Advancing the detection of short tandem repeats in health and disease

Andreas Halman

ORCID: 0000-0001-5248-4121

Doctor of Philosophy

March 2021

The Florey Department of Neuroscience and Mental Health

Faculty of Medicine, Dentistry and Health Sciences

The University of Melbourne

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

Abstract

Short tandem repeats (STRs) are repetitive DNA sequences composed of repeat units with 2–6 base pairs in length. STRs make up around 3% of the human genome and have higher mutation rate than single nucleotide variants. Mutations in STRs loci have been linked to around 60 genetic disorders known today, the majority of them being neurological. Genotyping STRs has been difficult due to technical limitations but advances in DNA sequencing methods and the emergence of bioinformatic analysis tools have led to the discoveries of many new disease-causing loci. In this thesis, we focused on exploring different methods of STR analysis, their performance on high-throughput sequencing data and improving these existing methods.

At first, we compared four genome-wide STR genotyping tools and determined their performance on whole exome sequencing data. We found that all tools have their merits and there is no clear winner. Difference between tools were observed when genotyping some repeat types with some tools having significantly higher error rates on two base-pair repeats. Next, we characterised STRs in whole genome sequencing data to determine genotyping accuracy of different types of repeats in various genomic settings and library preparation methods. We saw differences in accuracy between repeat unit sizes, repeat lengths and repeat composition. This demonstrates the need to use customised filtering strategies for different repeats that would improve quality of calls while retaining as much data as possible, which could be helpful detecting *de novo* mutations.

Finally, we focused on known disease-causing STRs to improve and extend the current method of analysis. A comprehensive list of disease-causing STR loci was collated based on the literature. This was used to create a variant catalogue for a robust genotyping tool ExpansionHunter allowing genotyping of more than twice as many loci than it currently allows. In addition, it is now possible to genotype alleles longer than the fragment length in over forty loci. This catalogue was validated with an extensive series of simulations. We have used this updated catalogue on our sequencing dataset of individuals with rare disease to detect expansions in any of these loci. Moreover, we developed our own computational tool, which has been made publicly available, to greatly simplify STR genotyping in sequencing data and enable researchers to genotype the highest number of known pathogenic STR loci compared to other existing methods.

Declaration

This is to certify that

i. the thesis comprises my original work towards the degree of Doctor of Philosophy except where indicated in the preface;

ii. due acknowledgement has been made in the text to all other material used;

iii. the thesis is fewer than 100,000 words in length.

Andreas Halman

Preface

Collaborations

I am the primary author of all work presented in this thesis. I performed all the analysis, software development and drafted all aspects of this work. The work presented in the thesis was done in collaboration with my supervisor Alicia Oshlack who guided all projects, advised on analyses, helped preparing the manuscript presented in Chapter 2 and 4 as well as revised other chapters in this thesis. Egor Dolzhenko provided advice on some of the methodology in Chapter 4 and reviewed the manuscript.

Publication status

Chapter 2: Published by F1000Research on 23 March 2020.

Halman & Oshlack. (2020). Accuracy of short tandem repeats genotyping tools in whole exome sequencing data. F1000Research, *9*(200). https://doi.org/10.12688/f1000research.22639.1

Roles of Andreas Halman: conceptualization, formal analysis, methodology, software, validation, visualization, writing – original draft preparation, writing – review & editing.

Roles of Alicia Oshlack: conceptualization, funding acquisition, resources, supervision, writing – original draft preparation, writing – review & editing.

Chapter 3: Unpublished material not submitted for publication.

Chapter 4: Part of the chapter is published by bioRxiv on 14 June 2021 as an unrefereed preprint and submitted for publication.

Halman, A., Dolzhenko, E. & Oshlack, A. (2021). STRipy: a graphical application for enhanced genotyping of pathogenic short tandem repeats in sequencing data. bioRxiv. https://doi.org/10.1101/2021.06.13.448220

Roles for Andreas Halman: conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, validation, visualization, writing – original draft preparation, writing – review & editing.

Roles for Egor Dolzhenko: conceptualization, methodology, software, writing – review & editing.

Roles for Alicia Oshlack: conceptualization, formal analysis, supervision, writing – original draft preparation, writing – review & editing.

Funding

My candidature was supported by the Melbourne Research Scholarship and The Florey Department of Neuroscience and Mental Health top-up stipend.

