be missed by each. LSPfam recalls only 159/850 SEC proteins, and is not sufficient to cover all LSPs either, and LSPblast can compare to the identity of high or medium confidence LSPs only. Therefore to classify as many LSP candidates tools as possible, an ensemble method of classification is needed.

4.4.3 Proposed ensemble workflow for secreted proteins in plants

A proposed computational workflow for identifying plant secreted proteins is shown in **Figure 4.10**. Firstly identify classically secreted proteins containing SPs, e.g using annotations, or prediction software such as SignalP (step 1) . These can be either removed from the workflow or assigned as classically secretory, depending on the purpose of the analysis. In **Figure 4.10** those are assigned secretory. Those that remain after step 1 can then be applied to any of the four tools described here: LSPpred, SPLpred, LSPfam and LSPblast, or the LSPDB itself.

Three of the computational tools (LSPpred, SPLpred, and LSPfam) are provided as command line programs. Note that automatically integrating BLAST databases is a direction for future work, however at the moment LSPblast is intended as a manual tool for small numbers of proteins. For whole proteomes or large sets of results, using either LSPDB or LSPblast may prove a bottleneck to efficient identification, and requires level

Accession	Description	Symbol	Whitebox
AT1G02500	S-adenosylmethionine synthetase 1	SAM1	-
AT1G07750	RmlC-like cupins superfamily protein	None	LSPfam
AT1G50480	10-formyltetrahydrofolate synthetase	THFS	-
AT1G79530	glyceraldehyde-3-phosphate dehydrogenase	GAPCP-1	LSPblast
AT2G30050	transducin family protein / WD-40 repeat family protein	SEC13B	-
AT2G36530	<i>Enolase</i>	<i>LOS2</i>	<i>LSPblast</i>
AT3G16410	nitrile specifier protein 4	NSP4	LSPblast
AT3G54660	<i>glutathione reductase</i>	<i>GR</i>	-
AT3G55800	sedoheptulose-bisphosphatase	SBPASE	-
AT3G57560	N-acetyl-l-glutamate kinase	NAGK	-
AT4G14960	<i>Tubulin/FtsZ family protein</i>	<i>TUA6</i>	<i>LSPblast</i>
AT4G26970	aconitase 2	ACO2	-
AT4G37870	phosphoenolpyruvate carboxykinase 1	PCK1	-
AT5G08300	Succinyl-CoA ligase, alpha subunit	None	-
AT5G09660	<i>peroxisomal NAD-malate dehydrogenase 2</i>	<i>PMDH2</i>	<i>LSPblast</i>
AT5G14040	phosphate transporter 3;1	PHT3;1	-
AT5G18170	glutamate dehydrogenase 1	GDH1	-
AT5G19350	RNA-binding (RRM/RBD/RNP motifs) family protein	None	LSPblast
AT5G37600	glutamine synthase clone R1	GSR 1	-
AT5G43940	GroES-like zinc-binding dehydrogenase family protein	HOT5	-
AT5G64380	Inositol monophosphatase family protein	None	-

Table 4.4: Consensus LSPpred and SPLpred predictions from low confidence LSPs. Proteins in italics in consensus only when predictions between the default and FPR controlled thresholds (0.5 - 0.68 for LSPpred, 0.5 to 0.70 for SPLpred) are taken as positive.

of judgement on behalf of the user. The speed and automation of software tools, with true or false results for LSP status, is better suited to larger scale analysis. The workflow therefore assigns LSPblast and using the LSPDB as optional steps.

Secondly, the non-SP proteins can be applied to each of LSPpred, SPLpred, and LSPfam in parallel (step 2). If any results in a positive, then a protein is assigned an LSP predicted status. Proteins that are not predicted in all three programs can then passed to the manual filter stage (step 3). For high throughput usage, this may not be feasible and then proteins assigned a negative LSP status. Otherwise, LSPblast can be used to compare the sequences against any of the high, medium and low LSPs to look for sequence similarity. Common sequence with other proteins that have previously been identified can be used to assign LSP status. Alternatively, further manual analysis can be made by using the LSPDB, noting that the LSPDB is currently only applicable to *Arabidopsis*. These remaining proteins could be analysed in the LSPDB by considering the experimental context of the proteins, including the local and whole network protein-protein interactions amongst them, and the previous studies the protein may have appeared in, to make a determination if they should be assigned as LSPs.

Depending on the size of the analysis, it may be preferred to rank the predicted LSPs in order to prioritise those for experimental validation. Each prediction tool could be assigned a weighted score, and the sum of scores used to identify the most confident proteins. This score could be added as step 2 of the workflow to collect the number of tools each protein predicts. For example, LSPpred could assigned a score of 2 (i.e since it is trained on unmodified putative LSPs) and SPLpred and LSPfam a score of 1 due to modification and use of secretory data as training. This method would allow consensus proteins (e.g. **Table 4.6**) to be further ranked.

4.4.4 Predictor feature importances

Random forests (RFs) were chosen as the machine learning model to ensure that some insight into how the models perform was available. Of the 21 low confidence proteins predicted by both LSPpred and SPLpred, 14 did not have a result with either LSPfam or LSPblast, and as such the origin for their selection is unknown: only the estimates

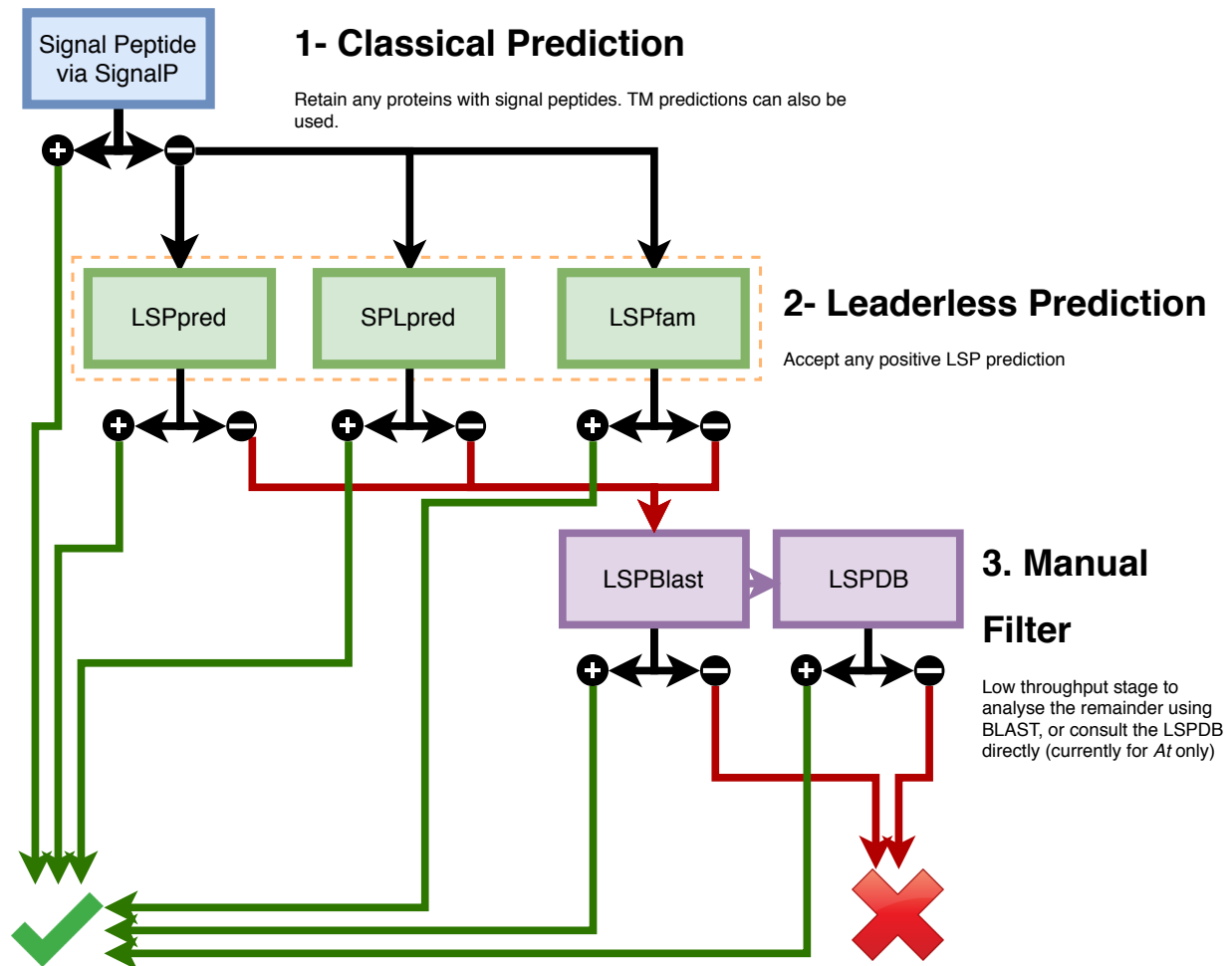


Figure 4.10: Recommended workflow for identifying secretory proteins including LSPs. 1) Identify classically secreted proteins, 2) identify LSPs using the combined predictions of LSPpred, LSPfam and SPLpred, 3) Manually inspect the remaining using LSPBlast, or the LSPDB if applicable

of accuracy during cross-validation and applied to the test data gives some measure of assurance as to their utility. This applies to any predictions made by these tools, and so to offer some insight into the actual features being considered and directions for further analysis, the RF importances can be extracted from the ML models and compared to suggest protein properties that contribute to the prediction.

Of the original models for LSPpred, there were 119 shared features between LSPpred1 and LSPpred2 (312 and 301 total features respectively). For SPLpred, SPLpred2 (309), SPLpred3 (289) and, SPLpred4 (292) shared a total of 114. Recall that each model was based on the upper quartile of important features from the original set of 1170. The RF importance can be viewed for each prediction tool, as shown for the top 20 features for LSPpred and SPLpred shown in **Figure 4.11** and **Figure 4.12**, respectively. In total SPLpred3 (289) and LSPpred2 (301) share 99 features. The Gini importance sums to 1, and the top ranked in each contributes the most to prediction. For each predictor, the feature with the highest Gini importance scores are still quite low. “Secondary Structure Transitions 23” contributed 3.3% in SPLpred and “G entropy” 1.5% to LSPpred, respectively. For each predictor, then the remainder Gini importance (96.7%, 98.5%) comes from a long tail of other features suggesting that the models are based on the combination of many protein elements. We consider then only the highest features of importance in each selected model, with the top 20 features for the alternative models shown in **Appendix Figures C.1, C.2, C.3 and C.4** for LSPpred1, SPLpred1, SPLpred2 and SPLpred4 respectively.

Three observations from the highest ranking features can be made. Firstly, a notable difference between SPLpred and LSPpred is the presence of particular amino acid features in LSPpred, including glycine entropy as the most important and the frequency of glycine is also amongst the most important. In the ProFET implementation, entropy is calculated in the information theory sense, related to the frequency and probability of observed an amino acid in a sequence. Amino acid features using compressed alphabets are also present in the top 20 features of LSPpred versions. Two reduced alphabets created for ProFET (Ofer et al., 2015) group amino acids together to reduce from 21 to either 14 in ofer14 or 8 in ofer8. The groups are as follows, with additional amino acids for ofer8 in parentheses: LIVM(A), RK(H), ST(NQ), (FYW) and (DE). The ofer8 asparagine entropy

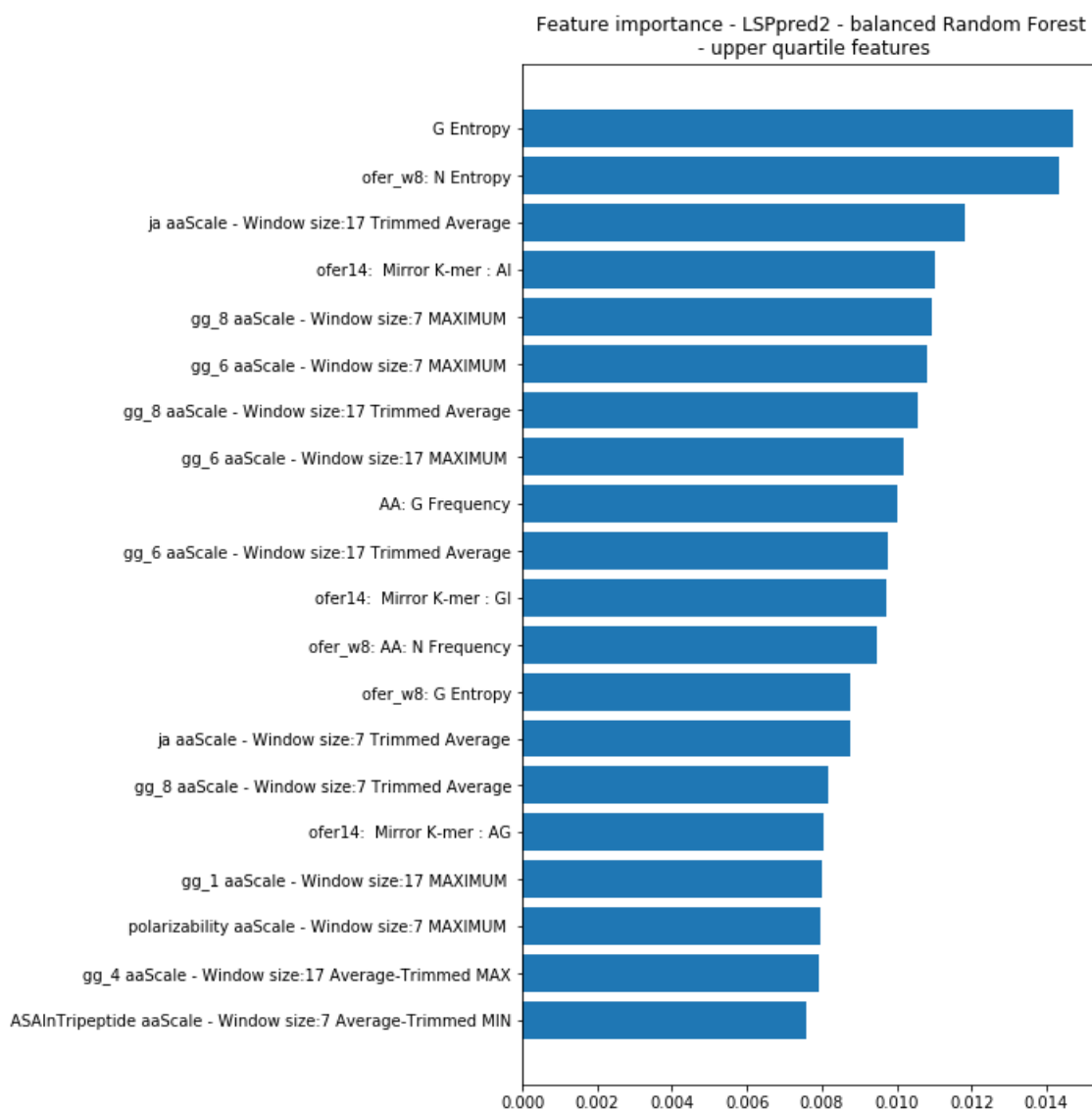


Figure 4.11: Top 20 features by importance for feature selected LSPpred version 2

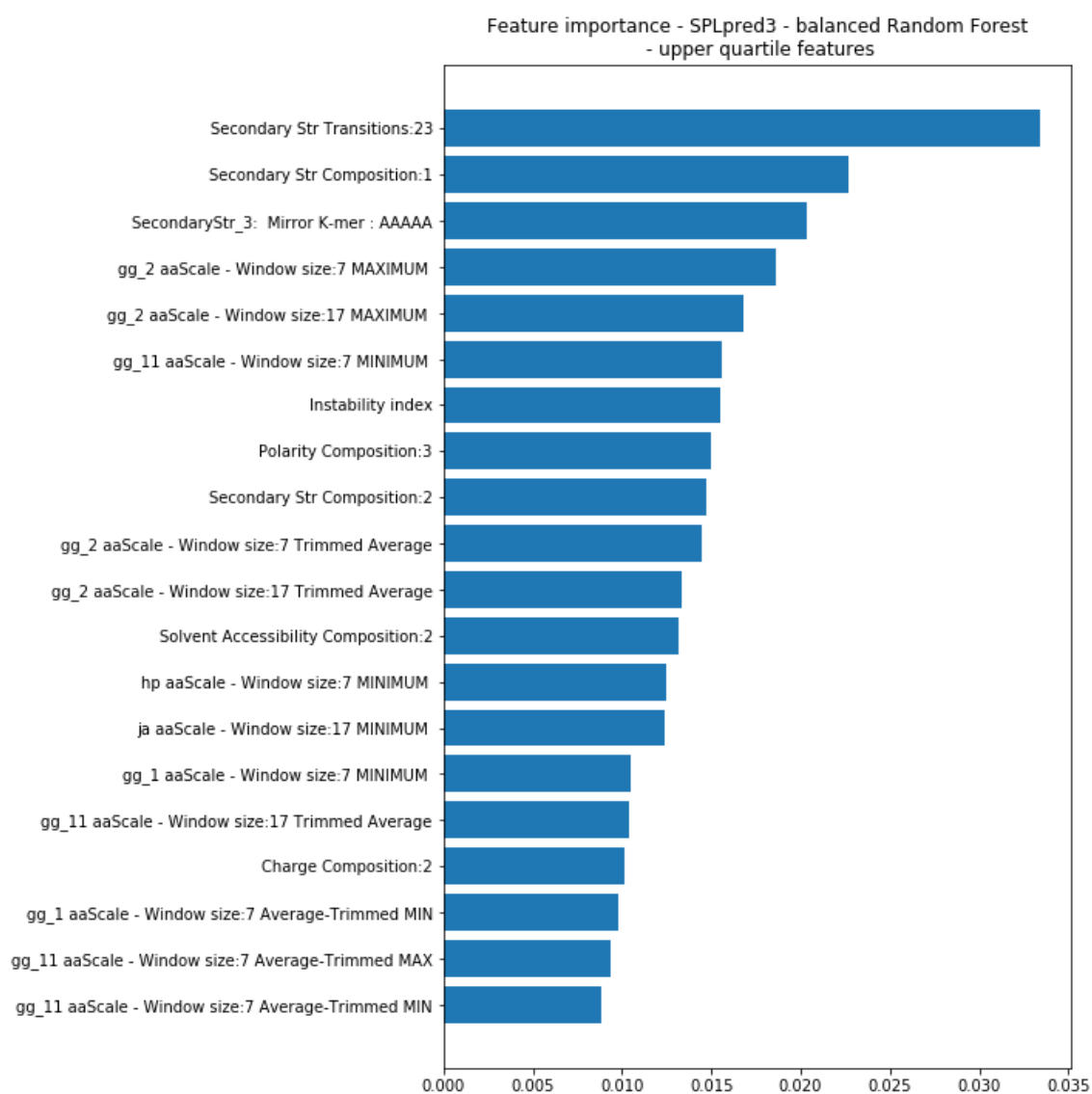


Figure 4.12: Top 20 features by importance for feature selected SPLpred version 3

is second in importance, and G entropy encoded under offer 8 is also highly ranked. Asparagine frequency is also featured under offer8 conditions. The offer14 alphabet also reveals mirror k-mers of (where either order of the combination AI, GI, AG occurs). For each of these, I represents one of leucine, isoleucine, valine or methionine.

That both entropy and frequency of glycine and asparagine features are ranked highly suggests that the presence and location with the sequence of these single amino acids is contributing to importance, however whether this is due to frequent or infrequent inclusion of these amino acids in the putative LSPs is known. As more LSPs are verified, it will be worth comparing the distribution of these two amino acids in LSPs and CSPs. Secondly, LSPpred and SPLpred both share importance features based on amino acid scales. In these scales, each amino acid is given a numerical value, and then they are often combined with a sliding window that calculates either the maximum, minimum or trimmed-mean score for that scale. The gg scales are from Georgiev (2009) which derived them to explain variance in the AAindex set of proteins, and the ja hydrophobicity scale is as defined by Janin (1979). For LSPpred, ja, gg-6,gg-8, occur more than once, whereas for SPLpred gg-2 and gg-11 occur multiple times. The gg scales often have comparisons to known indices, suggesting that for LSPpred β -sheet and C-terminal helix caps (bond patterns at the ends of helices which can include glycines) motifs are relevant, and for SPLpred α helices. The gg-11 scale has no correlated common index, demonstrating that despite using derived scale features not every one is informative and able to be linked to known structures.

Finally, whilst both predictors contain features of windowed scores based on scaled amino acid properties, only SPLpred alone contains in its top features those related to the summarised composition, transitions, and distribution (CTD) syntax (Dubchak et al., 1995) of reduced alphabets. This summary approach takes reduced alphabets that often consist of only two or three categories. It adds for each property encoded in CTD the proportion of the protein (one feature for each category) (Composition), the number of transitions from one to the next (Transitions), and the distribution using proportions across the 1%, 25%, 50%, 75% and 100% of the sequence. The most highly ranked CTD encoded alphabets represent protein properties such as polarity, polarizability, charge,

solvent accessibility and secondary structure. These alphabets can also be used to create reduced 3 letter alphabet (similar to ofer14 or ofer8) and subsequently mirror k-mer of them identified. Another non-CTD feature in SPLpred is the instability index as implemented is Expasy (Gasteiger et al., 2003). The most important three features for SPLpred are all related to secondary structure; the number of transitions from strand to coil amino acids, the number of helical amino acids and the AAAAA mirror k-mer representing five consecutive amino acids from the helical encoding (EALMQKRH). The number of amino acids over 10.4 polarity, strand structures, exposed amino acids and neutral charge proteins are also highly important. Taken together with the gg scales for α -helices, the most relevant features for SPLpred are all related to the physical properties of amino acids, and relate to the secondary structure as compared to the amino acid related features identified in LSPpred.

Without more verified LSPs, we can only speculate on implications of these different features between the two programs. Given that the training model for SPLpred explicitly removed the primary SP secretory mechanism, all that remain are largely functional components to the protein. We can hypothesise that SPLpred is recognising these components in the secondary structure. It is also possible to suggest that the contrast between the two models' features is informative. Parsimony would suggest that ML models will explain the data in the most efficient way possible. With the caveats that we have only taken the highest ranked features, and recalling the long tail of importances, we could expect simple motifs that are biologically relevant to be ranked higher over complex combinations of various functional or structural properties. For LSPpred, given a broad range of extracellular functions for the LSP candidates and a limited number of mechanisms for LSP secretion to occur, the inclusion of simple amino acid frequency and entropy based features of asparagine and glycine offers a direction for further work in how the LSP candidates may find their way into the secretory pathway. Both these amino acids are considered targets for the cleavage site of GPI-anchors (Pierleoni et al., 2008). Another secretory pathway aspect could be N-linked glycosylation, which conventionally begins in the ER and continues in the Golgi, with the glycosylation occurring on the asparagine protein. Given that one function of UPS proposed is as a secretory route that avoid PTMs, this and

other PTMs related to asparagine and glycine (e.g. acetylation, myristoylation (glycine)) could be an area of interest further investigation with respect to their relationship to LSPs.

4.4.5 Testing on plant LSPs

As has been reiterated throughout this **Thesis**, the lack of well characterised plant LSPs has motivated approaches to using either proxy data (such as CSPs) or putative data (such as the LSPDB confidence lists). The cross validated accuracy scores, control for FPR at 5% and test estimates provide some measure of the performance of LSPpred and SPLpred, but are not a substitute for independent data. With the caveats that it there are too few to be statistically confident, we can apply these tools to the best LSP candidates in plants. Importantly, this limited data is independent of the training data in these tools. As per Section 1.5, three of the candidates are not in *Arabidopsis*, and SAMS2 is in the low confidence list, and therefore unseen by both LSPpred and SPLpred.

We can apply LSPpred and SPLpred to these four protein sequences, along with other relevant tools for comparison as follows. Amongst the LSP prediction program published, SRTpred is one of the still few available online and functioning, along with SecretomeP. The tools compared are predictors for LSPs (LSPpred, SPLpred, SRTpred and SecretomeP), general sub-cellular localisation (DeepLoc, Almagro Armenteros et al. (2017)) and a plant apoplastic proteins from either plant or pathogenic source (ApoplastP from Sperschneider et al. (2018) and discussed further in Section 5.2). LSPpred predicts all four of mannitol dehydrogenase (MDH), hygromycin phosphotransferase (HYG^R), Helja and S-adenosylmethionine synthetase (SAMS2). SPLpred predicts 50%, noting that the FPR control requiring 0.70 confidence led to one negative result and can be considered a near miss. As expected given **Chapter 2**, SecretomeP does not perform well on these proteins. Interestingly Helja is predicted by both ApoplastP and SRTpred, noting that for ApoplastP in particular a plant effector model was used to train the program, independent of SP. It is therefore more likely to identify pathogen related LSPs. DeepLoc (Almagro Armenteros et al., 2017) is a general tool, and gives scores for multiple locations. None of the four were predicted apoplastic, and the individual extracellular prediction scores are all low.

Predicting all four plant LSPs by LSPpred is a welcome yet inconclusive result. Since the underlying model is based on putative LSPs, CSPs are not an appropriate proxy for testing its accuracy as was shown in **Chapter 2**. Ultimately, increasing numbers of verified plant LSPs from species other than *Arabidopsis* will be required to evaluate its accuracy, and false positive rate, however the ability to predict this independent test data combined with the internal cross validated TPR of 35% with a FPR at 5% indicates that LSPpred is able to identify some plant LSPs with acceptable estimates of error. Furthermore, since HYG^R is not truly a plant LSP, this protein highlights that the use LSPs from other systems could shape future experimental designs, testing which LSPs maintain unconventional secretion between organisms, and possibly revealing common secretion mechanisms. It is also another independent test, with evidence for protein secretion in *Arabidopsis* yet independent of the *Arabidopsis* proteome, and therefore separate from the data used to create LSPpred.

Tool →	SecretomeP (Mammalian)		DeepLoc			SRTpred		ApoplastP		LSPpred		SPLpred	
Protein	LSP	Score	Predicted	Score	Extracellular	LSP	Score	Status	Score	LSP	Score	LSP	Score
HYGR	No	0.513	Peroxisome, Soluble	0.386	0.007	Non-Secretory Protein	-0.362	Non-apoplastic	0.93	Yes	0.714	No	0.474
SAMS2	No	0.421	Cytoplasm, Soluble	0.730	0.0018	Non-Secretory Protein	-1.5007	Non-apoplastic	0.8	Yes	0.839	Yes	0.716
MTD	No	0.450	Cytoplasm, Soluble	0.554	0.0042	Non-Secretory Protein	-0.9236	Non-apoplastic	0.82	Yes	0.789	No	0.666
Helja	No	0.447	Cytoplasm, Soluble	0.742	0.0878	Secretory Protein	0.05923	Apoplastic	0.81	Yes	0.690	Yes	0.941

Figure 4.13: Prediction scores from six programs (LSPpred, SPLpred, SRTpred, SecretomeP, DeepLoc, ApoplastP) on the best candidate LSPs in plants; (HYG^R), SAMS2, MDH and Helja. Green indicates positive prediction, red negative and orange a near miss based on the recommended threshold for SPLpred

For the whitebox tools, LSPfam does not identify any domains in these four. LSPblast returns hits for all except HYG, perhaps unsurprisingly given its *E. coli* origins. The top 6 hits for the remaining three are shown in **Figure 4.14**. Applying the percentage approach for blast hits, SAMS2 has 7/7 (100%) hits to observed proteins either UNCLASSIFIED or in the confidence lists, however it is present in database and it matches to isoforms of itself. The remainder of the hits are also SAMS proteins (SAMS1, SAMS3 and SAMS4) which were identified previously in **Table 3.15** and shows how the network and BLAST methods have identified the same proteins. MTD has only 3/31 (9.68%) observed proteins in the blast hits, though notably the low confidence probable cinnamyl alcohol dehydrogenase 9 (CAD9, AT4G39330) is the second best hit, however whether it has mannitol dehydrogenase activity has been disputed (Kim et al., 2007).

Helja has 24/47 (51.06%) observed hits, though from the blast outputs and **Figure 4.14** there is a distinct pattern to the alignments. For SAMS2 and MTD the identity and pattern of alignment suggests close identity between the proteins, and a homologous relationship. Interestingly for Helja, the top hits and several more of the total of 47 hits shown multiple occurrences of the majority of Helja sequences, spaced throughout the proteins sequence. This pattern illustrates however that although automated tools to parse BLAST outputs could simplify analysis, this observation was only made from the visualisation. These proteins hits relate to 13 distinct genes, two unclassified proteins from the low and medium confidence lists, and 11 either UNCLASSIFIED (4) or NONSEC (7). All are annotated as carbohydrate binding and annotated as jacalin-related lectins (NONSEC: AT1G73040, AT1G19715, AT1G52000, AT1G52100, AT5G49870, AT1G52040, AT1G33790; UNCLASSIFIED: AT3G16460, AT3G16470, AT3G16450, AT3G16430) The low and medium are nitrile specifier protein 4 (AT3G16410) and PYK10-binding protein 1 (AT3G16420), respectively, though both are also annotated jacalin-related lectins. Inspection via the LSPDB reveals the observed have been seen collectively in several publications (H Nguyen-Kim et al. (2016), Herve et al. (2016), Cheng et al. (2009a) Kwon et al. (2005) Charmont et al. (2005) Borderies et al. (2003) Casasoli et al. (2008).

The ML based tools each has a model of LSPs based on its training data. This model is unclear to the end user, and only the prediction as to if new data is an LSP is visible.

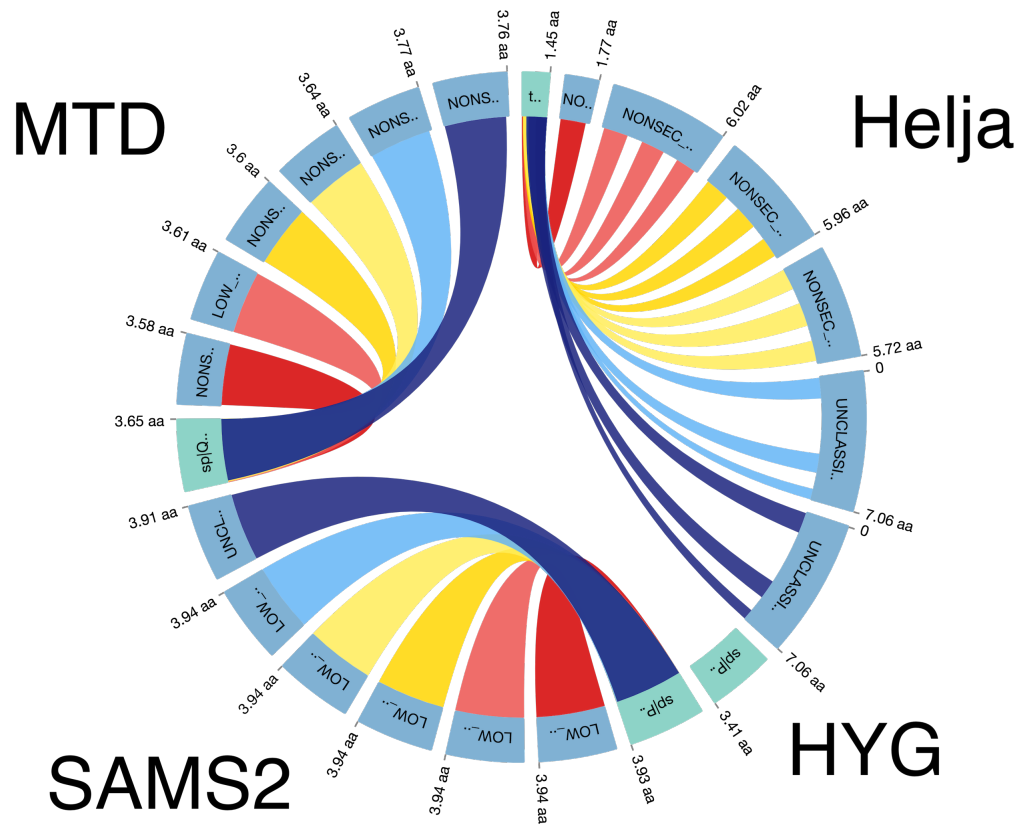


Figure 4.14: Circos plot of top LSPblast to four plant LSPs Helja, MTD, HYG and SAMS2 (each in green) to their top hits (blue)

This use of LSPblast for Helja shows how its model, in this case sequence similarity, is not only transparent, but how this allows the sequences it is based on to be identified, and potentially provide insight. Further work on these jacalin related lectins could design experiments around this insight, testing whether UPS was lost in the absence of one or more of the similar Helja sequence repeats.

4.4.6 Testing on LSPs from other organisms

As discussed in **Chapter 2**, using the same LSP tools for mammalian and plant species implies common features and mechanisms for leaderless secretion. The tools here are based on plant data, and intended to be plant-specific, however their application on the limited

number of LSPs in other organisms can suggest if assumption holds, or if the common features or extracellular spaces between organisms should be further investigated.

4.4.6.1 Human and disease related LSPs

Applying LSPpred and SPLpred to the same proteins indicates these tools do not reflect properties of mammalian LSPs. Only 2 of the original 13 human LSPs used by SecretomeP to evaluate its accuracy return a positive prediction from SPLpred only: Macrophage migration inhibitory factor (MIF) and galectin-3 (Gal-3). All other predictions were negative, though a related Galectin-1 (Gal-1) was just below the prediction threshold. Interestingly, one not predicted is the mammalian thioredoxin (Trx) though both predictors are on the positive side (LSPpred: 0.511, SPLpred:0.543). We also ran the tools on 11 additional LSPs related to human disease as collated by Kim et al. (2018b) not in the SecretomeP training data, including HIV TAT proteins, *Leishmania major* and *Plasmodium falciparum*. These additional results gave only one LSPpred prediction (0.69) for FAM3C though it is likely a false-positive as some isoforms are annotated with SP, and so it may be a secretory isoform, or misannotated as an LSP. Overall, few LSPs from humans are predicted by the plant trained tools, and these results do not suggest LSPpred and SPLpred have captured a shared mammalian-plant LSP signal, if one exists, and the tools should not be applied to human data.

4.4.6.2 Fungi

A fungal UPS protein, endochitinase Cts1 in *Ustilago maydis* gives positive predictions (LSpred:0.69, SPLpred: 0.72). An interesting experimental method was used for identifying UPS of this protein (Stock et al., 2012). A Gus (β -glucuronidase) reporter was used that is inactivated when passing through the conventional secretory pathway by N-glycosylation in the Golgi. Constructs with Gus and Cts1 were secreted with reporter active, indicating an unconventional secretion pathway.

Recent work on a potential motif for unconventional secretion was identified in yeast. A diacidic motif DE (aspartate , glutamine) when substituted with alanine resulted in a

reduction in unconventional secretion for both superoxide dismutase 1 (SOD1) and of acyl-CoA-binding proteins (Acb1) (Cruz-Garcia et al., 2017). Independent bioinformatics analysis (Padmanabhan et al., 2018) suggested that although DE is a common motif, the context in which it occurs may be relevant to UPS. The DE motif is present in all four plant LSPs, however it is indeed very common; for example it appears in 68.09% of *Arabidopsis* and 54.44% of rice protein sequences respectively. In the LSPDB combined confidence lists, using all isoforms, the motif occurs in 72.64% of sequences, and in 72.12% of the remaining UNCLASSIFIED. This suggests little difference and reiterates so more will be needed to be done to investigate this line of query.

Applying both these yeast proteins gives mixed prediction results; the superoxide dismutase gives consensus prediction results (LSPpred: 0.76; SPLpred:0.71), and ACB1 negative for both (LSPpred: 0.391; SPLpred:0.556). Notably, the DE77/78AA substitutions experimentally verified did not alter the positive prediction scores for SOD when the same change was made to the FASTa sequence, indicating that if DE is a true motif critical for plant LSP secretion, it is not captured in the tools created here. One limitation of the DE analysis work (Padmanabhan et al., 2018) was that it was applied only to 60 proteins with the DE motif. An avenue for future work could be to consider the distribution and context of this motif in LSP candidates compared to non-secreted proteins, including proteins both with and without DE.

4.4.6.3 Bacteria

As noted in **Chapter 1**, there are now over 12 secretion systems in bacteria, including those with conventional SP. There are several prediction programs for them and though the lack of availability of other LSP prediction programs was noted in **Chapter 2**, the Gram-positive and Gram-negative versions of SecretomeP (Bendtsen et al., 2005) are still available. As the revised analysis of StecomeP Nielsen et al. (2018) concerned only the eukaryotic version, and there are many more positive examples of LSPs, in bacteria, we have not considered its accuracy or applicability to plants. This increased number of LSPs in bacteria allows for direct ML approaches, rather than modifying secretory proteins under the common-feature hypothesis, and PenNGaRoo (Zhang et al., 2019) used

an ensemble of ML methods to predict Gram-positive LSPs, using as positive data experimentally verified bacterial LSPs.

The supposed conservation of properties between mammalian LSPs and plant LSPs motivated the earlier parts of this research, and the plant-based tools created here have so far shown little accuracy on human proteins. Considering instead a conservation with bacteria is motivated however by the curious case of the *E. coli* HYG resistance gene (HYG^R). The original gene is *hph* (*aphIV*) and is widely used as a marker for resistance (Miki and McHugh, 2004). The HYG^R protein sequence is not predicted to contain a SP. This was found to be an LSP when expressed in transgenic plants (Zhang et al., 2011a; Zhang and Li, 2016). It has generally been accepted as a plant LSP in several reviews (e.g. Ding et al. (2014); Davis et al. (2016)), yet the question remains if this is truly a plant LSP. One approach is to ask if it is recognised by bacterial LSP predictors.

To investigate this, we applied both Gram-positive and Gram-negative SecretomeP bacterial versions, as well PeNGaRoo. The expectation that PeNGaRoo would not predict this as an LSP, as *E. coli* is Gram-negative, was confirmed with a negative result (probability: 0.417). As noted previously, LSPpred predicts this as an LSP (0.71) and is the only one of the programs that did so (**Appendix Figure C.23**). As a comparison, this test included the other three plant LSPs, and the results were unexpected. PeNGaRoo predicted all three (Helja, MTD and SAMS12). SecretomeP Gram-negative predicted Helja, yet not HYG with its Gram-negative origins. As with our evaluation on SPLpred and LSPpred, results on three positive example is not enough to infer with confidence that the tool is accurate on plant proteins. There is also a need for an estimate of the false positives, for which we do not have sufficient data. Unlike those evaluations on limited plant data, we can explore the relationship between bacterial and plant LSPs in reverse. PeNGaRoo took data from bacterial genera *Bacillus*, *Listeria*, *Staphylococcus*, *Streptococcus*, *Mycobacterium*, *Lactobacillus*, and *Lactococcus*. After similar sequences were removed, there is 191 positive and 480 negative bacterial LSP data. We can use this to invert the bacterial tool on plant data test, and analyse if the plant tools are accurate on bacterial data.

The surprising result is shown in **Figure 4.15**. Recall that in a ROC curve, the diagonal line represent random chance, and the ROC path ideally tracks to the top left before veer-

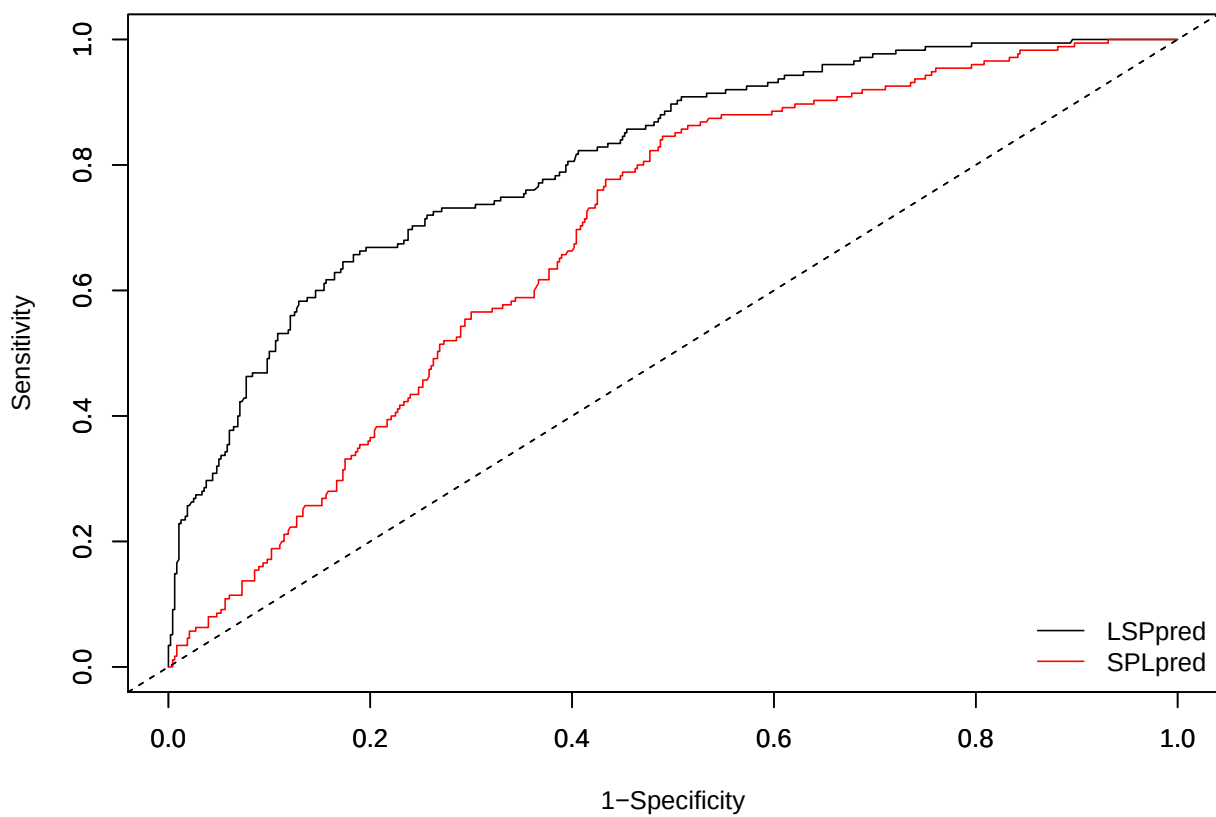


Figure 4.15: A ROC curve for PeNGaRoo training and test data when applied to LSPpred and SPLpred. A sensitivity of approximately 47% was achieved when False positives are 5%, using a threshold of 0.71. AUROC was 0.808 and 0.68

ing right, as the threshold decreases. Applying both SPLpred and LSPpred to positive and negative data, we can reconstruct an ROC curve. Both predictors are significantly better than chance (LSPpred p-value= 2.95×10^{-61} , SPLpred p-value= 3.31×10^{-18}) and LSPpred is in particularly clearly superior (AUROC: 0.81). Using the Generalized Youden Index (López-Ratón et al., 2014) implemented in the easyROC (Goksuluk et al., 2016) package to identify an optimal threshold score based on this ROC, we can estimate that for LSPpred using a threshold score of 0.78 identifies 46.3% of the bacteral LSPs, at a 7.7% FPR.

This result is by no means better than PeNGaRoo itself (TPR: 0.940 ± 0.008 , FPR: 0.0139 ± 0.029), however relative to random chance, and a prior expectation that the plant-specific tools should not work at all on bacterial data, this is suggestive of a commonality between bacterial and the putative plant LSPs selected from the LSPDB. Commentary from Rabouille (2016) suggested the possibility of UPS Type I (pore formation) and bacterial UPS having similarities , and this result could indicate that these kinds of UPS mechanisms have indeed been captured in LSPpred.

This preliminary result must be taken with an abundance of caution. Superficially, the results suggest that using curated putative LSPs from *Arabidopsis* in the LSPDB to build random forest models, and applying these models to experimentally validated Gram-positive LSPs produces accurate classifications significantly better than chance. A robust analysis must be applied to rule out if these results are due to the data being from only certain bacterial species, consider which of the ensemble sub-predictors in PeNGaRoo are most informative, and any other explanations for this result. Although it does not increase our confidence in the ability of these tools to identify plant LSPs, it is unexpected and is discussed further in the conclusion to this Thesis (**Chapter 6**) as an avenue for future work.

4.5 Conclusion

With the available data from the work presented in **Chapter 3**, we have produced four tools to aid in the discovery of LSPs. Taking the curated information from the LSPDB,

“white box” tools LSPfam and LSPblast are presented as low-throughput and low coverage methods that are transparent in assigning LSP status. Balancing these are the “black box” high-throughput methods SPLpred and LSPpred that have quantified accuracy and target false-positive rates. Importantly, each uses a different model and the agreement between each can be used to prioritise results in a pipeline.

The input data is drawn from the LSPDB classes. This leads to a more specific tool than just proteins with either SP or possible LSPs. SPLpred is based on common-features between CSPs that have been observed in the scope of the LSPDB (SEC) and LSPs, compared to negative data of proteins that have not been observed in the apoplast (NON-SEC). Using the entire *A. thaliana* proteome would not take into account the observations in the literature. The composition of the LSPDB allows a convenient division of protein data into the required training, test and evaluation data sets for building a tool. It also reduces (but does not eliminate) the inclusion of LSPs in the NONSEC negative data, thereby leaving these untouched from the training tool and an independent set of data.

The results from **Chapter 2** informed the design of SPLpred and estimating the inherent bias of using common features with secretory proteins avoided selecting the suspiciously accurate SPLpred1 as a model. The ML methods selected allowed for generating a hypothesis for common sequence profiles of plant LSPs. This speculation can be considered a dividend of the random forest models and feature importances they conveniently provide, and whether the speculated LSP mechanisms are shown in future work to be accurate remains to be seen. In practice, the SPLpred and LSPpred tools are accurate according to the bounds of the training approaches used to build them, and each controls for false positives to identify putative LSPs. Each tool also offers distinct hypotheses relating to emulating the biological context of candidates. SPLpred and LSPfam are based on finding proteins with similarity to known extracellular proteins, under the theory that these similarities imply similar suitability to the extracellular environment. Alternatively, LSPpred does not directly assume this, allowing for LSPs to differ from classically secreted proteins and typically inhabit a different subcellular environment, which may be relevant for moonlighting or stress-related secretion that occurs under specific conditions.

Future work using these approaches is dependent on the quantity and quality of data. This data availability is the major limitation of these “black box” tools, and the imbalance of scarce positive data to abundant negative data. The positive data for LSPpred was also selected from the LSPDB as putative LSPs, and would ideally be substituted for experimentally verified plant LSPs. This putative LSP data was, however, still quality-controlled, with restricted sequence similarity to avoid overconfidence on related proteins, and the exclusion of ribosomal annotated proteins as raised previously. Applying the prediction tools to the ribosomal-related proteins excluded from the training data (**Appendix Table B.3**) gives only one predicted positive (LSppred on AT2G18020) reinforcing the justification for their exclusion from the training data and indicating they likely don’t share common properties overall with other LSP candidates. Future work with LSPblast could include more nuanced analysis of the BLAST outputs, in particular the filtering of the partial hits, and specification of the regions that match putative LSP proteins.

Clearly, more needs to be done in protein prediction for LSPs in plants. Here, each of the tools presented contributes to improving the situation of it being better to refrain from prediction by offering plant specific tools for identifying LSPs. The combination of LSPblast, LSPfam, LSPpred, and SPLpred will each provide a different perspective. As tools suitable for identifying LSP candidates, each can contribute in prioritising proteins for functional confirmation of secretion with empirical testing.

Chapter 5

The Secret Secretome

5.1 Introduction

PLANT leaderless secretory proteins (LSPs) are hard to find using either experimental or bioinformatic approaches. The true number observed in proteomes has been obscured by the likelihood of intracellular protein contamination and distinguishing LSPs from contaminants remains a difficulty in analysing experimental results. This Thesis has used computational methods to address this challenge. In **Chapter 4** we produced a database (LSPDB), two prediction tools (LSPpred, SPLpred) and two classification tools (LSPblast, LSPfam). In this **Chapter**, the cumulative approaches to plant LSP discovery in this Thesis are critically evaluated, and the strengths, limitations and extensions to the work explored. In order to do so, three aspects of the Thesis need to be noted.

Firstly, the LSPDB and prediction tools have not relied on the limited specific examples of plant LSPs. Instead, the work has been built on methods to identify LSPs by their relationship to secretory proteins, either sharing protein features or through protein-protein interactions. Despite the small number, the results in **Figure 4.13** indicated that these independent methods were able to predict four putative plant LSPs (and curiously, bacterial LSPs). In the absence of experimental validation, which was beyond the scope of this Thesis, the independence of the predictions to either known or expected LSPs allowed these data to be used to assess how these approaches perform at predicting and identifying LSPs. More independent data can be used similarly and be divided into either *Arabidopsis* proteins from experiments not included in the LSPDB, or proteins from

other plant species.

Secondly, the classifications in the LSPDB, and inherited by the other LSP prediction tools, relate to the specific input data used to create them. Studies incorporated in the LSPDB include a decade of publications from cell wall proteomes (1997-2016), however, several studies published 2016 and 2017 were deliberately excluded for evaluation. These more recent publications between the completion of the LSPDB v1.0 and this **Chapter** allow a “time machine” technique (similar to Loctree Goldberg et al. (2012)), where we can assess how these computational predictions have held up in subsequent independent work, using these more recent excluded publications. These isolated data can be used to explore how the computational tools fit into an analysis.

Thirdly, a new prediction tool, ApoplastP, based on a combination of plant and fungal apoplastic proteins was published (Sperschneider et al., 2018). This program used a combination of apoplastic proteins from plants, and effectors from oomycetes and fungi. Since the data included SP-containing proteins, they also performed similar bias testing as we have shown here. In that work, they found no increase in FPR when SP was removed from effector proteins with SP that are secreted by pathogens and subsequently localise to the plant cytoplasm, and only a 2% FPR increase when the SP was removed from secreted plant proteins. As an orthogonal approach to plant secreted protein prediction, the results of applying ApoplastP on candidates identified by the new tools developed in **Chapter 4** is of interest. There may be expected to be some overlap with the predictions tools here, e.g. SPLpred and Apoplast, as both used secretory plant proteins as part of their positive data. Given though that the different training approaches, removing either the SP in SPLpred or complementing the data with pathogenic effectors in ApoplastP, complete consensus is not expected. Concordance prediction for LSPs from ApoplastP and LSPpred/SPLpred may identify effector-related predictions from the later tools.

To evaluate the tools in this context, in Section 5.2 we first apply the workflow outlined in **Figure 4.10** on a rice (*Oryza sativa*) proteome, and focus analysis on the results on the best characterised of these proteins. This includes a comparison of the outputs with that of ApoplastP.

In Section 5.3 the dynamic nature of the LSPDB is explored, by considering how protein classifications change with the addition of new observations from publications not currently included in the LSPDB. Similarly, in Section 5.4 we consider how the analysis on classifications in the LSPDB can change if additional data related to proteins is added to the LSPDB.

In Section 5.5 the results of using all the developed tools on several case studies is discussed, including remaining unclassified data in the LSPDB, external independent data, and previously published lists of LSPs. In particular, works citing **Chapter 2** are useful as they demonstrate an awareness of the leaderless secretory protein problem in plants, and make explicit decisions in the computation experimental method that affect the outcomes of their analysis. In these instances, we can compare the results of the new tools developed in this Thesis to what was actually undertaken, and explore the impact that the use of these new tools will have in practice.

An updated overview of *Arabidopsis* LSP candidates is presented in Section 5.6, including the biological insight into LSPs from these evaluations and case studies, followed by a conclusion in Section 5.7.

5.2 LSP workflow on the rice proteome

Looking at the agronomically important species of rice (*Oryza sativa subsp. japonica*) we took the reference proteome from UniProt (48907 proteins), and applied the combined workflow of methods in this Thesis (**Figure 4.10**). Before analysing the results on selected proteins, a kind precision check can indicate how the tools perform using rice proteins. The positive predictions for the limited *Arabidopsis* LSPs gives no indication of the specificity of the tools, i.e the false-positive rate. By applying the workflow to rice we calculate the percentage of the proteome predicted to be secreted as an indication of the overall size of positive predictions, which will consist of both true- and false-positives. The LSP workflow, taking the union of LSPfam, LSPpred and SPLpred results, identifies 8962 (18.3%) proteins as putative LSPs. As a comparison, SignalP predicts 4162 (8.5%) of the proteome as secretory. Given the hypothesis behind each of the tools, especially with

SPLpred designed to predict CSPs as well as LSPs, the LSP workflow may also identify secretory proteins. In the best case scenario there would be complete overlap between the SignalP and SPLpred predictions, and in this scenario the workflow would indicate ~10% of the rice proteome as LSPs. In the worst case there is no overlap, and so we can bound the workflow predictions between these two extremes of 8.5% and 26.8% of the rice proteome predicted as LSPs.

Without verified LSPs we cannot indicate how many of these predictions are false-positives, however the results do indicate a level of selectivity (irrespective of accuracy) in the tools, and that a limited number of candidates are selected from the proteome for prioritisation as LSP candidates. UniProt includes both manually curated and automatically annotated proteins. Since the full proteome may contain theoretical or redundant proteins, we used the manually curated subset of rice proteins (SwissProt: 4046, ~10% of the proteome) to further analyse the predictions of LSPs. The curation includes a review of the computational predictions, as well literature information, that gives a higher level of confidence in using the annotations of secretory features or sub-cellular localisations.

Of the these 4046 proteins, the UniProt annotations indicated 811 that are annotated as containing SP (512), TM (723), or both (88). Sub-cellular location includes 1252 annotated to the either nucleus or cytoplasm, or both. After filtering for manually curated proteins, and removing those without LSP workflow predictions due to sequence length or ambiguous amino acids, 4029 proteins remained. The individual tools initially classified a variable number of secreted protein candidates (SignalP (506), LSPpred (1055), SPLpred (382) and LSPfam (103)) that total 1553/4029 proteins, and the unique and common predictions between them are shown in **Figure 5.1**.

Firstly, we note the results of these tools on classically secreted proteins. SignalP predicts 506, though only 93.0% of these match the UniProt SP annotation. There is the possibility are false-positives in the annotation, or the similarity between TM and SP leading to false-positives from SignalP, however 93.0% concordance reinforces that SP prediction is still the standard in identifying secretory proteins.

SPLpred, our plant specific reimplement of the common-feature hypothesis, and SignalP agree on 212 classically secreted proteins, 41.8% and 55.5% of their predictions,

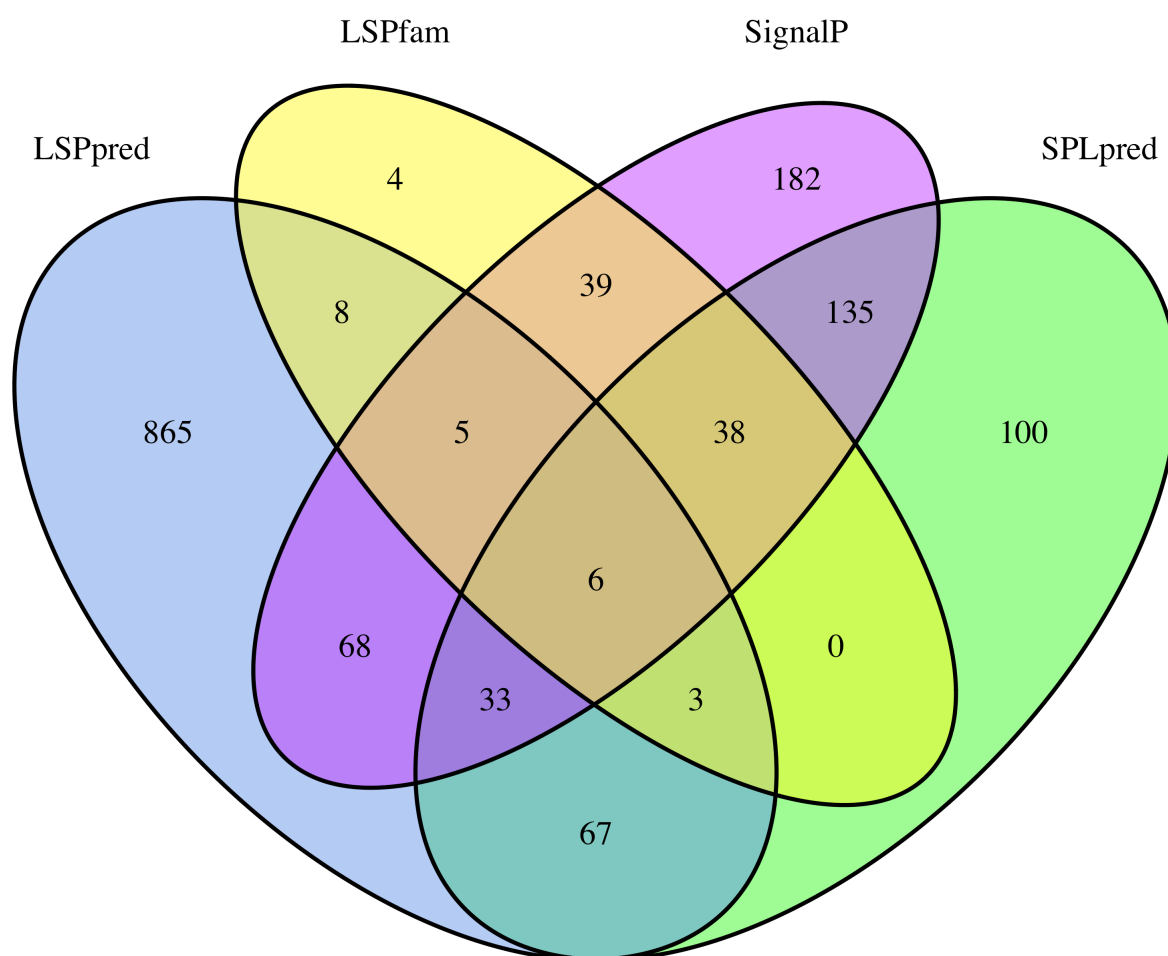


Figure 5.1: Venn diagram of the results of applying the automated (steps 1 and 2) stages of the recommended **Figure 4.10** on the reviewed subset of proteins in the UniProt (Swiss-Prot) reference proteome for *Oryza sativa subsp. japonica* (Rice).

respectively. SPLpred and the UniProt annotations similarly agree (216 in common, 42.7% TPR), indicating that SPLpred is not sufficient to identify all CSPs, let alone all the LSPs.

The recall of CSPs by LSPfam, estimated to be less than 20% based on *Arabidopsis* (see Section 4.4.2), is similar for rice, with 88 (17.4)% of the classically secreted predictions identified by LSPfam. Notably, only 15 other proteins are identified by LSPfam, 8 shared with LSPpred and 3 with SPLpred, indicating that predominantly identifies CSPs. LSPfam could reasonably be excluded from a secretory workflow based on this evidence, especially if probability estimates are required to rank candidates, which the “white box” approach does not provide. The transparency however may still be desirable, and knowledge of the PFAMs may be a warranted features, and there are several candidates unique to LSPfam.

LSPpred is training with putative LSPs as positive , and CSPs as negative data. Despite this 112 (10.6%) of the LSPpred predictions are also predicted by SignalP. Given the positive and negative classes used in training, these could be considered false-positives. In the workflow context, these FPs will not affect the list of secreted protein as they will already be classified as CSPs, however it does indicate that the FPR for LSPpred may be higher than the 5% estimated during model training. Initially, the inclusion of CSPs amongst the negative data in LSPpred attempted to force the model to penalise SPs, however the “black box” nature doesn’t reveal what features these proteins have that led to this classification, and it may be that these CSPs have features associated with LSPs by LSPpred. This result, as well as the observation that LSPpred had the most predictions (1055, 82.0% of them unique), suggests that results of LSPpred could be further scrutinised as outputs from the workflow, such as ranking them by confidence score.

After considers these CSPs, the remaining 1047 produced by the pipeline of secreted proteins are LSP candidates. The unique prediction for LSPfam are two probable L-ascorbate peroxidases, APX8 and APX3, as well as a cysteine proteinase inhibitor 2 (Os05g0494200) and the putative beta-glucosidase 15 (BGLU15). The majority of the proteins shared between LSPfam and LSPpred predictions are also members of the L-ascorbate peroxidase family (APX1,APX2,APX4,APX5,APX6,APX7) which are induced in plants by stress

including heat, cold, and salt (Pandey et al., 2017). Several of these are chloroplastic (APX5,APX6,APX7,APX8), and APX4 has a single TM domain. Another cytosolic beta-glucosidase 4 and similar beta-glucosidase-like SFR2 were identified. Thus in rice, LSP-Fam has primarily identified beta-glucosidases and L-ascorbate peroxidases. Generally beta-glucosidases may be expected to be secreted (e.g. BGLU7 in rice), and the annotation for BGLU15 lacks any SP, and suggests it could be a pseudogene product. This highlights one caveat regarding the workflow, and tools relating to features shared with CSPs. These similarities to CSPs are will lead to identification (for LSPfam) or prediction (SPLpred) of similar proteins. Without an experimental observations, these will be likely be false-positives, however if the workflow is applied post-experimentally, and such proteins are observed experimentally in the apoplast, then the similarity to CSPs can be used to prioritise validation.

Initially, 67 of these rice proteins are commonly predicted by LSPpred and SPLpred. Inspection of their annotations reveals two have SP that were not predicted by SignalP, 5 are single pass TM, and 13 are proteins with multiple TM domains. The workflow includes SP or TMHMM in step 1. This result confirms that there is an overlap between multi spanning TM proteins and the proteins identified by these predictions programs, and despite their absence from the training data, are required to ensure the remaining results of the pipeline are correctly annotated. The remaining 47 are listed in **Appendix Table D.1**, and several protein families are identified. Alcohol dehydrogenase are prominent, and LSPDB analysis identified cinnamyl alcohol dehydrogenase 9 (AT4G39330,CAD9) in the low confidence lists, as well as two GroES-like zinc-binding alcohol dehydrogenases (AT3G19450,AT5G51970). The protein list also contains four s-adenosylmethionine synthases related proteins are also present, recalling the SAMS2 LSP in *Arabidopsis*, and two superoxide dismutases, SODCC1 and SODCC2. In yeast superoxide dismutase 1 (SOD1) is a well characterised LSP (Dimou and Nickel, 2018).

There are 3 proteins identified by all three algorithms, LSPfam, LSPpred and SPLpred. Two are probable uridine nucleosidases (URH1,URH2) typically located in the cytoplasm, as well as Beta-fructofuranosidase, insoluble isoenzyme 5 (CIN5) which has a stress function. CIN is manually annotated as located in the extracellular space in

UniProt, despite the lack of an SP. Sequence analysis indicates the potential for glycosylation sites and disulfide bonds. The curation and assignment to a role as a cell-wall invertase is likely due to a study conducted by Cho et al. (2005) in which they did gene expression analysis and indicated a role for CIN5 in pathogen response. Since this analysis was transcript-based, there is no evidence of either sub-cellular location or secretion in this case.

This again highlights the caveat when using predictions prior to observation. Tools in this Thesis are primarily aimed at identifying LSPs, relative to an experimental observation. Predictions alone, without a matched experimental observation, have less weight. CIN5 is the only rice CIN gene without either an annotated or predicted SP. Tools here have likely identified shared features with the secreted members of the gene family, and despite the transcript evidence for upregulation after infection, more evidence and analysis is required before assigning it as an LSP. A version of the LSPDB for rice would assist in this analysis, a direction for future work discussed in **Chapter 6**.

In the interim, for observations of potential rice LSPs we return to the LSP review by Ding et al. (2012). This included a single proteomic study on rice, and putative root LSPs Shinano et al. (2011). The authors noted 60% of proteins had a SP, and the Ding et al. (2012) list identified 9 putative LSPs. Four of these are manually curated in SwissProt, and three are predicted to be LSPs by either one or both SPLpred and LSPpred **Table 5.1**.

Amongst these is salt stress-induced protein (LOC.Os01g24710, SALT), another example of alectin LSP candidate. Alignment with Helja indicates 32.42% identity (Clustal ω), and a salt induced protein in *Zea mays* was noted amongst sequences similar to Helja (Pinedo et al., 2015). In *Arabidopsis*, LSPblast returns a single hit, a jacalin-related lectin AT1G73040 that was noted on Section 4.7 to be a NONSEC protein with similarity to Helja. In that analysis, we noted the repetitive Helja sequences in several *Arabidopsis* proteins. With more LSP candidates, these kinds of comparisons become feasible. Alignment based approaches could be applied to investigate common amino acid sequences or motifs. With many LSPs from different species phylogenetic approaches can be used to explore the relationship between LSPs of different organisms, rather than merely identify them.

Accession	Description	LSPpred	Score	SPLpred	Score
LOC_Os07g30970 1	Nucleoside diphosphate kinase 1 (NDKR)	False	0.65	False	0.64
LOC_Os05g33380 2	Fructose-bisphosphate aldolase 1 (FBA1)	True	0.79	False	0.53
LOC_Os08g03290	Glyceraldehyde-3-phosphate dehydrogenase 1 (GAPC1)	True	0.78	True	0.73
LOC_Os01g24710	Salt stress-induced protein (SALT)	True	0.71	True	0.95

Table 5.1: LSPpred and SPLpred results for putative LSPs from Shinano et al. (2011) as selected by Ding et al. (2012) that are also manually curated in UniProt

Analysis of rice proteome allows us to compare the outputs of the tools developed here to ApoplastP. Applying the same curated subset of rice proteins, ApoplastP predicts 245 apoplastic proteins, of which 169 are predicted to have an SP. We therefore focus on the remaining 76 proteins, and the output of the tools developed here (**Figure 5.2**). Of these, 45/76 (59.21%) overlap with one or more of the tools developed in this Thesis, 14 are in agreement with both SPLpred and LSPpred, and 31 are uniquely predicted by ApoplastP.

The overlap between all three tools includes several LSP candidates noted previously including superoxide dismutases, a beta-glucosidase and three cinnamyl alcohol dehydrogenase proteins. The overlap also includes SALT, adding another prediction to this candidate observed by Shinano et al. (2011). The training data for ApoplastP however did include some secreted plant sequences, and may share some training data with SPLpred. The consensus between all three should take this similarity into consideration when ranking candidates. ApoplastP however also used distinct fungal effectors as training data, and the 31 unique predictions yield some distinct results.

The 31 proteins uniquely predicted by ApoplastP are a useful exploration of how this tool differs from those presented here, and can be used as a demonstration of the manual (step 3) part of the LSP workflow (**Figure 4.10**). Under this manual step, proteins not identified by any of LSPpred/SPLpred/LSPfam are then analysed using LSPblast, since rice is not currently present in the LSPDB. We use the 31 predicted solely by ApoplastP to note which do and do not bear resemblance to *Arabidopsis* proteins. Of the 31, 7 pro-

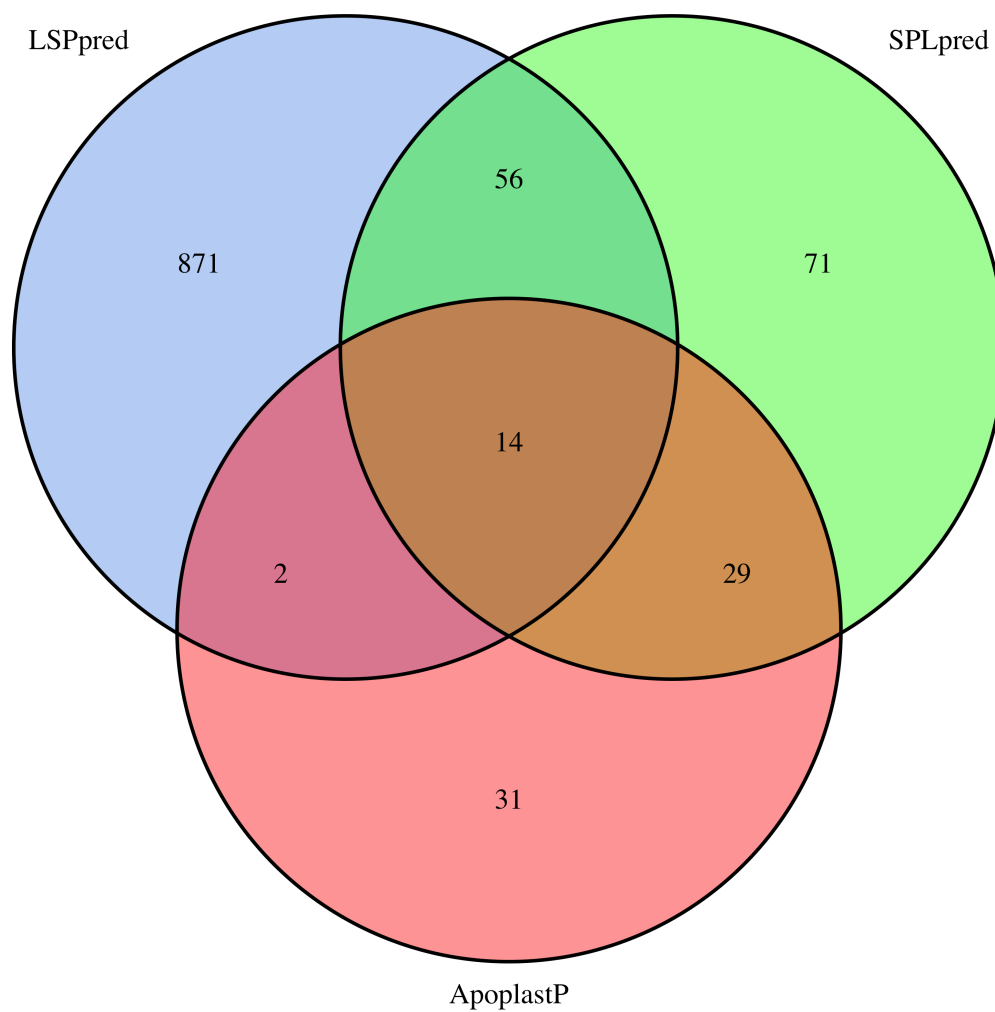


Figure 5.2: Venn diagram of positive LSP results (excluding SP or TM predicted) from LSPpred, SPLpred and ApoplastP on the 4029 subset of proteins in the UniProt (Swiss-Prot) reference proteome for *Oryza sativa subsp. japonica* (rice).

teins had no BLAST hits (**Appendix Table D.2**). From the annotation however, 5 of these represent protein with annotated TM domains. Only 6 have any BLAST hits to observed proteins in the LSPDB. A superoxide dismutase is predicted by ApoplastP solely, and a probable E3 ubiquitin-protein ligase (LOC_Os06g03800) is most similar to ankyrin repeat family proteins (AT2G03430, AT4G35450), each of which has been observed just once in separate *Arabidopsis* studies (Bayer et al. (2006) and Hervé et al. (2016)). A 60S ribosomal protein (LOC_Os01g16890) has 3 hits to proteins that are similar to L30 ribosomal proteins (NONSEC:AT2G34860, UNCLASSIFIED: AT1G77940, MEDIUM: AT3G18740). Given the conservative approach to ribosomal proteins, excluded from the training data, this orthogonal prediction of similar rice proteins could warrant further consideration of the plausibility of some of these proteins as LSPs further. AT3G18740 also appears in **Table 3.11** as amongst the frequently co-occurring PPIs, along with another ribosomal protein (AT2G18020). Hits of the two experimentally observed ribosomal proteins are either very lowly ranked compared to NONSEC, or blast to UNCLASSIFIED without further evidence in the LSPDB. The remaining 18 proteins (**Appendix D.2**) had hits only to NONSEC proteins. In contrast to those with LSPblast hits, ApoplastP predicts these with no discernible relation to contents of the LSPDB. If these are true LSPs, they would represent a gap in LSPpred and SPLpred. Several of the proteins identified are stress-associated zinc finger proteins (SAP1, SAP3, SAP5).

Therefore from this evaluation of the LSP workflow (**Figure 4.10**) we can make several adjustments based on the results from rice. Firstly, the presence of TM domains in some proteins predicted by SPLpred or LSPpred leads to making TM prediction or annotation at step 1 a requirement. If proteins contains classical features then they need not be run through the leaderless predictors. Secondly, the results from ApoplastP show that both unique and common proteins were identified, and gives it is trained on pathogen effectors, it complements the suite of tools in this Thesis. The updated workflow is shown in **Figure 5.3**. Returning to rice, the outcomes of this revised workflow in **Figure 5.3** on the curated subset of 4046 proteins (compared to the initial ~10% estimate on the entire proteome) are as follows. In step one, 816 proteins are diverted as classically secreted by SignalP or TMHMM (sourced from the UniProt annotations). Step two, applying

the high-throughput tools developed in this Thesis plus ApoplastP leads to 4 LSPfam, 26 ApoplastP, 871 LSPpred and 71 SPLpred unique predictions, with an additional 111 predictions by more than one tool. These 1083 (26.8%) total LSPs fall marginally in the range of 8.5 - 26.8% estimated from the entire proteome, yet couched with the observations regarding the accuracy of LSPpred could be further scrutinised for false-positives. However, the conflating of multiple LSP pathways in the prediction tools does allow for these totals to include LSPs that are only secreted under certain conditions, rather than constitutively secreted as in classically secreted proteins.}

5.3 LSPDB as a dynamic database

The classification of proteins in the LSPDB is dependent on both protein features and experimental observations. As more studies are added to the LSPDB, two transitions between classifications are possible. Proteins classified as SP_THEORY that are subsequently observed become SEC, and NONSEC that are observed become UNCLASSIFIED. After each addition, the criteria and classifications across the database could change; formerly SP_THEORY/NONSEC interactions, PFAM and GO terms may cascade other proteins to be reclassified. To assess how this change will affect the classifications, several studies from 2016 and 2017 were excluded Tsai et al. (2017); Duruflé et al. (2017b,a); Xu et al. (2016) so that we could simulate how the LSPDB will change over time, and offers a way to see if the inferences from the LSPDB can predict future experiments.

The four excluded studies identified 1715 proteins, both unique to a single experiment and common. A total of 1159 unique proteins in the N-Glycopeptide profile (183), cell wall proteomes (171 and 784) and mucilage proteins (21) were identified, with an additional 556 appear in up to three of the four studies. The distribution of the proteins is shown in **Figure 5.4**.

The approach of each of these papers to leaderless proteins varies. In the N-glycopeptide profiles (Xu et al., 2016), secreted proteins were typically found however the N-terminal glycosylation of AT3G52560, a cytosolic ubiquitin-conjugating enzyme, was a notable exception. If the computational tools available in this paper had been available, the in-

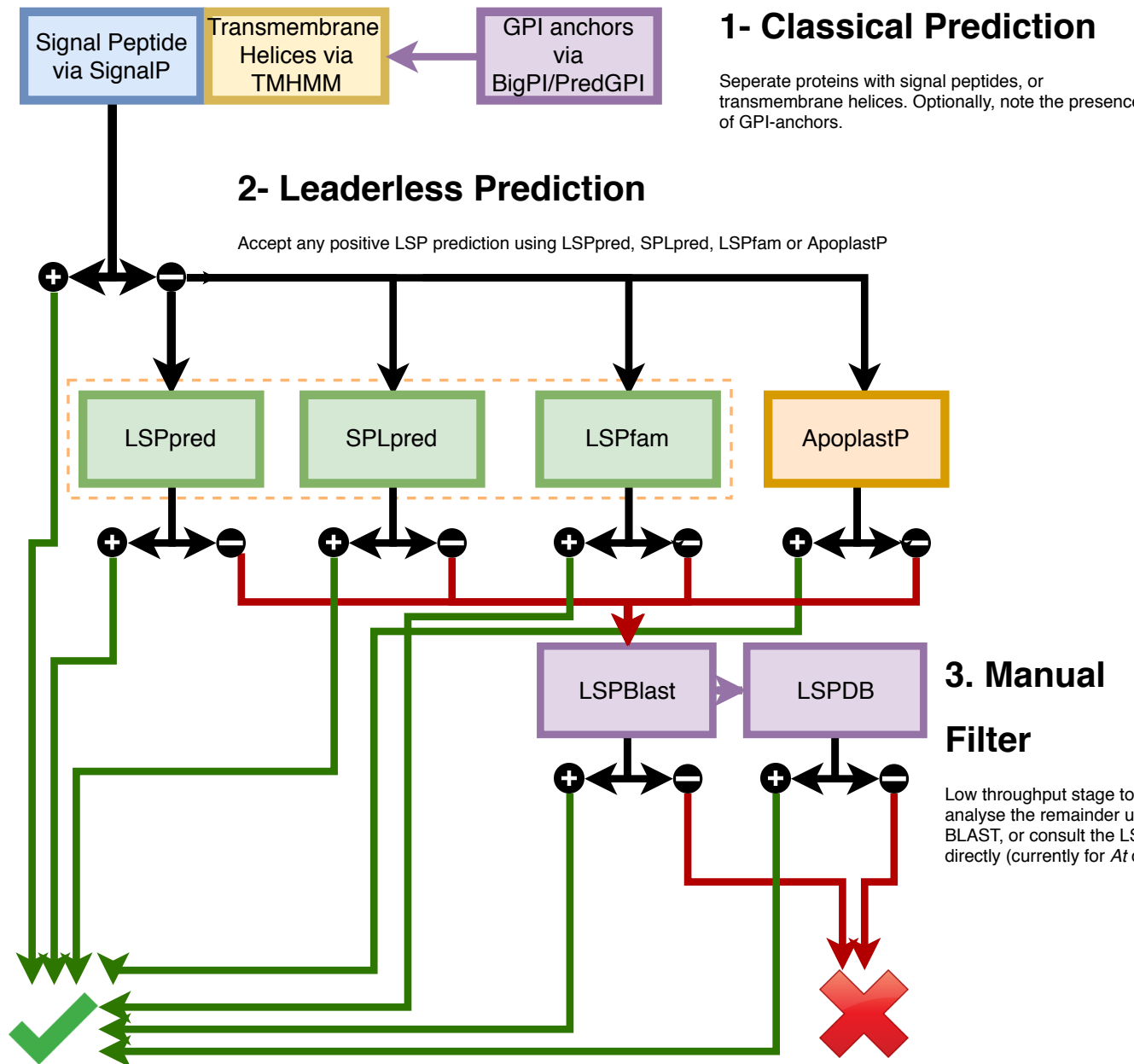


Figure 5.3: Updated workflow for identifying secretory proteins including LSPs. 1) Identify SP or TM containing proteins, 2) identify LSPs using the combined predictions of LSPpred, LSPfam, SPLpred and ApoplastP 3) Manually inspect the remaining using LSP-Blast, or the LSPDB if applicable to the organism of interest (currently only *Arabidopsis*)

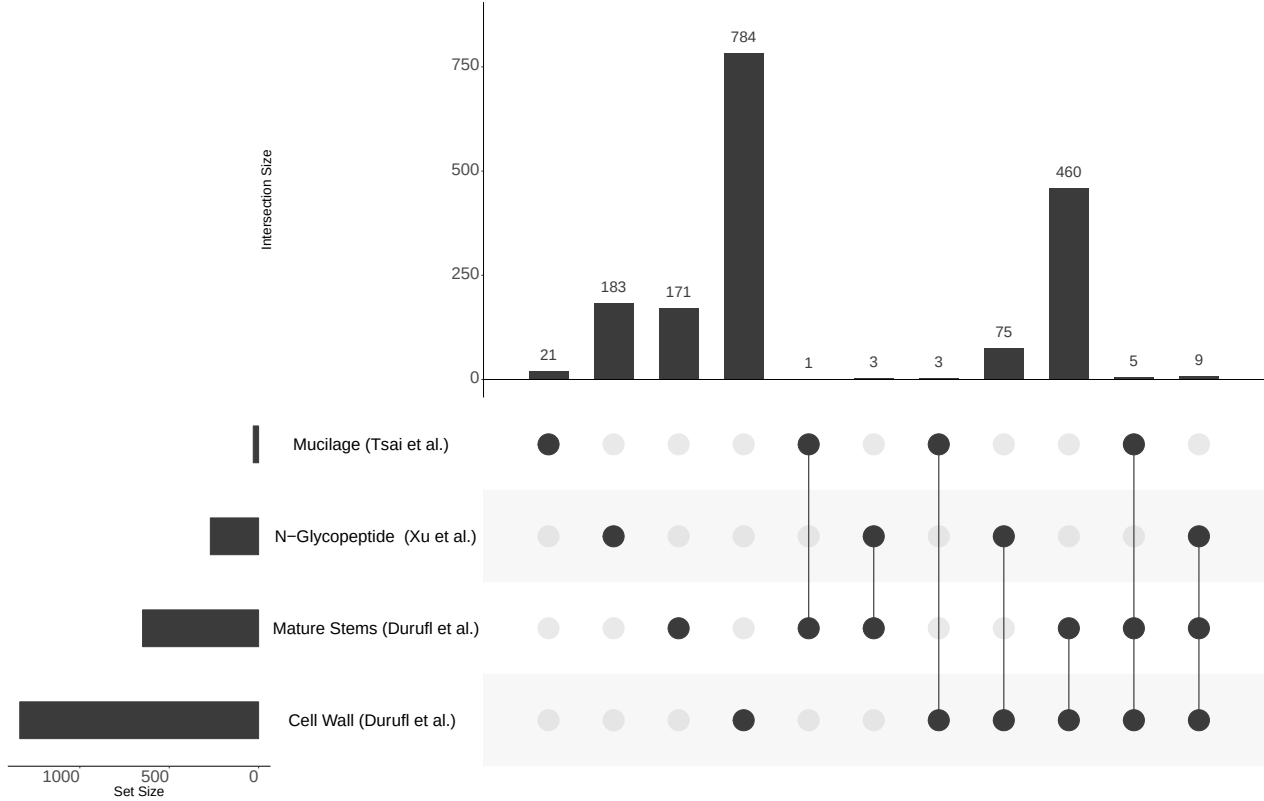


Figure 5.4: Upset plot of distributions of proteins from studies excluded from the LSPDB Tsai et al. (2017); Duruflé et al. (2017b,a); Xu et al. (2016).

interpretation would have been as follows. In the LSPDB this is a NONSEC protein, and so this observance would mark a transition to UNCLASSIFIED. Application of the black box prediction tools does not meet the FPR thresholds, yet both predictors have a probability > 0.5 for LSP status: LSPpred (0.56) and SPLpred (0.60). LSPfam does not match any domains. LSPblast returns hits to another ubiquitin-related protein that has previously been observed and labeled as yet unclassified (AT5G50870). These combined computational results provide secondary evidence to the observation and enhance analysis by allowing for quick reference to the similarity of potential LSPs to previous studies.

Both cell wall papers Duruflé et al. (2017b,a) excluded potential LSPs, and required either SP or predicted secreted status (using ProtAnnDB Clemente et al. (2009)), no retention signals and at most one TM domain. The mucilage study only found proteins with SP passed the proteomics criteria, though notably one SEC proteins is dismissed as lacking a signal peptide. The LSPDB approach of allowing SP and TM predictions in the N-terminal region to be grouped together overcame this, and multiple predictors concur that the protein is in the secretory pathway. Proteins without SP that are near to the criteria of multiple biological replicates in the study include AT5G17920, a methionine/cobalamin-independent synthase family protein. From **Table 3.3** earlier in the Thesis, we note this is the most frequently observed unclassified protein and so the status is unchanged by this additional paper. The LSPpred is the only tool to predict this protein as an LSP (0.76) and it also appears in Ding et al (2012) LSP list. Recent maize root analysis, comparing proteins in the apoplast between samples inoculated with *Trichoderma virens* spores and a control (Nogueira-Lopez et al., 2018) found the maize homolog of AT5G17920 secreted in response to the fungal infection. Application of the maize protein (UniProt: C0P5Y3.MAIZE) to LSPpred also yields a positive result (0.72). Therefore, by using these computational tools, studies from multiple species can be compared for LSP status, and cross-species analysis under certain stress/growth conditions, such as fungal infection, used to infer legitimacy. Of note is that methionine synthase (ATMS1) is not in the confidence lists for LSPs generated in **Chapter 3**, and only through the similarity to other putative LSPs built into the model has been selected.

In addition to specific additions either new or repeat putative LSPs as shown above,

the changes of the LSPDB over time can be collated, and a summary of how the LSPDB will change over time for each of the four papers is shown in **Table 5.2**.

Table 5.2: Counts of the repetitions and transitions of proteins classes in the LSPDB after the addition of 1715 proteins from excluded studies. Other includes secretory isoforms or multi-TM containing proteins.

Status	Update	Count
SEC	Repeat SEC observation	386
NONSEC	Transition to UNCLASSIFIED	205
SPT	Transition to SEC	147
UNCLASSIFIED	Repeat UNCLASSIFIED observation	773
Other	Repeat Other observation	204

The transition proteins represent 20.5% of the total from these studies, with the NON-SEC to UNCLASSIFIED transition of most interest for distinguishing between LSPs and contamination. The other transition (SPT to SEC) (**Table 5.6**) can be compared to the analysis of SPT proteins in **Chapter 3**. Repeat unclassified proteins are also worth investigating, particularly when comparing the experimental contexts for each repeated appearance. How these proteins are distributed across the studies, is shown below in **Table 5.3**, and an analysis of the most frequent shown.

Table 5.3: Distribution of UNCLASSIFIED (repeat and new observations) and SEC (new observations) proteins over the studies. No proteins in these categories appeared in all four excluded studies.

Observations	UNCLASSIFIED (repeat)	UNCLASSIFIED (new)	SEC (new)
1	353	191	142
2	413	14	5
3	7	0	0
4	0	0	0

For repeated UNCLASSIFIED proteins, the 7 that have occurred in 3 of the papers

are shown in **Table 5.4**. This table includes the high throughput results of the recommended LSP workflow, with LSPfam, LSPpred and SPLpred results. Of the 7 repeated proteins, 5 have a prediction from one of LSPfam, LSPpred and SPLpred, amongst them including methionine synthase (ATMS1) as noted in detail for mucilage study. Only one of these five proteins, bifunctional inhibitor (AT1G55260), appears in the medium confidence list however it is not predicted from either “black box” predictor, and has a positive LSPfam result due to the presence of the LTP2 probable lipid transfer domain (PF14368). One protein, reversibly glycosylated polypeptide 2 (AT5G15650) appeared on both the red low confidence LSPDB list and Ding et al list and has a positive SPLpred prediction. Fluorescence has shown this to be predominantly cytosolic, yet with an association with “Golgi-like” vesicles (Rautengarten et al., 2011). Lactate/malate dehydrogenase family protein (mMDH2, AT3G15020) has a positive LSPpred prediction, and via LSPblast can note sequence similarity to both lactate/malate dehydrogenase family protein (mMDH1, AT1G53240, low confidence LSP list), peroxisomal NAD-malate dehydrogenase 1 (PMDH1, AT2G22780, medium confidence list), and peroxisomal NAD-malate dehydrogenase 2 (PMDH2, AT5G09660, low confidence list). As noted in **Chapter 4**, PMDH2 notably achieves a positive prediction score from both SPLpred and LSPpred. In this list, the gene level identification of proteins is re-expanded to isoforms to demonstrate the differences in probability between isoforms. In this case, all isoforms share the same prediction results, though the possibility remains that isoforms of a protein can have different LSP predictions. A RuBisCO subunit, ribulose biphosphate carboxylase small chain 1A (AT1G67090) has SPLpred positive prediction for both isoforms, noting almost 0.09 difference between them. This unclassified protein is also frequently observed (**Table 3.3**), though since AT1G67090 and RuBisCO itself are highly abundant, the possibility of repeated appearances due to contamination is feasible. The relationship between protein abundance and observation is discussed further in Section 5.3. These five repeated proteins represent proteins with either similarities or features suggestive of secretion and should be priorities for experimental validation. The techniques used to identify them have not relied on this frequent observation. Negative validation results could be used to assign them as putative contamination, and prioritise a focus on other

targets for in depth sub-cellular localisation studies.

The four excluded studies had 205 proteins previously unobserved NONSEC proteins, including 14 that appear in two studies. These are shown in **Table 5.5**, with LSPpred and SPLpred predictions. No secretory protein domains were identified via LSPfam, and all have negative results from SPLpred. Four have a positive LSPpred prediction. Though GTP binding Elongation factor Tu family protein (AT1G07940) was previously unobserved, LSPblast identify AT4G02930 as identical from medium confidence list and so should be a repeated rather than a new appearance. Two other related proteins in the same family (AT1G06220, AT5G60390) are in the low confidence set. RmlC-like cupins superfamily protein (AT2G43120) is also predicted by LSPpred, and another similar RmlC-like cupins superfamily protein (AT1G07750) from the low confidence set has a positive LSPfam result, and consensus prediction from both predictors (**Table 4.4**). Similarly, inositol monophosphatase family protein (AT1G43670) has a related protein AT5G64380 in the low confidence set with predictions from both LSPpred and SPLpred. Hydroxymethylbilane synthase (HEMC, AT5G08280) has no results from LSPblast, or similar proteins in the candidate lists. The change in status for these proteins and related previously observed proteins demonstrates that as the database changes, proteins that individually had fewer features of interest can be revisited as related proteins are observed, and the families considered collectively.

Transitions of class from new observations can also cascade through the database to add more evidence. Five previously classified SPT proteins were observed in these studies. The senescence-associated gene 12 (SAG12, AT5G45890) was predicted explicitly from previous analysis (**Figure 3.15, Table 3.9**) as having a high number of SPT and SEC interactions and therefore likely to be eventually observed. This can be seen as justification of the lower confidence network criteria (Section 3.4.3) of SEC interacting SPT proteins as network evidence. Unclassified proteins that have PPI with these could now increase in score to medium network interactions, and the distribution of the categories in **Figure 3.23** could change. This in turn would change the composition of the training data, not only in the high/medium/low confidence sets used for LSPpred, but also in PFAM terms distributions, and the LSPfam list could change. This demonstrates the

Table 5.4: Prediction results for repeat unclassified proteins from excluded studies

Accession	Description	Symbol	LSPpred Score	SPLpred Score	LSPfam Count
AT1G55260	Bifunctional inhibitor/lipid-transfer protein	LTPG6	FALSE 0.472	FALSE 0.492	TRUE 1
AT1G55260	Bifunctional inhibitor/lipid-transfer protein	LTPG6	FALSE 0.480	FALSE 0.500	TRUE 1
AT5G17920	Cobalamin-independent synthase family protein	ATMS1	TRUE 0.763	FALSE 0.359	FALSE 0
AT1G07660	Histone superfamily protein		FALSE 0.504	FALSE 0.494	FALSE 0
AT1G07660	Histone superfamily protein		FALSE 0.564	FALSE 0.481	FALSE 0
AT3G15020	Lactate/malate dehydrogenase family protein	mMDH2	TRUE 0.734	FALSE 0.659	FALSE 0
AT3G15020	Lactate/malate dehydrogenase family protein	mMDH2	TRUE 0.808	FALSE 0.572	FALSE 0
AT4G11260	phosphatase-related	SGT1B	FALSE 0.444	FALSE 0.398	FALSE 0
AT5G15650	reversibly glycosylated polypeptide 2	RGP2	FALSE 0.509	TRUE 0.833	FALSE 0
AT1G67090	ribulose biphosphate carboxylase small chain 1A	RBCS1A	FALSE 0.464	TRUE 0.909	FALSE 0
AT1G67090	ribulose biphosphate carboxylase small chain 1A	RBCS1A	FALSE 0.536	TRUE 0.813	FALSE 0

Table 5.5: Prediction results for new unclassified proteins from excluded studies

Accession	Description	Symbol	LSPpred	Score	SPLpred	Score	LSPfam
AT4G36020	cold shock domain protein 1	CSDP1	FALSE	0.670	FALSE	0.664	FALSE
AT3G50970	dehydrin family protein	LTI30	FALSE	0.498	FALSE	0.573	FALSE
AT1G07940	GTP binding Elongation factor Tu family protein		TRUE	0.711	FALSE	0.597	FALSE
AT1G14900	high mobility group A	HMG A	FALSE	0.399	FALSE	0.333	FALSE
AT1G14900	high mobility group A	HMG A	FALSE	0.357	FALSE	0.326	FALSE
AT5G08280	hydroxymethylbilane synthase	HEMC	TRUE	0.698	FALSE	0.383	FALSE
AT1G43670	Inositol monophosphatase family protein		TRUE	0.705	FALSE	0.650	FALSE
AT2G43680	IQ-domain 14	IQD14	FALSE	0.292	FALSE	0.127	FALSE
AT2G43680	IQ-domain 14	IQD14	FALSE	0.290	FALSE	0.120	FALSE
AT4G23890	NAD(P)H-quinone oxidoreductase subunit S	ndHs	FALSE	0.477	FALSE	0.315	FALSE
AT5G45940	nudix hydrolase homolog 11	NUDT11	FALSE	0.357	FALSE	0.262	FALSE
AT5G45940	nudix hydrolase homolog 11	NUDT11	FALSE	0.285	FALSE	0.186	FALSE
AT1G03600	photosystem II family protein	PSB27	FALSE	0.485	FALSE	0.459	FALSE
AT2G46630	Putative extensin		FALSE	0.311	FALSE	0.240	FALSE
AT2G43120	RmlC-like cupins superfamily protein		TRUE	0.705	FALSE	0.654	FALSE
AT1G15270	Translation machinery associated TMA7		FALSE	0.406	FALSE	0.515	FALSE
AT5G08450	Zinc finger CCH domain protein		FALSE	0.286	FALSE	0.108	FALSE

tension between the LSPDB as a resource for placing new experimental data in context, and the LSPDB as input for researching prediction tools. The snapshot of the literature is necessary for stable results for analysis, and for the creation and accuracy estimates of the prediction tools but does preclude all available data being used. The size of this limitation depends on what extent the data used has captured properties of LSPs, and it predicts future ones. In later sections of this chapter, this restriction is used to advantage by treating subsequent data as case studies on which to judge the usefulness of the LSPDB and prediction tools, however the proportion of available data used is still a limitation to this work.

Table 5.6: New SEC proteins from held out studies

Accession	Description	Symbol
AT1G72970	Glucose-methanol-choline (GMC) oxidoreductase family protein	HTH
AT4G16660	Heat shock protein 70 family protein	HSP70
AT5G01930	Endo-beta-mannase 6	MAN6
AT5G09730	Beta-xylosidase 3	BXL3
AT5G45890	Senescence-associated gene 12	SAG12

5.4 Enhancing the LSPDB

Another clear limitation of the LSPDB is the choice of protein annotation included. As discussed in **Chapter 3**, GO terms, PFAM domains and PPI interaction data were added on top of the literature observations and secretory features. GO terms were shown to both have some redundancy and be influenced by manual curation (such as the cell wall terms being associated irrespective of whether the originating papers defined them as contamination). PFAM domains are probabilistic (present if part of a protein matches the model) and limited to only parts of proteins. Not all proteins have PFAM entries. Similarly, not all proteins have network data, and although interactions with other secretory proteins is indicative of some classically secreted proteins, this does not account for all such proteins, and so is of limited application when applied to UNCLASSIFIED proteins.

These present the pool of possible LSPs, and the presence of these protein features in these relative to the other LSPDB classes was used to identify similarities.