Acknowledgments

First and foremost, my deepest gratitude goes to my supervisor Alicia Oshlack, who has trained and guided me through this complex field of research. These years were full of growth and I learned a whole new set of skills with your help that are highly valuable for me, thank you for this. It has been my pleasure to be a member of your wonderful group and to work side-by-side with such bright minds in a top-class research facility. Thanks to all past and present members of the Oshlack lab at the Peter MacCallum Cancer Centre and the Murdoch Children's Research Institute for your help, especially Katrina and Simon for reading my work and giving feedback, as well as providing the samples I have worked on in this thesis. Thanks to Harriet, Sam, Breon, Andrew and others for your help and company, and my work-neighbour Caroline in MCRI for your everlasting shine and the positivity you brought in to work.

To my supervisor Anthony Hannan, thank you for providing all the advice, help and flexibility I needed at the early stage of my PhD. I really appreciate this and I would not have reached far without it. Thanks to all members of the Hannan lab at the Florey Institute of Neuroscience and Mental Health whom I had the joy to work during my studies. I am grateful to you, Terence, for paving the road for my PhD. Thank you, Shan and Shlomo for the fun environment you created and all the lengthy chats we had.

I am thankful to my advisory committee members Vicky, Noel and Simon for their guidance and tips they shared with me. Many thanks to Egor for the great collaboration and your leading work in the field of STRs analysis. Thank you, Jacquie for helping me to sort out all the issues a student can possibly encounter. Thank you to my reviewers for going through all my writings and providing feedback. Thank you, University of Melbourne and Florey for funding my studies which allowed me to pursue my dream.

Apart from all my wonderful colleagues, I am utterly grateful for the people closest to me whom I could share my lows and celebrate my highs at all times. That is, my loving family, including my always supportive parents, dearest friends and my cat 猫娘娘.

Finally, thank you to all friendly Melburnians and the shining Sun for brightening my days, and the excellent coffee around here to keep me awake during the gloomy ones.