Due to each features individual strengths and weaknesses, a hybrid approach to scoring likely candidates was used in **Chapter 3**, however with the limitations described above, other features could enhance the LSPDB. Considering the structure of the LSPDB, the combinations of secretory features and observations into the base classifications of the LSPDB (SEC/SPT/UNCLASSIFIED/NONSEC) can be considered the starting point for additional data about individual proteins, which can then be compared between classes in a similar manner to the current LSPDB protein information.

In the next section three external sources of data are combined with the LSPDB classification: protein co-expression data from ATTED-II (<http://atted.jp/>, Obayashi et al. (2018)), post translational modifications (PTMs) as curated by UniProt, and protein abundance from PaxDB (<https://pax-db.org/>, Wang et al. (2012a)). Each of these sources could be included directly into the LSPDB at a future point, and in general describes an approach to extend the LSPDB with other annotations not included in the first version.

5.4.1 Co-expression

In addition to mapping the LSPDB classifications to an interaction network, the network can be reinforced and refined with additional information. PPI networks imply a physical interaction between the proteins, whereas other networks such as co-expression may be less informative in the context of evidence of secretion. The presence of an unclassified protein that has co-expressed transcripts with secretory proteins doesn't necessarily imply its sub-cellular co-location; they may be co-expressed and perhaps in the same pathway, but this need not imply co-localisation. Therefore, we apply co-expression only to enhance a PPI network, rather than as a substitute. Note that in the co-occurrence we essentially "reveal" parts of the PPI network.

Using co-expression data from ATTED-II, for both microarray and RNA-seq, we can extend the network algorithms from **Chapter 3** by requiring that co-occurring PPI also have supporting co-expression evidence. Since all genes in ATTED-II are ranked accord-

ing to their relative co-expression, we use a threshold of the top 100 (also the default ranked genes displayed by default on the website) to count as co-expressed.

We can require co-expression in either both (**Figure 5.5**) or either (**Appendix Figure D.1**) RNA-seq and microarrays data, respectively, and focus networks of 3 or more proteins. As expected, this reduces the size of the network from 1431 co-occurring PPI to 234 (both microarray or RNA-seq co-expression) and 353 (either) nodes. The criticism of PPI that may be inferred from large scale studies is that they do not necessarily represent *in-vivo* interactions; these networks add the results from other unrelated gene expression experiments to include only interactions with multiple lines of evidence, as well as including co-occurrence in a publication. For the PPI network of 234 of these that require both co-expression experiments, several clusters remain. All pairs of proteins that meet this condition are in **Appendix Table D.5**, and we considered in depth the 18 clusters that include three or more proteins in **Figure 5.5**, with each protein listed in **Appendix Table D.6**.

There are several contributions that these co-expression enhanced networks add to the LSPDB. The existing network scoring mechanism in the LSPDB ignores UNCLASSIFIED interactions. The composition of several clusters are solely UNCLASSIFIED proteins (**Figures 5.5B, 5.5C, 5.5D, 5.5E, 5.5G, 5.5H, 5.5I, 5.5J, 5.5K, 5.5L and 5.5O**). In combination with other tools such the predictors, putative LSPs found in these clusters could then be added to an amended network scoring mechanism to include these UNCLASSIFIED interactions. This approach could also be used to weight the existing scores. In **Figure 5.5N**, a cluster of three prohibitin-related proteins is shown, with one a SEC and the other UNCLASSIFIED. In the current scheme, these SEC interactions are scored identically all others. Enhancing the PPI with co-expression and co-occurrence could be used to rank these SEC (or any PPI that are also co-expressed) interactions higher than others.

5.4.2 Post translational modifications (PTMs)

PTMs have been suspected to play a role in LSPs function in the apoplast (Agrawal et al., 2010), particularly glycosylation. PTM features are part of the SecretomePs features, specifically the output of ProP (Duckert et al., 2004) that predict cleavage sites for

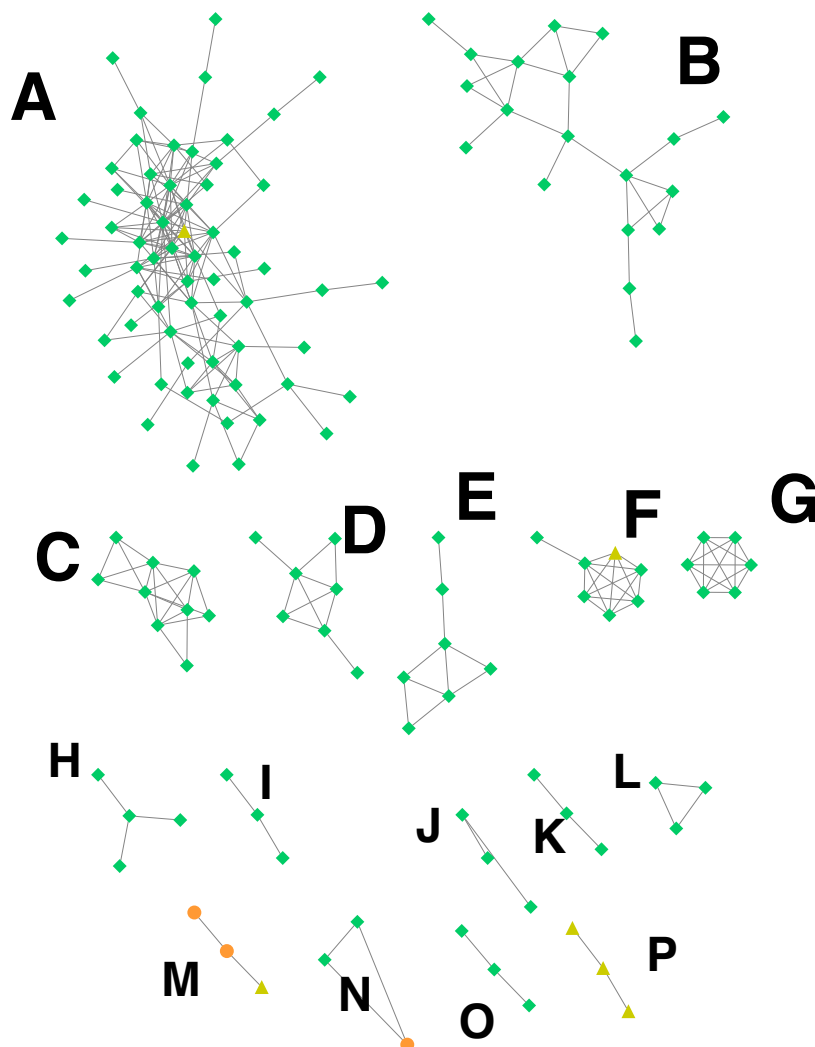


Figure 5.5: Clusters of 3 or more proteins identified from the LSPDB that co-occur in the same experiment and are co-expressed with both RNA-seq and microarray. Co-occurrence is required, and so only LSPDB observed proteins can be included from the UNCLASSIFIED (green diamond), SEC (orange circle) or Other (TM pr sectoroty isoform proteins, yellow triangle) classes. The clusters consist of some large groups discussed previously such as ribosomal proteins **A** and heat shock proteins **B**, as well as new smaller sets of proteins: ATPases and ubiquitin proteins **C**, histones **D**, tubulin proteins **E**, proteasome 20S or aminohydrolase **F**, chaperonin family proteins in two separate clusters **G** and **H**, citrate synthases **I**, dehydrogenase proteins **J**, nuclear antigen proteins **K**, beta-xylosidase **M**, prohibitin related proteins **N**, importin proteins **O**, and calnexin related proteins **P**.

proteolysis. To some extent we may also indirectly capture some PTMs in both GO term and PFAM domains, since these could relate to PTMs however this will not cover all PTM motifs or sequences. To complement the LSPDB with additional PTM data, we took a subset of LSPDB classes that have a verified entry in UniProt and applied PIEMAN (Nickchi et al., 2015) to calculate the over-representation of PTMs in each class. There are some caveats to this method. PIEMAN is limited to the UniProt version “October 2014”, and uses only the SwissProt subset of reviewed proteins. Therefore when applied to the LSPDB, the over-representation calculations are from the subsets rather than the entire LSPDB classes, and the subsets may not reflect the exact proportions of the original classes. The subset of proteins are SEC: 295, SP_THEORY: 813, NONSEC: 1883 and UNCLASSIFIED: 261.

Nonetheless, this allows a focused statistical measure of which PTMs are overrepresented in each base classification, and a means to prioritise unclassified proteins based on PTMs. Many PTMs occur either infrequently or are sub-terms of more general PTMs. To infer the most relevant, we set a threshold of 10 occurrences in Uniprot to filter out significant but low occurring PTMs from the curated results. For all that pass this threshold, after correction for multiple hypothesis testing (Bonferroni to 0.05) the remaining 60 PTM terms that are over-represented in one or more classes. The distribution amongst classes is shown in **Figure 5.6** and **Appendix Table D.3**.

Reassuringly, the SEC and SPT PTMs identified are consistent with our knowledge of the extracellular space; glycoprotein and proteoglycans, as well as GPI-anchored related modifications are known to be destined for secretion. For unclassified proteins we can then ask, which proteins have the SEC (or SP_THEORY) over represented PTMs? The overlaps between SEC and SP_THEORY of 9 PTMS indicate that many SP_THEORY may indeed be SP, consistent with the observations on network interactions of SP and SPT proteins with similar interaction networks.

As such we selected 15 PTMs and applied them to the UNCLASSIFIED proteins to identify 44 unclassified proteins (**Appendix Table D.4**). We also applied the other combinations of PTM profile to class, as shown in **Table 5.7**.

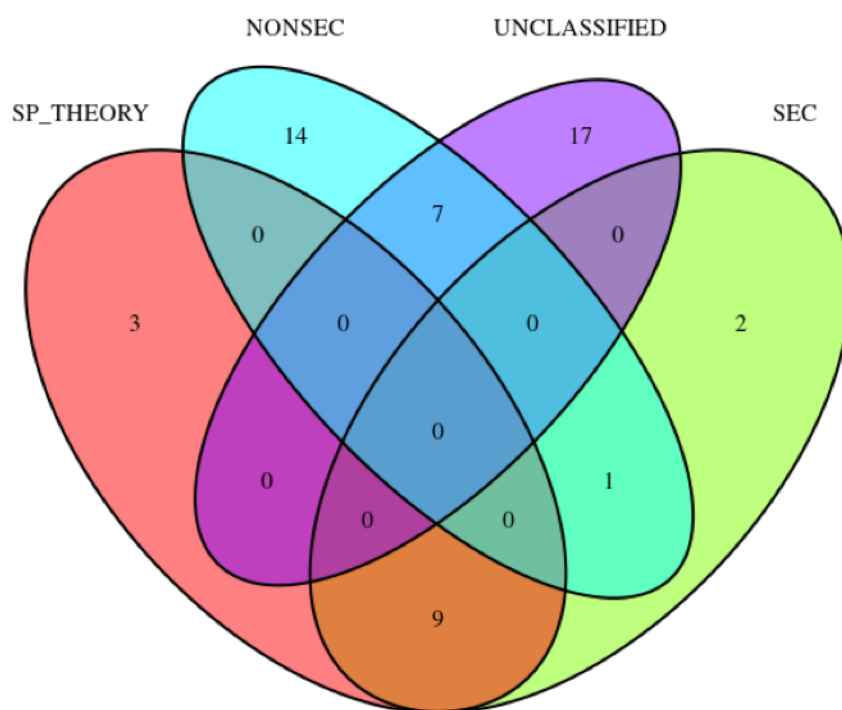


Figure 5.6: Venn diagram of PTMs in each class of the LSPDB

Table 5.7: Application of PTM profiles on LSPDB classes

PTM Profile	Class	Count
SEC_SPT	UNCLASS	44
SEC_SPT	NONSEC	259
NONSEC	UNCLASS	343
NONSEC	SEC	57
NONSEC	SPT	123
UNCLASS	NONSEC	470
UNCLASS	SEC	1
UNCLASS	SPT	8

Despite the caveats noted with the PIEMAN method, the results are still of interest. We have no evidence that any one of the four groups would be reduced more than the others, and so proportionally the overestimation is still of note. The 44 UNCLASSIFIED proteins that fit the SEC/SPT profile are listed in **Appendix D.4**, along with their LSPpred and SPLpred scores. A total of 19 of the 44 had a positive prediction, and a large number of the list are thioredoxin related proteins, indicating that PTMs are a driving feature behind their predictions, and as well as contrasting with the other methods used to identify them, such as PPI networks alone. This initial result with PTMs could be expanded in future work, particularly if the unequal PTM annotations between the protein classes could be addressed satisfactorily. The application of profiles is analogous to LSPfam, and so a potential LSPptm tool could be developed to identify putative LSPs.

5.4.3 Abundance

An aspect of the LSP versus contamination issue not yet considered is the abundance of proteins. Rather than focus on the properties of potential LSPs, we can consider what kinds of proteins might occur as cytosolic contamination. One hypothesis is that highly abundant proteins, already present in the cell, are going to be more likely to be in the contaminants when using proteomics compared to proteins of low abundance, particu-

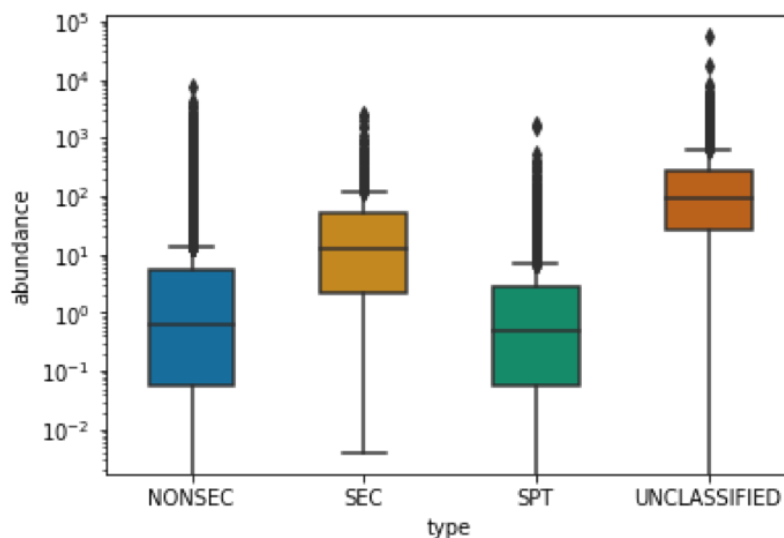


Figure 5.7: Boxplots of the abundance of LSPDB proteins using log scale of the protein abundances as calculated by PaxDb (in parts per million).

larly when sample preparation involves destructive methods. To consider this possibility, the LSPDB classifications were joined to *Arabidopsis* protein abundance data from PaxDB (Wang et al., 2012a).

The PaxDB abundance for each protein in each class is shown in a boxplot in **Figure 5.7**, and by kernel density estimate in **Figure 5.8**. Overall, the four classifications have different distributions though SPT and NONSEC are most similar. Then unobserved classes (NONSEC, SPT) have median abundance in the order of 1 part per million (ppm), compared to SEC (~ 10ppm) and UNCLASSIFIED (~ 100ppm). We can also compare mean abundance between SEC (68.24 ppm) and UNCLASSIFIED (320.56 ppm) via a t-test, suggesting a significant difference between them. All classes have outliers (diamonds shown in **Figure 5.7**), and either very high or very low abundant proteins occur in each class. The most abundant proteins are in the UNCLASSIFIED set, though there are notably highly abundant protein in both the SPT and NONSEC classes, which consist of proteins that have not been observed in the apoplast experimentally.

We can consider the overlap between the distributions SEC and UNCLASSIFIED abundance, and the placement of putative LSPs in these distributions for further analysis. In **Figure 5.8**, the properties of the SEC distribution is annotated, with 1 standard

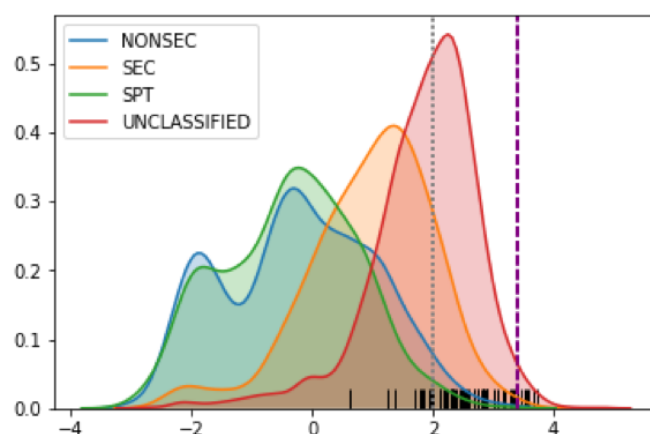


Figure 5.8: KDE distribution of LSPDB classes based on PaxDb protein abundances with markers for 1 standard deviation (grey) and maximum (purple) of the SEC class

deviation above mean expression of SEC (grey dotted line) and the maximum abundance of observed SEC (purple dashed line). This reinforces that UNCLASSIFIED proteins are more abundant than SEC. The putative LSPs from (Ding et al., 2012) are indicated in this figure by black lines; many are in this high abundance range between the grey and purple lines, and 7 from this list exceed the maximum SEC protein abundance. The complete list of UNCLASSIFIED proteins with abundance greater than maximum SEC abundance (2500ppm) are listed in **Table 5.8**. The overlap between these abundant proteins and the results drawn from LSPDB, and the LSPpred/SPLpred prediction tools is discussed further in section 5.5.

Table 5.8: UNCLASSIFIED proteins with a PaxDB abundance above 2500ppm

Accession	Description	Abundance (ppm)
ATCG00490	ribulose-bisphosphate carboxylases	53821
AT2G39730	rubisco activase	17456
AT4G38970	fructose-bisphosphate aldolase 2	8176
ATCG00480	ATP synthase subunit beta	7678
AT1G06680	photosystem II subunit P-1	5944
AT3G12780	phosphoglycerate kinase 1	5696
AT3G60750	Transketolase	5648

Accession	Description	Abundance (ppm)
AT1G07890	ascorbate peroxidase 1	5412
AT1G67090	ribulose biphosphate carboxylase small chain 1A	5015
AT4G20260	plasma-membrane associated cation-binding protein 1	4744
AT3G52590	ubiquitin extension protein 1	4302
AT5G35630	glutamine synthetase 2	4045
AT3G49010	breast basic conserved 1	4006
AT1G23410	Ribosomal protein S27a / Ubiquitin family protein	3675
AT2G47110	ubiquitin 6	3592
AT5G17920	Cobalamin-independent synthase family protein	3455
AT5G66190	ferredoxin-NADP(+)-oxidoreductase 1	3410
AT1G13440	glyceraldehyde-3-phosphate dehydrogenase C2	3353
AT1G02780	Ribosomal protein L19e family protein	3038
AT5G14740	carbonic anhydrase 2	3017
AT1G31340	related to ubiquitin 1	2884
AT2G35635	ubiquitin 7	2843
AT1G07920	GTP binding Elongation factor Tu family protein	2829
AT3G01500	carbonic anhydrase 1	2805
AT5G09660	peroxisomal NAD-malate dehydrogenase 2	2771
AT4G20360	RAB GTPase homolog E1B	2734
AT3G55440	triosephosphate isomerase	2541

There are several limitations to this abundance analysis. Since classifications are based on observations there is some uncertainty in the distribution and protein complement of each class. A mix of possible contaminants and LSPs are in UNCLASSIFIED, and NONSEC may also contain as yet unobserved LSPs. The use of *Arabidopsis* abundance markers is also of limited use to other plant species, as the abundance distribution for different species differs (**Figure 5.9**) and would require LSPDB classifications to determine the equivalent threshold for each species. There is also the issue that the abundance data in PaxDB is taken as the average from multiple experiments, and differences in protein

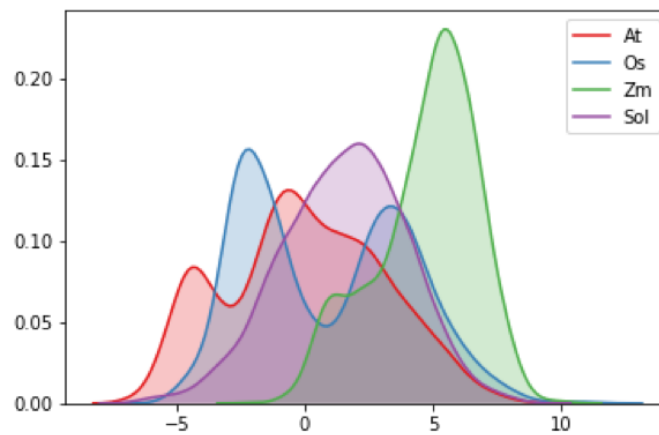


Figure 5.9: KDE distribution of the log10 PaxDb protein abundances for *Arabidopsis*, *Oryza sativa*, *Zea mays* and *Brachypodium distachyon*

abundance under experimental conditions is not taken into account.

It is difficult to draw conclusions from this data. That SEC and UNCLASSIFIED are typically more abundant than NONSEC and SPT suggests that the observed proteins in the LSPDB are more abundant than unobserved classes however the increased abundance in UNCLASSIFIED relative to SEC could be explained by an increased likelihood for contamination from highly abundant proteins. There is also the possibility that such a difference can be explained by stress-related UPS pathways, and that highly abundant proteins are the most suitable for the “fast track” (van de Meene et al., 2016) pathway of unconventional secretion. Therefore, in the absence of any other evidence (i.e. from tools developed in **Chapter 4**), repeated observed highly abundant LSP candidates could be considered a lower priority for validation, whereas lowly abundant LSP candidates that are repeatedly observed might be prioritised for experimental follow up if paired with other evidence.

Several of the high confidence LSP candidates such as ascorbate peroxidase are also highly abundant, however protein abundance or frequency of occurrence has not been directly used to train computational tools. Proteins found independent of their abundance, instead using only their properties or interactions with other proteins, can then be a useful filter when applied to the highly abundant candidates.

5.4.4 Effect of new protein data on the LSPDB

Combining evidence from each of the sections, we can look at only UNCLASSIFIED proteins and take results from each of the preceding chapters, as well as the current, and compare the different approaches to identify common LSP candidates. For each set of results, we define a single criteria similar to the individual high, medium and low confidence criteria in sections 3.4.2 and 3.4.3. We can also compare these to neutral categories related to protein expression or frequency of appearance in the database. The selected results and associated criteria are as follows:

- Frequency: UNCLASSIFIED proteins with more than 5 observations (**Chapter 3, Table 3.3**)
- GO: proteins with SEC + UNCLASS + SPT GO terms (**Chapter 3, high confidence set**)
- PFAM: proteins with SEC + UNCLASS + SPT PFAM terms (**Chapter 3, high confidence set**)
- Network: PPI interaction proteins with > 1 SEC interactions, ≤ 33 NONSEC (**Chapter 3, high confidence set**)
- Abundance: Protein higher than the most abundant SEC proteins (**Chapter 5, Table 5.8**)
- PTM : proteins with SEC PTM profile (**Chapter 5, Appendix Table D.4**)

The relationship between each set of results is shown in **Figure 5.10** using an UpSet plot to indicate the magnitude of overlaps between the sets. Looking at the results in this way reveals several combinations in distinct 15 distinct categories. Of 216 candidate proteins, only 19 proteins appear in more than one of the evidence categories indicating that each is relatively specific. Those that do occur in multiple lines are discussed below, noting the prediction tool results as well as their presence in the (Ding et al., 2012) review. Note that both frequently observed, and highly abundant proteins are included in this comparison.

For the frequent candidates, prediction tools have not used this information to build the model, so highly frequently observed with other features are good candidates for

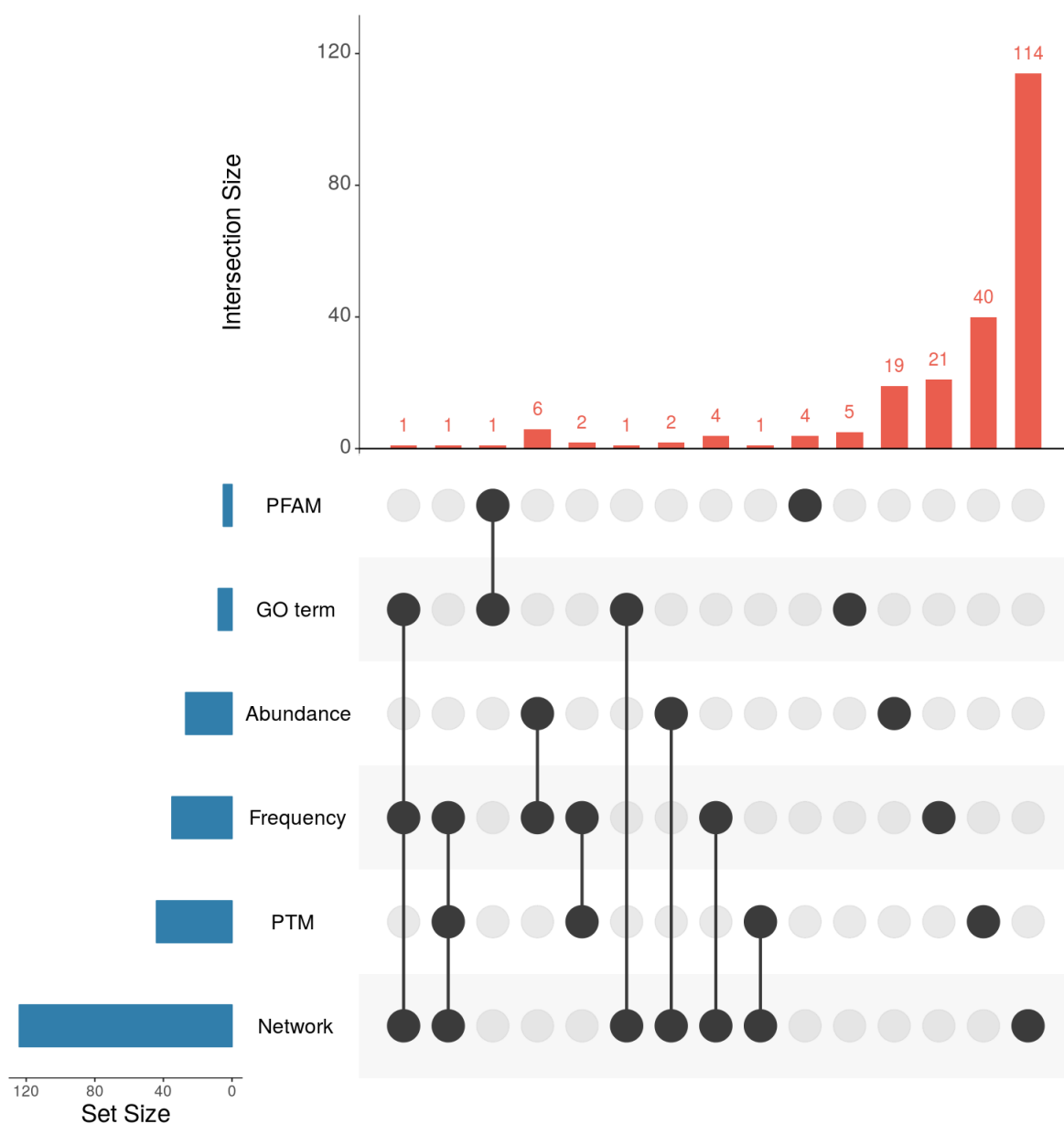


Figure 5.10: Upset plot of the overlaps of LSP candidates between evidence approaches of frequency of observation, GO terms, PFAM domains, PPI network, PTMs and abundance.

follow up. Abundance has also not been used to find candidates, however an overlap between abundant proteins and other results are less informative as an explanation that high abundance could contribute to a higher likelihood of contamination, since any experimental technique that captures cytosolic proteins may be more likely include these highly expressed proteins. Given these two categories that are ambiguous, instead of using an additive scoring mechanisms such as in **Figure 3.23**, we consider each of the combinations separately. Excluding those that have already been discussed in the LSPDB, i.e proteins occurring only between PFAM, GO term and Network approaches, 17 proteins across 7 category combinations remain.

Initially, we consider that the new categories contain the most frequent proteins. Combining GO term, frequency and network identifies NADH-cytochrome b5 reductase-like protein (AT5G20080), which was discussed in **Chapter 3** as a high confidence candidate. Four proteins are highlighted when combining only high confidence network with frequency: Superoxide dismutase (AT1G08830), two rotamase proteins (AT4G38740 CYP 1 and AT2G16600 CYP 3), and PYK10-binding protein 1 (AT3G16420, PBP1). All four appear in the LSPDB medium confidence list, indicating they also had either minor GO or PFAM evidence in addition to their network connections. This also means that they were all used as a positive training data for LSPpred and though they all have a true prediction as expected, they all also have positive predictions for SPLpred offering an independent positive prediction to support them as LSPs.

The rotamase pair, also known as peptidyl-prolyl *cis-trans* isomerases, are proteins that assist with protein folding and were identified along with heatshock-related proteins in the LSPDB medium confidence list (Appendix A.3). These two interact with each other and are amongst the proteins identified from **Figure 3.22** as PPI that frequently co-occur. Each also occurs in either the EV proteome (AT2G16600) or the secretory vesicle studies (AT4G38740 in wild-type). Two others have similarity to experimentally verified LSPs from other organisms: AT1G08830 is similar to the verified LSP in yeast, superoxide dismutase. The PYK10-binding protein 1(AT3G16420,PBP1), which as noted previously is in the (Ding et al., 2012) review, the LSPDB medium confidence list, nd the EV proteome, has sequence similarity to the sunflower LSP Helja.

The additional PTM approach, using only over-represented secretory PTMs on UNCLASSIFIED protein, identifies one additional protein to the network class, phototropin 1 (AT3G45780, PHOT1) which is also in the medium confidence list and has an experimental verified plasma membrane association (Sakamoto and Briggs, 2002). PTM and frequency yield three more proteins of interest, one of which also includes a network result. A pair of thioredoxin proteins, thioredoxin 3 with additional network result (AT5G42980, TRX3) and 2-Cys peroxiredoxin (AT3G11630, 2-Cys PrxA) join the growing list of thioredoxin proteins identified.

Adding abundance information contrasts the network classification of two proteins: the high LSPDB high confidence ascorbate peroxidase 1 (AT1G07890, APX1), which was also present in the EV proteome study, and the medium list ferredoxin-NADP(H)-oxidoreductase 1 (AT5G66190). In these two instances, the high abundance occurs in proteins with a least one other line of lower confidence evidence from PFAM or GO terms. Contrast this with the 6 proteins that have were frequently observed that were also highly abundant. None had a high confidence individual score from network, GO or PFAM, and only some appeared in low confidence results. Two are present in the low LSPDB list, glyceraldehyde-3-phosphate dehydrogenase (AT1G13440, GAPC2) and triose phosphate isomerase (AT3G55440, TPI) and also have positive LSPpred predictions (0.72 and 0.73, respectively). These two proteins, plus plasma-membrane associated cation-binding protein 1 (AT4G20260, PCAP1) and cobalamin-independent synthase family protein (AT5G17920, ATMS1) are all referenced as potential LSP in the original (Ding et al., 2012) list. ATMS1 is the most frequently occurring protein in the LSPDB, and was observed in the EV and secretory vesicle studies. Neither of the RuBisCO-related proteins (AT1G67090, ATCG00490) are in previous LSP lists or the LSPDB confidence lists, though interestingly each has prediction for their SPLpred (AT1G67090: 0.91) or LSPpred (ATCG00490: 0.86). Both were observed in the EV proteome, and AT1G67090 in subsequent studies excluded from the LSPDB (see Section 5.2).

In summary, the additional protein annotation expands and complements the LSPDB. Much of the data in the LSPDB has been focused on associating the UNCLASSIFIED with SEC, in order to identify LSPs. Abundance is of note as it can be used to associate UN-

CLASSIFIED with contamination, especially in the absence of an LSP status from the other sources of data. Initial results with PTMs indicate that may be suitable for a more rigorous statistical framework of over-representation, and gave some overlapping results as the LSPDB method with less curation than developing the scoring system. The reinforced network approaches add more nuance to the flat network structure used in the LSPDB. Combining candidates from additional sources, as with the LSPDB confidence lists, seeks to compensate for individual biases in the data. Overall, these additional protein data sources show that the core of the LSPDB, the combination of secretory features and observations, is extendable to different types of data.

5.5 LSP case studies

Despite these limitations, we can explore the utility of the tools developed in this Thesis by using new data including other species, to explore how the tools. In this section, the output of this Thesis is applied internally to data in the LSPDB not yet used in constructing tools and external data including known LSPs and experiments perturbing the secretory pathway.

5.5.1 Prediction tools on the remaining UNCLASSIFIED proteins

After the high, medium and low confidence proteins (384 total) were removed from the UNCLASSIFIED set, 1312 remain. Applying SPLpred and LSPpred were applied to the remaining proteins resulted in 232 that were predicted by either tools and 33 had a consensus from both SPLpred and LSPpred. (**Table 5.9**). The italicised proteins are those which are only positive in one of the predictors when the lower 0.5 threshold is used.

Since these UNCLASSIFIED proteins were not in the **Chapter 3** high, medium and low confidence lists, the majority of these have not previously been identified as candidate LSPs. Two proteins were however identified in the co-expression cluster approach in section 5.3, a tubulin alpha-4 chain protein (AT1G04820, TUA4, **Figure 5.5** cluster E), and a N-terminal nucleophile aminohydrolase (AT4G31300, PBA1, **Figure 5.5** cluster F). Identification of these proteins as positive could then be cascaded to con-

sider the whole cluster. Similarly to the comparisons in **Figure 4.13**, we also applied the plant apoplastic predictor ApoplastP and the general sub-cellular predictor DeepLoc (<http://www.cbs.dtu.dk/services/DeepLoc-1.0/>, Almagro Armenteros et al. (2017)). DeepLoc has an individual score for 10 localizations, and the rank and probability of the extracellular prediction score is shown in parentheses.

No proteins were predicted secreted by all four programs. However 9/33 proteins that are predicted by SPLpred, LSPpred and ApoplastP, similar in proportion to the rice results, and only one (glyoxylate reductase 2, AT1G17650) was in the lower 0.5 threshold category. PBA1 is also predicted by ApoplastP. DeepLoc agrees with two proteins from the LSPpred and SPLpred consensus set, the selenium-binding protein 1 (AT4G14030, SBP1) and a domain of unknown function (DUF946) protein (AT3G04350). Thus, multiple predictions tools, especially those that use different training data and hypothesis, can be combined to identify secretory targets for validation when there is no other experimental evidence available.

5.5.2 Secretory isoforms

For each gene in the LSPDB, every protein sequence was analysed and the number of SP, TM or GPI recorded. Any gene that did not maintain the same number of SP, TM or GPI in every one of its isoform was classified as secretory isoforms and handled separately in the LSPDB. As such they are another source of data that has not been incorporated into the input for the methods developed. Those secretory isoforms with and without SP are especially useful as they provide a natural source from alternative splicing to test the common feature hypothesis, and to test for bias in the presence or absence of a signal peptide. To test this, all genes from LSPDB with a SP_ISO status (767) were extracted. As a high confidence estimate of the SP impact, we pass these genes to Uniprot and compare LSPpred and SPLpred results only for the reviewed subset of UniProt (506) which have isoforms that are also reviewed. Of the remaining 129 genes that pass this filter (totalling 285 verified proteins isoforms), only 11 and 6 differ in positive predictions between the isoforms for SPLpred and LSPpred, respectively. In these cases, the change from positive to negative prediction is a difference in the prediction score crossing the threshold estab-

Accession	Description	AP	DL
AT1G04820	tubulin alpha-4 chain	No	No (7, 0.003)
AT1G17650	<i>glyoxylate reductase 2</i>	Yes	No (5, 0)
AT1G23740	Oxidoreductase, zinc-binding dehydrogenase family	No	No (2, 0)
AT1G52760	lysophospholipase 2	No	No (9, 0.003)
AT1G76680	12-oxophytodienoate reductase 1	No	No (5, 0.001)
AT2G13360	<i>alanine:glyoxylate aminotransferase</i>	No	No (6, 0.001)
AT2G33410	RNA-binding (RRM/RBD/RNP motifs) family protein	Yes	No (6, 0.012)
AT3G04350	<i>Plant protein of unknown function (DUF946)</i>	No	Yes (1, 0.457)
AT3G09810	<i>isocitrate dehydrogenase VI</i>	No	No (3, 0)
AT3G15010	RNA-binding (RRM/RBD/RNP motifs) family protein	Yes	No (6, 0)
AT3G16460	Mannose-binding lectin superfamily protein	Yes	No (3, 0.101)
AT3G16760	Tetratricopeptide repeat (TPR)-like protein	Yes	No (10, 0.003)
AT3G27430	N-terminal nucleophile aminohydrolases	No	No (6, 0.003)
AT3G44310	<i>nitrilase 1</i>	No	No (7, 0.033)
AT3G47040	<i>Glycosyl hydrolase family protein</i>	No	No (7, 0.003)
AT3G54640	<i>tryptophan synthase alpha chain</i>	No	No (2, 0)
AT3G55010	<i>phosphoribosylformylglycinamidine cyclo-ligase</i>	No	No (5, 0)
AT3G56460	GroES-like zinc-binding alcohol dehydrogenase	No	No (5, 0.02)
AT3G59970	<i>methylenetetrahydrofolate reductase 1</i>	No	No (6, 0.009)
AT4G13430	isopropyl malate isomerase large subunit 1	No	No (5, 0)
AT4G14030	selenium-binding protein 1	No	Yes (1, 0.317)
AT4G15040	Subtilisin-like serine endopeptidase family protein	Yes	No (2, 0.223)
AT4G31300	N-terminal nucleophile aminohydrolases	Yes	No (5, 0.015)
AT4G33070	Thiamine pyrophosphate dependent pyruvate decarboxylase	No	No (6, 0.011)
AT4G33680	LL-DAP-aminotransferase	No	No (2, 0)
AT5G01300	<i>phosphatidylethanolamine-binding protein</i>	No	No (2, 0.277)
AT5G02240	NAD(P)-binding Rossmann-fold superfamily protein	No	No (6, 0)
AT5G16970	alkenal reductase	No	No (4, 0)
AT5G39830	Trypsin family protein with PDZ domain	Yes	No (4, 0)
AT5G47870	Cobalt ion-binding protein	No	No (5, 0.009)
AT5G63570	glutamate-1-semialdehyde-2,1-aminomutase	No	No (5, 0.005)
AT5G65690	<i>phosphoenolpyruvate carboxykinase 2</i>	No	No (6, 0.007)
AT5G66570	PS II oxygen-evolving complex 1	Yes	No (3, 0.001)

Table 5.9: Consensus LSPpred2 and SPLpred3 predictions from the UNCLASSIFIED set. Proteins in italics in consensus only when predictions between the default and FPR controlled thresholds (0.5 - 0.68 for LSPpred, 0.5 to 0.70 for SPLpred) are taken as positive. ApoplastP (AP) and DeepLoc (DL) predictions are also noted. DeepLoc includes whether the protein is predicted as extracellular, as well as the rank and score of the extracellular prediction amongst the 10 sub-predictors.

lished to control for FPR, and the prediction tool has a majority (>0.5) probability. For example, endoglucanase 3 (AT1G22880) isoforms return prediction scores of 0.69 and 0.73, straddling the 0.70 cutoff. Each protein with results in this range are therefore differing in prediction outcomes only due to a need to control for false-positives, and considering all the isoform results collectively could address this. One strategy for such isoforms could be to take the maximum isoform prediction and use that against the threshold, as long as the minimum prediction score exceeds 0.5. This approach also serves as another measure of the bias in SPLpred, demonstrating again that most SPLpred results for known secretory proteins are unaffected by the presence of SP.

Secretory isoforms were also present in the original (Ding et al., 2012) LSP list; peptidase M1 family protein (AT1G63770), plant basic secretory protein family protein (AT2G15130) and beta-glucosidase 46 (BGLU46, AT1G61820) each appear, with versions of each without an SP. BGLU46 varies for SPLpred predictions, a positive prediction (0.83) for full length transcript with 12 exons, and negative prediction (0.67) for the shorter 7 exon variant, with an alternative start site that omits the SP. Even though this lowers the threshold to just below the cut-off, the shift is noticeable, and reflects the actual transcripts. Given a short protein (377 to 516 AA) the relative robustness of the score demonstrates that the loss of the 22 amino acid signal peptide in this sequence has made only a small contribution to the difference in prediction score. The secretory isoform labelling has been useful in this work to remove ambiguity from the database and highlight the need to consider LSP at gene, transcript and protein level. Further investigations and work can focus on determining which of the isoforms is being expressed for such proteins.

5.5.3 Secretory vesicles

A study of the secretory SYP121 and SYP122 vesicle contents Waghmare et al. (2018) can also be considered. This mass spectrometry-based experiment inferred the contents of SYP121 and SYP122 secretory vesicle, taken from fluids of *Arabidopsis* seed culture. It compared the secretomes of wild-type cultures to mutants affecting these vesicles to find unique and common protein types to the vesicles. The absence of proteins in specific vesicle related mutants was used to infer vesicle cargo specificity. The work cites the lack

of appropriate tools presented in **Chapter 2**, and avoids a prediction step for LSPs when determining the extracellular contents of these mutants. With the benefit of the tools presented in this Thesis we can perform this missing step. Notably numbers of both SP and non-SP containing proteins are reduced in the mutants (**Figure 5.11**). If a secretory pathway component is mutated that involves these vesicles, then if proteins without SP are contaminants or from dead cells, then the number observed between the wild-type and mutants should be conserved. That they follow the same reduced pattern indicates that these non-SP proteins may indeed be genuinely secreted via these vesicles. If so, it is an experiment that also gives insight into how these are secreted. The authors compare these results to other secretory studies, a cell wall proteome (Hervé et al., 2016) and an extracellular vesicle proteome (Rutter and Innes (2017), discussed in 5.5.4), and note where there is concordance of proteins. The advantage of the LSPDB is that we can compare to many more relevant studies to find which proteins have been experimentally observed previously. Taking the proteins without SP, we can compare their appearance in mutant or wild-type vesicles. 209 unique proteins are identified as potential LSPs after filtering through the LSPDB, 54 NONSEC and 155 UNCLASSIFIED. Applying our prediction workflow, 126 were predicted to be LSPs. Of these positive results, 95 were previously observed and 31 are new (current NONSEC in the LSPDB). Notable repetitions of UNCLASSIFIED proteins discussed previously include malate dehydrogenase (AT1G04410), glutamate dehydrogenase (AT5G07440), two jacalin-related lectins (AT2G25980, AT3G16460) and glutathione S-transferase PHI 2 (AT4G02520). Inositol monophosphatase family protein (AT1G43670) appears again, adding to the observations from the excluded studies in Section 5.1.

The criteria established for secretory proteins in Waghmare et al. (2018) was the identification of 2 unique peptides appearing in at least one of three replicates studies of WT, *syp122* and *syp121* samples respectively. Given the underlying issue of contamination, we can narrow these results by increasing the requirement to two replicates. In all three samples (WT, *syp122* and *syp121*) enolase (LOS2, AT2G36530) appears. As observed in **Tables 4.6, 3.12**, LOS2 is a low confidence, frequently observed LSP with consensus predictions from both prediction tools. It appears (three times!) in the (Ding et al., 2012)

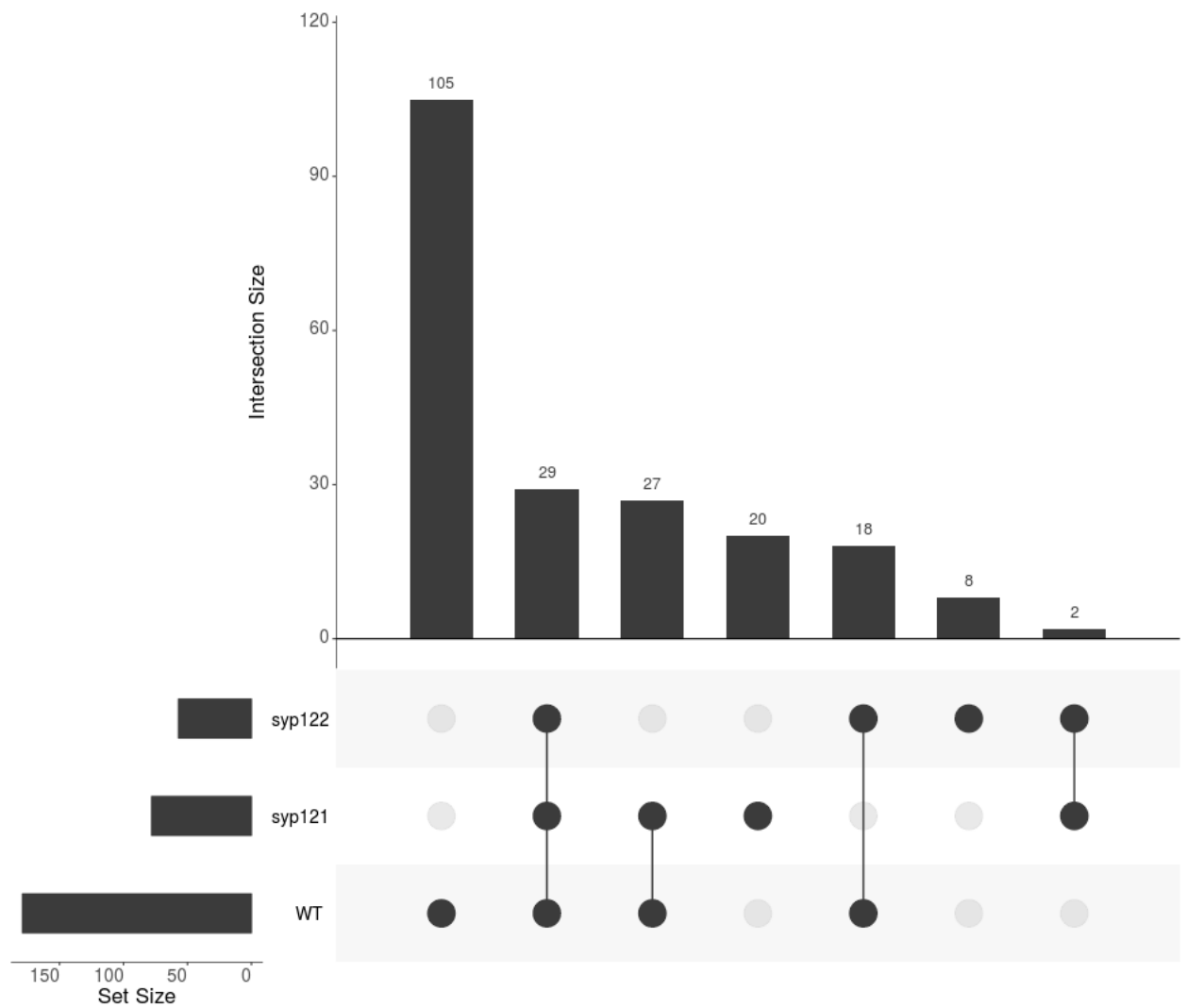


Figure 5.11: Upset plot of proteins without SP found in the WT, *syp122* and *syp121* samples from Waghmare et al. (2018).

list, as well as in the extracellular vesicle proteome (see 5.5.4). Similarly, the frequently observed glyceraldehyde-3-phosphate dehydrogenase C subunit 1 (AT3G04120, GAPC) is in two replicates each of WT and *syp122*, indicating a possible SYP121 vesicle association. Not appearing in the *syp122* mutant is chaperonin-60alpha (AT2G28000, CPN60A), which is an UNCLASSIFIED protein only noted in the co-expression cluster H (**Figure 5.5H**). The appearance of these proteins in this particular experiment demonstrates that well-constructed and executed experimental design to investigate the secretory pathway can be an interesting method to discover LSPs.

5.5.4 Extracellular vesicles (EVs)

The contents of purified plant EVs from apoplastic fluid have also recently been characterised (Rutter and Innes, 2017). After isolating EVs using a density gradient, mass spectrometry was performed and only 16% of the proteins found in their study had conventional SP. The authors note the hypothesis of EVs as an unconventional protein secretion pathway, especially in the presence of biotic stress. Applying 203 candidates from this study to the LSP workflow identified 50 that are predicted by one or more of SPLpred, LSPpred and LSPfam and lack classical secretory features. Using the LSPDB, 12 of these are NONSEC and 38 UNCLASSIFIED. Several previously unobserved NONSEC were found to be related to other UNCLASSIFIED proteins in the LSPDB, either by using gene annotation, or by inspecting sequence similarity with LSPBlast. This includes two jacalin-related lectin proteins (AT2G39330 in NONSEC, AT3G16470 in UNCLASSIFIED) as well proteins related to the nitrile specifier proteins (AT3G16400, AT3G16390, AT3G16410) and PYK10-binding protein 1 (AT3G16420). Glyceraldehyde-3-phosphate dehydrogenase GAPA1 (AT3G26650) is related to frequently observed GAPCP-1, which has a positive prediction in both SPLpred and LSPpred, and is in low confidence LSP set from the LSPDB. GAPCP-1 is also observed in this EV proteome, as well as the frequent GAPC2. Newly observed proteins include two protein kinase superfamily proteins (AT3G17410, AT1G48210). Additionally, the LSPDB identifies 8 secretory isoforms in the EV proteome with positive predictions for at least one isoform.

For the majority of proteins that have been observed previously, there are notable

repetitions. Bifunctional inhibitor/lipid-transfer protein (AT1G55260) is again found, as noted in the excluded studies. The frequently occurring malate dehydrogenase (AT1G04410) is found here as it was in the secretory vesicle study Waghmare et al. (2018). Interestingly, those secretory Qa-SNAREs including SYP121 and SYP122 were also found in the EV amongst other trafficking-related proteins, suggesting that the occurrence of LSP candidates in that study and in the EV are related.

The high confidence ascorbate peroxidase (AT1G07890, MEE6) is present, as noted in Section 3.4.5, two other low confidence proteins annexin1 (AT1G35720) and glutathione S-transferase PHI 2 (AT4G02520) are highlighted as oxidative stress response proteins, again linking EV and LSPs to biotic stress. The V-type proton ATPase catalytic subunit A protein (AT1G78900, VHA-A) which is in the medium confidence list appears. This was also amongst the top co-occurring co-expressing PPI interactions (appendix D.X), with vacuolar ATP synthase subunit VATC. Additional medium confidence list LSP proteins include phototropin 1, which has been shown to be PM-associated (Sakamoto and Briggs, 2002), with some present in the cytoplasm. This was also identified based on PTMs in **Appendix Table D.4**. The frequently observed PYK10-binding protein 1 (AT3G16420) and heat shock protein 70 (AT3G09440) also occurred. The singular ribosomal protein in the training data with a positive prediction (AT2G18020) also appears. As a high occurring PPI with co-occurrence (**Table 3.11**), it's partner is also ribosomal related (AT3G18740). This pair were also identified by similarity to ApoplastP rice results, suggesting that these two are the best candidates for resolving the large number of ribosomal results in the initial medium confidence LSP list.

The three RuBisCO related proteins (ATCG00490, AT2G39730, AT5G38410) are again present and the high abundance of all three must be noted in the context of contamination. The most frequently observed protein methionine/cobalamin-independent synthase family protein (AT5G17920, ATMS1) is again observed, as it was in the excluded studies from the LSPDB. As tools were not built using the frequency of occurrences, these repeated occurrences in different experimental designs combined with computation predictions independent of frequency suggest that this protein be targeted for experimental followup using antibody and/or fluorescent tag-based approaches.

5.6 LSP candidates

With the tools developed in this Thesis, and results from applying them to case studies, we can make several conclusions regarding potential LSP protein candidates that are secreted without a SP. In short, can we infer what makes up the secret secretome?

One perspective is to revisit the putative LSP list analysed in **Chapter 2** onwards, published by Ding et al. (2012). Noting that there is some redundancy in that original list, of the 65 unique genes identified the LSPDB reveals that only 54 are in the UNCLASSIFIED category, with the remained including misclassifications of SP, TM or secretory isoforms. Curation at this stage may account for inferences based on the use of SecretomeP in previous studies, and the LSPDB a convenient filter. Of those that passed, 30 of these correspond to the high and medium confidence lists created from the LSPDB and the remainder in the low confidence list. Unsurprisingly, as LSPpred was trained on these, after applying the prediction tools 33/54 had a positive prediction from one or both of the tools. Of interest though are 6 which are also positive for SPLpred, since the training of this tool was independent the confidence lists. This set includes pairs of jacalin-related lectins (AT3G16430, AT3G16450), glutamate dehydrogenases (AT5G07440, AT5G18170) as well as nitrile specifier protein (AT3G16410, NSP4) and S-adenosylmethionine synthetase 2 (AT4G01850, SAMS2). As mentioned in **Chapter 4**, the consensus predictions on the EXPO mediated LSP SAMS2 (Wang et al., 2010) is a positive note for the predictors, yet amongst too few LSPs to gauge the accuracy or establish a false-positive estimate.

The LSPDB also gave several candidate families, including chaperones, folding and heat shock-related proteins. Thioredoxin- and peroxiredoxin-related proteins were also identified, a family that has been observed in several instances, including the medium confidence LSPDB list (Section 3.4), the frequently observed proteins (**Table 3.3**), common co-occurring PPI (**Table 3.12**), and PTM based analysis (**Appendix Table D.4**). Calmodulin-related proteins were also identified in the medium confidence list, and have been experimentally observed by Cheng et al. (2009b) and in particular Cui et al. (2005), where immunoblotting was used to confirm an apoplasmic localisation.

The development of tools based on the outputs of LSPDB, including high and medium confidence LSPs for LSPpred, and SEC and NONSEC proteins for SPLpred, was applied

to the low confidence LSPDB proteins. These tools predicted several to be LSPs and of interest for experimental follow up. Predictions from both LSPpred and SPLpred, as well as a LSPfam positive result due to the presence of the overrepresented PF00190 Cupin 1, identified the RmlC-like cupins superfamily protein (AT1G07750). Another PFAM identified was the DUF538 proteins, and the corresponding PFAM domain (PF04398) that occurs predominantly in the SEC class of LSPDB proteins. Using LSPfam showed this was also present in two low confidence UNCLASSIFIED proteins, AT1G09310 and AT5G01610, both annotated as proteins of unknown function with DUF538. AT1G09310 was also found in the EV proteome Rutter and Innes (2017). Another protein with the DUF538 domain, secretory isoform serine protease (AT1G61667), was also found in an excluded study (Duruflé et al., 2017b). Interestingly, recent work on a mutant *glossy6* (*gl6*) maize line (Li et al., 2019), which contains the DUF538 domain, has shown a role for *gl6* in extracellular trafficking of wax. Another low confidence protein with this domain, AT5G46230, has shown to have increased expression in the suspensor during the early globular stage of development (Palovaara et al., 2017). Analysis by (Gholizadeh and Kohnhrouz, 2013) noted the similarity to proteins in the mammalian innate immune system, and linked to stress response in plants. This observation, as well as the overrepresentation of this PFAM domain in the SEC proteins, appearance in UNCLASSIFIED and relationship to development and involvement wax secretion suggest that a UPS function for DUF538 is worth investigating. Consensus between the two predictors (**Table 4.3**) identified several candidates, and this narrowing allows focus and comparison to other species. For example, enolase (AT2G36530, LOS2) was also found in all replicates of the secretory vesicle study (Waghmare et al. (2018)), as well as in EVs (Rutter and Innes, 2017). An enolase has been demonstrated to undergo UPS in yeast (Miura et al., 2012) in an analysis of the secretion of cytosolic glycolytic enzymes. These are of note, as several proteins identified here (triose phosphate isomerase, glyceraldehyde-3-phosphate dehydrogenase 1 and 2) are also canonically glycolytic and by narrowing the candidates, links between these plant studies and other organisms can be used to prompt further research.

From **Figure 5.7**, the relationship between abundance and leaderless secretion must

be considered, and the likelihood of observed contamination due to increased abundance cannot be dismissed. Using the maximum abundance of known secretory proteins as a marker, we've identified how candidates exceed this threshold overlap with the other positive evidence categories described in this Thesis. Only two high confidence LSP proteins from the network category overlapped with the abundance list. The singular categories also show no relationship for some genes; 19 that are abundant were not frequently observed, and 21 of the frequently observed are not amongst the most abundant. Several of the most frequent proteins observed are also abundant, lacking any other known features. That several of these had positive prediction scores indicates the protein are not completely dissimilar to those with more evidence for secretion. Ultimately then, the question is whether high abundance is a mechanism for some types of unconventional secretion, in that abundance determines availability for a rapid response to stress, or just increases the likelihood of a protein being present as contamination.

Therefore, based on these analyses and outstanding questions, highly abundant proteins should be scrutinised to a greater extent before being assigned as putative LSPs. One useful tool for this scrutiny is SPLpred. Since it was trained on secretory proteins only, the observance of abundant UNCLASSIFIED proteins in the LSPDB are not relevant to the training, and so abundant proteins with a positive secretory protein prediction or PTM profile, are useful orthogonal tools independent of the abundance parameter. This could also involve multiple prediction tools, i.e. only when there is consensus between ApoplastP, SPLpred and LSPpred, or using higher thresholds to decrease the chance of false-positives. It must be noted however that this approach is not without risk. Controlling for a FPR of 5% still suggests SPLpred may identify proteins incorrectly as secretory. As an example, of the most abundant proteins in **Table 5.8**, the only two with positive SPLpred prediction are malate dehydrogenase (AT5G09660) and RuBisCO-related proteins (AT1G67090) which also have predictions from LSPpred. The presence of both of these proteins in proteomics results have been used to measure cytosolic contamination (Alexandersson et al., 2013; Nogueira-Lopez et al., 2018), and so reclassification as LSPs would be a major shift in their use as markers for contaminants. Malate dehydrogenase was noted previously in secretory vesicle cargo which could suggest contamination

in that experimental approach, yet other jacalin-related lectins more closely related to known LSPs were also detected. Another marker for cytosolic contamination, glucose-6-phosphate dehydrogenase 6 (AT5G40760, G6PDH) (Chen et al., 2009b; Kwon, 2005) is not predicted to be an LSP by both SPLpred and LSPpred. The interpretation of these marker results illustrate the difficulty of balancing the results of the prediction tools, with the “known” contaminants. The lack of understanding of UPS mechanisms and functions leaves the possibility open that these markers for contamination could indeed undergo UPS, yet the tools are limited by the training data and/or hypotheses behind them, and each has the potential for false-positive results. Reliance on bioinformatics predictions alone cannot solve this dilemma, rather the combination of the tools as presented, along with literature evidence and ultimately experiments designed to investigate such unlikely candidates can attempt to resolve the issue. A summary of these LSP candidates, revising and expanding on the Ding et al. (2012) list, as well as possible contaminants is shown in **Appendix Table E.1**.

5.7 Conclusion

Evaluating these methods, like developing the methods themselves, is challenging without well verified plant LSPs. In this Chapter we have used new data to compensate by applying the tools to a new proteome (rice) and recent *Arabidopsis* studies excluded from the database, which included other secretory pathway experiments related to UPS. These applications have included a comparison to ApoplastP, and the results used to update the recommended workflow.

Each of the methods developed in this Thesis has limitations. This **Chapter** also compared and extended these methods to explore their complementary nature for predicting LSPs. The common candidates in the overlap between several approaches illustrates the usefulness of the tools provided, and the suggestions for improvement to overcome weaknesses are an avenue for future work. Developing methods for *Arabidopsis* has allowed other species to be used as test data, but the possibility that the methods are overly reliant on aspects unique to this species remains.

The similarity of lectin binding and thioredoxin proteins between mammals and plants indicates that at least for some classes of LSPs, there can be some commonality in the pathways for UPS but is likely to be a combination of conserved and plant-specific mechanisms. The relationship between the mammalian thioredoxin and thioredoxin family proteins identified previously, particular from the PTM analysis, is an area for further investigation.

Despite the prediction tools being built specifically with *Arabidopsis*-data, the ability to predict verified lectin LSPs in sunflower and superoxide dismutase in yeast shows they have some universality, though the tools were less effective at predicting mammalian LSPs. Several proteins recurred in the various analysis that were concordant with the putative LSPs in the literature including several jacalin-related lectins and S-adenosylmethionine synthetase. Importantly, the derivation of these candidates from the computational methods has been agnostic to these previous publications. That these generally (though not universally) accepted LSPs are predicted by these methods is a small measure of confidence that the tools developed here identify LSPs, and focused biochemical experiments to validate their secretion is required to confirm their status. The true LSP secretome is still something of a mystery, but the work presented here has gone some way to revealing its secrets.

Chapter 6

Conclusion

6.1 Thesis conclusion

This Thesis explored computational methods for identifying leaderless secretory proteins (LSPs) in *Arabidopsis thaliana* with applications to other species. In this final section, we review the contributions made, and discuss future computational refinements aided by targeted experimental validation that would assist in both identifying and elucidating the mechanism(s) of secretion of plant LSPs.

In **Chapter 2**, it was demonstrated that the mammalian trained SecretomeP, which is based on the theory that conventional and unconventional secreted proteins will share properties, did not meet the accuracy estimates known for human data when applied to plants. Therefore, its continued use for plant studies was not recommended, and previous inferences of LSP status based solely on SecretomeP need to be re-evaluated. This work has contributed to an awareness of the limitations of UPS prediction for plants and in more distantly related organisms in which this tool has been applied. For example, the perils of extrapolating the human LSP hypothesis wholesale into the trematode *Schistosoma* were recognised by Alan Wilson et al. (2017) who recommended avoiding the use of SecretomeP citing both this work and earlier warnings (DeMarco and Verjovski-Almeida, 2009). Alternative approaches have therefore been to either combine SecretomeP with alternative tools such as ApoplastP and to use the union of their predictions (de Lamo et al., 2018) or avoid prediction tools for LSPs altogether (Waghmare et al., 2018).

With the state of plant LSP prediction essentially reset, **Chapter 3**, detailed a Leaderless Secretory Protein Database (LSPDB) for *Arabidopsis*, taking the entire proteome

and applying a workflow of annotation, literature observation and experimental relationships between proteins to infer LSP candidates. The LSPDB is a resource available for all researchers to aid in the discovery of plant LSPs, and was also used as input for **Chapter 4**, where new prediction and classification tools for LSPs were developed. These included “white box” methods LSPfam and LSPblast that transparently identified elements present in new sequences within the LSPDB, and “black box” methods LSPpred and SPLpred consisting of random forest models created using LSP candidates or modified CSPs respectively as positive training data. These tool can then assign a LSP classification and prediction scores when applied to new data. All four tools as well as the LSPDB were then used in **Chapter 5** to identify more putative LSP candidates in *Arabidopsis* and rice.

6.2 Limitations

Each approach used in this Thesis to investigate LSPs has some limitations. Analysis of SecretomeP on plant data was limited (and also enabled) by the use of CSPs as a proxy to evaluate its accuracy, though subsequent analysis (Nielsen et al., 2018) has confirmed the results. The putative LSPs derived from the LSPDB are directly related to the confidence and scoring criteria. Changes in these criteria, or with the addition of data into the database, would change the putative lists. The bioinformatics tools developed are all dependent on their input data and are similarly limited by changes in the criteria. That only three plant LSPs are generally accepted as verified prevents thorough investigation into the accuracy of the tools.

The limitations of the input data of networks, GO terms and PFAM domains in the LSPDB were explored. PFAM domains are not present in all proteins and the GO terms were shown to have annotations that relied on the very experimental observations the database is designed to clarify. Network data was also shown to be incomplete as not all proteins have interactions. Several additional sources of data were identified for inclusion into the LSPDB. Adding co-expression, PTM and protein abundance data enabled additional insights into potential LSPs. These extra approaches also have limitations, however. Abundance data joined to the LSPDB classifications was derived from PaxDB a

database that gives an abundance measure for a protein based on its average expression across multiple studies, yet specific protein expression patterns under certain conditions or in specific tissues may be obscured. Similarly, the PTM results are also based on a subset of publicly available protein data and so some information may be lost due to the composition of the subset.

The criteria for selecting the LSPDB outputs are also limited. The confidence criteria for the GO terms and PFAM domains were determined on either exclusivity or majority rather than statistical enrichment in those classes. The possibility that the NONSEC group may contain unobserved LSPs, as well as the uneven PFAM distributions and hierarchy GO terms in those classes led to these simplified criteria. Evaluating the PPI network by establishing the cut-off for SEC interactions amongst SEC proteins was included as comparison for the network criteria, though the absence of PPI in many proteins meant that since the threshold is not sufficient to identify all SEC proteins, it is expected to also be limited when applied to the other classes. A more formal statistical enrichment approach is discussed in **Chapter 5** with PTMs. Applying a similar statistical framework to GO terms, PPI and PFAM domains in order to compare baselines in secretory proteins to LSP candidates is an avenue for future iterations of establishing database criteria.

Lack of ‘gold standard’ plant LSP data is the biggest issue for predicting LSPs presented in this work. Modifying training data to compensate for missing data is a double-edged sword. We have shown its danger when applied without testing for bias. SPLpred tempers bias thorough training data selection, and imbalance of data is also an issue in machine learning tools. In this case, training data were mostly NONSEC proteins. LSPpred and SPLpred compensate for this imbalance, but there are trade-offs. This lack of positive examples has led to pooled data from cell wall proteomes of different tissues, cell suspension cultures, roots, apoplastic fluid, etc being used, aggregating data at the cost of tissue specificity.

Due to these limitations, the ‘black box’ prediction tools are also explicit in what they do: LSPpred predicts similarity to previously observed unclassified proteins with relevant PPI, PFAM domains or GO terms related to secretion as curated in the LSPDB. SPLpred predicts sequence similarity of proteins to plant secretory proteins with a SP

(excluding secretory isoforms) with their SP removed. Each has prediction threshold guidance aimed at keeping false positives below 5%. Used with these limitations, they are new useful tools for narrowing down potential LSP targets. Consensus between the two “black box” tools LSPpred and SPLpred is also interesting given the different models each tool is built upon. Proteins predicted by both tools therefore contain features similar to secretory proteins, as well as features similar to other observed potential leaderless proteins, noting that when training LSPpred, the secretory proteins were an explicit negative class. While not guaranteed, the differing features between the predictors can indicate that each prediction is identifying a separate signal within the proteins.

6.3 Future work

Future work can be divided into extending the methods, validation of results, and application of the tools. Three additional sources of data (abundance, PTM and co-expression networks) were used in Section 5.3 to demonstrate the effect of adding new data to the LSPDB. Each source of new data was paired with the observation classes in the LSPDB, and the distribution of the data in those classes analysed. This illustrates the template for extending these methods, and the utility of this approach means data could be added to the LSPDB and used in an updated scoring system for selecting confidence sets in the future.

The networking layer drew on aggregate network information, and based inferences on the number of directed (or near direct) interactions. Further consideration of the local network structure, looking systematically for clusters to link LSPs to a secretory role may give additional insight. Network methods could also be extended to look at either pathway or curated gene sets present in the networks, rather than the individual gene interactions.

As discussed in **Chapter 3**, our identification of secretory isoforms has attempted to remove ambiguous evidence for secretion from the LSPDB. Further work at the level of mass-spectrometry raw data, analysing peptides results rather than the reported proteins, could be used to rescue some excluded observations when it is clear that the secre-

tory isoform is present, and that the secretory version has been observed. These rescued isoforms could then be included in the database. Secretory isoform information is also of use for other these species during secretome analysis, and standalone tools to identify them from reference proteomes could be created.

Having established that the database is useful for LSP analysis in *Arabidopsis*, we can also apply the database method to genomes/proteomes of other plants, or indeed any organism. Future work also requires reflection on the scalability of the pipeline - to scale to such a large number, existing annotations rather than predictions may be required, and be sourced directly from larger databases such as UniProt or a species-specific database, rather than running multiple prediction tools. Predictions were selected to be extensible to other species that do not have an annotation, and a combination approach may be required.

Repeating the database creation methods outlined in this Thesis for many species may not be feasible due to the intensive literature search and custom database design, therefore other approaches may be needed. The work in **Chapter 5** has shown that joining the classification scheme from the LSPDB to external data facilitated extended analysis. Therefore, a minimal approach that links genome and protein information to observations in the literature could serve to replicate a sufficient minimum amount of the LSPDB methods presented here, as it would produce the required SEC, UNCLASSIFIED, SPT and NONSEC categories. Initial investigation indicates a valid approach could be Inter-mine (Kalderimis et al., 2014), a framework for such integration that is in use for a variety of organisms including yeast, *Drosophila*, mouse, humans as well as plants such as wheat, maize, and of course, *Arabidopsis* (Araport/Thalemine). A rapid bioinformatics pipeline would also alleviate issues relating to analysis and the dynamic nature of the LSPDB, as it would allow new literature observations to propagate through the analysis. Repeating the LSPDB approach using this alternative framework on *Arabidopsis* would be the first step to validate the same classifications can be made using the expedited method. Expanding to more species with a common framework would also open up the possibility of a cross-species approach to the LSPDB, built on protein observations from multiple plant species.

The four tools (LSPpred, SPLpred, LSPblast and LSPfam) and the recommended LSP workflow (**Figure 5.3**) could be applied to find LSPs across all plant and algae, and investigate the difference in LSPs between species. Our initial work in *Oryza sativa subsp. japonica* identified LSPs, and essentially inverts the work in this Thesis. Rather than using the literature to build a dataset of LSPs, with new computational methods we can first analyse proteomes for candidate LSPs, and then either search for experimental evidence of secretion, or indeed propose new experiments to do so. These tools when applied to other species such as rice can also be complemented in the same way as illustrated in **Chapter 5** for example by comparing reference protein abundance information or comparing predicted candidates to the results of other experiments separate to secretomes, such as the proteomes of extracellular vesicles.

Notably though, there is an underlying assumption that LSPs are able to be predicted! Since UPS has several suggested pathways in plants (**Figure 1.4**), there may be sequence motifs or other features that are specific to each pathway and a single prediction model may not ultimately be appropriate. As noted in **Chapter 1.5** the pathways may also differ in terms of either active or passive selection of proteins. The uptake of proteins (and small RNAs Robinson et al. (2015)) into MVBs could be non-specific and contribute to the contaminants observed. The approach in this Thesis, however, has been to take observed data agnostic to whether these mechanisms are passive or active, and based only on observation as, in the absence of understanding LSPs, whether they are actually passively or actively selected is unknown. With the benefit of more candidate proteins as LSPs, it may lead to the identification of either motifs or mechanisms, which would truly allow for LSP candidates to be differentiated as passive or active. This would also allow the true positive data associated with a pathway to be collated and used individually rather than the combined LSP classification that the lack of 'gold standard' data has necessitated in this work. In the future, tools trained on each of possible distinct type of unconventional secretion could be created, e.g. UPSpred-Leaderless, UPSpred-MVB, UPSpred-EXPO, UPSpred-EV etc, and subsequently combined to create an ensemble predictor for LSPs. For example, a hypothetical UPSpred-Leaderless tool could be used to identify proteins that are unconventionally secreted across the PM. As such a tool be-

comes more accurate with additional training data; it effectively could mimic the biological recognition of a motif or secondary structure that facilitates the UPS. The goal to build specific tools for specific pathways (or parts thereof) would eventually negate the need for overarching LSP prediction tools as the exact mechanisms become known and ML can be substituted for recognition. Until there are sufficient experimental examples to develop such tools however, a combined LSP class and subsequent unified LSP tool is the practical alternative.

An alternative to the combined LSP class could be to invert the approach adopted in this Thesis. That is, rather than conflate potential LSP mechanisms together in order to identify candidates that are similar to either CSPs or other potential LSPs, using the same LSPDB data we could consider which observed proteins that lack secretory features are less likely to be LSPs, that is, contaminants. In this approach, features and interactions of unobserved, non-secretory proteins could be used to identify similar proteins, and these identified as likely contaminants. The requisite data for such an approach is already captured in some part in both the PFAM and GO terms associated with either NONSEC or NONSEC+UNCLASSIFIED, in **Chapter 3.7**. A more statistical view of the PTMs associated with each class, presented in **Chapter 5.4.2**, outlined which PTMs were over-represented in the SEC class and then applied to UNCLASSIFIED. Similarly, **Appendix Table D.3** also contains those PTMs associated with the NONSEC and UNCLASSIFIED sets. The same method of applying the SEC profile to UNCLASSIFIED, resulting in candidates shown in **Appendix Table D.4**, could equally be applied to the profile of NONSEC proteins, which as shown in **Table 5.7** overlapped with 343 proteins. Proteins identified in this way could then be subtracted from potential candidates, and the remainder considered LSPs. However, a major caveat to this approach is that the similarity to NONSEC proteins does not preclude that a protein may be an LSP, either due to an additional mechanism or the possibility that NONSEC contain LSPs that are as yet unobserved in an *Arabidopsis* proteome. Although this approach could reduce the number of potential LSP candidates by filtering the observed UNCLASSIFIED, and complement the approaches outlined in this Thesis, it would only produce LSP candidates by subtraction, and is akin to 'proving a negative' in terms of secretion. It does not assist in prioritising

which of those remaining members should be validated experimentally.

Similar future work could expand original or ensemble predictors to consider either strong evidence or predictions for specific subcellular locations apart from the extracellular space. Predictions for other subcellular compartments could also be used as input features to prediction tools, although the influence of whole prediction programs as inputs should be taken with caution, as shown in **Chapter 2.2**. The hierarchy of support vector machines as used in Loctree (Goldberg et al. 2012) similarly cascades through individual predictors for certain subcellular locations. Applying this logic to potential LSPs for either weak or missing prediction for other locations could be used to pre-filter UNCLASSIFIED proteins. However, this approach shares the disadvantage with contamination prediction in that features of other localisations do not preclude unconventional signals, especially as these are largely unknown in plants. The possibility of moonlighting proteins, that is proteins with multiple functions in multiple subcellular locations, means that strong prediction for those localisations doesn't preclude LSPs, and the suggested stress-response role that LSPs play could involve the repurposing and relocation of such proteins. Excluding proteins on this basis could remove potential LSPs, and so reduce true positives along with false positives. The trade-off of such future work would be dependent on comparing the cost and convenience of validation by bench experiments to reduce the number of candidates with this approach.

This need for experimental validation is the core issue with plant LSPs. Other research has allowed previously published lists of potential LSPs to be refined. Several putative plant LSPs were shown to have been erroneously assigned due to inaccurate prediction methods, are secretory isoforms, or were duplicates in the original publication of 76 proteins (Ding et al., 2012). After removing these duplicate LSPs candidates, 65 proteins remained (**Appendix Table A.2**). As summarised in **Appendix Table E.1** the original list can be further reduced to 54 once the inaccurately predicted CSPs and secretory isoforms are removed based on LSPDB annotations. Of these, more than half (30) were observed in the LSPDB confidence lists, with 13 of the remaining 24 predicted by either LSPpred or SPLpred. Eleven of the Ding proteins are not identified by either database or predictions tools. Yet despite this overlap, these are still computational approaches,

and consensus between the putative list and these methods (or indeed between the LSP-pred and SPLpred methods as we have noted) does not provide evidence they are true LSPs. The possibility of contamination, either from the technique used to isolate the cell wall/apoplastic fluid, cell lysis (especially in pathogen-related experiments) or due to the system being unrepresentative of *in planta* biology (i.e. cell suspension cultures) means that appearances in secretomes, even when combined with computational methods, are not enough to demonstrate UPS.

Biochemical evidence, as has been reiterated by Alberne et al. (2013) and others, is the ultimate requirement, yet is variable in its strength depending on the technique used. Fluorescent protein tagging of putative LSPs has been utilised as a high-throughput approach, yet can impact protein stability and/or turnover, protein dynamics and targeting location, among other things (Tanz et al., 2013). Other instances, even for established plant LSPs like SAMS2, rely on a two-step inference, such as GFP localisation with EXPO, and then evidence of EXPO impacting on the PM. Microscopic evidence for the subcellular localisation of LSPs using protein-specific antibodies is needed truly confirm extracellular localisations for LSPs, yet this is an intensive process. The inclusion of the *E. coli* resistance gene HYG in plant LSP literature may in fact be due to evidence for UPS of this gene in plants reaching the higher bar. The study by Zhang et al. (2011a) combines the observance of a HYG-GFP construct in the cell wall, continued secretion after inhibition of classical pathways with brefeldin A, and critically immunogold labelling of HYG to the cell wall. This combination of techniques exemplifies what is biochemically required for validation as an LSP. However, as outlined in **Chapter 1**, elucidation and evidence of specific mechanisms and conditions for secretion are the ultimate goals for accurate validation.

The tools developed in this Thesis are aids in prioritising candidates for such validation work. The LSPDB identified 1696 proteins in *Arabidopsis* secretomes without secretory features, and 384 of were identified as either high, medium or low confidence LSPs. These represent an additional set of 330 proteins that did not appear in the LSP list published by (Ding et al., 2012) (54 proteins). These proteins should be considered a “hit-list” of proteins for further analysis, with the computational tools developed in **Chapter 4** able

to be used to rank the list, as well as identify proteins in other plant species. If additional experimental data can be used to increase confidence in these tools, then larger scale analysis is also feasible. The presence of LSPs across plants could be investigated in reference genomes, or even in a broader evolutionary sense using data such as the OneKP (One Thousand Plant Transcriptomes Initiative) (Leebens-Mack et al., 2019). The distribution of LSPs in green plants could be investigated using these 1124 plant transcriptomes, applying the tools to the predicted protein sequences to investigate if LSPs are present in all plant (and algal) species. Alternatively, such an approach could be used to query if the protein features identified in these tools are overly specific to *Arabidopsis*, and explore if there are unexpected differences in predicted LSPs in species relative to experimental evidence of UPS. An InterMine-based approach to the LSPDB could also address the issue of *Arabidopsis* specificity, and be a source of data for prediction tools based on multiple plant species.

Future tools based on LSPpred or SPLpred approaches could be used on these other organisms, especially if new species are added to the LSPDB. The application of tools trained on mammalian data to plant proteins was a catalyst for much of the work in this Thesis, and highlighted the danger of extrapolating hypotheses about LSPs to other organisms beyond which a bioinformatics tool is trained on. Unexpected results in **Chapter 4**, showed a fair predictive ability of LSPpred on bacterial (Gram-positive) LSPs. The reciprocal ability of PeNGaRoo (Zhang et al., 2019) to identify 3/4 plant LSPs was also of note, although this comparison lacks an estimation of false positives when applied to plants. More plant LSPs will be needed to see if these existing programs already are able to predict both TP and TN proteins, and a critical analysis of the initial good performance of LSPpred on bacteria is required with the same skepticism of using SecretomeP on plants. If this accuracy is shown to be genuine, it could form the basis of a hypothesis implying that a conserved mechanism/s or extracellular environments between plants, fungi and bacteria are able to be exploited for prediction and discovery of LSPs across multiple species. The poor performance of a mammalian secretory protein predictor, SecretomeP, on plant proteins facilitated the creation of new plant-specific tools. To observe that plant-specific tools, LSPpred in particular, appears to be significantly better

than chance at distinguishing LSPs from non-secreted proteins in Gram-positive bacteria is not without irony.

Desperate times call for desperate measures, and the scarcity of well characterised plant LSP data has led to intriguing approaches to predict LSPs, for example, by the modification of CSPs in the broader literature and within the work presented in this Thesis. The curation stages of the LSPDB established criteria for UNCLASSIFIED proteins to be LSP candidates, given plausible features, and considered as putatively secreted. These proteins were then used as training data for prediction models, yet like common features within secreted proteins, this is a placeholder method. As more proteins are found to be genuine LSPs, machine learning approaches only using these verified proteins as training data will lead to the development of a more confident method to predict them. Again, however, future work based on these computational methods must be complemented with lab-based validation experiments.

Validation work can also include functional assays, that can be used to infer mechanisms. As more LSPs become empirically verified, experiments designed to increase or perturb their unconventional secretion are needed to further elucidate LSP mechanisms. Response to pathogens can be used to compare LSPs between normal and induced conditions, for example. In Nogueira-Lopez et al. (2018) observed an increase in LSPs from 26% to 55% of the maize apoplastic proteins in response to the fungus *Trichoderma virens*. Approaches similar to the modification of a diacidic DE motif that led to a modulation of secretion of yeast LSPs (Cruz-Garcia et al., 2017) demonstrate that if a cellular mechanism for leaderless secretion can be identified, mutations and a subsequent loss of observed LSPs can be used as further evidence of UPS.

Dimerisation, lectin domains, and disulfide bonds were all indicated in this Thesis as potential protein features that aid secretion. The frequency of asparagine and glycine residues in protein sequences were also indicated as important features from LSPpred. Mutation of these amino acid residues and loss of UPS secretion could provide orthogonal evidence of a LSP, as could the prevention of PTM additions to those proteins identified as PTM targets. Using these approaches will both validate and enhance computational methods of LSP prediction. Every validated LSP adds another example for com-

putational methods, increasing the sizes of the datasets that can then facilitate improved accuracy and different machine learning methods. As these tools improve, they can identify more LSPs for validation and start a virtuous cycle that leads to an increase in knowledge regarding the mechanisms and biology behind leaderless protein secretion in plants. With each additional LSP confirmed, this knowledge moves towards defining them by what features they may have that lead to secretion, rather than those they lack.

References

- Agrawal, G. K., Jwa, N.-S., Lebrun, M.-H., Job, D., and Rakwal, R. (2010). Plant secretome: unlocking secrets of the secreted proteins. *Proteomics* 10, 799–827
- Alan Wilson, R., Li, X. H., and Castro-Borges, W. (2017). Schistosome vaccines: problems, pitfalls and prospects. *Emerging Topics in Life Sciences* 1, 641–650
- Albenne, C., Canut, H., and Jamet, E. (2013). Plant cell wall proteomics: the leadership of arabidopsis thaliana. *Front. Plant Sci.* 4, 111
- Alexandersson, E., Ali, A., Resjö, S., and Andreasson, E. (2013). Plant secretome proteomics. *Front. Plant Sci.* 4, 9
- Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H., and Winther, O. (2017). DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* 33, 3387–3395
- Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
- Armenteros, J. J. A., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., et al. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* , 1
- Balakireva, A. V. and Zamyatnin, A. A. (2018). Indispensable role of proteases in plant innate immunity. *Int. J. Mol. Sci.* 19
- Basu, U., Francis, J. L., Whittall, R. M., Stephens, J. L., Wang, Y., Zaiane, O. R., et al. (2006).

- Extracellular proteomes of arabidopsis thaliana and brassica napus roots: Analysis and comparison by MudPIT and LC-MS/MS. *Plant Soil* 286, 357–376
- Batemen, A. (2007). EDITORIAL. *Nucleic Acids Res.* 35, D1
- Bayer, E. M., Bottrill, A. R., Walshaw, J., Vigouroux, M., Naldrett, M. J., Thomas, C. L., et al. (2006). Arabidopsis cell wall proteome defined using multidimensional protein identification technology. *Proteomics* 6, 301–311
- Bellincampi, D., Cervone, F., and Lionetti, V. (2014). Plant cell wall dynamics and wall-related susceptibility in plant-pathogen interactions. *Front. Plant Sci.* 5, 228
- Bellucci, M., De Marchis, F., and Pompa, A. (2017). The endoplasmic reticulum is a hub to sort proteins toward unconventional traffic pathways and endosymbiotic organelles. *J. Exp. Bot.*
- Bendtsen, J. D., Jensen, L. J., Blom, N., von Heijne, G., and Brunak, S. (2004). Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng. Des. Sel.* 17, 349–356
- Bendtsen, J. D., Kiemer, L., Fausbøll, A., and Brunak, S. (2005). Non-classical protein secretion in bacteria. *BMC Microbiol.* 5, 58
- Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., et al. (2015). The arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome. *Genesis* 53, 474–485
- Bernal-Bayard, P., Ojeda, V., Hervás, M., Cejudo, F. J., Navarro, J. A., Velázquez-Campoy, A., et al. (2014). Molecular recognition in the interaction of chloroplast 2-cys peroxiredoxin with NADPH-thioredoxin reductase C (NTRC) and thioredoxin x. *FEBS Lett.* 588, 4342–4347
- Bhushan, D., Jaiswal, D. K., Ray, D., Basu, D., Datta, A., Chakraborty, S., et al. (2011). Dehydration-Responsive Reversible and Irreversible Changes in the Extracellular Matrix: Comparative Proteomics of Chickpea Genotypes with Contrasting Tolerance. *J. Proteome Res.* 10, 2027–2046

-
- Borderies, G., Jamet, E., Lafitte, C., Rossignol, M., Jauneau, A., Boudart, G., et al. (2003). Proteomics of loosely bound cell wall proteins of *Arabidopsis thaliana* cell suspension cultures: A critical analysis. *Electrophoresis* 24, 3421–3432
- Borner, G. H. H. (2003). Identification of Glycosylphosphatidylinositol-Anchored proteins in *Arabidopsis*. a proteomic and genomic analysis. *Plant Physiol.* 132, 568–577
- Boudart, G., Jamet, E., Rossignol, M., Lafitte, C., Borderies, G., Jauneau, A., et al. (2005). Cell wall proteins in apoplastic fluids of *Arabidopsis thaliana* rosettes: identification by mass spectrometry and bioinformatics. *Proteomics* 5, 212–221
- Brandão, M. M., Dantas, L. L., and Silva-Filho, M. C. (2009). AtPIN: *Arabidopsis thaliana* protein interaction network. *BMC Bioinformatics* 10, 454
- Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32
- Brettner, L. M. and Masel, J. (2012). Protein stickiness, rather than number of functional protein-protein interactions, predicts expression noise and plasticity in yeast. *BMC Syst. Biol.* 6, 128
- Canut, H., Albenne, C., and Jamet, E. (2017). Isolation of the cell wall. In *Isolation of Plant Organelles and Structures: Methods and Protocols*, eds. N. L. Taylor and A. H. Millar (New York, NY: Springer New York). 171–185
- Carroll, A. J. (2013). The *Arabidopsis* cytosolic ribosomal proteome: From form to function. *Front. Plant Sci.* 4, 32
- Carvalho, C. M., Santos, A. A., Pires, S. R., Rocha, C. S., Saraiva, D. I., Machado, J. P. B., et al. (2008). Regulated nuclear trafficking of rpL10A mediated by NIK1 represents a defense strategy of plant cells against virus. *PLoS Pathog.* 4, e1000247
- Casasoli, M., Spadoni, S., Lilley, K. S., Cervone, F., De Lorenzo, G., and Mattei, B. (2008). Identification by 2-D DIGE of apoplastic proteins regulated by oligogalacturonides in *Arabidopsis thaliana*. *Proteomics* 8, 1042–1054

- Ceballos-Laita, L., Gutierrez-Carbonell, E., Lattanzio, G., Vázquez, S., Contreras-Moreira, B., Abadía, A., et al. (2015). Protein profile of *Beta vulgaris* leaf apoplastic fluid and changes induced by Fe deficiency and Fe resupply 6
- Chapple, C. E., Robisson, B., Spinelli, L., Guien, C., Becker, E., and Brun, C. (2015). Extreme multifunctional proteins identified from a human protein interaction network. *Nat. Commun.* 6, 7412
- Charmont, S., Jamet, E., Pont-Lezica, R., and Canut, H. (2005). Proteomic analysis of secreted proteins from *arabidopsis thaliana* seedlings: improved recovery following removal of phenolic compounds. *Phytochemistry* 66, 453–461
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and others (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Organs*
- Chen, C., Zabad, S., Liu, H., Wang, W., and Jeffery, C. (2018). MoonProt 2.0: an expansion and update of the moonlighting proteins database. *Nucleic Acids Res.* 46, D640–D644
- Chen, T. Y., Ho, J. W. K., Liu, H., and Xie, X. (2009a). An innovative approach for testing bioinformatics programs using metamorphic testing. *BMC Bioinformatics* 10, 24
- Chen, X.-Y., Kim, S. T., Cho, W. K., Rim, Y., Kim, S., Kim, S.-W., et al. (2009b). Proteomics of weakly bound cell wall proteins in rice calli. *J. Plant Physiol.* 166, 675–685
- Chen, Y., Ye, D., Held, M. A., Cannon, M. C., Ray, T., Saha, P., et al. (2015). Identification of the abundant Hydroxyproline-Rich glycoproteins in the root walls of Wild-Type *arabidopsis*, an ext3 mutant line, and its phenotypic revertant. *Plants* 4, 85–111
- Cheng, F.-Y., Blackburn, K., Lin, Y.-M., Goshe, M. B., and Williamson, J. D. (2009a). Absolute Protein Quantification by LC/MS Efor Global Analysis of Salicylic Acid-Induced Plant Protein Secretion Responses. *J. Proteome Res.* 8, 82–93
- Cheng, F.-Y., Blackburn, K., Lin, Y.-M., Goshe, M. B., and Williamson, J. D. (2009b). Absolute protein quantification by LC/MS(E) for global analysis of salicylic acid-induced plant protein secretion responses. *J. Proteome Res.* 8, 82–93

- Cheng, F.-Y., Zamski, E., Guo, W.-W., Pharr, D. M., and Williamson, J. D. (2009c). Salicylic acid stimulates secretion of the normally symplastic enzyme mannitol dehydrogenase: a possible defense against mannitol-secreting fungal pathogens. *Planta* 230, 1093–1103
- Chivasa, S., Ndimba, B. K., Simon, W. J., Robertson, D., Yu, X.-L., Knox, J. P., et al. (2002). Proteomic analysis of the arabidopsis thaliana cell wall. *Electrophoresis* 23, 1754–1765
- Cho, J.-I., Lee, S.-K., Ko, S., Kim, H.-K., Jun, S.-H., Lee, Y.-H., et al. (2005). Molecular cloning and expression analysis of the cell-wall invertase gene family in rice (*oryza sativa* l.). *Plant Cell Rep.* 24, 225–236
- Cho, W. K., Chen, X. Y., Chu, H., Rim, Y., Kim, S., Kim, S. T., et al. (2009). Proteomic analysis of the secretome of rice calli. *Physiol. Plant.* 135, 331–341
- Clemente, H. S., Pont Lezica, R., and Jamet, E. (2009). Bioinformatics as a tool for assessing the quality of sub-cellular proteomic strategies and inferring functions of proteins: plant cell wall proteomics as a test case. *Bioinform. Biol. Insights* 3, 15–28
- Cruz-Garcia, D., Brouwers, N., Duran, J. M., Mora, G., Curwin, A. J., and Malhotra, V. (2017). A diacidic motif determines unconventional secretion of wild-type and ALS-linked mutant SOD1. *J. Cell Biol.*
- Cui, S., Guo, X., Chang, F., Cui, Y., Ma, L., Sun, Y., et al. (2005). Apoplastic calmodulin receptor-like binding proteins in suspension-cultured cells of arabidopsis thaliana. *J. Biol. Chem.* 280, 31420–31427
- Dafoe, N. J. and Constabel, C. P. (2009). Proteomic analysis of hybrid poplar xylem sap. *Phytochemistry* 70, 856–863
- Davis, D. J., Kang, B.-H., Heringer, A. S., Wilkop, T. E., and Drakakaki, G. (2016). Unconventional protein secretion in plants. In *Unconventional Protein Secretion: Methods and Protocols*, eds. A. Pompa and F. De Marchis (New York, NY: Springer New York). 47–63
- Davis, E. J., Chen, D., DeWald, D. B., Shope, J., MacAdam, J., and Wu, Y. (2009). Searching for new cell wall protein genes in Arabidopsis. *Plant Physiol. Biochem.* 47, 81–85

- Davletova, S., Rizhsky, L., Liang, H., Shengqiang, Z., Oliver, D. J., Coutu, J., et al. (2005). Cytosolic ascorbate peroxidase 1 is a central component of the reactive oxygen gene network of arabidopsis. *Plant Cell* 17, 268–281
- de Lamo, F. J., Constantin, M. E., Fresno, D. H., Boeren, S., Rep, M., and Takken, F. L. W. (2018). Xylem sap proteomics reveals distinct differences between R gene- and Endophyte-Mediated resistance against fusarium wilt disease in tomato. *Front. Microbiol.* 9, 2977
- De Maio, A. and Vazquez, D. (2013). Extracellular heat shock proteins: a new location, a new function. *Shock* 40, 239–246
- Del Rio, M., de la Canal, L., Pinedo, M., and Regente, M. (2018). Internalization of a sunflower mannose-binding lectin into phytopathogenic fungal cells induces cytotoxicity. *J. Plant Physiol.* 221, 22–31
- Delaunais, B., Colby, T., Belloy, N., Conreux, A., Harzen, A., Baillieul, F., et al. (2013). Large-scale proteomic analysis of the grapevine leaf apoplastic fluid reveals mainly stress-related proteins and cell wall modifying enzymes. *BMC Plant Biol.* 13, 24
- DeMarco, R. and Verjovski-Almeida, S. (2009). Schistosomes—proteomics studies for potential novel vaccines and drug targets. *Drug Discov. Today* 14, 472–478
- Demarty, M., Morvan, C., and Thellier, M. (1984). Calcium and the cell wall. *Plant Cell Environ.* 7, 441–448
- Dimou, E. and Nickel, W. (2018). Unconventional mechanisms of eukaryotic protein secretion. *Curr. Biol.* 28, R406–R410
- Ding, Y., Robinson, D. G., and Jiang, L. (2014). Unconventional protein secretion (UPS) pathways in plants. *Curr. Opin. Cell Biol.* 29, 107–115
- Ding, Y., Wang, J., Wang, J., Stierhof, Y.-D., Robinson, D. G., and Jiang, L. (2012). Unconventional protein secretion. *Trends Plant Sci.* 17, 606–615
- Doblin, M. S., Pettolino, F., and Bacic, A. (2010). Evans review: Plant cell walls: the skeleton of the plant world. *Funct. Plant Biol.* 37, 357

-
- Dubchak, I., Muchnik, I., Holbrook, S. R., and Kim, S. H. (1995). Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. U. S. A.* 92, 8700–8704
- Duckert, P., Brunak, S., and Blom, N. (2004). Prediction of proprotein convertase cleavage sites. *Protein Eng. Des. Sel.* 17, 107–112
- Duncan, O., Taylor, N. L., Carrie, C., Eubel, H., Kubiszewski-Jakubiak, S., Zhang, B., et al. (2011). Multiple lines of evidence localize signaling, morphology, and lipid biosynthesis machinery to the mitochondrial outer membrane of arabidopsis. *Plant Physiol.* 157, 1093–1113
- Duruflé, H., Clemente, H. S., Balliau, T., Zivy, M., Dunand, C., and Jamet, E. (2017a). Cell wall proteome analysis of arabidopsis thaliana mature stems. *Proteomics* 17
- Duruflé, H., Hervé, V., Ranocha, P., Balliau, T., Zivy, M., Chourré, J., et al. (2017b). Cell wall modifications of two arabidopsis thaliana ecotypes, col and sha, in response to sub-optimal growth conditions: An integrative study. *Plant Sci.* 263, 183–193
- Eisenhaber, B., Bork, P., and Eisenhaber, F. (1999). Prediction of potential GPI-modification sites in proprotein sequences. *J. Mol. Biol.* 292, 741–758
- Feiz, L., Irshad, M., Pont-Lezica, R. F., Canut, H., and Jamet, E. (2006). Evaluation of cell wall preparations for proteomics: a new procedure for purifying cell walls from arabidopsis hypocotyls. *Plant Methods* 2, 10
- Fernández, M. B., Pagano, M. R., Daleo, G. R., and Guevara, M. G. (2012). Hydrophobic proteins secreted into the apoplast may contribute to resistance against *Phytophthora infestans* in potato. *Plant Physiol. Biochem.* 60, 59–66
- Foissner, I., Sommer, A., Hoefftberger, M., Hoepflinger, M. C., and Absolonova, M. (2016). Is Wortmannin-Induced reorganization of the trans-golgi network the key to explain charasome formation? *Front. Plant Sci.* 7, 756
- Galperin, M. Y., Fernández-Suárez, X. M., and Rigden, D. J. (2017). The 24th annual