Table of contents

ABSTRACT **3**

DECLARATION **4**

PREFACE **5**

ACKNOWLEDGMENTS **6**

LIST OF TABLES **10**

LIST OF FIGURES **11**

CHAPTER 1: INTRODUCTION **14**

1.1 GENETIC VARIATION AND CLINICAL DIAGNOSIS FROM SEQUENCING DATA 14

1.1.1 HUMAN GENETICS 14

1.1.2 TYPES OF GENETIC VARIATION 15

1.1.3 DETECTION OF VARIANTS FROM SEQUENCING DATA 17

1.1.3.1 History of sequencing methods development 17

1.1.3.2 Types of high-throughput sequencing 18

1.1.3.3 Methods of high-throughput sequencing 19

1.1.3.4 Variant calling from high-throughput sequencing data 20

1.2 SHORT TANDEM REPEATS AS A SOURCE OF GENETIC VARIATION 23

1.2.1 OVERVIEW 23

1.2.2 DISTRIBUTION 24

1.2.3 MUTATIONS 25

1.2.3.1 Mutation models 25

1.2.3.2 Mutation mechanisms 25

1.2.3.3 Magnitude of mutation and mutation rates 27

1.3 IMPORTANCE OF SHORT TANDEM REPEATS 27

1.3.1 REPEATS IN CODING REGIONS 28

1.3.1.1 Polyglutamine disorders 28

1.3.1.2 Polyalanine disorders 29

1.3.2 REPEATS IN NON-CODING REGIONS 30

1.3.3 SOMATIC MOSAICISM OF SHORT TANDEM REPEATS 31

1.4 METHODS OF SHORT TANDEM REPEAT ANALYSIS 32

1.5 AIMS AND OUTLINE OF THE THESIS 35

CHAPTER 2: ACCURACY OF SHORT TANDEM REPEATS GENOTYPING TOOLS IN WHOLE EXOME SEQUENCING DATA **37**

2.1 INTRODUCTION 38

2.1.1 OVERVIEW OF SHORT TANDEM REPEATS (STRS) AND METHODS OF ANALYSIS 38

2.1.2 COMPUTATIONAL TOOLS TO GENOTYPE STRS FROM HTS DATA 40

2.1.2.1 LobSTR 41
2.1.2.2 RepeatSeq 42
2.1.2.3 GangSTR 42
2.1.2.4 HipSTR 43
2.1.2.5 GATK HaplotypeCaller 43
2.2 RESULTS 44
2.2.1 DATASET 44
2.2.2 EFFECTS OF REPEAT UNIT AND COVERAGE ON ACCURACY 47
2.2.3 EFFECTS OF REPEAT UNIT AND QUALITY SCORES ON ACCURACY 51
2.2.4 ACCURACY OF GANGSTR BY LOOKING AT ONLY THE ENCLOSING CLASS OF READS 53
2.2.5 RUNNING TIME OF TOOLS 55
2.3 DISCUSSION 56
2.4 METHODS 59
2.4.1 DATASET 59
2.4.2 COMPUTATIONAL STEPS 59
2.4.2.1 SRA to FASTQ conversion 59
2.4.2.2 Defining STR regions 60
2.4.2.3 FASTQ alignment and calculating BAM coverage 60
2.4.2.4 Calling STRs and genotyping 61
2.4.2.5 Data extraction from variant calling files (VCFs) and analysis 63
2.4.2.6 Running time and multithreading 64
2.4.3 DATA AVAILABILITY 64
2.4.4 SOFTWARE AVAILABILITY 64

CHAPTER 3: ANALYSIS OF SHORT TANDEM REPEATS IN WHOLE GENOME SEQUENCING DATA 65

3.1 INTRODUCTION 65
3.2 RESULTS 67
3.2.1 DATASET 67
3.2.2 INITIAL ANALYSIS OF SEVEN WGS SAMPLES 69
3.2.3 RECALL OF STRS 70
3.2.4 ACCURACY OF STRS 71
3.2.5 EFFECT OF REPEAT UNIT SIZE ON ACCURACY OF STRS 75
3.2.6 EFFECTS OF REPEAT LENGTH ON ACCURACY OF STRS 78
3.2.7 MAGNITUDE OF STR GENOTYPE DIFFERENCES 82
3.2.8 EFFECTS OF REPEAT UNITS ON ACCURACY OF STRS 86
3.3 DISCUSSION 91
3.4 METHODS 94

CHAPTER 4: DETECTING DISEASE-CAUSING SHORT TANDEM REPEAT MUTATIONS IN KNOWN PATHOGENIC LOCI 96

4.1 INTRODUCTION 96

4.2 RESULTS 99
4.2.1 COMPILING THE LIST OF DISEASES CAUSED BY MUTATIONS OF SHORT TANDEM REPEATS 99
4.2.1.1 Standard type of repeats 100
4.2.2 IMPERFECT GCN TYPE OF REPEATS 112
4.2.3 REPLACED AND NESTED TYPES OF REPEATS 116
4.2.4 SUMMARY OF KNOWN DISEASE-CAUSING SHORT TANDEM REPEATS 121
4.2.5 VALIDATION OF GENOTYPING OF THE NEW VARIANT CATALOGUE 122
4.2.5.1 Overview of the method 122
4.2.5.2 Genotyping samples up to the fragment length 125
4.2.5.3 Genotyping replaced and nested STR expansions 129
4.2.5.4 Genotyping alleles longer than the fragment length 132
4.2.5.5 Ability to genotype STRs into all pathogenic ranges 135
4.2.5.6 Summary of improvements made to the pathogenic STR catalogue 137
4.2.6 EFFECTS OF EXPANSIONS IN OFF-TARGET REGIONS ON GENOTYPE ESTIMATION 138
4.2.7 GENOTYPING PCR SAMPLES WITH EXPANSIONHUNTER 142
4.2.8 ANALYSIS OF A REAL SET OF SAMPLES BY USING THE NEW CATALOGUE 144
4.2.9 DEVELOPMENT A NEW COMPUTATIONAL TOOL FOR STR ANALYSIS – STRIPY 146
4.3 DISCUSSION 157
4.4 METHODS 161
4.4.1 SIMULATING SAMPLES WITH VARIABLE LENGTH OF STRS 161
4.4.2 TEMPLATE GENERATION 161
4.4.3 SIMULATING READS AND ALIGNMENT 163
4.4.4 GENOTYPING SAMPLES WITH EXPANSIONHUNTER 164
4.4.5 DETERMINING OFF-TARGET REGIONS AND CREATION OF NEW VARIANT CATALOGUE 165
4.4.6 ANALYSIS OF THE UNDIAGNOSED SAMPLES DATASET 165
4.4.7 VALIDATION OF STRIPY 165
4.4.8 POPULATION DATA ANALYSIS OF STR LOCI FOR STRIPY 166