- nucleic acids research database issue: a look back and upcoming changes. *Nucleic Acids Res.* 45, D1–D11
- Gao, C., Cai, Y., Wang, Y., Kang, B.-H., Aniento, F., Robinson, D. G., et al. (2014). Retention mechanisms for ER and golgi membrane proteins. *Trends Plant Sci.* 19, 508–515
- Garg, A. and Raghava, G. P. S. (2008). A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search. *In Silico Biol.* 8, 129–140
- Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., and Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* 31, 3784–3788
- Ge, W., Song, Y., Zhang, C., Zhang, Y., Burlingame, A. L., and Guo, Y. (2011). Proteomic analyses of apoplastic proteins from germinating arabidopsis thaliana pollen. *Biochim. Biophys. Acta* 1814, 1964–1973
- Gee, H. Y., Kim, J., and Lee, M. G. (2018). Unconventional secretion of transmembrane proteins. *Semin. Cell Dev. Biol.* 83, 59–66
- Georgiev, A. G. (2009). Interpretable numerical descriptors of amino acid space. *J. Comput. Biol.* 16, 703–723
- Gholizadeh, A. and Kohnhrouz, S. B. (2013). DUF538 protein super family is predicted to be the potential homologue of bactericidal/permeability-increasing protein in plant system. *Protein J.* 32, 163–171
- Giannoulatou, E., Park, S.-H., Humphreys, D. T., and Ho, J. W. K. (2014). Verification and validation of bioinformatics software without a gold standard: a case study of BWA and bowtie. *BMC Bioinformatics* 15 Suppl 16, S15
- Goksuluk, D., Korkmaz, S., Zararsiz, G., and Karaagaoglu, A. E. (2016). easyROC: an interactive web-tool for ROC curve analysis using R language environment. *RJ* 8, 213–230

-
- Goldberg, T., Hamp, T., and Rost, B. (2012). LocTree2 predicts localization for all domains of life. *Bioinformatics* 28, i458–i465
- Gray-Mitsumune, M. and Matton, D. P. (2005). The Egg apparatus 1 gene from maize is a member of a large gene family found in both monocots and dicots. *Planta* 223, 618–625
- Green, E. R. and Meccas, J. (2016). Bacterial secretion systems: An overview. *Microbiol Spectr* 4
- Guerra-Guimarães, L., Pinheiro, C., Chaves, I., Barros, D. R., and Ricardo, C. P. (2016). Protein dynamics in the plant extracellular space. *Proteomes* 4
- Guerra-Guimarães, L., Vieira, A., Chaves, I., Pinheiro, C., Queiroz, V., Renaut, J., et al. (2014). Effect of greenhouse conditions on the leaf apoplastic proteome of *Coffea arabica* plants. *J. Proteomics* 104, 128–139
- Gupta, S., Wardhan, V., Verma, S., Gayali, S., Rajamani, U., Datta, A., et al. (2011). Characterization of the Secretome of Chickpea Suspension Culture Reveals Pathway Abundance and the Expected and Unexpected Secreted Proteins. *J. Proteome Res.* 10, 5006–5015
- Haslam, R. P., Downie, A. L., M, R., K, G., D, J., Pallett, K. E., et al. (2003). The assessment of enriched apoplastic extracts using proteomic approaches. *Ann. Appl. Biol.* 143, 81–91
- Hatsugai, N., Iwasaki, S., Tamura, K., Kondo, M., Fuji, K., Ogasawara, K., et al. (2009). A novel membrane fusion-mediated plant immunity against bacterial pathogens. *Genes Dev.* 23, 2496–2506
- Heard, W., Sklenář, J., Tomé, D. F. A., Robatzek, S., and Jones, A. M. E. (2015). Identification of regulatory and cargo proteins of endosomal and secretory pathways in *arabidopsis thaliana* by proteomic dissection. *Mol. Cell. Proteomics* 14, 1796–1813
- Helmy, M., Crits-Christoph, A., and Bader, G. D. (2016). Ten simple rules for developing public biological databases. *PLoS Comput. Biol.* 12, e1005128

- Hervé, V., Duruflé, H., San Clemente, H., Albenne, C., Balliau, T., Zivy, M., et al. (2016). An enlarged cell wall proteome of arabidopsis thaliana rosettes. *Proteomics* 16, 3183–3187
- Höfte, H. and Voxeur, A. (2017). Plant cell walls. *Curr. Biol.* 27, R865–R870
- Huang, W.-L. (2012). Ranking gene ontology terms for predicting non-classical secretory proteins in eukaryotes and prokaryotes. *J. Theor. Biol.* 312, 105–113
- Hwang, J. E., Hong, J. K., Lim, C. J., Chen, H., Je, J., Yang, K. A., et al. (2010). Distinct expression patterns of two arabidopsis phytocystatin genes, AtCYS1 and AtCYS2, during development and abiotic stresses. *Plant Cell Rep.* 29, 905–915
- Irshad, M., Canut, H., Borderies, G., Pont-Lezica, R., and Jamet, E. (2008). A new picture of cell wall protein dynamics in elongating cells of arabidopsis thaliana: Confirmed actors and newcomers. *BMC Plant Biol.* 8, 94
- Ito, J., Batth, T. S., Petzold, C. J., Redding-Johanson, A. M., Mukhopadhyay, A., Verboom, R., et al. (2011). Analysis of the arabidopsis cytosolic proteome highlights subcellular partitioning of central plant metabolism. *J. Proteome Res.* 10, 1571–1582
- Jamet, E., Albenne, C., Boudart, G., Irshad, M., Canut, H., and Pont-Lezica, R. (2008). Recent advances in plant cell wall proteomics. *Proteomics* 8, 893–908
- Jamet, E., Canut, H., Boudart, G., and Pont-Lezica, R. F. (2006). Cell wall proteins: a new insight through proteomics. *Trends Plant Sci.* 11, 33–39
- Janin, J. (1979). Surface and inside volumes in globular proteins. *Nature* 277, 491–492
- Jaquinod, M., Villiers, F., Kieffer-Jaquinod, S., Hugouvieux, V., Bruley, C., Garin, J., et al. (2007). A proteomics dissection of arabidopsis thaliana vacuoles isolated from cell culture. *Mol. Cell. Proteomics* 6, 394–412
- Jiang, Y., Yang, B., Harris, N. S., and Deyholos, M. K. (2007). Comparative proteomic analysis of NaCl stress-responsive proteins in arabidopsis roots. *J. Exp. Bot.* 58, 3591–3607

-
- Juarez-Diaz, J. A., McClure, B., Vazquez-Santana, S., Guevara-Garcia, A., Leon-Mejia, P., Marquez-Guzman, J., et al. (2006). A Novel Thioredoxin h Is Secreted in *Nicotiana glauca* and Reduces S-RNase in Vitro. *J. Biol. Chem.* 281, 3418–3424
- Jung, B., Hoffmann, C., and Möhlmann, T. (2011). Arabidopsis nucleoside hydrolases involved in intracellular and extracellular degradation of purines: Arabidopsis nucleoside hydrolase family. *Plant J.* 65, 703–711
- Kalderimis, A., Lyne, R., Butano, D., Contrino, S., Lyne, M., Heimbach, J., et al. (2014). InterMine: extensive web services for modern biology. *Nucleic Acids Res.* 42, W468–72
- Kandaswamy, K. K., Pugalenth, G., Hartmann, E., Kalies, K.-U., Möller, S., Suganthan, P. N., et al. (2010). SPRED: A machine learning approach for the identification of classical and non-classical secretory proteins in mammalian genomes. *Biochem. Biophys. Res. Commun.* 391, 1306–1311
- Kaneko, T. S., Sato, M., Osumi, M., Muroi, M., and Takatsuki, A. (1996). Two isoforms of acid phosphatase secreted by tobacco protoplasts: differential effect of brefeldin A on their secretion. *Plant Cell Rep.* 15, 409–413
- Keegstra, K. (2010). Plant cell walls. *Plant Physiol.* 154, 483–486
- Kim, J., Gee, H. Y., and Lee, M. G. (2018a). Unconventional protein secretion - new insights into the pathogenesis and therapeutic targets of human diseases. *J. Cell Sci.* 131
- Kim, J., Gee, H. Y., and Lee, M. G. (2018b). Unconventional protein secretion - new insights into the pathogenesis and therapeutic targets of human diseases. *J. Cell Sci.* 131
- Kim, S.-J., Kim, K.-W., Cho, M.-H., Franceschi, V. R., Davin, L. B., and Lewis, N. G. (2007). Expression of cinnamyl alcohol dehydrogenases and their putative homologues during *Arabidopsis thaliana* growth and development: lessons for database annotations? *Phytochemistry* 68, 1957–1974
- Krause, C., Richter, S., Knöll, C., and Jürgens, G. (2013). Plant secretome — from cellular

- process to biological activity. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1834, 2429–2441
- Krause, M. and Rodrigues-Krause, J. d. C. (2011). Extracellular heat shock proteins (eHSP70) in exercise: Possible targets outside the immune system and their role for neurodegenerative disorders treatment. *Med. Hypotheses* 76, 286–290
- Krogh, A., Larsson, B., Heijne, G., and Sonnhammer, E. L. L. (2001). Predicting trans-membrane protein topology with a hidden markov model: Application to complete genomes. *available online at <http://www.idealibrary.com on J. Mol. Biol>* 305, 567–580
- Kusumawati, L., Imin, N., and Djordjevic, M. A. (2008). Characterization of the Secretome of Suspension Cultures of Medicago Species Reveals Proteins Important for Defense and Development. *J. Proteome Res.* 7, 4508–4520
- Kwon, H. K. (2005). A proteomic approach to apoplastic proteins involved in cell wall regeneration in protoplasts of arabidopsis suspension-cultured cells. *Plant Cell Physiol.* 46, 843–857
- Kwon, H.-K., Yokoyama, R., and Nishitani, K. (2005). A proteomic approach to apoplastic proteins involved in cell wall regeneration in protoplasts of arabidopsis suspension-cultured cells. *Plant Cell Physiol.* 46, 843–857
- Laohavisit, A., Mortimer, J. C., Demidchik, V., Coxon, K. M., Stancombe, M. A., Macpherson, N., et al. (2009). Zea mays Annexins Modulate Cytosolic Free Ca²⁺ and Generate a Ca²⁺-Permeable Conductance. *THE PLANT CELL ONLINE* 21, 479–493
- Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., et al. (2006). Machine learning in bioinformatics. *Brief. Bioinform.* 7, 86–112
- Leebens-Mack, J. H., Barker, M. S., Carpenter, E. J., Deyholos, M. K., Gitzendanner, M. A., Graham, S. W., et al. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*
- Lehtonen, M. T., Takikawa, Y., Rönholm, G., Akita, M., Kalkkinen, N., Ahola-Iivarinen,

-
- E., et al. (2014). Protein Secretome of Moss Plants (*Physcomitrella patens*) with Emphasis on Changes Induced by a Fungal Elicitor. *J. Proteome Res.* 13, 447–459
- Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* 18, 1–5
- Li, L., Du, Y., He, C., Dietrich, C. R., Li, J., Ma, X., et al. (2019). Maize glossy6 is involved in cuticular wax deposition and drought tolerance. *J. Exp. Bot.* 70, 3089–3099
- Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659
- Lonsdale, A., Davis, M. J., Doblin, M. S., and Bacic, A. (2016). Better than nothing? limitations of the prediction tool SecretomeP in the search for leaderless secretory proteins (LSPs) in plants. *Front. Plant Sci.* 7
- López-Ratón, M., Rodríguez-Álvarez, M., Cadarso-Suárez, C., and Gude-Sampedro, F. (2014). OptimalCutpoints: An R package for selecting optimal cutpoints in diagnostic tests. *Journal of Statistical Software, Articles* 61, 1–36
- Lum, G., Meinken, J., Orr, J., and Frazier, S. (2014). PlantSecKB: the plant secretome and subcellular proteome KnowledgeBase. *Computational Molecular*
- Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21, 3448–3449
- Marmagne, A., Ferro, M., Meinnel, T., Bruley, C., Kuhn, L., Garin, J., et al. (2007). A high content in lipid-modified peripheral proteins and integral receptor kinases features in the arabidopsis plasma membrane proteome. *Mol. Cell. Proteomics* 6, 1980–1996
- Marmagne, A., Rouet, M.-A., Ferro, M., Rolland, N., Alcon, C., Joyard, J., et al. (2004). Identification of new intrinsic proteins in arabidopsis plasma membrane proteome. *Mol. Cell. Proteomics* 3, 675–691

- Merret, R., Carpentier, M.-C., Favory, J.-J., Picart, C., Descombin, J., Bousquet-Antonelli, C., et al. (2017). Heat shock protein HSP101 affects the release of ribosomal protein mRNAs for recovery after heat shock. *Plant Physiol.* 174, 1216–1225
- Meyer, D., Pajonk, S., Micali, C., O’Connell, R., and Schulze-Lefert, P. (2009). Extracellular transport and integration of plant secretory proteins into pathogen-induced cell wall compartments. *Plant J.* 57, 986–999
- Mi, H. and Thomas, P. (2009). PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol. Biol.* 563, 123–140
- Michaelis, S. and Barrowman, J. (2012). Biogenesis of the *saccharomyces cerevisiae* pheromone α -factor, from yeast mating to human disease. *Microbiol. Mol. Biol. Rev.* 76, 626–651
- Miki, B. and McHugh, S. (2004). Selectable marker genes in transgenic plants: applications, alternatives and biosafety. *J. Biotechnol.* 107, 193–232
- Minic, Z., Jamet, E., Negroni, L., Arsene der Garabedian, P., Zivy, M., and Jouanin, L. (2007). A sub-proteome of *arabidopsis thaliana* mature stems trapped on concanavalin a is enriched in cell wall glycoside hydrolases. *J. Exp. Bot.* 58, 2503–2512
- Mitra, S. K., Gantt, J. A., Ruby, J. F., Clouse, S. D., and Goshe, M. B. (2007). Membrane proteomic analysis of *arabidopsis thaliana* using alternative solubilization techniques. *J. Proteome Res.* 6, 1933–1950
- Miura, N., Kirino, A., Endo, S., Morisaka, H., Kuroda, K., Takagi, M., et al. (2012). Tracing putative trafficking of the glycolytic enzyme enolase via SNARE-driven unconventional secretion. *Eukaryot. Cell* 11, 1075–1082
- Nativel, B., Planesse, C., Gasque, P., Da Silva, C. R., Meihac, O., and Viranaïcken, W. (2019). Biology of extracellular HSP60. In *Chaperokine Activity of Heat Shock Proteins*, eds. A. A. A. Asea and P. Kaur (Cham: Springer International Publishing). 57–80
- Nguyen-Kim, H., San Clemente, H., Balliau, T., Zivy, M., Dunand, C., Albenne, C., et al. (2016). *Arabidopsis thaliana* root cell wall proteomics: Increasing the proteome cover-

- age using a combinatorial peptide ligand library and description of unexpected hyp in peroxidase amino acid sequences. *Proteomics* 16, 491–503
- Nickchi, P., Jafari, M., and Kalantari, S. (2015). PEIMAN 1.0: Post-translational modification enrichment, integration and matching ANalysis. *Database* 2015, bav037
- Nickel, W. and Rabouille, C. (2018). Unconventional protein secretion: Diversity and consensus. *Semin. Cell Dev. Biol.*
- Nielsen, H., Petsalaki, E. I., Zhao, L., and Stühler, K. (2018). Predicting eukaryotic protein secretion without signals. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*
- Nikolovski, N., Rubtsov, D., Segura, M. P., Miles, G. P., Stevens, T. J., Dunkley, T. P. J., et al. (2012). Putative glycosyltransferases and other plant golgi apparatus proteins are revealed by LOPIT proteomics. *Plant Physiol.* 160, 1037–1051
- Nogueira-Lopez, G., Greenwood, D. R., Middleditch, M., Winefield, C., Eaton, C., Steyaert, J. M., et al. (2018). The apoplastic secretome of trichoderma virens during interaction with maize roots shows an inhibition of plant defence and scavenging oxidative stress secreted proteins. *Front. Plant Sci.* 9, 409
- Nombela, C., Gil, C., and Chaffin, W. L. (2006). Non-conventional protein secretion in yeast. *Trends Microbiol.* 14, 15–21
- Novaković, L., Guo, T., Bacic, A., Sampathkumar, A., and Johnson, K. L. (2018). Hitting the Wall-Sensing and signaling pathways involved in plant cell wall remodeling in response to abiotic stress. *Plants* 7
- Obayashi, T., Aoki, Y., Tadaka, S., Kagaya, Y., and Kinoshita, K. (2018). ATTED-II in 2018: A plant coexpression database based on investigation of the statistical property of the mutual rank index. *Plant Cell Physiol.* 59, e3
- Oehme, D. P., Doblin, M. S., Wagner, J., Bacic, A., Downton, M. T., and Gidley, M. J. (2015). Gaining insight into cell wall cellulose microfibril organisation by simulating microfibril adsorption. *Cellulose* 22, 3501–3520

- Ofer, D. and Linial, M. (2015). ProFET: Feature engineering captures high-level protein functions. *Bioinformatics* 31, 3429–3436
- Oh, I. S. (2005). Secretome analysis reveals an arabidopsis lipase involved in defense against alternaria brassicicola. *THE PLANT CELL ONLINE* 17, 2832–2847
- Padmanabhan, S., Biswal, M. R., Manjithaya, R., and Prakash, M. K. (2018). Exploring the context of diacidic motif D-E as a signal for unconventional protein secretion in eukaryotic proteins
- Palm, D., Streit, D., Ruprecht, M., Simm, S., Scharf, C., and Schleiff, E. (2018). Late ribosomal protein localization in arabidopsis thaliana differs to that in saccharomyces cerevisiae. *FEBS Open Bio* 8, 1437–1444
- Palovaara, J., Saiga, S., Wendrich, J. R., van 't Wout Hofland, N., van Schayck, J. P., Hater, F., et al. (2017). Transcriptome dynamics revealed by a gene expression atlas of the early arabidopsis embryo. *Nat Plants* 3, 894–904
- Pandey, S., Fartyal, D., Agarwal, A., Shukla, T., James, D., Kaul, T., et al. (2017). Abiotic stress tolerance in plants: Myriad roles of ascorbate peroxidase. *Front. Plant Sci.* 8, 581
- Parsons, H. T., Christiansen, K., Knierim, B., Carroll, A., Ito, J., Batth, T. S., et al. (2012). Isolation and proteomic characterization of the arabidopsis golgi defines functional and novel components involved in plant cell wall biosynthesis. *Plant Physiol.* 159, 12–26
- Pauwels, L., Ritter, A., Goossens, J., Durand, A. N., Liu, H., Gu, Y., et al. (2015). The RING E3 ligase KEEP ON GOING modulates JASMONATE ZIM-DOMAIN12 stability. *Plant Physiol.* 169, 1405–1417
- Pechanova, O., Hsu, C.-Y., Adams, J. P., Pechan, T., Vandervelde, L., Drnevich, J., et al. (2010). Apoplast proteome reveals that extracellular matrix contributes to multistress response in poplar. *BMC Genomics* 11, 674
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830

-
- Pierleoni, A., Martelli, P. L., and Casadio, R. (2008). PredGPI: a GPI-anchor predictor. *BMC Bioinformatics* 9, 392
- Pinedo, M., Orts, F., Carvalho, A. d. O., Regente, M., Soares, J. R., Gomes, V. M., et al. (2015). Molecular characterization of helja, an extracellular jacalin-related protein from *helianthus annuus*: Insights into the relationship of this protein with unconventionally secreted lectins. *J. Plant Physiol.* 183, 144–153
- Pinedo, M., Regente, M., Elizalde, M., Quiroga, I., Pagnussat, L., Jorin-Novo, J., et al. (2012a). Extracellular sunflower proteins: Evidence on non-classical secretion of a Jacalin-Related lectin. *Protein Pept. Lett.* 19, 270–276
- Pinedo, M., Regente, M., Elizalde, M., Quiroga, I., Pagnussat, L., Jorin-Novo, J., et al. (2012b). Extracellular Sunflower Proteins: Evidence on Non-classical Secretion of a Jacalin-Related Lectin. *Protein Pept. Lett.* 19, 270–276
- Podgórska, A., Burian, M., and Szal, B. (2017). Extra-Cellular but Extra-Ordinarily important for cells: Apoplastic reactive oxygen species metabolism. *Front. Plant Sci.* 8, 1353
- Popa, S. J., Stewart, S. E., and Moreau, K. (2018). Unconventional secretion of annexins and galectins. *Semin. Cell Dev. Biol.*
- Poulsen, C. P., Dilokpimol, A., and Geshi, N. (2015). Arabinogalactan biosynthesis: Implication of AtGALT29A enzyme activity regulated by phosphorylation and co-localized enzymes for nucleotide sugar metabolism in the compartments outside of the golgi apparatus. *Plant Signal. Behav.* 10, e984524
- Poulsen, C. P., Dilokpimol, A., Mouille, G., Burow, M., and Geshi, N. (2014). Arabino-galactan glycosyltransferases target to a unique subcellular compartment that may function in unconventional secretion in plants. *Traffic* 15, 1219–1234
- Priyam, A., Woodcroft, B. J., Rai, V., Moghul, I., Mungala, A., Ter, F., et al. (2019). Sequenceserver: a modern graphical user interface for custom BLAST databases. *Mol. Biol. Evol.*

- Prudovsky, I., Kumar, T. K. S., Sterling, S., and Neivandt, D. (2013). Protein-phospholipid interactions in nonclassical protein secretion: problem and methods of study. *Int. J. Mol. Sci.* 14, 3734–3772
- Rabouille, C. (2016). Pathways of unconventional protein secretion. *Trends Cell Biol.*
- Rautengarten, C., Ebert, B., Herter, T., Petzold, C. J., Ishii, T., Mukhopadhyay, A., et al. (2011). The interconversion of UDP-arabinopyranose and UDP-arabinofuranose is indispensable for plant development in arabidopsis. *Plant Cell* 23, 1373–1390
- Reddy, V. S., Madala, S. K., Trinath, J., and Reddy, G. B. (2018). Extracellular small heat shock proteins: exosomal biogenesis and function. *Cell Stress Chaperones* 23, 441–454
- Restrepo-Montoya, D., Pino, C., Nino, L. F., Patarroyo, M. E., and Patarroyo, M. A. (2011). NClassG+: A classifier for non-classically secreted gram-positive bacterial proteins. *BMC Bioinformatics* 12
- Ribeiro, D. M., Briere, G., Bely, B., Spinelli, L., and Brun, C. (2019). MoonDB 2.0: an updated database of extreme multifunctional and moonlighting proteins. *Nucleic Acids Res.* 47, D398–D402
- Robertson, D., Mitchell, G. P., Gilroy, J. S., Gerrish, C., Bolwell, G. P., and Slabas, A. R. (1997). Differential extraction and protein sequencing reveals major differences in patterns of primary cell wall proteins from plants. *J. Biol. Chem.* 272, 15841–15848
- Robinson, D. G., Ding, Y., and Jiang, L. (2015). Unconventional protein secretion in plants: a critical assessment. *Protoplasma*
- Rossi, F. R., Marina, M., and Pieckenstain, F. L. (2015). Role of arginine decarboxylase (ADC) in arabidopsis thaliana defence against the pathogenic bacterium pseudomonas viridiflava. *Plant Biol.* 17, 831–839
- Rutter, B. D. and Innes, R. W. (2017). Extracellular vesicles isolated from the leaf apoplast carry Stress-Response proteins. *Plant Physiol.* 173, 728–741
- Sakamoto, K. and Briggs, W. R. (2002). Cellular and subcellular localization of phototropin 1. *Plant Cell* 14, 1723–1735

-
- San-Clemente, H. and Jamet, E. (2015). WallProtDB, a database resource for plant cell wall proteomics. *Plant Methods* 11, 2
- Scarafoni, A., Ronchi, A., Prinsi, B., Espen, L., Assante, G., Venturini, G., et al. (2013). The proteome of exudates from germinating *Lupinus albus* seeds is secreted through a selective dual-step process and contains proteins involved in plant defence. *FEBS J.* 280, 1443–1459
- Schultz, C. J. (2004). Post-translational modifications of arabinogalactan-peptides of *Arabidopsis thaliana*: ENDOPLASMIC RETICULUM AND GLYCOSYLPHOSPHATIDYLINOSITOL-ANCHOR SIGNAL CLEAVAGE SITES AND HYDROXYLATION OF PROLINE. *J. Biol. Chem.* 279, 45503–45511
- Seguel, A., Jelenska, J., Herrera-Vásquez, A., Marr, S. K., Joyce, M. B., Gagesch, K. R., et al. (2018). PROHIBITIN3 forms complexes with ISOCHORISMATE SYNTHASE1 to regulate Stress-Induced salicylic acid biosynthesis in *Arabidopsis*. *Plant Physiol.* 176, 2515–2531
- Sehrawat, A. and Deswal, R. (2014). S-Nitrosylation Analysis in *Brassica juncea* Apoplast Highlights the Importance of Nitric Oxide in Cold-Stress Signaling. *J. Proteome Res.* 13, 2599–2619
- Sevilla, F., Camejo, D., Ortiz-Espín, A., Calderón, A., Lázaro, J. J., and Jiménez, A. (2015). The thioredoxin/peroxiredoxin/sulfiredoxin system: current overview on its redox function in plants and regulation by reactive oxygen and nitrogen species. *J. Exp. Bot.* 66, 2945–2955
- Shahri, M. P., Srinivasan, M., Reynolds, G., Bimczok, D., Kahanda, I., and Kanewala, U. (2019). Metamorphic testing for quality assurance of protein function prediction tools
- Shinano, T., Komatsu, S., Yoshimura, T., Tokutake, S., Kong, F.-J., Watanabe, T., et al. (2011). Proteomic analysis of secreted proteins from aseptically grown rice. *Phytochemistry* 72, 312–320

- Song, Y., Zhang, C., Ge, W., Zhang, Y., Burlingame, A. L., and Guo, Y. (2011). Identification of NaCl stress-responsive apoplastic proteins in rice shoot stems by 2D-DIGE. *J. Proteomics* 74, 1045–1067
- Sperschneider, J., Dodds, P. N., Singh, K. B., and Taylor, J. M. (2018). ApoplastP: prediction of effectors and plant proteins in the apoplast using machine learning. *New Phytol.* 217, 1764–1778
- Srivastava, A. C., Palanichelvam, K., Ma, J., Steele, J., Blancaflor, E. B., and Tang, Y. (2010). Collection and analysis of expressed sequence tags derived from laser capture microdissected switchgrass (*panicum virgatum* l. alamo) vascular tissues. *Bioenergy Res.* 3, 278–294
- Stock, J., Sarkari, P., Kreibich, S., Brefort, T., Feldbrügge, M., and Schipper, K. (2012). Applying unconventional secretion of the endochitinase cts1 to export heterologous proteins in *ustilago maydis*. *J. Biotechnol.* 161, 80–91
- Tanz, S. K., Castleden, I., Small, I. D., and Millar, A. H. (2013). Fluorescent protein tagging as a tool to define the subcellular distribution of proteins in plants. *Front. Plant Sci.* 4, 214
- Tran, H. T. and Plaxton, W. C. (2008). Proteomic analysis of alterations in the secretome of *arabidopsis thaliana* suspension cells subjected to nutritional phosphate deficiency. *Proteomics* 8, 4317–4326
- Trentin, A. R., Pivato, M., Mehdi, S. M. M., Barnabas, L. E., Giarretta, S., Fabrega-Prats, M., et al. (2015). Proteome readjustments in the apoplastic space of *arabidopsis thaliana* ggt1 mutant leaves exposed to UV-B radiation. *Front. Plant Sci.* 6, 128
- Tsai, A. Y.-L., Kunieda, T., Rogalski, J., Foster, L. J., Ellis, B. E., and Haughn, G. W. (2017). Identification and characterization of *arabidopsis* seed coat mucilage proteins. *Plant Physiol.* 173, 1059–1074
- Turek, I., Marondedze, C., Wheeler, J. I., Gehring, C., and Irving, H. R. (2014). Plant na-

- triuretic peptides induce proteins diagnostic for an adaptive response to stress. *Front. Plant Sci.* 5, 661
- Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Proteomics. tissue-based map of the human proteome. *Science* 347, 1260419
- Ushijima, K., Nakano, R., Bando, M., Shigezane, Y., Ikeda, K., Namba, Y., et al. (2011). Isolation of the floral morph-related genes in heterostylous flax (*Linum grandiflorum*): the genetic polymorphism and the transcriptional and post-transcriptional regulations of the S locus. *Plant J.* 69, 317–331
- van de Meene, A. M. L., Doblin, M. S., and Bacic, A. (2016). The plant secretory pathway seen through the lens of the cell wall. *Protoplasma*
- Waghmare, S., Lileikyte, E., Karnik, R., Goodman, J. K., Blatt, M. R., and Jones, A. M. E. (2018). SNAREs SYP121 and SYP122 mediate the secretion of distinct cargo subsets. *Plant Physiol.* 178, 1679–1688
- Wang, J., Ding, Y., Wang, J., Hillmer, S., Miao, Y., Lo, S. W., et al. (2010). EXPO, an exocyst-positive organelle distinct from multivesicular endosomes and autophagosomes, mediates cytosol to cell wall exocytosis in arabidopsis and tobacco cells. *The Plant Cell Online* 22, 4009–4030
- Wang, M., Weiss, M., Simonovic, M., Haertinger, G., Schrimpf, S. P., Hengartner, M. O., et al. (2012a). PaxDb, a database of protein abundance averages across all three domains of life. *Mol. Cell. Proteomics* 11, 492–500
- Wang, Y., Wu, J., Lee, D. Y., Kim, Y. J., Kim, Y. C., Park, S.-R., et al. (2012b). Comparative secretome analysis of differentially-induced proteins in rice lesion mimic mutant spotted leaf 11 (spl11) 5, 567
- Wegehingel, S., Zehe, C., and Nickel, W. (2008). Rerouting of fibroblast growth factor 2 to the classical secretory pathway results in post-translational modifications that block binding to heparan sulfate proteoglycans. *FEBS Lett.* 582, 2387–2392

- Xie, X., Ho, J. W. K., Murphy, C., Kaiser, G., Xu, B., and Chen, T. Y. (2011). Testing and validating machine learning classifiers by metamorphic testing. *J. Syst. Softw.* 84, 544–558
- Xu, S.-L., Medzihradsky, K. F., Wang, Z.-Y., Burlingame, A. L., and Chalkley, R. J. (2016). N-Glycopeptide profiling in arabidopsis inflorescence. *Mol. Cell. Proteomics* 15, 2048–2054
- Yu, L., Guo, Y., Li, Y., Li, G., Li, M., Luo, J., et al. (2010a). SecretP: Identifying bacterial secreted proteins by fusing new features into Chous pseudo-amino acid composition. *J. Theor. Biol.* 267, 1–6
- Yu, L., Guo, Y., Zhang, Z., Li, Y., Li, M., Li, G., et al. (2010b). SecretP: A new method for predicting mammalian secreted proteins. *Peptides* 31, 574–578
- Yu, L., Luo, J., Guo, Y., Li, Y., Pu, X., and Li, M. (2013). In silico identification of gram-negative bacterial secreted proteins from primary sequence. *Comput. Biol. Med.* 43, 1177–1181
- Yu, T.-Y., Shi, D.-Q., Jia, P.-F., Tang, J., Li, H.-J., Liu, J., et al. (2016). The arabidopsis receptor kinase ZAR1 is required for zygote asymmetric division and its daughter cell fate. *PLoS Genet.* 12, e1005933
- Zhang, H. and Li, J. (2016). From cytosol to the apoplast: The hygromycin phosphotransferase (HYGR) model in arabidopsis. In *Unconventional Protein Secretion: Methods and Protocols*, eds. A. Pompa and F. De Marchis (New York, NY: Springer New York). 81–90
- Zhang, H., Zhang, L., Gao, B., Fan, H., Jin, J., Botella, M. A., et al. (2011a). Golgi apparatus-localized synaptotagmin 2 is required for unconventional secretion in arabidopsis. *PLoS One* 6, e26477
- Zhang, Y., Giboulot, A., Zivy, M., Valot, B., Jamet, E., and Albenne, C. (2011b). Combining various strategies to increase the coverage of the plant cell wall glycoproteome. *Phytochemistry* 72, 1109–1123

- Zhang, Y., Yu, S., Xie, R., Li, J., Leier, A., Marquez-Lago, T. T., et al. (2019). PeNGaRoo, a combined gradient boosting and ensemble learning framework for predicting non-classical secreted proteins. *Bioinformatics*
- Zhang, Z., Xin, W., Wang, S., Zhang, X., Dai, H., Sun, R., et al. (2014). Xylem sap in cotton contains proteins that contribute to environmental stress response and cell wall development. *Funct. Integr. Genomics* 15, 17–26
- Zhou, L., Bokhari, S. A., Dong, C.-J., and Liu, J.-Y. (2011). Comparative Proteomics Analysis of the Root Apoplasts of Rice Seedlings in Response to Hydrogen Peroxide. *PLoS One* 6, e16723
- Zhou, X., Liao, W.-J., Liao, J.-M., Liao, P., and Lu, H. (2015). Ribosomal proteins: functions beyond the ribosome. *J. Mol. Cell Biol.* 7, 92–104
- Zhu, J., Chen, S., Alvarez, S., Asirvatham, V. S., Schachtman, D. P., Wu, Y., et al. (2006). Cell wall proteome in the maize primary root elongation zone. i. extraction and identification of water-soluble and lightly ionically bound proteins. *Plant Physiol.* 140, 311–325

Appendices

Appendix A

Appendix

Table A.1: Updated citation counts for Table 2.1 for November 2019

Tool	Publication	2016 Citations	2019 Citations
SecretomeP (v1)	Bendtsen et al. (2004)	642	975
SecretomeP (v2)	Bendtsen et al. (2005)	342	541
SecretP (v1)	Yu et al. (2010a)	17	25
SecretP(v2)	Yu et al. (2010b)	76	101
SecretP (v2.1)	Yu et al. (2013)	3	11
SPRED	Kandaswamy et al. (2010)	15	26
SRTpred	Garg and Raghava (2008)	35	48
Sec-GO	Huang (2012)	7	13
NClassG+	Restrepo-Montoya et al. (2011)	6	12

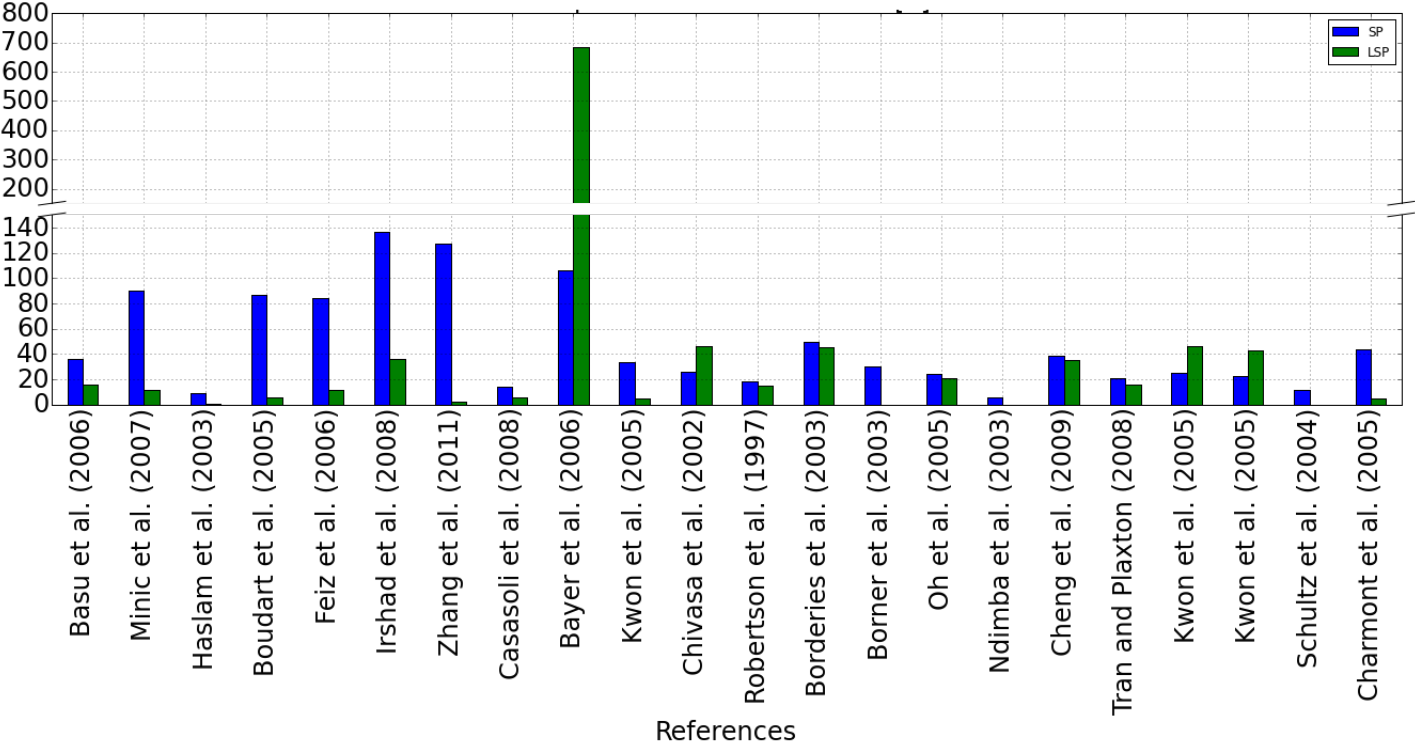


Figure A.1: Summary of the number of SP containing and LSP proteins in selected plant apoplastic proteomics studies, adapted from data by (Albenne et al., 2013)

Table A.2: List of putative LSPs from Ding et al. (2012) with duplicate entries removed

gene	protein
AT1G02930	Glutathione S-transferase F6
AT1G07890	MEE6
AT1G08110	Lactoylglutathione lyase
AT1G08830	Superoxide dismutase
AT1G11580	methylesterase PCR A
AT1G12080	Vacuolar calcium-binding protein-like protein
AT1G13440	Glyceraldehyde-3-phosphate dehydrogenase GAPC2, cytosolic
AT1G19570	Glutathione S-transferase DHAR1, mitochondrial
AT1G23190	Probable phosphoglucomutase, cytoplasmic 1
AT1G23410	Ubiquitin-40S ribosomal protein S27a-1
AT1G25490	Serine/threonine-protein phosphatase 2A 65 kDa regulatory subunit A alpha isoform
AT1G35720	Annexin
AT1G45145	Thioredoxin H5
AT1G56070	Elongation factor 2
AT1G61820	Beta-glucosidase 46
AT1G62380	1-aminocyclopropane-1-carboxylate oxidase 2
AT1G63770	Peptidase M1 family protein
AT1G78380	Glutathione S-transferase U19 (GSTU19)

gene	protein
AT1G79550	Phosphoglycerate kinase
AT2G15130	Plant basic secretory protein (BSP) family protein
AT2G16060	Non-symbiotic hemoglobin 1
AT2G24200	Leucine aminopeptidase 1
AT2G32520	alpha/beta-Hydrolases superfamily protein
AT2G35635	UBQ7
AT2G36460	Fructose-bisphosphate aldolase 6, cytosolic
AT2G36530	LOS2
AT3G01280	Mitochondrial outer membrane protein porin 1
AT3G02230	RGP1
AT3G03780	5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase 2
AT3G03910	Glutamate dehydrogenase
AT3G04120	Glyceraldehyde-3-phosphate dehydrogenase GAPC1, cytosolic
AT3G06860	Peroxisomal fatty acid beta-oxidation multifunctional protein MFP2
AT3G08580	ADP/ATP carrier 1 (AAC1)
AT3G08900	Glycosyltransferase (Fragment)
AT3G10920	Superoxide dismutase
AT3G16410	Nitrile-specifier protein 4
AT3G16420	PYK10-binding protein 1

gene	protein
AT3G16430	PYK10-binding protein 2
AT3G16450	Jacalin-related lectin 33
AT3G28940	Protein AIG2 B
AT3G32980	Peroxidase 32
AT3G43810	Calmodulin 7
AT3G52930	Fructose-bisphosphate aldolase
AT3G54660	glutathione reductase
AT3G55440	Triosephosphate isomerase, cytosolic
AT4G01850	S-adenosylmethionine synthase
AT4G09320	Nucleoside diphosphate kinase
AT4G20260	plasma-membrane associated cation-binding protein 1
AT4G23590	Probable aminotransferase TAT4
AT4G25100	Superoxide dismutase
AT4G29680	Alkaline-phosphatase-like family protein
AT4G30910	Cytosol aminopeptidase family protein
AT4G30920	Leucine aminopeptidase 2, chloroplastic
AT5G07440	Glutamate dehydrogenase
AT5G10450	G-box regulating factor 6
AT5G15490	UDP-glucose 6-dehydrogenase

gene	protein
AT5G15650	reversibly glycosylated polypeptide 2 (RGP2)
AT5G16510	Glycosyltransferase (Fragment)
AT5G17920	5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase 1
AT5G18100	Superoxide dismutase
AT5G18170	Glutamate dehydrogenase
AT5G19440	NAD(P)-binding Rossmann-fold superfamily protein
AT5G53560	CB5-E
AT5G55730	Fasciclin-like arabinogalactan protein 1
AT5G59880	Actin-depolymerizing factor 3

Appendix B

Appendix

B.1 Secretory Isoform Observations

Table B.1: Summary of combinations of LSPDB proteins with secretory isoform features, stratified by observation

Feature Combination	Observation	Count
SP=Yes,TM=Yes,GPI=Yes,Retention=No	unobs	31
SP=Yes,TM=Yes,GPI=Yes,Retention=No	obs	3
SP=Yes,TM=Yes,GPI=No,Retention=Yes	unobs	52
SP=Yes,TM=Yes,GPI=No,Retention=Yes	obs	7
SP=Yes,TM=Yes,GPI=No,Retention=No	unobs	933
SP=Yes,TM=Yes,GPI=No,Retention=No	obs	43
SP=Yes,TM=Yes,GPI=No,Retention=Iso	unobs	9
SP=Yes,TM=Yes,GPI=Iso,Retention=No	unobs	11
SP=Yes,TM=Yes,GPI=Iso,Retention=Iso	unobs	1
SP=Yes,TM=No,GPI=Yes,Retention=Yes	unobs	1
SP=Yes,TM=No,GPI=Yes,Retention=No	unobs	240
SP=Yes,TM=No,GPI=Yes,Retention=No	obs	82
SP=Yes,TM=No,GPI=No,Retention=Yes	unobs	47
SP=Yes,TM=No,GPI=No,Retention=Yes	obs	15
SP=Yes,TM=No,GPI=No,Retention=No	unobs	2478
SP=Yes,TM=No,GPI=No,Retention=No	obs	543

Feature Combination	Observation	Count
SP=Yes,TM=No,GPI=No,Retention=Iso	unobs	20
SP=Yes,TM=No,GPI=No,Retention=Iso	obs	7
SP=Yes,TM=No,GPI=Iso,Retention=No	unobs	37
SP=Yes,TM=No,GPI=Iso,Retention=No	obs	13
SP=Yes,TM=No,GPI=Iso,Retention=Iso	obs	2
SP=Yes,TM=No,GPI=Iso,Retention=Iso	unobs	1
SP=Yes,TM=Iso,GPI=Yes,Retention=No	unobs	2
SP=Yes,TM=Iso,GPI=No,Retention=No	unobs	31
SP=Yes,TM=Iso,GPI=No,Retention=No	obs	6
SP=Yes,TM=Iso,GPI=No,Retention=Iso	unobs	6
SP=Yes,TM=Iso,GPI=No,Retention=Iso	obs	1
SP=Yes,TM=Iso,GPI=Iso,Retention=No	unobs	7
SP=No,TM=Yes,GPI=Yes,Retention=Yes	unobs	2
SP=No,TM=Yes,GPI=Yes,Retention=No	unobs	51
SP=No,TM=Yes,GPI=Yes,Retention=No	obs	1
SP=No,TM=Yes,GPI=No,Retention=Yes	unobs	116
SP=No,TM=Yes,GPI=No,Retention=Yes	obs	4
SP=No,TM=Yes,GPI=No,Retention=No	unobs	2824
SP=No,TM=Yes,GPI=No,Retention=No	obs	94
SP=No,TM=Yes,GPI=No,Retention=Iso	unobs	23
SP=No,TM=Yes,GPI=No,Retention=Iso	obs	1
SP=No,TM=Yes,GPI=Iso,Retention=No	unobs	27
SP=No,TM=No,GPI=Yes,Retention=Yes	unobs	1
SP=No,TM=No,GPI=Yes,Retention=No	unobs	97
SP=No,TM=No,GPI=Yes,Retention=No	obs	10
SP=No,TM=No,GPI=No,Retention=Yes	unobs	289
SP=No,TM=No,GPI=No,Retention=Yes	obs	26
SP=No,TM=No,GPI=No,Retention=No	unobs	16473
SP=No,TM=No,GPI=No,Retention=No	obs	1664

Feature Combination	Observation	Count
SP=No,TM=No,GPI=No,Retention=Iso	unobs	88
SP=No,TM=No,GPI=No,Retention=Iso	obs	6
SP=No,TM=No,GPI=Iso,Retention=No	unobs	72
SP=No,TM=No,GPI=Iso,Retention=No	obs	10
SP=No,TM=Iso,GPI=No,Retention=Yes	unobs	7
SP=No,TM=Iso,GPI=No,Retention=No	unobs	303
SP=No,TM=Iso,GPI=No,Retention=No	obs	22
SP=No,TM=Iso,GPI=No,Retention=Iso	unobs	8
SP=No,TM=Iso,GPI=No,Retention=Iso	obs	1
SP=No,TM=Iso,GPI=Iso,Retention=No	unobs	11
SP=Iso,TM=Yes,GPI=No,Retention=Yes	unobs	6
SP=Iso,TM=Yes,GPI=No,Retention=No	unobs	201
SP=Iso,TM=Yes,GPI=No,Retention=No	obs	11
SP=Iso,TM=Yes,GPI=No,Retention=Iso	unobs	4
SP=Iso,TM=Yes,GPI=Iso,Retention=No	unobs	13
SP=Iso,TM=No,GPI=Yes,Retention=No	unobs	10
SP=Iso,TM=No,GPI=Yes,Retention=No	obs	5
SP=Iso,TM=No,GPI=No,Retention=Yes	unobs	5
SP=Iso,TM=No,GPI=No,Retention=Yes	obs	3
SP=Iso,TM=No,GPI=No,Retention=No	unobs	311
SP=Iso,TM=No,GPI=No,Retention=No	obs	79
SP=Iso,TM=No,GPI=No,Retention=Iso	unobs	2
SP=Iso,TM=No,GPI=Iso,Retention=No	unobs	12
SP=Iso,TM=No,GPI=Iso,Retention=No	obs	9
SP=Iso,TM=Iso,GPI=Yes,Retention=No	unobs	5
SP=Iso,TM=Iso,GPI=No,Retention=Yes	unobs	2
SP=Iso,TM=Iso,GPI=No,Retention=No	unobs	71
SP=Iso,TM=Iso,GPI=No,Retention=No	obs	8
SP=Iso,TM=Iso,GPI=No,Retention=Iso	unobs	2