CHAPTER 5: GENERAL DISCUSSION 167

5.1 SCIENTIFIC CONTRIBUTIONS 168
5.2 DISCUSSION 173

LIST OF REFERENCES 177

APPENDICES 204

List of tables

Table 2.1. Feature comparison of short tandem repeats (STRs) specific genotyping tools used in our analysis 41
Table 3.1. All seven samples and their amount of DNA input, fragmentation size, PCR cycles, library yield and calculated sequencing coverage 67
Table 3.2. General parameters of samples after downsampling. Table is sorted by the median insert size in descending order 69
Table 3.3. STR recall compared to median insert size and coverage of all loci 71
Table 4.1. Types of short tandem repeats 99
Table 4.2. All known disease-causing STR loci made of standard type of repeats 100
Table 4.3. All known disease-causing STR loci made of imperfect GCN type of repeats 113
Table 4.4. All known disease-causing STR loci made of nested and replaced types of repeats 117
Table 4.5. Information about control samples with their true and detected genotype by ExpansionHunter 139
Table 4.6. True and detected genotype by ExpansionHunter for set of simulated samples with expansions in up to three loci 140
Table 4.7. True and detected genotype by ExpansionHunter for set of simulated samples with expansions in up to six loci 141
Table 4.8. Comparison between ExpansionHunter and STRipy 156
Table 4.9. Formulas used to simulate alleles 162

List of figures

Figure 1.1. The general workflow from HTS data to results 21
Figure 1.2. Example of STR structure and mutations 24
Figure 1.3. DNA replication slippage mechanism 26
Figure 1.4. Genomic location of some disorders caused by STR expansions in humans 28
Figure 1.5. Technical difficulties in genotyping STRs from HTS data 33
Figure 1.6. Expected STR coverage depending on the read length and sequencing depth 34
Figure 2.1. STR loci and repeat unit distribution in the X chromosome 45
Figure 2.2. Comparison of different short tandem repeat genotyping tools 46
Figure 2.3. Comparison of different short tandem repeat genotyping tools on making heterozygous calls with and without filters.....47
Figure 2.4. Percentage of heterozygous calls per minimum number of reads for each short tandem repeat genotyping tool 50
Figure 2.5. Percentage of heterozygous calls per minimum number of reads for dinucleotide repeats.....51
Figure 2.6. Percentage of heterozygous calls in different bins of quality score separated by each repeat unit length.....53
Figure 2.7. Percentage of heterozygous calls per minimum number of reads and in different quality score ranges for GangSTR (enc.) 54
Figure 3.1. STR recall and accuracy of the dna_109 sample downsampled in 10 steps.68
Figure 3.2. Frequency of repeat units in the genome as reported by TRF 69
Figure 3.3. Median insert size relative to total number of genotyped alleles longer than the read length and mean coverage of STR loci that are longer than the read length. Number of calls is the sum of calls where at least one of alleles is longer than the read length. 70
Figure 3.4. PCR-free sample compared with the sample that had the highest STR recall (dna_109) 73
Figure 3.5. All samples with no filtering and minimum quality score at least 0.9 or 1.0 compared to the PCR-free sample.....74
Figure 3.6. Differences in accuracy in different repeat unit sizes at three filtering strategies.....76

Figure 3.7. Accuracy of dinucleotide repeat alleles of different repeat length 79
Figure 3.8. Accuracy of trinucleotide and tetranucleotide repeat alleles of different repeat length 80
Figure 3.9. Accuracy of pentanucleotide and hexanucleotide repeat alleles of different repeat length 81
Figure 3.10. Frequencies of 2 bp repeat unit's differences between samples and the PCR-free one 83
Figure 3.11. Frequencies of 3 bp and 4 bp repeat unit's differences between samples and the PCR-free one 84
Figure 3.12. Frequencies of 5 bp and 6 bp repeat unit's differences between samples and the PCR-free one 85
Figure 3.13. Accuracy and amount of dinucleotide and trinucleotide motifs 87
Figure 3.14. Accuracy and amount of tetranucleotide motifs 88
Figure 3.15. Accuracy and amount of pentanucleotide motifs 89
Figure 3.16. Accuracy and amount of hexanucleotide motifs.....90
Figure 4.1. Genotyping accuracy for simulated heterozygous and homozygous samples of each gene.....124
Figure 4.2. Genotyping results of increasing number of repeats in the locus in DMPK gene known to cause myotonic dystrophy type 1 disease 125
Figure 4.3. Genotyping results for three problematic loci (AR, TCF4 and ATXN1) that are made of standard type of repeats 126
Figure 4.4. Results of genotyping AR and TCF4 loci with smaller length of reads and fragments 128
Figure 4.5. Results of genotyping SBMA (AR locus) and FECD3 (TCF4 locus) loci by fixed version (4.0.2) of the tool.....129
Figure 4.6. Results of genotyping CANVAS (RFC1 locus) and FAME6 (TNRC6A locus) loci representing replaced and nested types of repeats 131
Figure 4.7. Genotyping accuracy for simulated heterozygous and homozygous samples of each gene including results after specifying off-target regions 133
Figure 4.8. Genotyping results of heterozygous samples when using off-target regions in the CNBP and DMPK locus.....134
Figure 4.9. Genotyping results of homozygous samples when using off-target regions in the CNBP and DMPK locus.....135