Feature Combination	Observation	Count
SP=Iso,TM=Iso,GPI=Iso,Retention=No	unobs	7
SP=Iso,TM=Iso,GPI=Iso,Retention=No	obs	1

B.2 LSPDB low confidence list

Table B.2: Putative LSPs from the low (red) category

Accession	Description
AT1G01800	NAD(P)-binding Rossmann-fold superfamily protein
AT1G02500	S-adenosylmethionine synthetase 1
AT1G02690	importin alpha isoform 6
AT1G02920	glutathione S-transferase 7
AT1G02930	glutathione S-transferase 6
AT1G04170	eukaryotic translation initiation factor 2 gamma subunit
AT1G04410	Lactate/malate dehydrogenase family protein
AT1G04510	MOS4-associated complex 3A
AT1G06220	Ribosomal protein S5/Elongation factor G/III/V family protein
AT1G07370	proliferating cellular nuclear antigen 1
AT1G07750	RmlC-like cupins superfamily protein
AT1G08110	lactoylglutathione lyase family protein / glyoxalase I family protein
AT1G08410	P-loop containing nucleoside triphosphate hydrolases superfamily protein
AT1G08880	Histone superfamily protein
AT1G09270	importin alpha isoform 4
AT1G09310	Protein of unknown function, DUF538
AT1G09640	Translation elongation factor EF1B, gamma chain
AT1G09760	U2 small nuclear ribonucleoprotein A

Accession	Description
AT1G10390	Nucleoporin autopeptidase
AT1G12000	Phosphofructokinase family protein
AT1G13440	glyceraldehyde-3-phosphate dehydrogenase C2
AT1G14170	RNA-binding KH domain-containing protein
AT1G14980	chaperonin 10
AT1G15390	peptide deformylase 1A
AT1G16030	heat shock protein 70B
AT1G16340	Aldolase superfamily protein
AT1G16470	proteasome subunit PAB1
AT1G18070	Translation elongation factor EF1A/initiation factor IF2gamma family protein
AT1G19120	Small nuclear ribonucleoprotein family protein
AT1G19730	Thioredoxin superfamily protein
AT1G19920	Pseudouridine synthase/archaeosine transglycosylase-like family protein
AT1G20010	tubulin beta-5 chain
AT1G20220	Alba DNA/RNA-binding protein
AT1G20260	ATPase, V1 complex, subunit B protein
AT1G20630	catalase 1
AT1G20693	high mobility group B2
AT1G20960	U5 small nuclear ribonucleoprotein helicase, putative

Accession	Description
AT1G21630	Calcium-binding EF hand family protein
AT1G22780	Ribosomal protein S13/S18 family
AT1G22920	COP9 signalosome 5A
AT1G23190	Phosphoglucomutase/phosphomannomutase family protein
AT1G24360	NAD(P)-binding Rossmann-fold superfamily protein
AT1G25490	ARM repeat superfamily protein
AT1G26630	Eukaryotic translation initiation factor 5A-1 (eIF-5A 1) protein
AT1G26880	Ribosomal protein L34e superfamily protein
AT1G27450	adenine phosphoribosyl transferase 1
AT1G27970	nuclear transport factor 2B
AT1G29150	non-ATPase subunit 9
AT1G29990	prefoldin 6
AT1G30580	GTP binding
AT1G34030	Ribosomal protein S13/S18 family
AT1G35720	annexin 1
AT1G43170	ribosomal protein 1
AT1G43190	polypyrimidine tract-binding protein 3
AT1G45000	AAA-type ATPase family protein
AT1G48920	nucleolin like 1

Accession	Description
AT1G49240	actin 8
AT1G49760	poly(A) binding protein 8
AT1G50200	Alanyl-tRNA synthetase
AT1G50480	10-formyltetrahydrofolate synthetase
AT1G50920	Nucleolar GTP-binding protein
AT1G51060	histone H2A 10
AT1G51710	ubiquitin-specific protease 6
AT1G52500	MUTM homolog-1
AT1G53240	Lactate/malate dehydrogenase family protein
AT1G53280	Class I glutamine amidotransferase-like superfamily protein
AT1G54270	eif4a-2
AT1G54290	Translation initiation factor SUI1 family protein
AT1G56110	homolog of nucleolar protein NOP56
AT1G56580	Protein of unknown function, DUF538
AT1G57720	Translation elongation factor EF1B, gamma chain
AT1G58684	Ribosomal protein S5 family protein
AT1G59900	pyruvate dehydrogenase complex E1 alpha subunit
AT1G60070	Adaptor protein complex AP-1, gamma subunit
AT1G60740	Thioredoxin superfamily protein

Accession	Description
AT1G61580	R-protein L3 B
AT1G62380	ACC oxidase 2
AT1G62800	aspartate aminotransferase 4
AT1G63460	glutathione peroxidase 8
AT1G64230	ubiquitin-conjugating enzyme 28
AT1G65970	thioredoxin-dependent peroxidase 2
AT1G65980	thioredoxin-dependent peroxidase 1
AT1G66410	calmodulin 4
AT1G69740	Aldolase superfamily protein
AT1G70310	spermidine synthase 2
AT1G70410	beta carbonic anhydrase 4
AT1G72560	ARM repeat superfamily protein
AT1G73230	Nascent polypeptide-associated complex NAC
AT1G74310	heat shock protein 101
AT1G74560	NAP1-related protein 1
AT1G74920	aldehyde dehydrogenase 10A8
AT1G75510	Transcription initiation factor IIF, beta subunit
AT1G75950	S phase kinase-associated protein 1
AT1G76860	Small nuclear ribonucleoprotein family protein

Accession	Description
AT1G77180	chromatin protein family
AT1G78300	general regulatory factor 2
AT1G78380	glutathione S-transferase TAU 19
AT1G79500	Aldolase-type TIM barrel family protein
AT1G79530	glyceraldehyde-3-phosphate dehydrogenase of plastid 1
AT1G79930	heat shock protein 91
AT1G80050	adenine phosphoribosyl transferase 2
AT1G80500	SNARE-like superfamily protein
AT2G01250	Ribosomal protein L30/L7 family protein
AT2G01290	ribose-5-phosphate isomerase 2
AT2G02560	cullin-associated and neddylation dissociated
AT2G03430	Ankyrin repeat family protein
AT2G04390	Ribosomal S17 family protein
AT2G04750	Actin binding Calponin homology (CH) domain-containing protein
AT2G05710	aconitase 3
AT2G07690	Minichromosome maintenance (MCM2/3/5) family protein
AT2G14120	dynamamin related protein
AT2G16060	hemoglobin 1
AT2G16570	GLN phosphoribosyl pyrophosphate amidotransferase 1

Accession	Description
AT2G17265	homoserine kinase
AT2G17510	ribonuclease II family protein
AT2G19730	Ribosomal L28e protein family
AT2G19750	Ribosomal protein S30 family protein
AT2G19760	profilin 1
AT2G20580	26S proteasome regulatory subunit S2 1A
AT2G20890	photosystem II reaction center PSB29 protein
AT2G21170	triosephosphate isomerase
AT2G21250	NAD(P)-linked oxidoreductase superfamily protein
AT2G21540	SEC14-like 3
AT2G21660	cold, circadian rhythm, and rna binding 2
AT2G22230	Thioesterase superfamily protein
AT2G24200	Cytosol aminopeptidase family protein
AT2G25080	glutathione peroxidase 1
AT2G27020	20S proteasome alpha subunit G1
AT2G27680	NAD(P)-linked oxidoreductase superfamily protein
AT2G29550	tubulin beta-7 chain
AT2G29570	proliferating cell nuclear antigen 2
AT2G30050	transducin family protein / WD-40 repeat family protein

Accession	Description
AT2G30410	tubulin folding cofactor A (KIESEL)
AT2G30860	glutathione S-transferase PHI 9
AT2G30870	glutathione S-transferase PHI 10
AT2G30880	Pleckstrin homology (PH) domain-containing protein
AT2G33040	gamma subunit of Mt ATP synthase
AT2G33210	heat shock protein 60-2
AT2G33340	MOS4-associated complex 3B
AT2G33800	Ribosomal protein S5 family protein
AT2G36160	Ribosomal protein S11 family protein
AT2G36530	Enolase
AT2G37600	Ribosomal protein L36e family protein
AT2G37620	actin 1
AT2G38870	Serine protease inhibitor, potato inhibitor I-type family protein
AT2G39460	ribosomal protein L23AA
AT2G39730	rubisco activase
AT2G40290	Eukaryotic translation initiation factor 2 subunit 1
AT2G40490	Uroporphyrinogen decarboxylase
AT2G40660	Nucleic acid-binding, OB-fold-like protein
AT2G40830	RING-H2 finger C1A

Accession	Description
AT2G41090	Calcium-binding EF-hand family protein
AT2G41530	S-formylglutathione hydrolase
AT2G42590	general regulatory factor 9
AT2G43030	Ribosomal protein L3 family protein
AT2G43460	Ribosomal L38e protein family
AT2G43810	Small nuclear ribonucleoprotein family protein
AT2G44610	Ras-related small GTP-binding family protein
AT2G45300	RNA 3'-terminal phosphate cyclase/enolpyruvate transferase, alpha/beta
AT2G46280	TGF-beta receptor interacting protein 1
AT2G47220	DOMAIN OF UNKNOWN FUNCTION 724 5
AT2G47580	spliceosomal protein U1A
AT2G47730	glutathione S-transferase phi 8
AT3G01390	vacuolar membrane ATPase 10
AT3G01420	Peroxidase superfamily protein
AT3G01500	carbonic anhydrase 1
AT3G02080	Ribosomal protein S19e family protein
AT3G02090	Insulinase (Peptidase family M16) protein
AT3G02230	reversibly glycosylated polypeptide 1
AT3G02870	Inositol monophosphatase family protein

Accession	Description
AT3G03250	UDP-GLUCOSE PYROPHOSPHORYLASE 1
AT3G03920	H/ ACA ribonucleoprotein complex, subunit Gar1/Naf1 protein
AT3G04120	glyceraldehyde-3-phosphate dehydrogenase C subunit 1
AT3G04790	Ribose 5-phosphate isomerase, type A protein
AT3G05060	NOP56-like pre RNA processing ribonucleoprotein
AT3G06050	peroxiredoxin IIF
AT3G06130	Heavy metal transport/detoxification superfamily protein
AT3G06720	importin alpha isoform 1
AT3G06860	multifunctional protein 2
AT3G07880	Immunoglobulin E-set superfamily protein
AT3G08530	Clathrin, heavy chain
AT3G09200	Ribosomal protein L10 family protein
AT3G09790	ubiquitin 8
AT3G09820	adenosine kinase 1
AT3G09840	cell division cycle 48
AT3G10090	Nucleic acid-binding, OB-fold-like protein
AT3G10920	manganese superoxide dismutase 1
AT3G11130	Clathrin, heavy chain
AT3G11400	eukaryotic translation initiation factor 3G1

Accession	Description
AT3G11910	ubiquitin-specific protease 13
AT3G11940	ribosomal protein 5A
AT3G12390	Nascent polypeptide-associated complex (NAC), alpha subunit family protein
AT3G12780	phosphoglycerate kinase 1
AT3G12860	NOP56-like pre RNA processing ribonucleoprotein
AT3G13110	serine acetyltransferase 2;2
AT3G13300	Transducin/WD40 repeat-like superfamily protein
AT3G13610	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein
AT3G14390	Pyridoxal-dependent decarboxylase family protein
AT3G15360	thioredoxin M-type 4
AT3G15650	alpha/beta-Hydrolases superfamily protein
AT3G16390	nitrile specifier protein 3
AT3G16410	nitrile specifier protein 4
AT3G16810	pumilio 24
AT3G16857	response regulator 1
AT3G17390	S-adenosylmethionine synthetase family protein
AT3G17590	transcription regulatory protein SNF5, putative (BSH)
AT3G18030	HAL3-like protein A
AT3G19450	GroES-like zinc-binding alcohol dehydrogenase family protein

Accession	Description
AT3G19760	eukaryotic initiation factor 4A-III
AT3G22110	20S proteasome alpha subunit C1
AT3G22320	Eukaryotic rpb5 RNA polymerase subunit family protein
AT3G22330	putative mitochondrial RNA helicase 2
AT3G22890	ATP sulfurylase 1
AT3G23990	heat shock protein 60
AT3G24170	glutathione-disulfide reductase
AT3G24830	Ribosomal protein L13 family protein
AT3G25070	RPM1 interacting protein 4
AT3G25230	rotamase FKBP 1
AT3G25520	ribosomal protein L5
AT3G26060	Thioredoxin superfamily protein
AT3G28710	ATPase, V0/A0 complex, subunit C/D
AT3G28730	high mobility group
AT3G29350	histidine-containing phosphotransmitter 2
AT3G42160	Pectin lyase-like superfamily protein
AT3G43810	calmodulin 7
AT3G45030	Ribosomal protein S10p/S20e family protein
AT3G45300	isovaleryl-CoA-dehydrogenase

Accession	Description
AT3G46030	Histone superfamily protein
AT3G46560	Tim10/DDP family zinc finger protein
AT3G48000	aldehyde dehydrogenase 2B4
AT3G48530	SNF1-related protein kinase regulatory subunit gamma 1
AT3G49680	branched-chain aminotransferase 3
AT3G49910	Translation protein SH3-like family protein
AT3G51260	20S proteasome alpha subunit PAD1
AT3G52380	chloroplast RNA-binding protein 33
AT3G52590	ubiquitin extension protein 1
AT3G52880	monodehydroascorbate reductase 1
AT3G52930	Aldolase superfamily protein
AT3G52960	Thioredoxin superfamily protein
AT3G52980	Zinc finger (CCCH-type) family protein / RNA recognition motif (RRM)-containing protein
AT3G53230	ATPase, AAA-type, CDC48 protein
AT3G53750	actin 3
AT3G53880	NAD(P)-linked oxidoreductase superfamily protein
AT3G53890	Ribosomal protein S21e
AT3G54470	uridine 5'-monophosphate synthase / UMP synthase (PYRE-F) (UMPS)
AT3G54660	glutathione reductase

Accession	Description
AT3G55440	triosephosphate isomerase
AT3G55605	Mitochondrial glycoprotein family protein
AT3G55620	Translation initiation factor IF6
AT3G55800	sedoheptulose-bisphosphatase
AT3G56090	ferritin 3
AT3G56490	HIS triad family protein 3
AT3G57150	homologue of NAP57
AT3G57560	N-acetyl-l-glutamate kinase
AT3G58660	Ribosomal protein L1p/L10e family
AT3G59760	O-acetylserine (thiol) lyase isoform C
AT3G60240	eukaryotic translation initiation factor 4G
AT3G60750	Transketolase
AT3G60770	Ribosomal protein S13/S15
AT3G62870	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein
AT3G63490	Ribosomal protein L1p/L10e family
AT4G00100	ribosomal protein S13A
AT4G01370	MAP kinase 4
AT4G01690	Flavin containing amine oxidoreductase family
AT4G01850	S-adenosylmethionine synthetase 2

Accession	Description
AT4G01900	GLNB1 homolog
AT4G02510	translocon at the outer envelope membrane of chloroplasts 159
AT4G02520	glutathione S-transferase PHI 2
AT4G04885	PCF11P-similar protein 4
AT4G05320	polyubiquitin 10
AT4G05420	damaged DNA binding protein 1A
AT4G08390	stromal ascorbate peroxidase
AT4G08900	arginase
AT4G09010	ascorbate peroxidase 4
AT4G10040	cytochrome c-2
AT4G11010	nucleoside diphosphate kinase 3
AT4G11600	glutathione peroxidase 6
AT4G13660	pinorexinol reductase 2
AT4G13940	S-adenosyl-L-homocysteine hydrolase
AT4G14570	acylaminoacyl-peptidase-related
AT4G14960	Tubulin/FtsZ family protein
AT4G15000	Ribosomal L27e protein family
AT4G15802	heat shock factor binding protein
AT4G16420	homolog of yeast ADA2 2B

Accession	Description
AT4G16830	Hyaluronan / mRNA binding family
AT4G17260	Lactate/malate dehydrogenase family protein
AT4G18480	P-loop containing nucleoside triphosphate hydrolases superfamily protein
AT4G19210	RNAse I inhibitor protein 2
AT4G19750	Glycosyl hydrolase family protein with chitinase insertion domain
AT4G20850	tripeptidyl peptidase ii
AT4G22690	cytochrome P450, family 706, subfamily A, polypeptide 1
AT4G23570	phosphatase-related
AT4G24130	Protein of unknown function, DUF538
AT4G24770	31-kDa RNA binding protein
AT4G24830	arginosuccinate synthase family
AT4G25100	Fe superoxide dismutase 1
AT4G25630	fibrillarin 2
AT4G26530	Aldolase superfamily protein
AT4G26840	small ubiquitin-like modifier 1
AT4G26970	aconitase 2
AT4G27090	Ribosomal protein L14
AT4G27270	Quinone reductase family protein
AT4G29010	Enoyl-CoA hydratase/isomerase family

Accession	Description
AT4G29310	Protein of unknown function (DUF1005)
AT4G29510	arginine methyltransferase 11
AT4G30330	Small nuclear ribonucleoprotein family protein
AT4G30690	Translation initiation factor 3 protein
AT4G30920	Cytosol aminopeptidase family protein
AT4G31580	serine/arginine-rich 22
AT4G32720	La protein 1
AT4G33090	aminopeptidase M1
AT4G33650	dynamin-related protein 3A
AT4G34050	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein
AT4G34200	D-3-phosphoglycerate dehydrogenase
AT4G34490	cyclase associated protein 1
AT4G34870	rotamase cyclophilin 5
AT4G35090	catalase 2
AT4G37870	phosphoenolpyruvate carboxykinase 1
AT4G39200	Ribosomal protein S25 family protein
AT4G39260	cold, circadian rhythm, and RNA binding 1
AT4G39330	cinnamyl alcohol dehydrogenase 9
AT4G39660	alanine:glyoxylate aminotransferase 2

Accession	Description
AT5G01610	Protein of unknown function, DUF538
AT5G02500	heat shock cognate protein 70-1
AT5G02530	RNA-binding (RRM/RBD/RNP motifs) family protein
AT5G03850	Nucleic acid-binding, OB-fold-like protein
AT5G04800	Ribosomal S17 family protein
AT5G06060	NAD(P)-binding Rossmann-fold superfamily protein
AT5G07350	TUDOR-SN protein 1
AT5G08300	Succinyl-CoA ligase, alpha subunit
AT5G08570	Pyruvate kinase family protein
AT5G08670	ATP synthase alpha/beta family protein
AT5G09620	Octicosapeptide/Phox/Bem1p family protein
AT5G09660	peroxisomal NAD-malate dehydrogenase 2
AT5G09810	actin 7
AT5G10810	enhancer of rudimentary protein, putative
AT5G10920	L-Aspartase-like family protein
AT5G11170	DEAD/DEAH box RNA helicase family protein
AT5G11200	DEAD/DEAH box RNA helicase family protein
AT5G11260	Basic-leucine zipper (bZIP) transcription factor family protein
AT5G11670	NADP-malic enzyme 2

Accession	Description
AT5G11900	Translation initiation factor SUI1 family protein
AT5G13120	cyclophilin 20-2
AT5G13850	nascent polypeptide-associated complex subunit alpha-like protein 3
AT5G14040	phosphate transporter 3;1
AT5G14250	Proteasome component (PCI) domain protein
AT5G14800	pyrroline-5- carboxylate (P5C) reductase
AT5G15450	casein lytic proteinase B3
AT5G15650	reversibly glycosylated polypeptide 2
AT5G16050	general regulatory factor 5
AT5G16840	binding partner of acd11 1
AT5G16980	Zinc-binding dehydrogenase family protein
AT5G17380	Thiamine pyrophosphate dependent pyruvate decarboxylase family protein
AT5G18100	copper/ zinc superoxide dismutase 3
AT5G18170	glutamate dehydrogenase 1
AT5G19150	pfkB-like carbohydrate kinase family protein
AT5G19300	(1 of 1) PTHR12150:SF13 - DUF171 FAMILY PROTEIN
AT5G19320	RAN GTPase activating protein 2
AT5G19350	RNA-binding (RRM/RBD/RNP motifs) family protein
AT5G19510	Translation elongation factor EF1B/ribosomal protein S6 family protein

Accession	Description
AT5G19760	Mitochondrial substrate carrier family protein
AT5G19820	ARM repeat superfamily protein
AT5G19990	regulatory particle triple-A ATPase 6A
AT5G20010	RAS-related nuclear protein-1
AT5G20290	Ribosomal protein S8e family protein
AT5G20890	TCP-1/cpn60 chaperonin family protein
AT5G20960	aldehyde oxidase 1
AT5G22100	RNA cyclase family protein
AT5G22440	Ribosomal protein L1p/L10e family
AT5G23540	Mov34/MPN/PAD-1 family protein
AT5G24400	NagB/RpiA/CoA transferase-like superfamily protein
AT5G24490	30S ribosomal protein, putative
AT5G26830	Threonyl-tRNA synthetase
AT5G27450	mevalonate kinase
AT5G27670	histone H2A 7
AT5G27770	Ribosomal L22e protein family
AT5G28060	Ribosomal protein S24e family protein
AT5G37600	glutamine synthase clone R1
AT5G37640	ubiquitin 9

Accession	Description
AT5G38410	Ribulose biphosphate carboxylase (small chain) family protein
AT5G38470	Rad23 UV excision repair protein family
AT5G38480	general regulatory factor 3
AT5G38640	NagB/RpiA/CoA transferase-like superfamily protein
AT5G39740	ribosomal protein L5 B
AT5G39950	thioredoxin 2
AT5G40760	glucose-6-phosphate dehydrogenase 6
AT5G42190	E3 ubiquitin ligase SCF complex subunit SKP1/ASK1 family protein
AT5G42950	GYF domain-containing protein
AT5G42970	Proteasome component (PCI) domain protein
AT5G43010	regulatory particle triple-A ATPase 4A
AT5G43280	delta(3,5),delta(2,4)-dienoyl-CoA isomerase 1
AT5G43940	GroES-like zinc-binding dehydrogenase family protein
AT5G44340	tubulin beta chain 4
AT5G46070	Guanylate-binding family protein
AT5G46230	Protein of unknown function, DUF538
AT5G47720	Thiolase family protein
AT5G47760	2-phosphoglycolate phosphatase 2
AT5G47770	farnesyl diphosphate synthase 1

Accession	Description
AT5G47930	Zinc-binding ribosomal protein family protein
AT5G49510	prefoldin 3
AT5G50850	Transketolase family protein
AT5G50920	CLPC homologue 1
AT5G51970	GroES-like zinc-binding alcohol dehydrogenase family protein
AT5G52640	heat shock protein 90.1
AT5G53480	ARM repeat superfamily protein
AT5G54110	membrane-associated mannitol-induced
AT5G54770	thiazole biosynthetic enzyme, chloroplast (ARA6) (THI1) (THI4)
AT5G55160	small ubiquitin-like modifier 2
AT5G55990	calcineurin B-like protein 2
AT5G56350	Pyruvate kinase family protein
AT5G58070	temperature-induced lipocalin
AT5G58230	Transducin/WD40 repeat-like superfamily protein
AT5G58290	regulatory particle triple-A ATPase 3
AT5G58690	phosphatidylinositol-specific phospholipase C5
AT5G59950	RNA-binding (RRM/RBD/RNP motifs) family protein
AT5G60160	Zn-dependent exopeptidases superfamily protein
AT5G60390	GTP binding Elongation factor Tu family protein

Accession	Description
AT5G60980	Nuclear transport factor 2 (NTF2) family protein with RNA binding (RRM-RBD-RNP motifs) domain
AT5G61410	D-ribulose-5-phosphate-3-epimerase
AT5G61770	PETER PAN-like protein
AT5G63400	adenylate kinase 1
AT5G63680	Pyruvate kinase family protein
AT5G64380	Inositol monophosphatase family protein
AT5G65750	2-oxoglutarate dehydrogenase, E1 component
AT5G65780	branched-chain amino acid aminotransferase 5 / branched-chain amino acid transaminase 5 (BCAT5)
AT5G66760	succinate dehydrogenase 1-1
ATCG00480	ATP synthase subunit beta
ATCG00490	ribulose-bisphosphate carboxylases
ATCG00830	ribosomal protein L2
AT2G41950	T6D20_26
AT3G48860	(1 of 3) PTHR31762:SF4 - EXPRESSED PROTEIN
AT3G56510	RNA-binding (RRM/RBD/RNP motifs) family protein
AT5G22650	histone deacetylase 2B
AT5G24460	RING-H2 zinc finger protein

B.3 LSPDB orange - ribosomal

AT3G05560 Ribosomal L22e protein family
AT2G37270 ribosomal protein 5B
AT2G37190 Ribosomal protein L11 family protein
AT1G04480 Ribosomal protein L14p/L23e family protein
AT1G14320 Ribosomal protein L16p/L10e family protein
AT2G34480 Ribosomal protein L18ae/LX family protein
AT1G70600 Ribosomal protein L18e/L15 superfamily protein
AT4G02230 Ribosomal protein L19e family protein
AT1G67430 Ribosomal protein L22p/L17e family protein
AT4G17390 Ribosomal protein L23/L15e family protein
AT2G18020 Ribosomal protein L2 family
AT4G36130 Ribosomal protein L2 family
AT5G56710 Ribosomal protein L31e family protein
AT5G46430 Ribosomal protein L32e
AT3G09630 Ribosomal protein L4/L1 family
AT1G05190 Ribosomal protein L6 family
AT4G10450 Ribosomal protein L6 family
AT1G74050 Ribosomal protein L6 family protein
AT3G18740 Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein
AT2G42740 ribosomal protein large subunit 16A
AT3G47370 Ribosomal protein S10p/S20e family protein
AT3G11510 Ribosomal protein S11 family protein
AT1G07770 ribosomal protein S15A
AT4G29390 Ribosomal protein S30 family protein
AT4G34670 Ribosomal protein S3Ae
AT2G09990 Ribosomal protein S5 domain 2-like superfamily protein
AT5G18380 Ribosomal protein S5 domain 2-like superfamily protein
AT1G48830 Ribosomal protein S7e family protein
AT5G59850 Ribosomal protein S8 family protein

AT5G52650 RNA binding Plectin/S10 domain-containing protein
AT3G52150 RNA-binding (RRM/RBD/RNP motifs) family protein
AT2G45710 Zinc-binding ribosomal protein family protein
AT3G23390 Zinc-binding ribosomal protein family protein
AT3G60245 Zinc-binding ribosomal protein family protein
AT3G49010 60S ribosomal protein L13-1 (breast basic conserved 1)
AT1G66580 60S ribosomal protein L10-3 (senescence associated gene 24)
AT1G01100 60S acidic ribosomal protein family
AT3G48930 40S ribosomal protein S11-1

B.4 LSPDB orange - heat shock

AT3G12580 heat shock protein 70
AT3G09440 Heat shock protein 70 (Hsp 70) family protein
AT5G02490 Heat shock protein 70 (Hsp 70) family protein
AT5G56030 heat shock protein 81-2
AT5G56000 HEAT SHOCK PROTEIN 81.4
AT3G07770 HEAT SHOCK PROTEIN 89.1 *
AT4G37910 mitochondrial heat shock protein 70-1
AT5G09590 mitochondrial HSO70 2
AT4G02450 HSP20-like chaperones superfamily protein
AT2G04030 Chaperone protein htpG family protein
AT4G38740 rotamase CYP 1
AT2G16600 rotamase CYP 3

B.5 Homodimers

Accession	Description
AT1G02500	S-adenosylmethionine synthetase 1
AT1G02690	importin alpha isoform 6
AT1G02920	glutathione S-transferase 7
AT1G04170	eukaryotic translation initiation factor 2 gamma subunit
AT1G04410	Lactate/malate dehydrogenase family protein
AT1G04510	MOS4-associated complex 3A
AT1G07370	proliferating cellular nuclear antigen 1
AT1G08110	lactoylglutathione lyase family protein / glyoxalase I family protein
AT1G08410	P-loop containing nucleoside triphosphate hydrolases superfamily protein
AT1G09270	importin alpha isoform 4
AT1G09640	Translation elongation factor EF1B, gamma chain
AT1G09760	U2 small nuclear ribonucleoprotein A
AT1G10390	Nucleoporin autopeptidase
AT1G12230	Aldolase superfamily protein
AT1G14170	RNA-binding KH domain-containing protein
AT1G14320	Ribosomal protein L16p/L10e family protein
AT1G14980	chaperonin 10
AT1G15390	peptide deformylase 1A
AT1G16030	heat shock protein 70B
AT1G16470	proteasome subunit PAB1
AT1G18070	Translation elongation factor EF1A/initiation factor IF2gamma family protein
AT1G19120	Small nuclear ribonucleoprotein family protein
AT1G19730	Thioredoxin superfamily protein
AT1G19920	Pseudouridine synthase/archaeosine transglycosylase-like family protein
AT1G20260	ATPase, V1 complex, subunit B protein
AT1G20630	catalase 1
AT1G20693	high mobility group B2

Accession	Description
AT1G20960	U5 small nuclear ribonucleoprotein helicase, putative
AT1G21630	Calcium-binding EF hand family protein
AT1G22920	COP9 signalosome 5A
AT1G23190	Phosphoglucomutase/phosphomannomutase family protein
AT1G24360	NAD(P)-binding Rossmann-fold superfamily protein
AT1G26880	Ribosomal protein L34e superfamily protein
AT1G27450	adenine phosphoribosyl transferase 1
AT1G27970	nuclear transport factor 2B
AT1G29150	non-ATPase subunit 9
AT1G30580	GTP binding
AT1G43190	polypyrimidine tract-binding protein 3
AT1G45000	AAA-type ATPase family protein
AT1G48920	nucleolin like 1
AT1G49240	actin 8
AT1G49760	poly(A) binding protein 8
AT1G50200	Alanyl-tRNA synthetase
AT1G50480	10-formyltetrahydrofolate synthetase
AT1G50920	Nucleolar GTP-binding protein
AT1G51710	ubiquitin-specific protease 6
AT1G53240	Lactate/malate dehydrogenase family protein
AT1G53280	Class I glutamine amidotransferase-like superfamily protein
AT1G54270	eif4a-2
AT1G54290	Translation initiation factor SUI1 family protein
AT1G59900	pyruvate dehydrogenase complex E1 alpha subunit
AT1G60070	Adaptor protein complex AP-1, gamma subunit
AT1G60740	Thioredoxin superfamily protein
AT1G64230	ubiquitin-conjugating enzyme 28
AT1G65970	thioredoxin-dependent peroxidase 2
AT1G66240	homolog of anti-oxidant 1

Accession	Description
AT1G69740	Aldolase superfamily protein
AT1G71750	Hypoxanthine-guanine phosphoribosyltransferase
AT1G72560	ARM repeat superfamily protein
AT1G72730	DEA(D/H)-box RNA helicase family protein
AT1G74030	enolase 1
AT1G74310	heat shock protein 101
AT1G74560	NAP1-related protein 1
AT1G75510	Transcription initiation factor IIF, beta subunit
AT1G75950	S phase kinase-associated protein 1
AT1G76860	Small nuclear ribonucleoprotein family protein
AT1G77180	chromatin protein family
AT1G79530	glyceraldehyde-3-phosphate dehydrogenase of plastid 1
AT1G79930	heat shock protein 91
AT1G80500	SNARE-like superfamily protein
AT2G01250	Ribosomal protein L30/L7 family protein
AT2G01290	ribose-5-phosphate isomerase 2
AT2G02560	cullin-associated and neddylation dissociated
AT2G07690	Minichromosome maintenance (MCM2/3/5) family protein
AT2G09990	Ribosomal protein S5 domain 2-like superfamily protein
AT2G14120	dynamain related protein
AT2G17190	ubiquitin family protein
AT2G17200	ubiquitin family protein
AT2G17265	homoserine kinase
AT2G17510	ribonuclease II family protein
AT2G19750	Ribosomal protein S30 family protein
AT2G20580	26S proteasome regulatory subunit S2 1A
AT2G21170	triosephosphate isomerase
AT2G21250	NAD(P)-linked oxidoreductase superfamily protein
AT2G21660	cold, circadian rhythm, and rna binding 2

Accession	Description
AT2G24200	Cytosol aminopeptidase family protein
AT2G27020	20S proteasome alpha subunit G1
AT2G27170	Structural maintenance of chromosomes (SMC) family protein
AT2G28720	Histone superfamily protein
AT2G29570	proliferating cell nuclear antigen 2
AT2G30050	transducin family protein / WD-40 repeat family protein
AT2G30410	tubulin folding cofactor A (KIESEL)
AT2G30860	glutathione S-transferase PHI 9
AT2G33340	MOS4-associated complex 3B
AT2G34480	Ribosomal protein L18ae/LX family protein
AT2G36160	Ribosomal protein S11 family protein
AT2G36310	uridine-ribohydrolase 1
AT2G37620	actin 1
AT2G37760	NAD(P)-linked oxidoreductase superfamily protein
AT2G40290	Eukaryotic translation initiation factor 2 subunit 1
AT2G40490	Uroporphyrinogen decarboxylase
AT2G40660	Nucleic acid-binding, OB-fold-like protein
AT2G41530	S-formylglutathione hydrolase
AT2G42590	general regulatory factor 9
AT2G43460	Ribosomal L38e protein family
AT2G43810	Small nuclear ribonucleoprotein family protein
AT2G44510	CDK inhibitor P21 binding protein
AT2G45300	RNA 3'-terminal phosphate cyclase/enolpyruvate transferase, alpha/beta
AT2G46280	TGF-beta receptor interacting protein 1
AT2G47220	DOMAIN OF UNKNOWN FUNCTION 724 5
AT2G47580	spliceosomal protein U1A
AT2G47730	glutathione S-transferase phi 8
AT3G01390	vacuolar membrane ATPase 10
AT3G02870	Inositol monophosphatase family protein

Accession	Description
AT3G03250	UDP-GLUCOSE PYROPHOSPHORYLASE 1
AT3G03920	H/ACA ribonucleoprotein complex, subunit Gar1/Naf1 protein
AT3G04120	glyceraldehyde-3-phosphate dehydrogenase C subunit 1
AT3G04790	Ribose 5-phosphate isomerase, type A protein
AT3G05060	NOP56-like pre RNA processing ribonucleoprotein
AT3G05530	regulatory particle triple-A ATPase 5A
AT3G05560	Ribosomal L22e protein family
AT3G06720	importin alpha isoform 1
AT3G07880	Immunoglobulin E-set superfamily protein
AT3G08530	Clathrin, heavy chain
AT3G09200	Ribosomal protein L10 family protein
AT3G09630	Ribosomal protein L4/L1 family
AT3G09820	adenosine kinase 1
AT3G09840	cell division cycle 48
AT3G10920	manganese superoxide dismutase 1
AT3G11130	Clathrin, heavy chain
AT3G11400	eukaryotic translation initiation factor 3G1
AT3G11910	ubiquitin-specific protease 13
AT3G12390	Nascent polypeptide-associated complex (NAC), alpha subunit family protein
AT3G12580	heat shock protein 70
AT3G12860	NOP56-like pre RNA processing ribonucleoprotein
AT3G13110	serine acetyltransferase 2;2
AT3G13300	Transducin/WD40 repeat-like superfamily protein
AT3G13920	eukaryotic translation initiation factor 4A1
AT3G14390	Pyridoxal-dependent decarboxylase family protein
AT3G15650	alpha/beta-Hydrolases superfamily protein
AT3G16810	pumilio 24
AT3G16857	response regulator 1
AT3G17590	transcription regulatory protein SNF5, putative (BSH)

Accession	Description
AT3G18030	HAL3-like protein A
AT3G19450	GroES-like zinc-binding alcohol dehydrogenase family protein
AT3G19760	eukaryotic initiation factor 4A-III
AT3G22110	20S proteasome alpha subunit C1
AT3G22320	Eukaryotic rpb5 RNA polymerase subunit family protein
AT3G22890	ATP sulfurylase 1
AT3G24170	glutathione-disulfide reductase
AT3G24830	Ribosomal protein L13 family protein
AT3G25230	rotamase FKBP 1
AT3G25520	ribosomal protein L5
AT3G28730	high mobility group
AT3G29350	histidine-containing phosphotransmitter 2
AT3G43810	calmodulin 7
AT3G44110	DNAJ homologue 3
AT3G45300	isovaleryl-CoA-dehydrogenase
AT3G45780	phototropin 1
AT3G46030	Histone superfamily protein
AT3G46560	Tim10/DDP family zinc finger protein
AT3G47370	Ribosomal protein S10p/S20e family protein
AT3G48000	aldehyde dehydrogenase 2B4
AT3G48930	Nucleic acid-binding, OB-fold-like protein
AT3G49010	breast basic conserved 1
AT3G49680	branched-chain aminotransferase 3
AT3G49910	Translation protein SH3-like family protein
AT3G51260	20S proteasome alpha subunit PAD1
AT3G52150	RNA-binding (RRM/RBD/RNP motifs) family protein
AT3G52590	ubiquitin extension protein 1
AT3G53230	ATPase, AAA-type, CDC48 protein
AT3G53750	actin 3

Accession	Description
AT3G53880	NAD(P)-linked oxidoreductase superfamily protein
AT3G53890	Ribosomal protein S21e
AT3G54470	uridine 5'-monophosphate synthase / UMP synthase (PYRE-F) (UMPS)
AT3G54660	glutathione reductase
AT3G55440	triosephosphate isomerase
AT3G55605	Mitochondrial glycoprotein family protein
AT3G55620	Translation initiation factor IF6
AT3G55800	sedoheptulose-bisphosphatase
AT3G56090	ferritin 3
AT3G56490	HIS triad family protein 3
AT3G57150	homologue of NAP57
AT3G57560	N-acetyl-l-glutamate kinase
AT3G58660	Ribosomal protein L1p/L10e family
AT3G59760	O-acetylserine (thiol) lyase isoform C
AT3G60240	eukaryotic translation initiation factor 4G
AT3G60750	Transketolase
AT3G62870	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein
AT3G63490	Ribosomal protein L1p/L10e family
AT4G01370	MAP kinase 4
AT4G01690	Flavin containing amine oxidoreductase family
AT4G01900	GLNB1 homolog
AT4G02450	HSP20-like chaperones superfamily protein
AT4G02510	translocon at the outer envelope membrane of chloroplasts 159
AT4G04885	PCF11P-similar protein 4
AT4G05320	polyubiquitin 10
AT4G05420	damaged DNA binding protein 1A
AT4G10040	cytochrome c-2
AT4G10480	Nascent polypeptide-associated complex (NAC), alpha subunit family protein
AT4G11010	nucleoside diphosphate kinase 3

Accession	Description
AT4G13940	S-adenosyl-L-homocysteine hydrolase
AT4G14570	acylaminoacyl-peptidase-related
AT4G15802	heat shock factor binding protein
AT4G16420	homolog of yeast ADA2 2B
AT4G17260	Lactate/malate dehydrogenase family protein
AT4G17620	glycine-rich protein
AT4G19210	RNAse I inhibitor protein 2
AT4G19750	Glycosyl hydrolase family protein with chitinase insertion domain
AT4G20850	tripeptidyl peptidase ii
AT4G23570	phosphatase-related
AT4G24770	31-kDa RNA binding protein
AT4G24830	arginosuccinate synthase family
AT4G25550	Cleavage/polyadenylation specificity factor, 25kDa subunit
AT4G25630	fibrillarin 2
AT4G26530	Aldolase superfamily protein
AT4G26840	small ubiquitin-like modifier 1
AT4G26910	Dihydrolipoamide succinyltransferase
AT4G26970	aconitase 2
AT4G27090	Ribosomal protein L14
AT4G29510	arginine methyltransferase 11
AT4G30330	Small nuclear ribonucleoprotein family protein
AT4G30920	Cytosol aminopeptidase family protein
AT4G31580	serine/arginine-rich 22
AT4G32720	La protein 1
AT4G33650	dynamamin-related protein 3A
AT4G34200	D-3-phosphoglycerate dehydrogenase
AT4G34490	cyclase associated protein 1
AT4G34870	rotamase cyclophilin 5
AT4G35090	catalase 2

Accession	Description
AT4G37870	phosphoenolpyruvate carboxykinase 1
AT4G38740	rotamase CYP 1
AT4G39260	cold, circadian rhythm, and RNA binding 1
AT4G39660	alanine:glyoxylate aminotransferase 2
AT5G02500	heat shock cognate protein 70-1
AT5G02530	RNA-binding (RRM/RBD/RNP motifs) family protein
AT5G03340	ATPase, AAA-type, CDC48 protein
AT5G06060	NAD(P)-binding Rossmann-fold superfamily protein
AT5G06290	2-cysteine peroxiredoxin B
AT5G08570	Pyruvate kinase family protein
AT5G08670	ATP synthase alpha/beta family protein
AT5G09590	mitochondrial HSO70 2
AT5G09620	Octicosapeptide/Phox/Bem1p family protein
AT5G09660	peroxisomal NAD-malate dehydrogenase 2
AT5G10450	G-box regulating factor 6
AT5G10810	enhancer of rudimentary protein, putative
AT5G10920	L-Aspartase-like family protein
AT5G11170	DEAD/DEAH box RNA helicase family protein
AT5G11200	DEAD/DEAH box RNA helicase family protein
AT5G11260	Basic-leucine zipper (bZIP) transcription factor family protein
AT5G11900	Translation initiation factor SUI1 family protein
AT5G11920	6-&1-fructan exohydrolase
AT5G13120	cyclophilin 20-2
AT5G13850	nascent polypeptide-associated complex subunit alpha-like protein 3
AT5G14250	Proteasome component (PCI) domain protein
AT5G14800	pyrroline-5- carboxylate (P5C) reductase
AT5G15450	casein lytic proteinase B3
AT5G16840	binding partner of acd11 1
AT5G16980	Zinc-binding dehydrogenase family protein

Accession	Description
AT5G17380	Thiamine pyrophosphate dependent pyruvate decarboxylase family protein
AT5G18100	copper/zinc superoxide dismutase 3
AT5G18170	glutamate dehydrogenase 1
AT5G19090	Heavy metal transport/detoxification superfamily protein
AT5G19150	pfkB-like carbohydrate kinase family protein
AT5G19300	(1 of 1) PTHR12150:SF13 - DUF171 FAMILY PROTEIN
AT5G19320	RAN GTPase activating protein 2
AT5G19510	Translation elongation factor EF1B/ribosomal protein S6 family protein
AT5G19760	Mitochondrial substrate carrier family protein
AT5G19990	regulatory particle triple-A ATPase 6A
AT5G22100	RNA cyclase family protein
AT5G22440	Ribosomal protein L1p/L10e family
AT5G23540	Mov34/MPN/PAD-1 family protein
AT5G24490	30S ribosomal protein, putative
AT5G26830	Threonyl-tRNA synthetase
AT5G27450	mevalonate kinase
AT5G27670	histone H2A 7
AT5G27770	Ribosomal L22e protein family
AT5G37830	oxoprolinase 1
AT5G38470	Rad23 UV excision repair protein family
AT5G38480	general regulatory factor 3
AT5G38640	NagB/RpiA/CoA transferase-like superfamily protein
AT5G39740	ribosomal protein L5 B
AT5G39950	thioredoxin 2
AT5G40760	glucose-6-phosphate dehydrogenase 6
AT5G42190	E3 ubiquitin ligase SCF complex subunit SKP1/ASK1 family protein
AT5G42950	GYF domain-containing protein
AT5G42970	Proteasome component (PCI) domain protein
AT5G42980	thioredoxin 3

Accession	Description
AT5G43010	regulatory particle triple-A ATPase 4A
AT5G43280	delta(3,5),delta(2,4)-dienoyl-CoA isomerase 1
AT5G43940	GroES-like zinc-binding dehydrogenase family protein
AT5G46070	Guanylate-binding family protein
AT5G47760	2-phosphoglycolate phosphatase 2
AT5G47770	farnesyl diphosphate synthase 1
AT5G47930	Zinc-binding ribosomal protein family protein
AT5G48870	Small nuclear ribonucleoprotein family protein
AT5G50850	Transketolase family protein
AT5G51970	GroES-like zinc-binding alcohol dehydrogenase family protein
AT5G52640	heat shock protein 90.1
AT5G53480	ARM repeat superfamily protein
AT5G54110	membrane-associated mannitol-induced
AT5G54770	thiazole biosynthetic enzyme, chloroplast (ARA6) (THI1) (THI4)
AT5G55160	small ubiquitin-like modifier 2
AT5G55990	calcineurin B-like protein 2
AT5G56000	HEAT SHOCK PROTEIN 81.4
AT5G56030	heat shock protein 81-2
AT5G56350	Pyruvate kinase family protein
AT5G58230	Transducin/WD40 repeat-like superfamily protein
AT5G58290	regulatory particle triple-A ATPase 3
AT5G58690	phosphatidylinositol-specific phospholipase C5
AT5G59950	RNA-binding (RRM/RBD/RNP motifs) family protein
AT5G60160	Zn-dependent exopeptidases superfamily protein
AT5G60980	Nuclear transport factor 2 (NTF2) family protein with RNA binding (RRM-RBD-RNP motifs) d
AT5G61410	D-ribulose-5-phosphate-3-epimerase
AT5G61770	PETER PAN-like protein
AT5G63310	nucleoside diphosphate kinase 2
AT5G63680	Pyruvate kinase family protein

Accession	Description
AT5G64370	beta-ureidopropionase
AT5G64380	Inositol monophosphatase family protein
AT5G65750	2-oxoglutarate dehydrogenase, E1 component
AT5G65780	branched-chain amino acid aminotransferase 5 / branched-chain amino acid transaminase
AT5G67500	voltage dependent anion channel 2
AT3G48860	(1 of 3) PTHR31762:SF4 - EXPRESSED PROTEIN
AT3G56510	RNA-binding (RRM/RBD/RNP motifs) family protein

B.6 Common GO and PFAM

Table B.4: Most common GO terms in the SEC class of the LSPDB

GO ID	Description
GO:0004553	hydrolase activity, hydrolyzing O-glycosyl compounds
GO:0005576	extracellular region
GO:0005618	cell wall
GO:0005773	vacuole
GO:0005783	endoplasmic reticulum
GO:0005975	carbohydrate metabolic process
GO:0006508	proteolysis
GO:0008152	metabolic process
GO:0008233	peptidase activity
GO:0009505	plant-type cell wall
GO:0009506	plasmodesma
GO:0016020	membrane
GO:0016491	oxidoreductase activity
GO:0016787	hydrolase activity
GO:0016798	hydrolase activity, acting on glycosyl bonds
GO:0046872	metal ion binding

GO ID	Description
GO:0048046	apoplast
GO:0055114	oxidation-reduction process
GO:0071555	cell wall organization