Figure 4.10. Genotyping results of pathogenic samples of all diseases 137
Figure 4.11. Comparison of simulated heterozygous PCR-free and PCR+ samples with stutter noise 144
Figure 4.12. Genotyping results of WGS samples of 37 undiagnosed patients 145
Figure 4.13. Overview of STRipy method describing the main steps 146
Figure 4.14. STRipy's method for defining off target regions and extracting reads 148
Figure 4.15. Screen capture of the front screen in STRipy v0.9 149
Figure 4.16. Screen capture of the results screen in STRipy 151
Figure 4.17. Screen capture of the alignment visualiser screen in STRipy 152
Figure 4.18. Comparison of accuracy by using STRipy and ExpansionHunter + catalogue method 154
Figure 4.19. Comparison of sensitivity for pathogenic alleles using STRipy and ExpansionHunter + catalogue method 155

Chapter 1: Introduction

1.1 Genetic variation and clinical diagnosis from sequencing data

1.1.1 Human genetics

Humans are eukaryotic organisms that are made up of over a trillion cells. Each eukaryotic cell (with the exception of red blood cells) includes a nucleus where deoxyribonucleic acid (DNA) is found. DNA is a molecule that contains genetic information responsible for determining the structure and function of each cell, together with information from other cells, organs and tissues (Campbell, 2010). DNA has two helical chains that are built with pyrimidine or purine bases. Pyrimidines are thymine (T) and cytosine (C), and purines are adenine (A) and guanine (G). In the double-helix, A's are bonded with T's with hydrogen bonds, whereas C's are bonded with G's (Watson & Crick, 1953). According to the central dogma of molecular biology, a gene is a nucleotide sequence of DNA that contains information for protein synthesis in a cell. All cells in an organism contain the same genetic information (genome), with the exception of somatic mutations, and are potentially able to make all proteins encoded by a given genome (Campbell, 2010). However, as cells are specialised, different proteins are expressed in cells of different tissues (Campbell, 2010), largely via epigenetic mechanisms that can regulate gene expression through alternations to the chromatin structure that regulates gene expression instead of changing the DNA sequence (Berger et al., 2009).

The instructions for making proteins are first synthesised from a DNA template to form a ribonucleic acid (RNA) molecule via a process called transcription. Transcription starts when RNA polymerase binds to DNA (specifically on a promoter site that is in front of the gene to be transcribed) and stops when it reaches to the terminator site. The newly synthesised RNA undergoes modifications, including removing introns (sequences that do not contain protein-coding instructions) from the RNA (known as splicing) to form a molecule called messenger RNA (mRNA). Nucleotides in the mRNA molecule form a series of triplets (codons) and each codon corresponds to a specific amino acid (Campbell, 2010).

In the cell, the mRNA molecule is transported to the cytoplasm where ribosomes, specialised molecules for translation, synthesise protein. Translation is a process where a protein is being synthesised based on the sequence of the mRNA molecule. As each codon

represents a specific amino acid, the outcome of the process is a polypeptide chain of amino acids in the order corresponding to the order of codons in the mRNA. The formed polypeptide chain undergoes modifications as well before it folds into the mature protein (Campbell, 2010). Mutations in nucleotide sequences and misregulation of expression can have deleterious effects and cause diseases (Lee & Young, 2013; Tahmasebi et al., 2018).