Table B.5: Most common PFAM domains in the SEC class of the LSPDB

PFAM	Symbol	Description
PF00082	Peptidase_S8	Subtilase family
PF00112	Peptidase_C1	Papain family cysteine protease
PF00141	peroxidase	Peroxidase
PF00190	Cupin_1	Cupin
PF00232	Glyco_hydro.1	Glycosyl hydrolase family 1
PF00295	Glyco_hydro.28	Glycosyl hydrolases family 28
PF00394	Cu-oxidase	Multicopper oxidase
PF00450	Peptidase_S10	Serine carboxypeptidase
PF00657	Lipase_GDSL	GDSL-like Lipase/ Acylhydrolase
PF01357	Pollen_allerg.1	Pollen allergen
PF01565	FAD_binding.4	FAD binding domain
PF03330	DPBB_1	Lytic transglycolase
PF04043	PMEI	Plant invertase/pectin methylesterase inhibitor
PF05922	Inhibitor_I9	Peptidase inhibitor I9
PF07731	Cu-oxidase.2	Multicopper oxidase
PF08031	BBE	Berberine and berberine like
PF08263	LRRNT.2	Leucine rich repeat N-terminal domain
PF14541	TAXi_C	Xylanase inhibitor C-terminal
PF14543	TAXi_N	Xylanase inhibitor N-terminal

Table B.6: Most common GO terms in the SPT class of the LSPDB

GO ID	Description
GO:0003674	molecular_function
GO:0005506	iron ion binding
GO:0005576	extracellular region
GO:0005783	endoplasmic reticulum
GO:0005794	Golgi apparatus
GO:0006952	defense response
GO:0008150	biological_process
GO:0016020	membrane
GO:0016021	integral component of membrane
GO:0016491	oxidoreductase activity
GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of mole
GO:0016740	transferase activity
GO:0016787	hydrolase activity
GO:0020037	heme binding
GO:0031640	killing of cells of other organism
GO:0046872	metal ion binding
GO:0050832	defense response to fungus
GO:0055114	oxidation-reduction process
GO:0071555	cell wall organization

Table B.7: Most common PFAM domains in the SPT class of the LSPDB

PFAM	Symbol	Description
PF00067	p450	Cytochrome P450
PF00069	Pkinase	Protein kinase domain
PF00295	Glyco_hydro_28	Glycosyl hydrolases family 28
PF00450	Peptidase_S10	Serine carboxypeptidase

PFAM	Symbol	Description
PF00657	Lipase_GDSL	GDSL-like Lipase/Acylhydrolase
PF01095	Pectinesterase	Pectinesterase
PF01501	Glyco_transf_8	Glycosyl transferase family 8
PF01657	Stress-antifung	Salt stress response/antifungal
PF02485	Branch	Core-2/I-Branching enzyme
PF03080	Neprosin	Neprosin
PF03141	Methyltransf_29	Putative S-adenosyl-L-methionine-dependent methyltransferase
PF04043	PMEI	Plant invertase/pectin methylesterase inhibitor
PF05498	RALF	Rapid ALkalinization Factor (RALF)
PF05617	Prolamin_like	Prolamin-like
PF05938	Self-incomp_S1	Plant self-incompatibility protein S1
PF07333	SLR1-BP	S locus-related glycoprotein 1 binding pollen coat protein (SLR1-BP)
PF07983	X8	X8 domain
PF13855	LRR_8	Leucine rich repeat
PF14365	Neprosin_AP	Neprosin activation peptide

Table B.8: Most common GO terms in the UNCLASSIFIED class of the LSPDB

GO ID	Description
GO:0000166	nucleotide binding
GO:0005515	protein binding
GO:0005524	ATP binding
GO:0005634	nucleus
GO:0005737	cytoplasm
GO:0005829	cytosol
GO:0005840	ribosome
GO:0005886	plasma membrane
GO:0006412	translation
GO:0009506	plasmodesma

GO ID	Description
GO:0009507	chloroplast
GO:0009536	plastid
GO:0009570	chloroplast stroma
GO:0016020	membrane
GO:0016740	transferase activity
GO:0016787	hydrolase activity
GO:0030529	intracellular ribonucleoprotein complex
GO:0046872	metal ion binding
GO:0055114	oxidation-reduction process

Table B.9: PFAM domains most occurring in the UNCLASSIFIED group

PFAM	Symbol	Description
PF00004	AAA	ATPase family associated with various cellular activities (AAA)
PF00009	GTP_EFTU	Elongation factor Tu GTP binding domain
PF00076	RRM_1	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
PF00085	Thioredoxin	Thioredoxin
PF00118	Cpn60_TCP1	TCP-1/cpn60 chaperonin family
PF00125	Histone	Core histone H2A/H2B/H3/H4
PF00155	Aminotran_1_2	Aminotransferase class I and II
PF00227	Proteasome	Proteasome subunit
PF00240	ubiquitin	Ubiquitin family
PF00244	14-3-3	14-3-3 protein
PF00270	DEAD	DEAD/DEAH box helicase
PF00271	Helicase_C	Helicase conserved C-terminal domain
PF00400	WD40	WD domain, G-beta repeat
PF00582	Usp	Universal stress protein family
PF01419	Jacalin	Jacalin-like lectin domain
PF01423	LSM	LSM domain

PFAM	Symbol	Description
PF03144	GTP_EFTU_D2	Elongation factor Tu domain 2
PF07992	Pyr_redox_2	Pyridine nucleotide-disulphide oxidoreductase
PF13499	EF-hand_7	EF-hand domain pair

Table B.10: Most common GO terms in the NONSEC class of the LSPDB

GO ID	Description
GO:0000166	nucleotide binding
GO:0003674	molecular_function
GO:0003677	DNA binding
GO:0003700	transcription factor activity, sequence-specific DNA binding
GO:0003723	RNA binding
GO:0005515	protein binding
GO:0005524	ATP binding
GO:0005634	nucleus
GO:0005737	cytoplasm
GO:0005829	cytosol
GO:0005886	plasma membrane
GO:0006351	transcription, DNA-templated
GO:0006355	regulation of transcription, DNA-templated
GO:0008150	biological_process
GO:0009507	chloroplast
GO:0016020	membrane
GO:0016740	transferase activity
GO:0016787	hydrolase activity
GO:0046872	metal ion binding

Table B.11: Most common PFAM domains in the NONSEC class of the LSPDB

PFAM	Symbol	Description
PF00010	HLH	Helix-loop-helix DNA-binding domain
PF00069	Pkinase	Protein kinase domain
PF00076	RRM_1	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
PF00249	Myb_DNA-binding	Myb-like DNA-binding domain
PF00271	Helicase_C	Helicase conserved C-terminal domain
PF00319	SRF-TF	SRF-type transcription factor (DNA-binding and dimerisation domain)
PF00400	WD40	WD domain, G-beta repeat
PF00646	F-box	F-box domain
PF00847	AP2	AP2 domain
PF00931	NB-ARC	NB-ARC domain
PF01535	PPR	PPR repeat
PF01582	TIR	TIR domain
PF07714	Pkinase_Tyr	Protein tyrosine kinase
PF07723	LRR_2	Leucine Rich Repeat
PF07725	LRR_3	Leucine Rich Repeat
PF07734	FBA_1	F-box associated
PF12854	PPR_1	PPR repeat
PF13041	PPR_2	PPR repeat family
PF13639	zf-RING_2	Ring finger domain

Appendix C

Appendix

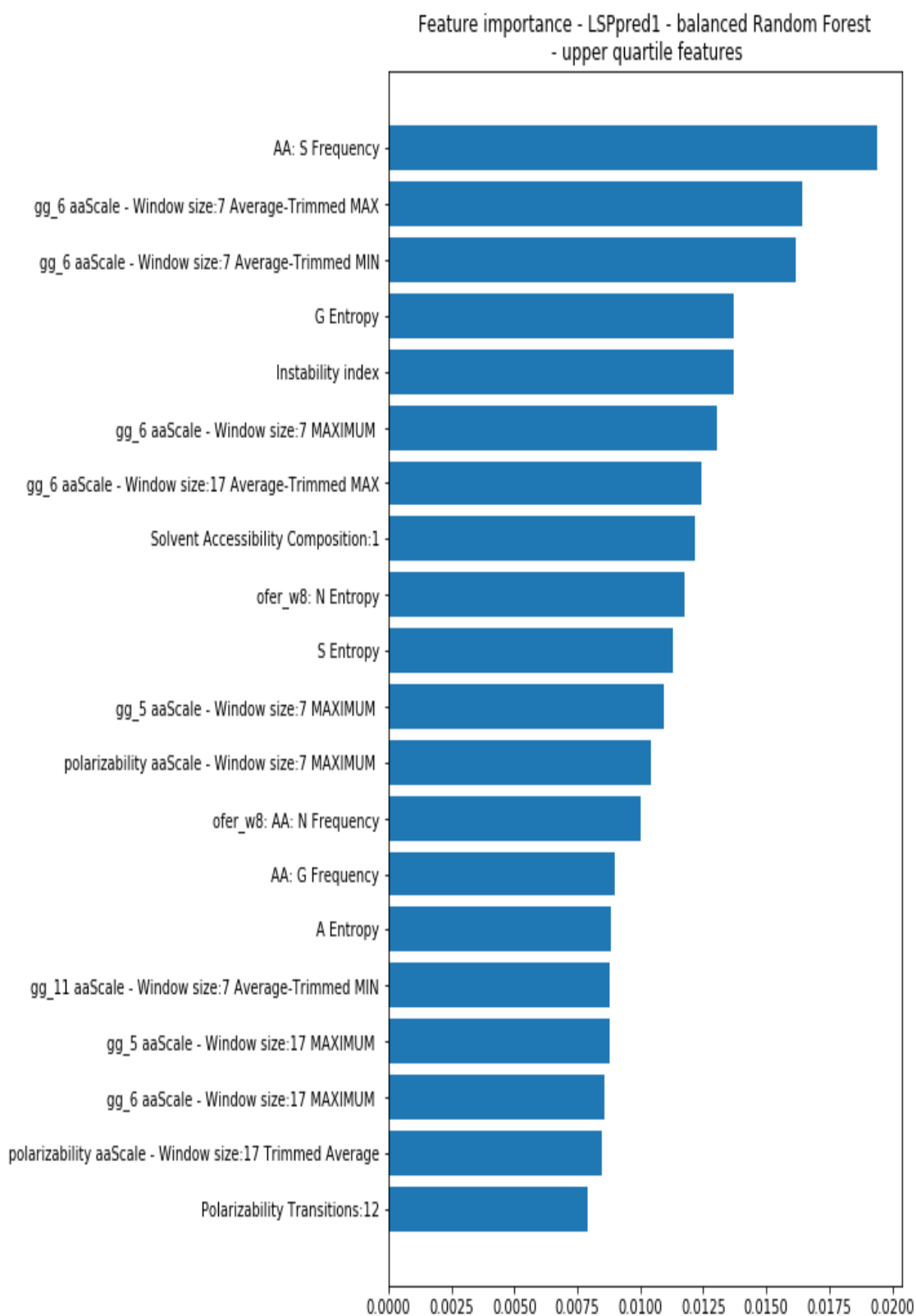


Figure C.1: Top 20 features by importance for feature selected LSPpred version 1

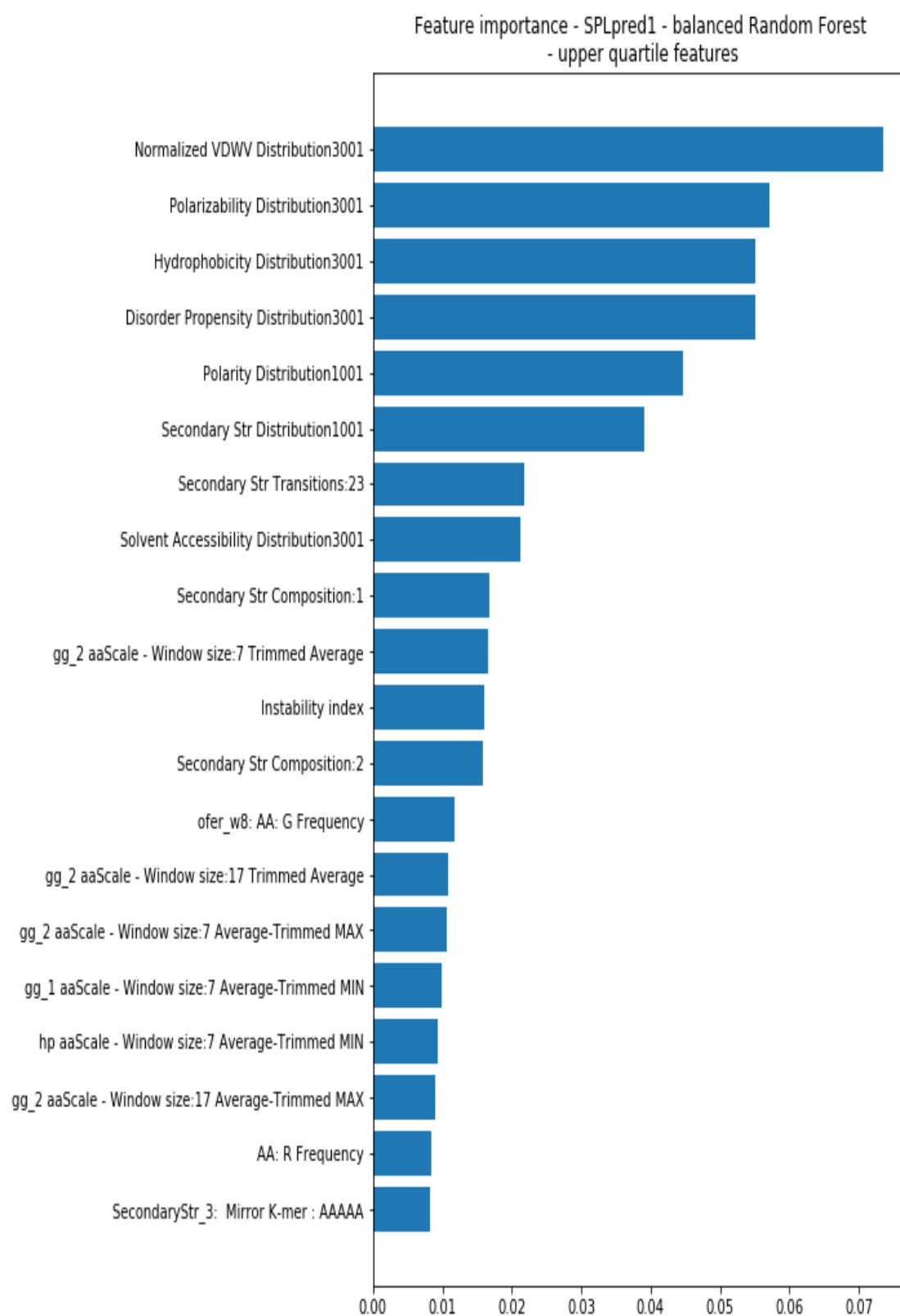


Figure C.2: Top 20 features by importance for feature selected SPLpred version 1

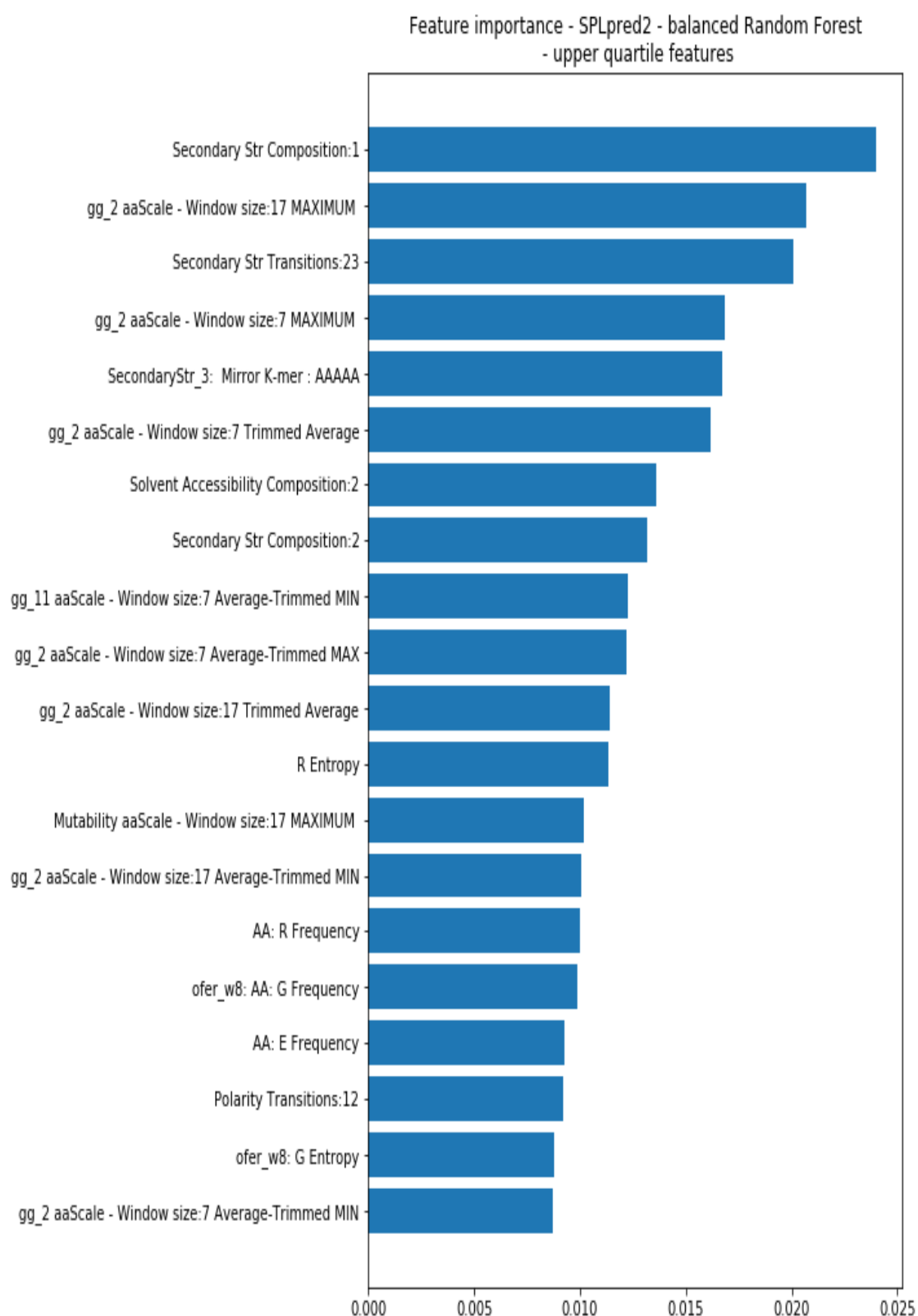


Figure C.3: Top 20 features by importance for feature selected SPLpred version 2

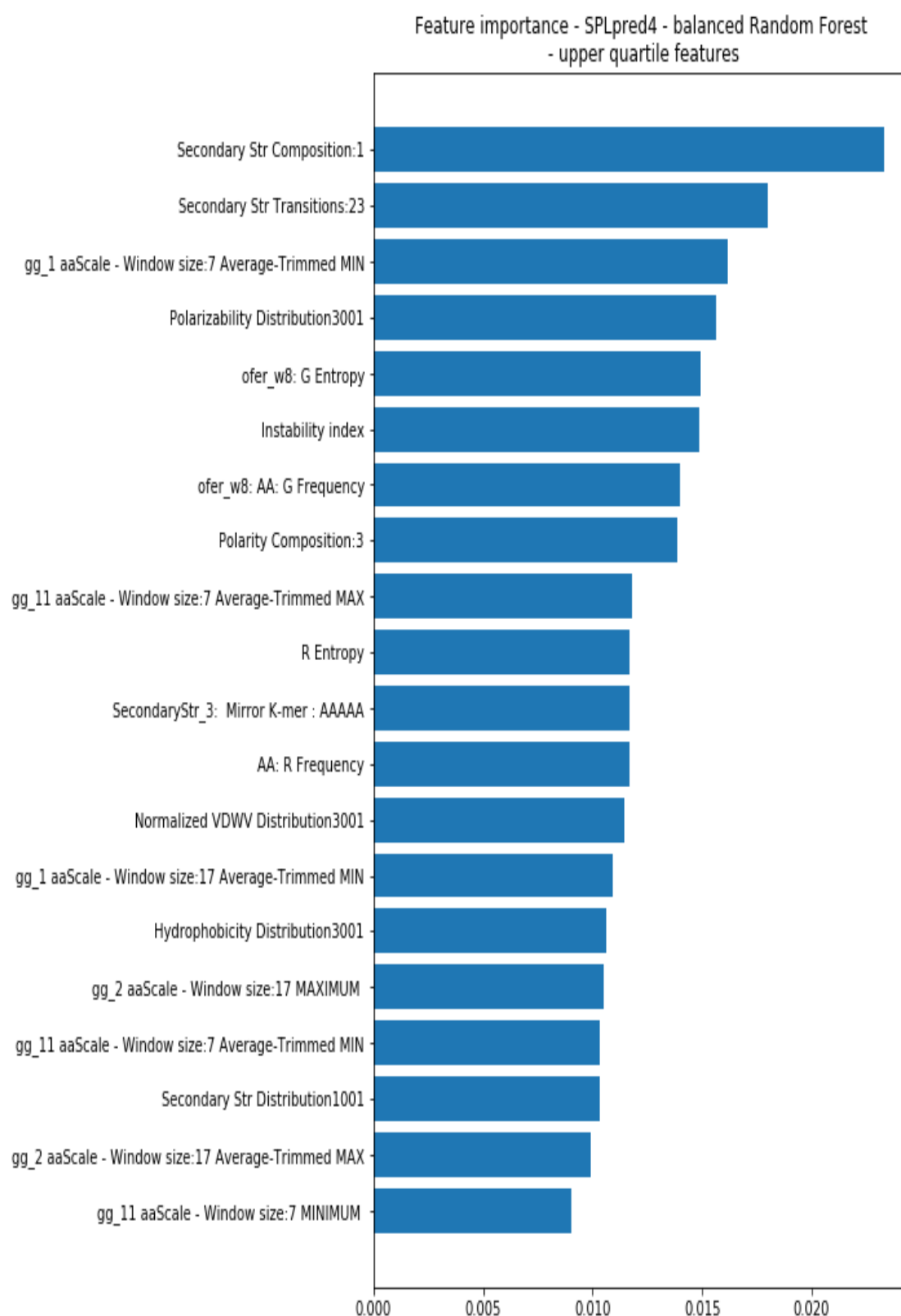


Figure C.4: Top 20 features by importance for feature selected SPLpred version 4

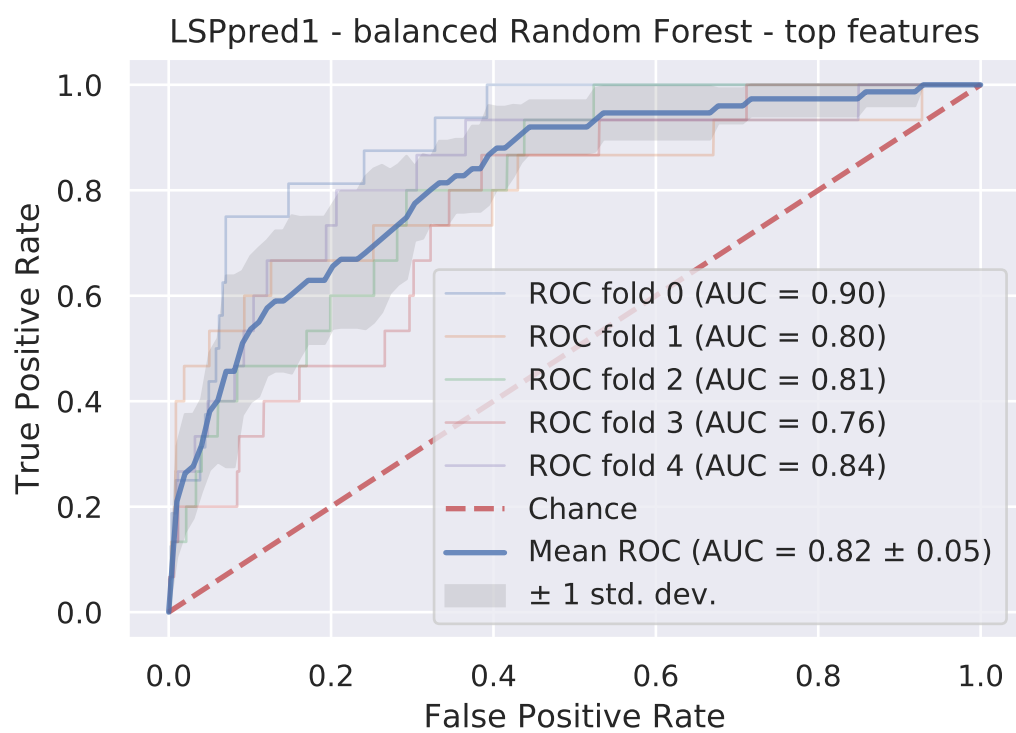


Figure C.5: ROC curve for LSPpred version 1

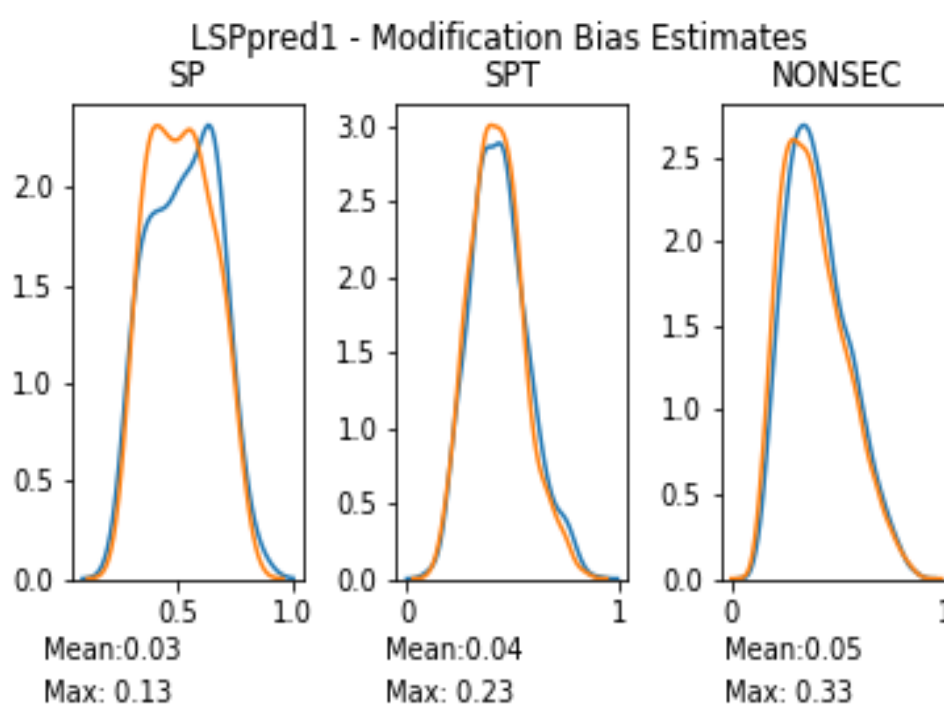


Figure C.6: Score distribution for LSPpred 1 with (blue) and without (orange) modifications on SEC, SPT and NONSEC proteins

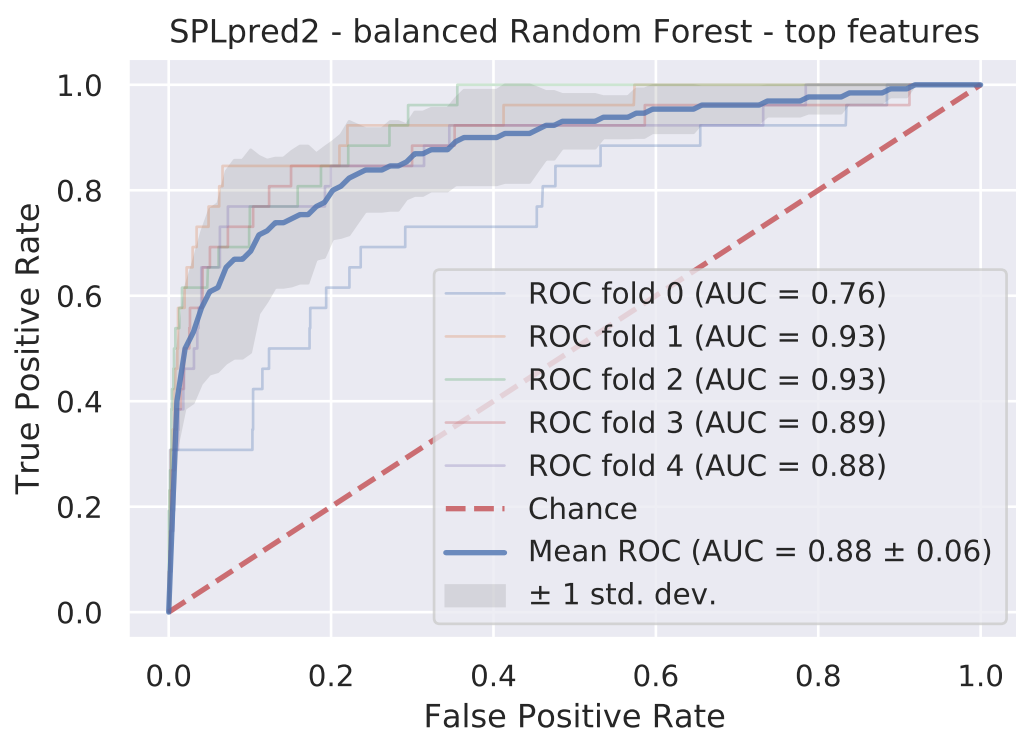


Figure C.7: ROC curve for SPLpred version 2

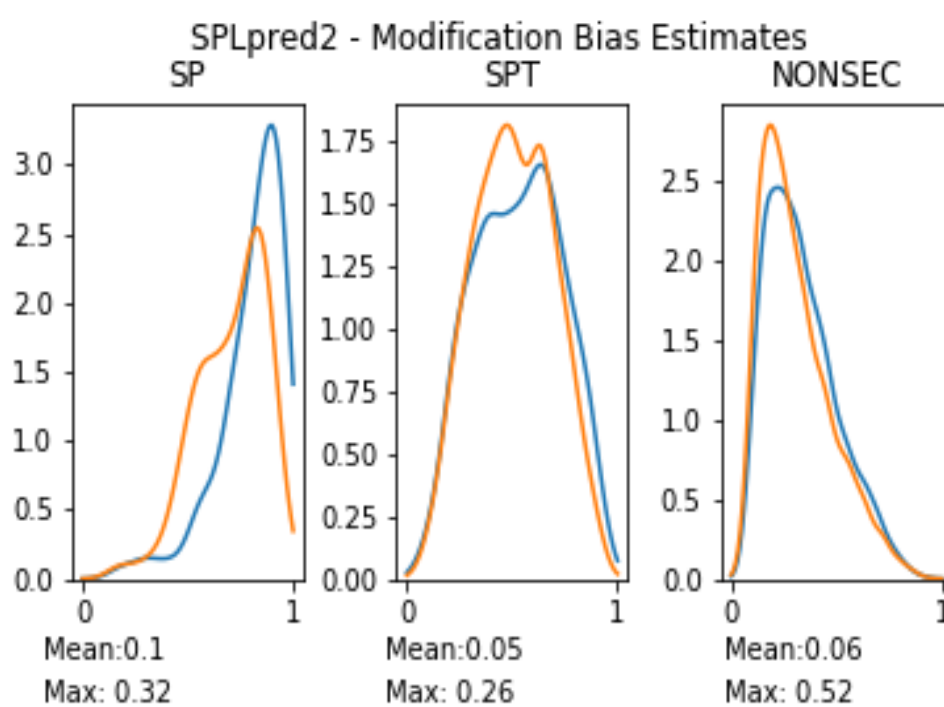


Figure C.8: Score distribution for SPLpred 2 with (blue) and without (orange) modifications on SEC, SPT and NONSEC proteins

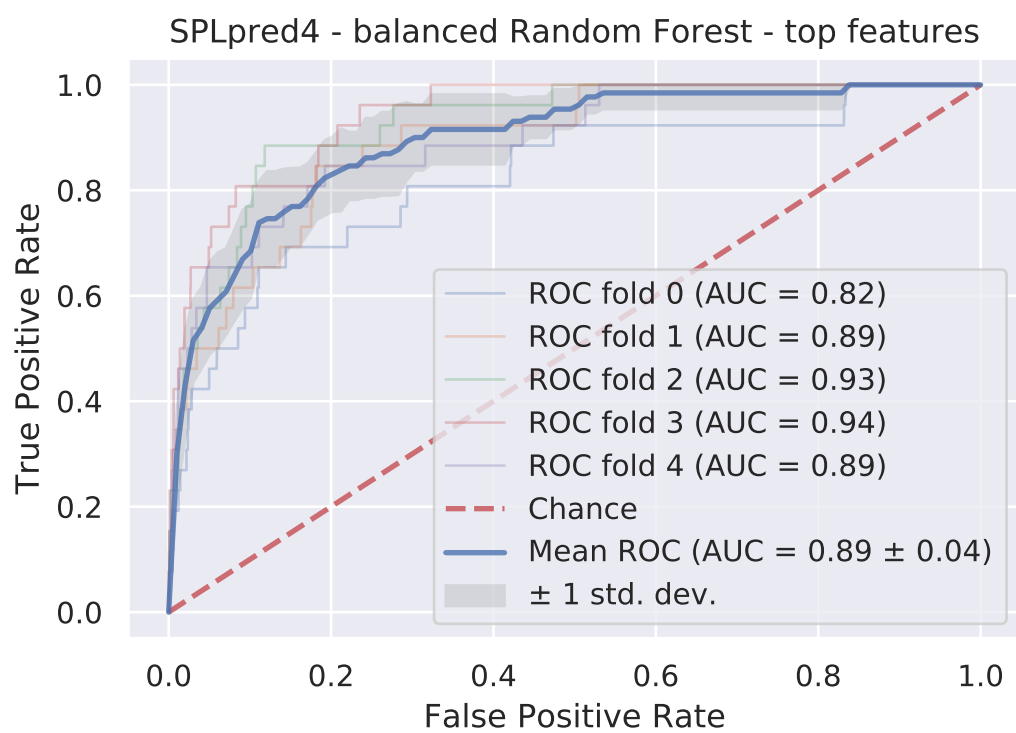


Figure C.9: ROC curve for SPLpred version 4

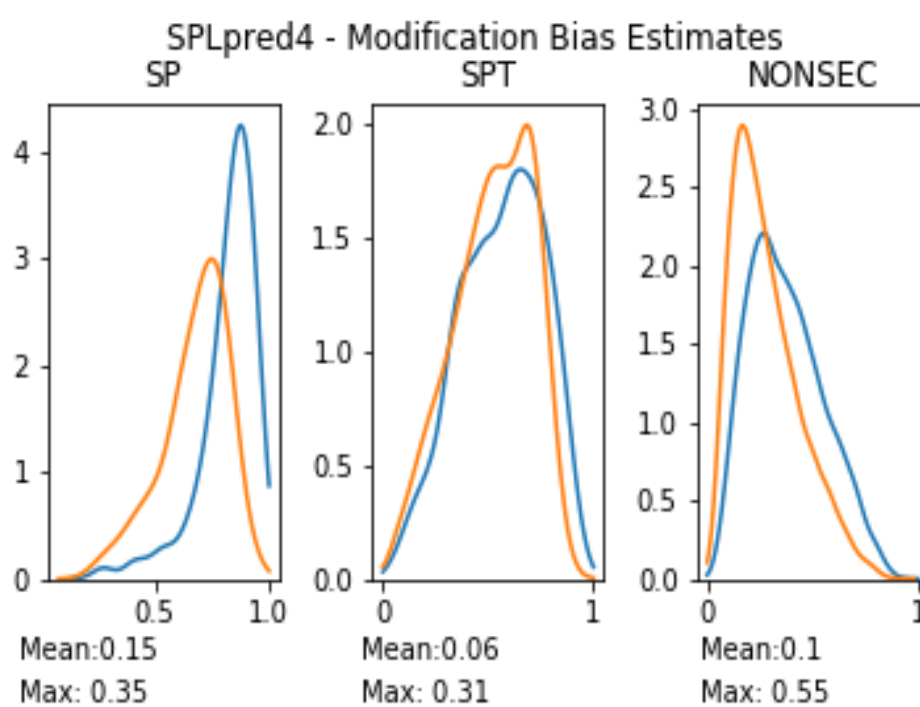


Figure C.10: Score distribution for SPLpred 4 with (blue) and without (orange) modifications on SEC, SPT and NONSEC proteins

Decision tree - LSPpred1 - default threshold
Balanced accuracy: 0.71 - Geometric mean 0.70

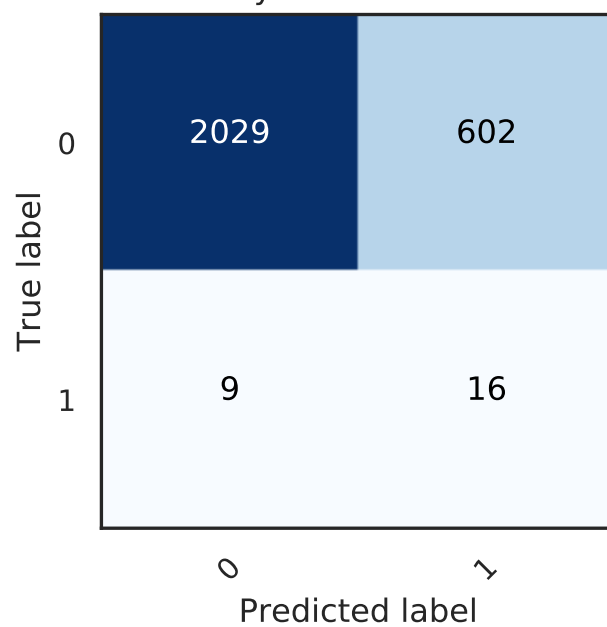


Figure C.11: Confusion matrix for LSPpred version 1 at 0.5 threshold

Decision tree - LSPpred1 - 0.05 FPR threshold
Balanced accuracy: 0.60 - Geometric mean 0.48

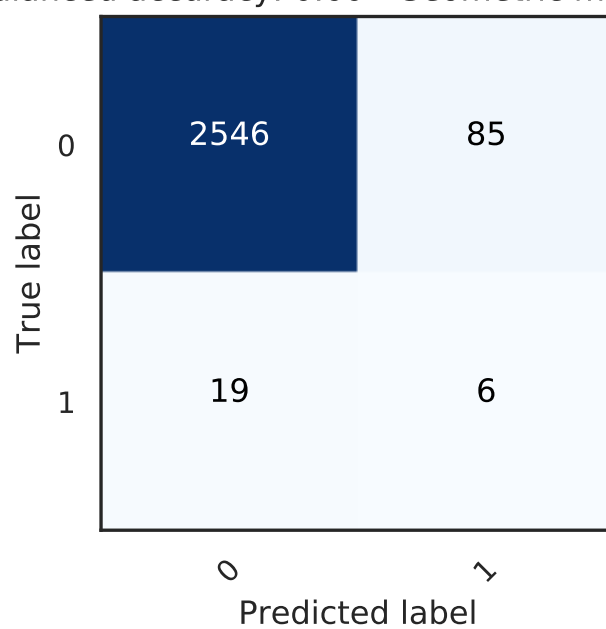


Figure C.12: Confusion matrix for LSPpred version 1 at 0.645 threshold

Decision tree - LSPpred2 - default threshold
Balanced accuracy: 0.65 - Geometric mean 0.64

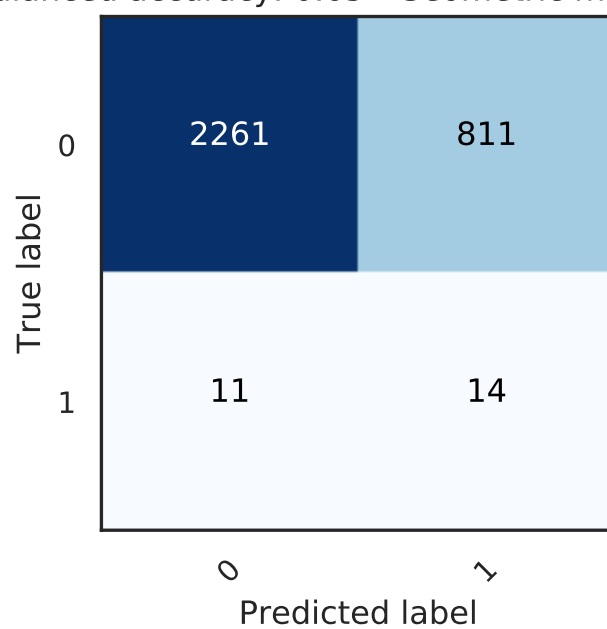


Figure C.13: Confusion matrix for LSPpred version 2 at 0.5 threshold

Decision tree - LSPpred2 - 0.05 FPR threshold
Balanced accuracy: 0.63 - Geometric mean 0.55

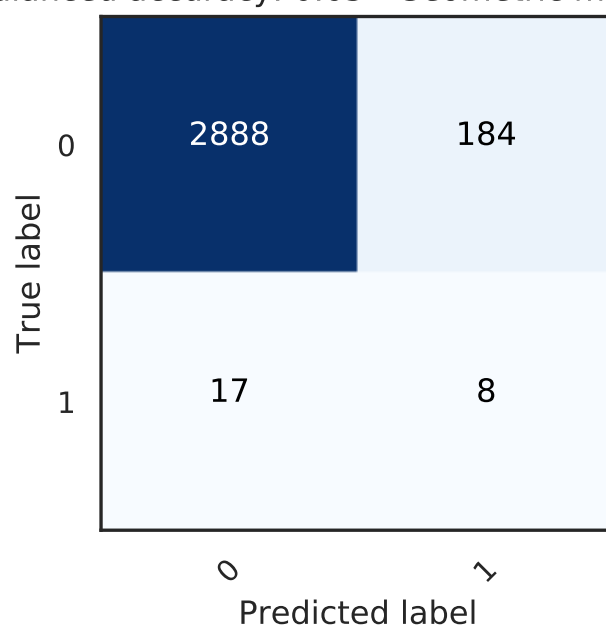


Figure C.14: Confusion matrix for LSPpred version 2 at 0.68 threshold

Decision tree - SPLpred1 - default threshold
Balanced accuracy: 0.94 - Geometric mean 0.94

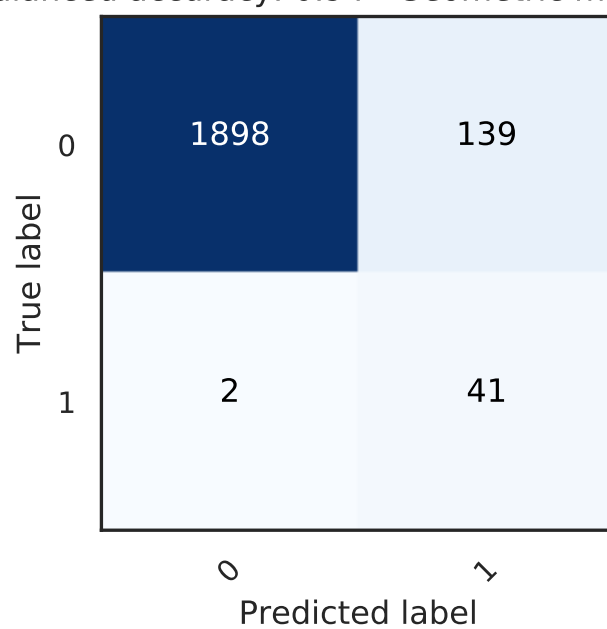


Figure C.15: Confusion matrix for SPLpred version 1 at 0.5 threshold

Decision tree - SPLpred1 - 0.05 FPR threshold
Balanced accuracy: 0.95 - Geometric mean 0.95

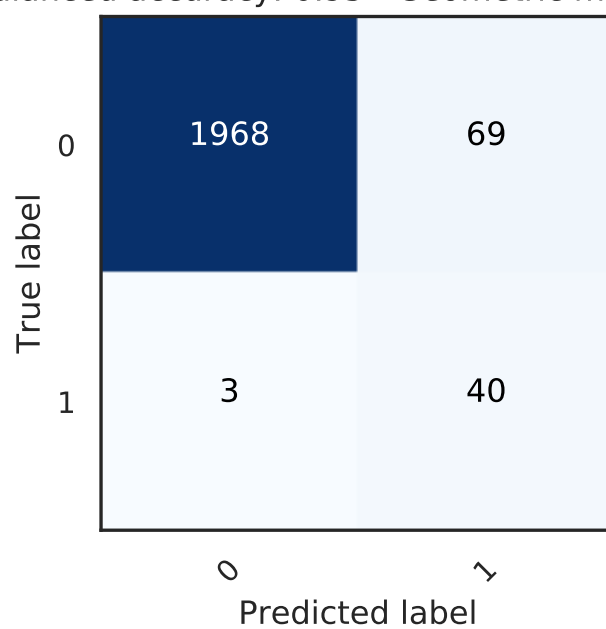


Figure C.16: Confusion matrix for SPLpred version 1 at 0.61 threshold

Decision tree - SPLpred2 - default threshold
Balanced accuracy: 0.81 - Geometric mean 0.81

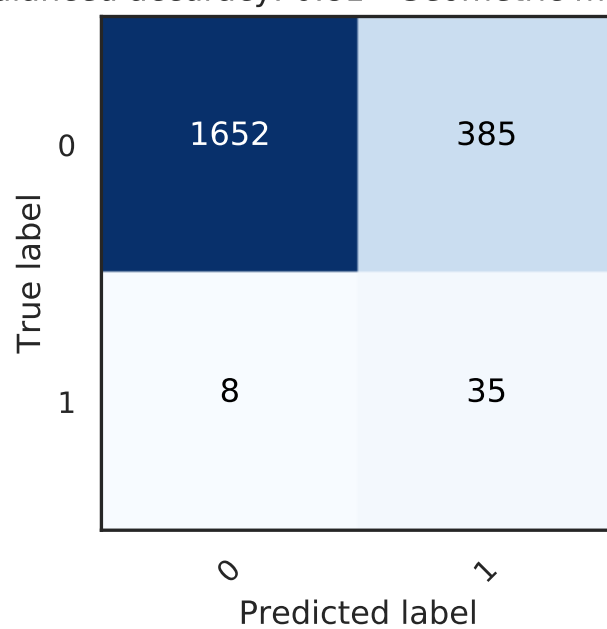


Figure C.17: Confusion matrix for SPLpred version 2 at 0.5 threshold

Decision tree - SPLpred2 - 0.05 FPR threshold
Balanced accuracy: 0.72 - Geometric mean 0.68

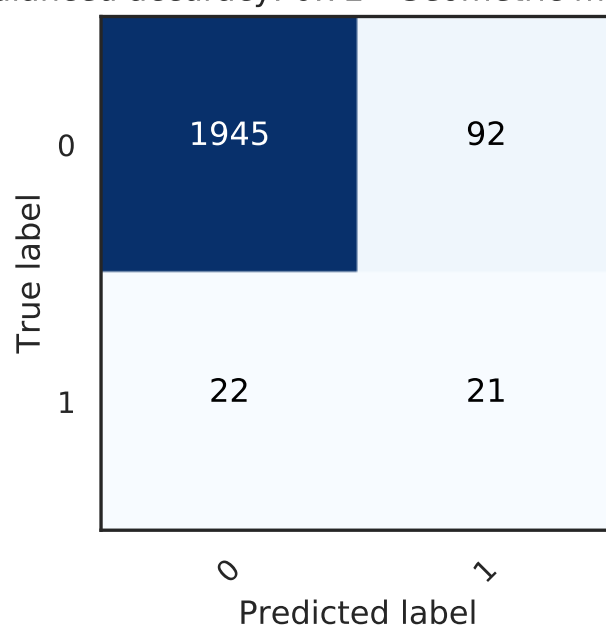


Figure C.18: Confusion matrix for SPLpred version 2 at 0.70 threshold

Decision tree - SPLpred3 - default threshold
Balanced accuracy: 0.82 - Geometric mean 0.82

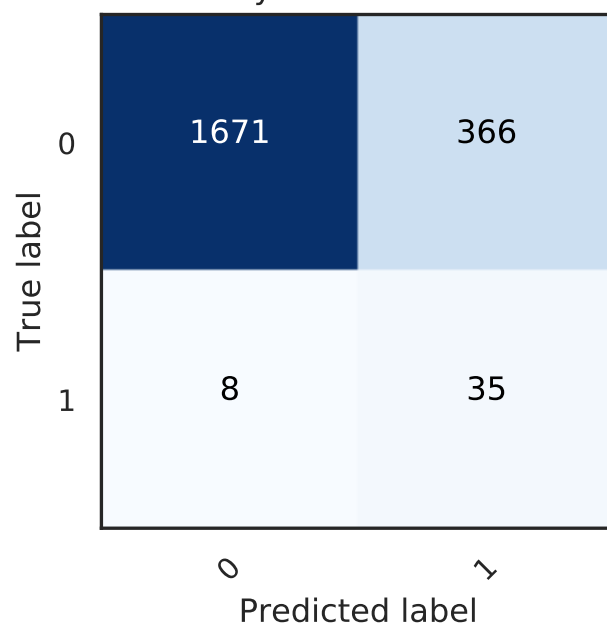


Figure C.19: Confusion matrix for SPLpred version 3 at 0.5 threshold

Decision tree - SPLpred3 - 0.05 FPR threshold
Balanced accuracy: 0.76 - Geometric mean 0.73

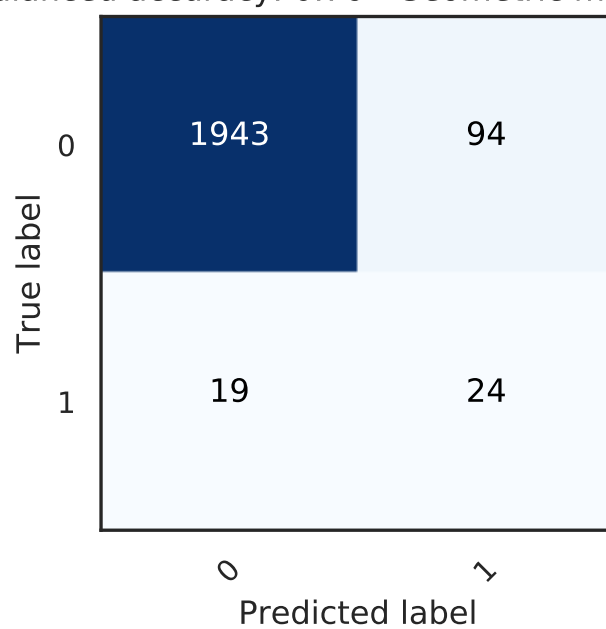


Figure C.20: Confusion matrix for SPLpred version 3 at 0.70 threshold

Decision tree - SPLpred4 - default threshold
Balanced accuracy: 0.84 - Geometric mean 0.84

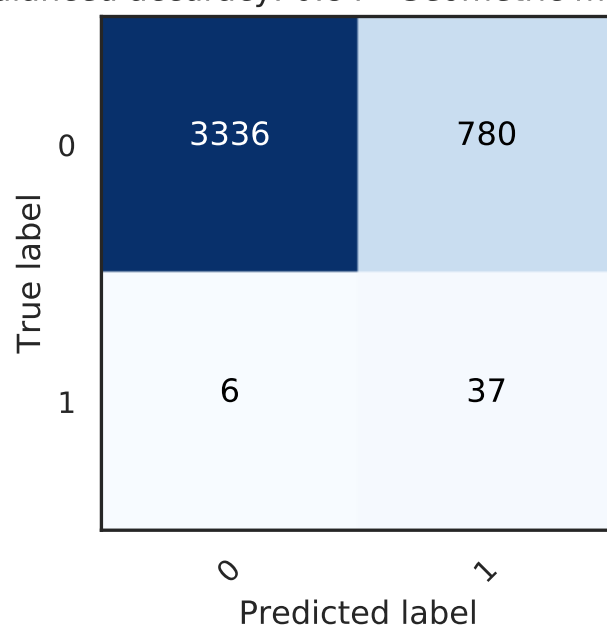


Figure C.21: Confusion matrix for SPLpred version 4 at 0.5 threshold

Decision tree - SPLpred4 - 0.05 FPR threshold
Balanced accuracy: 0.83 - Geometric mean 0.82

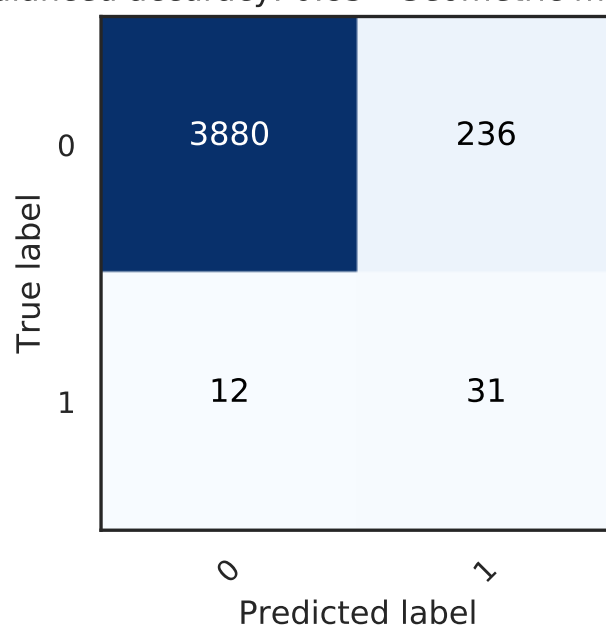


Figure C.22: Confusion matrix for SPLpred version 4 at 0.66 threshold

Tool →	SecretomeP (Gram-)		SecretomeP (Gram+)		PeNGaRoo (Gram+)		LSPpred		SPLpred	
Protein	LSP	Score	LSP	Score	LSP	Score	LSP	Score	LSP	Score
HYGR	No	0.078	No	0.087	No	0.417	Yes	0.714	No	0.474
SAMS2	No	0.173	No	0.153	Yes	0.575	Yes	0.839	Yes	0.716
MTD	No	0.108	No	0.086	Yes	0.626	Yes	0.789	No	0.666
Helja	Yes	0.55	No	0.402	Yes	0.582	Yes	0.690	Yes	0.941

Figure C.23: Prediction scores from bacterial programs on the best candidate LSPs in plants; (HYG^R), SAMS2, MDH and Helja. The tools are predictors for LSPs developed in this work (LSPpred, SPLpred), the bacterial versions SecretomeP, and a Gram-positive predictor for UPS (PeNGaRoo).