1.1.2 Types of genetic variation

The sequence of human genome is variable between individuals and is responsible for phenotypic variation, which together with environmental factors makes each individual unique. These differences in the genome are known as genetic variation. Most genetic variation is benign but sometimes variants can be pathogenic, depending on the functional impact of the variation. Several types of sequence variations are known that lead to either substitution of bases or structural changes to DNA. Sequence variations can be either single-nucleotide variants (SNVs) or multi-nucleotide variants (MNVs) where base-pairs of a sequence are substituted so the length of a sequence does not change (Haraksingh & Snyder, 2013). SNV is a common type of genetic variation, commonly totalling 4–5 million mutational events in a diploid human genome (Eichler, 2019). SNVs replace one base-pair with another and MNVs replace two or more base-pairs in a sequence (Haraksingh & Snyder, 2013; Wang et al., 2020). SNVs and MNVs that are present in more than a 1% of a population are known as single- or multinucleotide polymorphisms (SNPs and MNPs, respectively). Another type of variant involves small insertions and deletions (INDELs) that are usually up to 50 bp in size (Haraksingh & Snyder, 2013) and have a prevalence of about 700–800 thousand mutational events in a human genome (Eichler, 2019). INDELs in coding regions can lead to pathogenic frameshift mutations (Haraksingh & Snyder, 2013).

Structural variation (SV) is less frequent but still account for 23–28 thousand mutational events in human genome (Eichler, 2019). SVs comprises of changes in the genome such as copy number variations (CNVs), inversions, translocations and chromosomal aneuploidies. SVs are large changes in a genome that range from hundreds to up to millions of base-pairs. CNV, a type of SV, is a deletion (either homo- or heterozygous) or insertion (tandem or dispersed in genome) of a DNA stretch. Another class of SV are mutations where the length of DNA does not change, such as inversions (sequence of

DNA that change orientation) or balanced chromosomal translocations (chromosomes exchange equal length of DNA with each other) (Haraksingh & Snyder, 2013). In the case of unbalanced chromosomal translocations, the exchanged DNA is unequal and resulting in either gain or loss of the genetic material (Harewood & Fraser, 2014). A specific type of variation involves tandem repeats (TR), which are defined as DNA sequences repeated adjacent to each other (Lander et al., 2001). Tandem repeats can be classified as micro- or minisatellites, depending on repeat length. This name originates from the satellite bands that were seen when genomic DNA was separated with density-gradient centrifugal methods. Microsatellites are also known by other names, such as simple sequence repeats (SSRs) or short tandem repeats (STRs) (Gemayel et al., 2010).

There are multiple ways that sequence variations can have a functional impact on an organism. For example, in the case of SNVs in a coding region, change of one nucleotide in a DNA sequence can either have no impact on the amino acid (synonymous mutation) or it can cause a change in the amino acid sequence – either missense or nonsense (non-synonymous mutation). For instance, nonsense mutations can result in shorter peptides due to a premature stop codon, whereas missense mutations could change the encoded amino acid sequence or expand it when converting a stop codon into a regular amino acid codon. In addition, SNVs in the initial codon can make transcripts untranslatable (Haraksingh & Snyder, 2013).

INDELs in the coding sequence can cause frameshift or non-frameshift mutations. INDELs that are not divisible by three can change the coding frame and therefore the amino-acid sequence from the location of the mutation until the end of the 3′ end. On the other hand, non-frameshift mutations only alter a part of an amino-acid sequence (Haraksingh & Snyder, 2013). Similarly, tandem repeats, such as STRs, can cause frameshift or non-frameshift mutations as well, depending on the repeat-unit size and number of repeat units causing an expansion or contraction (Metzgar et al., 2000). Mutations in intronic regions can result in alternative splicing events (Li et al., 2004). As the mutation rate of STRs (~10^{-4} mutations per generation per site) is significantly higher than SNVs (10^{-8} mutations per generation per site) (Tang et al., 2017) they are an important source of genetic variation that are not routinely analysed in sequencing studies due to technical challenges (Gymrek et al., 2012). Moreover, STRs are important as they show both, somatic and germline instability and mutations can arise in both of these types

of cells (Zoghbi & Orr, 2000). Somatic mutations refer to alternations in DNA after conception (in somatic cells). These changes are usually not transmitted to offspring (except in case of some post-zygotic mutations) which is in contrast to germline mutations that are transmitted to progeny (Oota, 2020). Pathogenic mutations can occur in both types of those cells (McMurray, 2010).