Appendix D

Appendix

Table D.1: Consensus proteins between LSPpred and SPLpred in results the of applying LSP workflow, excluding those with SP or TM, to the UniProt curated subset of rice proteins.

Accession	Description
LOC_Os07g44430	1-Cys peroxiredoxin A ()
LOC_Os02g38340	Actin-related protein 3 (ARP3)
LOC_Os11g10480	Alcohol dehydrogenase 1 (ADH1)
LOC_Os11g10510	Alcohol dehydrogenase 2 (ADH2)
LOC_Os02g57040	Alcohol dehydrogenase class-3 ()
LOC_Os04g02780	Amidase 1 (AMI1)
LOC_Os01g19450	ATP-citrate synthase beta chain protein 1 (ACLB-1)
LOC_Os12g04240	Calcium-binding protein CBP (CBP)
LOC_Os02g50240	Glutamine synthetase cytosolic isozyme 1-1 (GLN1-1)
LOC_Os03g12290	Glutamine synthetase cytosolic isozyme 1-2 (GLN1-2)
LOC_Os03g50490	Glutamine synthetase cytosolic isozyme 1-3 (GLN1-3)
LOC_Os04g56400	Glutamine synthetase, chloroplastic (GLN2)
LOC_Os08g03290	Glyceraldehyde-3-phosphate dehydrogenase 1, cytosolic (GAPC1)
LOC_Os01g51170	Mitochondrial outer membrane protein porin 4 (VDAC4)
LOC_Os03g10510	Mitochondrial outer membrane protein porin 5 (VDAC5)
LOC_Os09g39380	Monodehydroascorbate reductase 3, cytosolic (MDAR3)
LOC_Os03g57680	Probable aldehyde oxidase 3 ()
LOC_Os07g18154	Probable aldehyde oxidase 4 ()

Accession	Description
LOC_Os10g11810	Probable cinnamyl alcohol dehydrogenase 1 (CAD1)
LOC_Os08g16910	Probable cinnamyl alcohol dehydrogenase 5 (CAD5)
LOC_Os09g23530	Probable cinnamyl alcohol dehydrogenase 8A (CAD8A)
LOC_Os09g23560	Probable cinnamyl alcohol dehydrogenase 8D (CAD8D)
LOC_Os03g12270	Probable cinnamyl alcohol dehydrogenase 9 (CAD9)
LOC_Os03g18810	Probable LL-diaminopimelate aminotransferase, chloroplastic (AGD2)
LOC_Os03g60090	Probable methylenetetrahydrofolate reductase ()
LOC_Os02g17650	Probable tocopherol cyclase, chloroplastic (VTE1)
OsJ_20711	Putative 12-oxophytodienoate reductase 11 (OPR11)
LOC_Os08g09200	Putative aconitate hydratase, cytoplasmic ()
LOC_Os05g39310	Pyruvate decarboxylase 1 (PDC1)
LOC_Os03g18220	Pyruvate decarboxylase 2 (PDC2)
LOC_Os07g49250	Pyruvate decarboxylase 3 (PDC3)
LOC_Os02g42230	Replication protein A 32 kDa subunit B (RPA2B)
LOC_Os06g47830	Replication protein A 32 kDa subunit C (RPA2C)
LOC_Os03g21040	Ricin B-like lectin R40C1 (R40C1)
LOC_Os07g48500	Ricin B-like lectin R40G3 (R40G3)
LOC_Os04g42090	S-adenosylmethionine decarboxylase proenzyme (SAMDC)
LOC_Os05g04510	S-adenosylmethionine synthase 1 (SAM1)
LOC_Os01g22010	S-adenosylmethionine synthase 2 (SAM2)
LOC_Os01g18860	S-adenosylmethionine synthase 3 (METK3)
LOC_Os01g24710	Salt stress-induced protein (SALT)
LOC_Os03g22810	Superoxide dismutase [Cu-Zn] 1 (SODCC1)
LOC_Os07g46990	Superoxide dismutase [Cu-Zn] 2 (SODCC2)
LOC_Os07g38730	Tubulin alpha-1 chain (TUBA1)
LOC_Os11g14220	Tubulin alpha-2 chain (TUBA)
LOC_Os09g35800	UDP-glucose 4-epimerase 3 (UGE-3)
LOC_Os02g45480	Zinc finger CCCH domain-containing protein 17 ()

Accession	Description
LOC_Os01g15460	Zinc finger CCCH domain-containing protein 6 ()

Table D.2: LSPblast results on proteins predicted solely by ApoplastP

Accession	Protein names	LSPblast Hits	Observed	Notes
LOC_Os01g03840	Zinc finger protein STAMENLESS 1 (SL1)	6	0	
LOC_Os01g14870	Zinc finger CCCH domain-containing protein 3	7	0	
LOC_Os01g16890	60S ribosomal protein L30 (RPL30)	3	66.67	
LOC_Os01g56040	Zinc finger A20 and AN1 domain-containing stress-associated protein 3 (SAP3)	19	0	
LOC_Os01g66590	LOB domain-containing protein 6 (LBD6)	49	0	
LOC_Os01g69830	Squamosa promoter-binding-like protein 2 (SPL2)	28	0	
LOC_Os02g17380	Protein FERTILITY RESTORER RF2, mitochondrial (RF2)	0	0	
LOC_Os02g32840	Zinc finger A20 and AN1 domain-containing stress-associated protein 5 (SAP5)	19	0	
LOC_Os02g49440	Dof zinc finger protein 4 (DOF4)	52	0	
LOC_Os03g05510	LOB domain-containing protein CRL1 (CRL1)	51	0	
LOC_Os03g30430	Protein SPIRAL1-like 1	12	0	
LOC_Os03g46650	Guanine nucleotide-binding protein subunit beta (RGB1)	292	6.85	
LOC_Os04g36670	Protein AUXIN-REGULATED GENE INVOLVED IN ORGAN SIZE (ARGOS)	0	0	TM
LOC_Os04g56780	WUSCHEL-related homeobox 1 (WOX1)	20	0	
LOC_Os05g46370	Transcription factor LATE FLOWERING (LF)	119	0	
LOC_Os05g50270	GATA transcription factor 15 (GATA15)	31	0	

Accession	Protein names	LSPblast Hits	Observed	Notes
LOC.Os06g03800	Probable E3 ubiquitin-protein ligase XBOS36 (XBOS36)	30	40	
LOC.Os06g45230	Pollen allergen Ory s 2-A	0	0	
LOC.Os06g51330	Photosystem II stability/assembly factor HCF136, chloroplastic (HCF136)	2	0	
LOC.Os07g04670	Protein G1 (G1)	11	0	
LOC.Os08g36140	Protein PHOTOSYSTEM I ASSEMBLY 2, chloroplastic (PSA2)	2	100	
LOC.Os08g44770	Superoxide dismutase [Cu-Zn], chloroplastic (SODCP)	5	40	
LOC.Os09g31200	Zinc finger A20 and AN1 domain-containing stress-associated protein 1 (SAP1)	19	0	
LOC.Os10g20250	CASP-like protein 5A2	0	0	TM
LOC.Os11g26790	Water stress-inducible protein Rab21 (RAB21)	6	50	
LOC.Os11g41150	Protein SPIRAL1-like 2	12	0	
LOC.Os11g42960	CASP-like protein 1U3	0	0	TM
LOC.Os11g42970	CASP-like protein 1U2	0	0	TM
LOC.Os11g47809	Metallothionein-like protein 1A (MT1A)	3	0	
LOC.Os12g04270	Cytochrome c-type biogenesis ccda-like chloroplastic protein 2 (CCDA2)	0	0	TM
Os07g0472200	Protein LOL1 (LOL1)	9	0	

Table D.3: Significantly over-represented PTMs for each LSPDB classification, for PTMs occurring in more than 10 proteins. PTMs were labelled according to the UniProt controlled vocabulary.

Post translational modification	SEC	SPT	UNCLASS	NONSEC
Disulfide bond	✓	✓		

Post translational modification	SEC	SPT	UNCLASS	NONSEC
Glycoprotein	✓	✓		
GPI-anchor	✓	✓		
GPI-anchor amidated serine	✓	✓		
Lipoprotein	✓			✓
Pyrrolidone carboxylic acid	✓	✓		
Proteoglycan	✓	✓		
Zymogen	✓			
Hydroxylation	✓	✓		
GPI-anchor amidated asparagine	✓	✓		
GPI-anchor amidated glycine	✓			
4-hydroxyproline	✓	✓		
Hydroxyproline		✓		
Sulfation		✓		
Sulfotyrosine		✓		
Acetylation			✓	
Glycyl lysine isopeptide			✓	
Isopeptide bond			✓	
N-acetylalanine			✓	
interchain with G-Cter in ubiquitin			✓	
Glycyl lysine isopeptide (Lys-Gly)			✓	
N6-acetyllysine			✓	
Pyridoxal phosphate			✓	✓
N6-(pyridoxal phosphate)lysine			✓	✓
Phosphohistidine			✓	
Cysteine sulfenic acid (-SOH)			✓	
Lipoyl			✓	
N6-lipoyllysine			✓	
N-acetyls erine			✓	
Pyruvate			✓	

Post translational modification	SEC	SPT	UNCLASS	NONSEC
N6-methyllysine			✓	
Phosphoserine			✓	✓
N6,N6-dimethyllysine			✓	
N-acetylmethionine			✓	
Methylation			✓	✓
FAD			✓	✓
Methionine sulfoxide			✓	✓
N6,N6,N6-trimethyllysine			✓	
Nucleotide-binding			✓	✓
Myristate				✓
N-myristoyl glycine				✓
Prenylation				✓
Ubl conjugation				✓
S-geranylgeranyl cysteine				✓
Palmitate				✓
Cysteine methyl ester				✓
Phosphocysteine				✓
FMN				✓
S-farnesyl cysteine				✓
Cysteine persulfide				✓
Chromophore				✓
Phosphoprotein				✓
S-palmitoyl cysteine				✓

Table D.4: UNCLASSIFIED proteins with post-translational modifications identified in SEC proteins

Accession	Name	Symbol	LSPpred Score	SPLpred Score
AT4G04640	ATPase	ATPC1	False 0.555250	False 0.279083

Accession	Name	Symbol	LSPpred Score	SPLpred Score
AT3G11630	Thioredoxin superfamily protein	AT3G11630	False 0.662917	False 0.583250
AT5G06290	2-cysteine peroxiredoxin B	2-Cys Prx B	True 0.759333	False 0.695238
AT4G23650	calcium-dependent protein kinase 6	CDPK6	False 0.602083	False 0.209321
AT4G18480	P-loop containing nucleoside triphosphate hydrolases superfamily protein	CHLI1	False 0.642000	False 0.344625
AT5G45930	magnesium chelatase i2	CHLI2	False 0.608250	False 0.189167
AT5G55990	calcineurin B-like protein 2	CBL2	False 0.443000	False 0.388417
AT3G62410	CP12 domain-containing protein 2	CP12-2	False 0.444167	False 0.403095
AT4G18930	RNA ligase/cyclic nucleotide phosphodiesterase family protein	AT4G18930	False 0.420333	False 0.423417
AT3G17240	lipoamide dehydrogenase 2	mtLPD2	True 0.866667	False 0.554250
AT3G54050	high cyclic electron flow 1	HCEF1	True 0.760869	False 0.669917
AT4G28730	Glutaredoxin family protein	GrxC5	False 0.494619	True 0.745917
AT3G24170	glutathione-disulfide reductase	GR1	True 0.844500	False 0.470250
AT3G54660	glutathione reductase	GR	True 0.845000	False 0.690750
AT4G18040	eukaryotic translation initiation factor 4E	EIF4E	False 0.562833	False 0.499500
AT5G35620	Eukaryotic initiation factor 4E protein	LSP1	False 0.517083	False 0.396583
AT5G56950	nucleosome assembly protein 1;3	NAP1;3	False 0.325667	False 0.346500
AT5G47890	NADH-ubiquinone oxidoreductase B8 subunit	AT5G47890	False 0.288750	False 0.471917

Accession	Name	Symbol	LSPpred Score	SPLpred Score
AT3G45780	phototropin 1	PHOT1	True	0.706750
AT4G16155	dihydrolipoamide dehydrogenase	AT4G16155	True	0.851500
AT5G13120	cyclophilin 20-2	Pnsl5	False	0.630333
AT3G52960	Thioredoxin superfamily protein	AT3G52960	False	0.677250
AT3G06050	peroxiredoxin IIF	PRXIIF	True	0.697500
AT3G46830	RAB GTPase homolog A2C	RABA2c	False	0.620417
AT4G35860	GTP-binding 2	GB2	False	0.641333
AT4G17170	RAB GTPase homolog B1C	RABB1C	True	0.680083
AT3G46060	RAB GTPase homolog 8A	RABE1c	False	0.535250
AT3G18820	RAB GTPase homolog G3F	RAB7B	False	0.482167
AT5G15650	reversibly glycosylated polypeptide 2	RGP2	False	0.509167
AT3G08900	reversibly glycosylated polypeptide 3	RGP3	False	0.535917
AT5G16510	Alpha-1	RGP5	False	0.644583
AT3G55800	sedoheptulose-bisphosphatase	SBPASE	True	0.796250
AT5G18100	copper/zinc superoxide dismutase 3	CSD3	False	0.640750
AT5G42850	Thioredoxin superfamily protein	AT5G42850	False	0.265000
AT3G16990	heme oxygenase-like	AT3G16990	False	0.519167
AT3G46560	Tim10/DDP family zinc finger protein	TIM9	False	0.357167
AT4G26160	atypical CYS HIS rich thioredoxin 1	ACHT1	False	0.372417

Accession	Name	Symbol	LSPpred Score		SPLpred Score	
AT4G29670	atypical CYS HIS rich thioredoxin 2	ACHT2	False	0.453911	False	0.527500
AT4G35460	NADPH-dependent thioredoxin reductase B	NTRB	True	0.846500	False	0.675167
AT5G16400	thioredoxin F2	TRXF2	False	0.639083	True	0.835750
AT3G51030	thioredoxin H-type 1	TRX1	False	0.558417	False	0.547667
AT5G39950	thioredoxin 2	TRX2	False	0.562083	False	0.407917
AT5G42980	thioredoxin 3	TRX3	True	0.807417	False	0.532000
AT3G15360	thioredoxin M-type 4	TRX-M4	False	0.565667	False	0.484500

Table D.5: LSPDB proteins pairs that co-occur in the same experiment, and co-express within the top 100 mutual rank,

Edge	Descriptions		Symbols	Weight
AT1G75040 AT2G14610	(-) pathogenesis-related gene 5	(-) pathogenesis-related gene 1	PR5 (-) PR1	6
AT4G28520 AT5G44120	(-) cruciferin 3	(-) RmlC-like cupins superfamily protein	CRU3 (-) CRA1	6
AT3G45970 AT4G38400	(-) expansin-like A1	(-) expansin-like A2	EXLA1 (-) EXLA2	4
AT4G27160 AT4G27170	(-) seed storage albumin 3	(-) seed storage albumin 4	SESA3 (-) SESA4	4
AT3G13920 AT1G54270	(-) eukaryotic translation initiation factor 4A1	(-) eif4a-2	EIF4A1 (-) EIF4A-2	3
AT1G20960 AT1G80070	(-) U5 small nuclear ribonucleoprotein helicase, putative	(-) Pre-mRNA-processing-splicing factor	emb1507 (-) SUS2	3

Edge		Descriptions	Symbols	Weight
AT4G14680	(-)	Pseudouridine synthase/archaeosine	APS3 (-) APS1	3
AT3G22890		transglycosylase-like family protein (-) ATP sulfurylase 1		
AT4G35260	(-)	isocitrate dehydrogenase 1 (-) isocitrate dehy-	IDH1 (-) IDH2	2
AT2G17130		drogenase subunit 2		
AT2G04030	(-)	Chaperone protein htpG family protein (-)	CR88 (-) Hsp89.1	2
AT3G07770		HEAT SHOCK PROTEIN 89.1		
AT2G05710	(-)	aconitase 3 (-) aconitase 2	ACO3 (-) ACO2	2
AT4G26970				
AT3G29360	(-)	UDP-glucose 6-dehydrogenase family protein		2
AT5G15490		(-) UDP-glucose 6-dehydrogenase family pro- tein		
AT1G25490	(-)	ARM repeat superfamily protein (-) protein	RCN1 (-) PP2AA3	2
AT1G13320		phosphatase 2A subunit A3		
AT1G02500	(-)	S-adenosylmethionine synthetase 1 (-) S-	SAM1 (-) SAM-2	2
AT4G01850		adenosylmethionine synthetase 2		
AT2G30870	(-)	glutathione S-transferase PHI 10 (-) glutathione	GSTF10 (-) GSTF9	2
AT2G30860		S-transferase PHI 9		
AT3G09840	(-)	cell division cycle 48 (-) ATPase, AAA-type,		2
AT5G03340		CDC48 protein		
AT4G18480	(-)	P-loop containing nucleoside triphosphate hy-	CHLI1 (-) CHLI2	2
AT5G45930		drolases superfamily protein (-) magnesium chelataase i2		
AT1G52050	(-)	Mannose-binding lectin superfamily protein (-)		2
AT1G52060		Mannose-binding lectin superfamily protein		
AT1G74000	(-)	strictosidine synthase 3 (-) Calcium-dependent		2
AT1G74010		phosphotriesterase superfamily protein		
AT5G16050	(-)	general regulatory factor 5 (-) general regula-	GRF5 (-) GRF7	2
AT3G02520		tory factor 7		

Edge		Descriptions	Symbols	Weight
AT5G10450	(-)	G-box regulating factor 6 (-) general regulatory	GRF6 (-) GRF8	2
AT5G65430		factor 8		
AT2G33040	(-)	gamma subunit of Mt ATP synthase (-) phos-	ATP3 (-) PHT3;1	2
AT5G14040		phate transporter 3;1		
AT5G13980	(-)	Glycosyl hydrolase family 38 protein (-) Glyco-		2
AT3G26720		syl hydrolase family 38 protein		
AT2G38530	(-)	lipid transfer protein 2 (-) lipid transfer protein	LTP2 (-) LP1	2
AT2G38540		1		
AT2G43570	(-)	chitinase, putative (-) basic chitinase	CHI (-) HCHIB	2
AT3G12500				
AT2G28790	(-)	Pathogenesis-related thaumatin superfamily		2
AT1G73620		protein (-) Pathogenesis-related thaumatin su-		
		perfamily protein		
AT1G49240	(-)	actin 8 (-) actin 2	ACT8 (-) ACT2	2
AT3G18780				
AT1G77120	(-)	alcohol dehydrogenase 1 (-) pyruvate	ADH1 (-) PDC2	2
AT5G54960		decarboxylase-2		
AT2G40290	(-)	Eukaryotic translation initiation factor 2 sub-		2
AT5G05470		unit 1 (-) eukaryotic translation initiation factor		
		2 alpha subunit		
AT5G53480	(-)	ARM repeat superfamily protein (-) ARM re-		2
AT3G59020		peat superfamily protein		
AT4G23630	(-)	VIRB2-interacting protein 1 (-) VIRB2-	BTI1 (-) BTI2	2
AT4G11220		interacting protein 2		
AT2G19760	(-)	profilin 1 (-) profilin 2	PRF1 (-) PFN2	2
AT4G29350				
AT4G34870	(-)	rotamase cyclophilin 5 (-) rotamase CYP 3	ROC5 (-) ROC3	2
AT2G16600				

Edge		Descriptions	Symbols	Weight
AT4G14360	(-)	S-adenosyl-L-methionine-dependent methyl-		2
AT1G04430		transferases superfamily protein (-) S-adenosyl-		
		L-methionine-dependent methyltransferases		
		superfamily protein		
AT1G15750	(-)	Transducin family protein / WD-40 repeat fam-	TPL (-) TPR2	2
AT3G16830		ily protein (-) TOPLESS-related 2		
AT5G42970	(-)	Proteasome component (PCI) domain protein (-	COP8 (-) COP13	2
AT5G14250		) Proteasome component (PCI) domain protein		
AT4G02150	(-)	ARM repeat superfamily protein (-) importin	MOS6 (-) IMPA-6	2
AT1G02690		alpha isoform 6		
AT2G32920	(-)	PDI-like 2-3 (-) PDI-like 2-2	PDIL2-3 (-) PDIL2-2	2
AT1G04980				
AT5G50850	(-)	Transketolase family protein (-) Thiamin	MAB1 (-) IAR4	2
AT1G24180		diphosphate-binding fold (THDP-binding)		
		superfamily protein		
AT1G78900	(-)	vacuolar ATP synthase subunit A (-) vacuolar	VHA-A (-) DET3	2
AT1G12840		ATP synthase subunit C (VATC) / V-ATPase		
		C subunit / vacuolar proton pump C subunit		
		(DET3)		
AT3G61110	(-)	ribosomal protein S27 (-) Zinc-binding riboso-		2
AT5G47930		mal protein family protein		
AT3G02090	(-)	Insulinase (Peptidase family M16) protein (-)		2
AT1G51980		Insulinase (Peptidase family M16) protein		
AT3G12780	(-)	phosphoglycerate kinase 1 (-) Phosphoglycer-		2
AT1G56190		ate kinase family protein		
AT3G52930	(-)	Aldolase superfamily protein (-) Aldolase su-		2
AT2G01140		perfamily protein		

Edge	Descriptions		Symbols	Weight
AT3G17390	(-)	S-adenosylmethionine synthetase family pro-	MTO3 (-) MAT3	2
AT2G36880		tein (-) methionine adenosyltransferase 3		

Table D.6: LSPDB proteins that co-occur in the same experiment, and co-express within the top 100 mutual rank

Accession	Description	Cluster
AT1G04270	cytosolic ribosomal protein S15	A
AT1G14320	Ribosomal protein L16p/L10e family protein	A
AT1G26880	Ribosomal protein L34e superfamily protein	A
AT1G34030	Ribosomal protein S13/S18 family	A
AT1G48830	Ribosomal protein S7e family protein	A
AT1G50920	Nucleolar GTP-binding protein	A
AT1G56110	homolog of nucleolar protein NOP56	A
AT1G61580	R-protein L3 B	A
AT1G66580	senescence associated gene 24	A
AT1G67430	Ribosomal protein L22p/L17e family protein	A
AT1G72370	40s ribosomal protein SA	A
AT1G77940	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein	A
AT2G01250	Ribosomal protein L30/L7 family protein	A
AT2G17360	Ribosomal protein S4 (RPS4A) family protein	A
AT2G20450	Ribosomal protein L14	A
AT2G34480	Ribosomal protein L18ae/LX family protein	A
AT2G36160	Ribosomal protein S11 family protein	A
AT2G44120	Ribosomal protein L30/L7 family protein	A

Accession	Description	Cluster
AT2G47610	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein	A
AT3G02560	Ribosomal protein S7e family protein	A
AT3G03920	H/ACA ribonucleoprotein complex, subunit Gar1/Naf1 protein	A
AT3G05060	NOP56-like pre RNA processing ribonucleoprotein	A
AT3G05560	Ribosomal L22e protein family	A
AT3G07050	GTP-binding family protein	A
AT3G09200	Ribosomal protein L10 family protein	A
AT3G09630	Ribosomal protein L4/L1 family	A
AT3G11250	Ribosomal protein L10 family protein	A
AT3G11510	Ribosomal protein S11 family protein	A
AT3G12860	NOP56-like pre RNA processing ribonucleoprotein	A
AT3G16810	pumilio 24	A
AT3G18740	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein	A
AT3G23390	Zinc-binding ribosomal protein family protein	A
AT3G25520	ribosomal protein L5	A
AT3G49010	breast basic conserved 1	A
AT3G49910	Translation protein SH3-like family protein	A
AT3G52140	tetratricopeptide repeat (TPR)-containing protein	A
AT3G53430	Ribosomal protein L11 family protein	A

Accession	Description	Cluster
AT3G57150	homologue of NAP57	A
AT3G62870	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein	A
AT4G10450	Ribosomal protein L6 family	A
AT4G17390	Ribosomal protein L23/L15e family protein	A
AT4G25630	fibrillarin 2	A
AT4G27090	Ribosomal protein L14	A
AT4G29390	Ribosomal protein S30 family protein	A
AT4G31700	ribosomal protein S6	A
AT4G34670	Ribosomal protein S3Ae	A
AT4G39200	Ribosomal protein S25 family protein	A
AT5G07090	Ribosomal protein S4 (RPS4A) family protein	A
AT5G09510	Ribosomal protein S19 family protein	A
AT5G10360	Ribosomal protein S6e	A
AT5G15200	Ribosomal protein S4	A
AT5G15520	Ribosomal protein S19e family protein	A
AT5G16130	Ribosomal protein S7e family protein	A
AT5G19510	Translation elongation factor EF1B/ribosomal protein S6 family protein	A
AT5G20290	Ribosomal protein S8e family protein	A
AT5G22330	P-loop containing nucleoside triphosphate hydrolases superfamily protein	A

Accession	Description	Cluster
AT5G27850	Ribosomal protein L18e/L15 superfamily protein	A
AT5G28060	Ribosomal protein S24e family protein	A
AT5G35620	Eukaryotic initiation factor 4E protein	A
AT5G39740	ribosomal protein L5 B	A
AT5G52470	fibrillarin 1	A
AT5G56710	Ribosomal protein L31e family protein	A
AT5G59850	Ribosomal protein S8 family protein	A
AT1G14980	chaperonin 10	B
AT1G16030	heat shock protein 70B	B
AT1G74310	heat shock protein 101	B
AT2G33210	heat shock protein 60-2	B
AT2G36530	Enolase	B
AT2G44350	Citrate synthase family protein	B
AT3G09440	Heat shock protein 70 (Hsp 70) family protein	B
AT3G12580	heat shock protein 70	B
AT3G23990	heat shock protein 60	B
AT3G44110	DNAJ homologue 3	B
AT4G02930	GTP binding Elongation factor Tu family protein	B
AT4G23570	phosphatase-related	B

Accession	Description	Cluster
AT4G37910	mitochondrial heat shock protein 70-1	B
AT5G02490	Heat shock protein 70 (Hsp 70) family protein	B
AT5G02500	heat shock cognate protein 70-1	B
AT5G09590	mitochondrial HSO70 2	B
AT5G26860	lon protease 1	B
AT5G52640	heat shock protein 90.1	B
AT5G56030	heat shock protein 81-2	B
AT1G45000	AAA-type ATPase family protein	C
AT1G51710	ubiquitin-specific protease 6	C
AT1G64520	regulatory particle non-ATPase 12A	C
AT2G20580	26S proteasome regulatory subunit S2 1A	C
AT3G05530	regulatory particle triple-A ATPase 5A	C
AT4G29040	regulatory particle AAA-ATPase 2A	C
AT5G38470	Rad23 UV excision repair protein family	C
AT5G43010	regulatory particle triple-A ATPase 4A	C
AT5G58290	regulatory particle triple-A ATPase 3	C
AT1G07790	Histone superfamily protein	D
AT1G08880	Histone superfamily protein	D
AT1G51060	histone H2A 10	D

Accession	Description	Cluster
AT2G37470	Histone superfamily protein	D
AT4G27230	histone H2A 2	D
AT5G54640	Histone superfamily protein	D
AT5G59870	histone H2A 6	D
AT1G04820	tubulin alpha-4 chain	E
AT1G20010	tubulin beta-5 chain	E
AT1G75780	tubulin beta-1 chain	E
AT4G14960	Tubulin/FtsZ family protein	E
AT5G12250	beta-6 tubulin	E
AT5G23860	tubulin beta 8	E
AT5G44340	tubulin beta chain 4	E
AT1G16470	proteasome subunit PAB1	F
AT1G56450	20S proteasome beta subunit G1	F
AT3G22110	20S proteasome alpha subunit C1	F
AT3G51260	20S proteasome alpha subunit PAD1	F
AT3G60820	N-terminal nucleophile aminohydrolases (Ntn hydrolases) superfamily protein	F
AT4G14800	20S proteasome beta subunit D2	F
AT4G31300	N-terminal nucleophile aminohydrolases (Ntn hydrolases) superfamily protein	F
AT3G11830	TCP-1/cpn60 chaperonin family protein	G

Accession	Description	Cluster
AT3G18190	TCP-1/cpn60 chaperonin family protein	G
AT3G20050	T-complex protein 1 alpha subunit	G
AT5G16070	TCP-1/cpn60 chaperonin family protein	G
AT5G20890	TCP-1/cpn60 chaperonin family protein	G
AT5G26360	TCP-1/cpn60 chaperonin family protein	G
AT1G55490	chaperonin 60 beta	H
AT2G28000	chaperonin-60alpha	H
AT3G13470	TCP-1/cpn60 chaperonin family protein	H
AT5G20720	chaperonin 20	H
AT2G42790	citrate synthase 3	I
AT3G06860	multifunctional protein 2	I
AT4G29010	Enoyl-CoA hydratase/isomerase family	I
AT1G64190	6-phosphogluconate dehydrogenase family protein	J
AT3G47520	malate dehydrogenase	J
AT5G41670	6-phosphogluconate dehydrogenase family protein	J
AT1G07370	proliferating cellular nuclear antigen 1	K
AT2G29570	proliferating cell nuclear antigen 2	K
AT5G10390	Histone superfamily protein	K
AT1G62020	Coatomer, alpha subunit	L

Accession	Description	Cluster
AT2G21390	Coatomer, alpha subunit	L
AT5G05010	clathrin adaptor complexes medium subunit family protein	L
AT1G02640	beta-xylosidase 2	M
AT5G49360	beta-xylosidase 1	M
AT5G64570	beta-D-xylosidase 4	M
AT1G03860	prohibitin 2	N
AT3G27280	prohibitin 4	N
AT5G40770	prohibitin 3	N
AT3G06720	importin alpha isoform 1	O
AT4G16143	importin alpha isoform 2	O
AT5G19820	ARM repeat superfamily protein	O
AT2G01720	Ribophorin I	P
AT5G61790	calnexin 1	P
AT5G66680	dolichyl-diphosphooligosaccharide-protein glycosyltransferase 48kDa subunit family protein	P

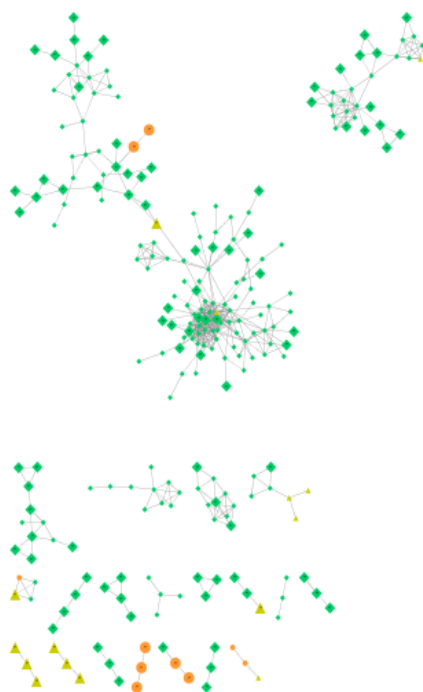


Figure D.1: Clusters of 3 or more proteins identified from the LSPDB that co-occur in the same experiment and are co-expressed with either RNA-seq or microarray

Appendix E

Appendix

Table E.1: Revised LSP candidate list indicating accession, if included in the Ding et. al (2012) LSP list, whether considered an LSP in this Thesis, the LSPDB classification and confidence lists, SPLpred and LSPpred results, if protein exceeds abundance threshold and protein description

Accession	Ding List	Thesis LSP	LSPDB	Confidence Lists	SPLpred	LSPpred	High Abund.	Description
AT1G02930	Yes	Yes	UNCLASSIFIED	Low	No	No		Glutathione S-transferase F6
AT1G07890	Yes	Yes	UNCLASSIFIED	High	No	Yes	Yes	MEE6
AT1G08110	Yes	Yes	UNCLASSIFIED	Low	No	No		Lactoylglutathione lyase
AT1G08830	Yes	Yes	UNCLASSIFIED	Medium	Yes	Yes		Superoxide dismutase
AT1G11580	Yes	No	Secretory	n/a	No	No		methylesterase PCR A
AT1G12080	Yes	Yes	UNCLASSIFIED	None	No	No		Vacuolar calcium-binding protein-like protein

Accession	Ding List	Thesis LSP	LSPDB	Confidence Lists	SPLpred	LSPpred	High Abund.	Description
AT1G13440	Yes	Yes	UNCLASSIFIED	Low	No	Yes		Glyceraldehyde-3-phosphate dehydrogenase GAPC2, cytosolic
AT1G19570	Yes	Yes	UNCLASSIFIED	None	No	No		Glutathione S-transferase DHAR1, mitochondrial
AT1G23190	Yes	Yes	UNCLASSIFIED	Low	No	Yes		Probable phosphoglucomu- tase, cytoplasmic 1
AT1G23410	Yes	Yes	UNCLASSIFIED	None	No	No	Yes	Ubiquitin-40S ribosomal pro- tein S27a-1
AT1G25490	Yes	Yes	UNCLASSIFIED	Low	No	No		Serine/threonine-protein phosphatase 2A 65 kDa regulatory subunit A alpha isoform

Accession	Ding List	Thesis LSP	LSPDB	Confidence Lists	SPLpred	LSPpred High Abund.	Description
AT1G35720	Yes	Yes	UNCLASSIFIED	Low	No	No	Annexin
AT1G45145	Yes	Yes	UNCLASSIFIED	None	No	No	Thioredoxin H5
AT1G56070	Yes	Yes	UNCLASSIFIED	None	No	Yes	Elongation factor 2
AT1G61820	Yes	No	Secretory Isoform	n/a	Yes	No	Beta-glucosidase 46
AT1G62380	Yes	Yes	UNCLASSIFIED	Low	No	No	1-aminocyclopropane-1- carboxylate oxidase 2
AT1G63770	Yes	No	Secretory Isoform	n/a	No	No	Peptidase M1 family protein
AT1G78380	Yes	Yes	UNCLASSIFIED	Low	No	No	Glutathione S-transferase U19 (GSTU19)

Accession	Ding List	Thesis LSP	LSPDB	Confidence Lists	SPLpred	LSPpred	High Abund.	Description
AT1G79550	Yes	Yes	UNCLASSIFIED	None	No	Yes		Phosphoglycerate kinase
AT2G15130	Yes	No	Secretory Isoform	n/a	Yes	No		Plant basic secretory protein (BSP) family protein
AT2G16060	Yes	Yes	UNCLASSIFIED	Low	No	No		Non-symbiotic hemoglobin 1
AT2G24200	Yes	Yes	UNCLASSIFIED	Low	No	Yes		Leucine aminopeptidase 1
AT2G32520	Yes	Yes	UNCLASSIFIED	None	No	Yes		alpha/beta-Hydrolases superfamily protein
AT2G35635	Yes	Yes	UNCLASSIFIED	None	No	No	Yes	UBQ7
AT2G36460	Yes	Yes	UNCLASSIFIED	None	No	Yes		Fructose-bisphosphate al- dolase 6, cytosolic

Accession	Ding List	Thesis LSP	LSPDB	Confidence Lists	SPLpred	LSPpred	High Abund.	Description
AT2G36530	Yes	Yes	UNCLASSIFIED	Low	No	Yes		LOS2
AT3G01280	Yes	No	NONSEC	n/a	Yes	No		Mitochondrial outer mem- brane protein porin 1
AT3G02230	Yes	Yes	UNCLASSIFIED	Low	Yes	No		RGP1
AT3G03780	Yes	Yes	UNCLASSIFIED	None	No	Yes		5-methyltetra- hydropteroyltriglutamate-- homocysteine methyltrans- ferase 2
AT3G03910	Yes	Yes	UNCLASSIFIED	None	No	Yes		Glutamate dehydrogenase
AT3G04120	Yes	Yes	UNCLASSIFIED	Low	No	Yes		Glyceraldehyde-3-phosphate dehydrogenase GAPC1, cytosolic

Accession	Ding List	Thesis LSP	LSPDB	Confidence Lists	SPLpred	LSPpred	High Abund.	Description
AT3G06860	Yes	Yes	UNCLASSIFIED	Low	No	Yes		Peroxisomal fatty acid beta-oxidation multifunctional protein MFP2
AT3G08580	Yes	No	Secretory	n/a	No	Yes		ADP/ATP carrier 1 (AAC1)
AT3G08900	Yes	Yes	UNCLASSIFIED	None	Yes	No		Glycosyltransferase (Frag-ment)
AT3G10920	Yes	Yes	UNCLASSIFIED	Low	No	Yes		Superoxide dismutase
AT3G16410	Yes	Yes	UNCLASSIFIED	Low	Yes	Yes		Nitrile-specifier protein 4
AT3G16420	Yes	Yes	UNCLASSIFIED	Medium	Yes	Yes		PYK10-binding protein 1
AT3G16430	Yes	Yes	UNCLASSIFIED	None	Yes	Yes		PYK10-binding protein 2

Accession	Ding List	Thesis LSP	LSPDB	Confidence Lists	SPLpred	LSPpred	High Abund.	Description
AT3G16450	Yes	Yes	UNCLASSIFIED	None	Yes	Yes		Jacalin-related lectin 33
AT3G28940	Yes	Yes	UNCLASSIFIED	None	No	No		Protein AIG2 B
AT3G32980	Yes	No	SEC	n/a	Yes	No		Peroxidase 32
AT3G43810	Yes	Yes	UNCLASSIFIED	Low	No	No		Calmodulin 7
AT3G52930	Yes	Yes	UNCLASSIFIED	Low	No	Yes		Fructose-bisphosphate al- dolase
AT3G54660	Yes	Yes	UNCLASSIFIED	Low	No	Yes		glutathione reductase
AT3G55440	Yes	Yes	UNCLASSIFIED	Low	No	Yes	Yes	Triosephosphate isomerase, cytosolic

Accession	Ding List	Thesis LSP	LSPDB	Confidence Lists	SPLpred	LSPpred	High Abund.	Description
AT4G01850	Yes	Yes	UNCLASSIFIED	Low	Yes	Yes		S-adenosylmethionine syn- thase
AT4G09320	Yes	Yes	UNCLASSIFIED	None	Yes	No		Nucleoside diphosphate ki- nase
AT4G20260	Yes	Yes	UNCLASSIFIED	None	No	No	Yes	plasma-membrane associ- ated cation-binding protein 1
AT4G23590	Yes	Yes	UNCLASSIFIED	None	No	No		Probable aminotransferase TAT4
AT4G25100	Yes	Yes	UNCLASSIFIED	Low	No	No		Superoxide dismutase
AT4G29680	Yes	No	Secretory	n/a	Yes	No		Alkaline-phosphatase-like family protein

Accession	Ding List	Thesis LSP	LSPDB	Confidence Lists	SPLpred	LSPpred High Abund.	Description
AT4G30910	Yes	Yes	UNCLASSIFIED	None	No	Yes	Cytosol aminopeptidase family protein
AT4G30920	Yes	Yes	UNCLASSIFIED	Low	No	Yes	Leucine aminopeptidase 2, chloroplastic
AT5G07440	Yes	Yes	UNCLASSIFIED	None	No	Yes	Glutamate dehydrogenase
AT5G10450	Yes	Yes	UNCLASSIFIED	Medium	No	No	G-box regulating factor 6
AT5G15490	Yes	No	Secretory	n/a	No	Yes	UDP-glucose 6-dehydrogenase
AT5G15650	Yes	Yes	UNCLASSIFIED	Low	Yes	No	reversibly glycosylated polypeptide 2 (RGP2)

Accession	Ding List	Thesis LSP	LSPDB	Confidence Lists	SPLpred	LSPpred	High Abund.	Description
AT5G16510	Yes	Yes	UNCLASSIFIED	None	No	No		Glycosyltransferase (Frag- ment)
AT5G17920	Yes	Yes	UNCLASSIFIED	None	No	Yes	Yes	5- methyltetrahydropteroyltriglutamate- -homocysteine methyltrans- ferase 1
AT5G18100	Yes	Yes	UNCLASSIFIED	Low	Yes	No		Superoxide dismutase
AT5G18170	Yes	Yes	UNCLASSIFIED	Low	Yes	Yes		Glutamate dehydrogenase
AT5G19440	Yes	Yes	UNCLASSIFIED	None	No	No		NAD(P)-binding Rossmann- fold superfamily protein
AT5G53560	Yes	No	Secretory	n/a	No	No		CB5-E

Accession	Ding List	Thesis LSP	LSPDB	Confidence Lists	SPLpred	LSPpred	High Abund.	Description
AT5G55730	Yes	No	SEC	n/a	No	No		Fasciclin-like arabinogalactan protein 1
AT5G59880	Yes	Yes	UNCLASSIFIED	None	No	No		Actin-depolymerizing factor 3