1.1.3 Detection of variants from sequencing data

1.1.3.1 History of sequencing methods development

DNA sequencing is a combination of molecular biology and nucleotide chemistry, both which blossomed in 1950s. In the early 1970s, after Frederick Sanger's laboratory published a method for RNA sequencing they started to look for a method for DNA sequencing (Mardis, 2013). That endeavour was successful and later in the same decade a paper was published describing a method for DNA sequencing (Sanger et al., 1977). This original sequencing protocol was based on chemically modified nucleotides (dideoxynucleotides – ddNTPs) which were incorporated to the end of a DNA sequence by DNA polymerase. Dideoxynucleotides have 3′ hydroxyl group replaced with a hydrogen atom that inhibits DNA chain elongation and terminates replication. Native deoxynucleotides were mixed with one of their ddNTP analogues (either ddATP, ddTTP, ddCTP or ddGTP). That resulted all the possible fragments with different length that were terminated at the added ddNTP nucleotide. The DNA sequence was determined by labelling ddNTPs with a radioactive ^{32}P, followed by separating fragments by size on a polyacrylamide gel and then exposing it to X-ray film (Mardis, 2013). This method underwent many improvements over the years until 1986 when a new method of sequencing was introduced that replaced the radiolabelled nucleotides with fluorescently labelled DNA. The new method increased the throughput of sequencing data as it required significantly less manual labour compared to the previous method (Smith et al., 1986). Instead of using and developing X-ray film and reading autoradiographs, the sequencing instrument used laser beam to scan gel plates and reading DNA sequence by exciting fluorescent primers and detecting the emitted wavelength (Mardis, 2013). Many improvements were made over the years that greatly improved on the sequencing technology, such as incorporating polymerase chain reaction (PCR), improvements to chemistry and increasing per-run throughput (Mardis, 2013).

In 1999, a bigger change in sequencing technology took place when capillary sequencing instruments were introduced. This method increased sequencing throughput even more by removing many rate-limiting steps by introducing improvements such as automatically injecting a matrix into capillaries instead of manually loading gels, simplifying downstream processes and by reducing run times due to better heat dissipation of capillaries (Mardis, 2013).

Still, the most revolutionary change was ahead. In 2004 a program was initiated by the National Human Genome Research Institute (NHGRI) which aimed to lower the cost of sequencing the human genome multiple orders of magnitude and bring it down to 1000 USD in 10 years (Schloss, 2008). That motivated the development of several new technologies and soon after, the high-throughput sequencing (HTS) technology (also known as the next-generation sequencing) was introduced (Mardis, 2013). HTS has many key differences compared to previous sequencing technologies. For example, a library of DNA fragments are constructed instead of bacterial cloning, and fragments are amplified on a solid surface instead of conducting sequencing reactions in a plate well (Mardis, 2013). Also, the read length in HTS, which is typically in the range of 100–400 bp, is shorter than produced by Sanger sequencing which reads are between 500 and 1000 bp in length (Caspar, S.M. et al., 2018). However, HTS yields millions of these short reads providing a significantly higher sequencing throughput than Sanger sequencing (Abel & Duncavage, 2013). As an impact, HTS has greatly reduced the cost of sequencing while also increasing sequencing throughput (Mardis, 2013). HTS can be used to sequence whole genomes, whole exomes or perform a targeted sequencing to capture specific regions in a genome (Abel & Duncavage, 2013).

1.1.3.2 Types of high-throughput sequencing

Targeted sequencing (TS) is a method where specific genes or regions of interest are enriched and sequenced by HTS. Commonly, specific targets (a set of genes or regions) are enriched using specific probes that hybridizes regions of interest which are then isolated before sequencing (Haas et al., 2014). This is a relatively fast and less costly method compared to whole-exome or whole-genome sequencing. However, targeted sequencing is not a suitable method for identification of clinically relevant mutations in the human genome when the mutation is in the region that is not in the gene panel (Caspar, S.M. et al., 2018).

Whole-exome sequencing (WES) is used to sequence all known exons in a genome (exome) which captures only ~1% of the human genome, but contain is thought to around 85% of rare disease-causing sequence variants (Caspar, S.M. et al., 2018; Choi et al., 2009). In contrast to TS, WES includes all genes (including the ones that are not known to be associated with diseases) and therefore it allows the discovery of novel mutations and identification of new genes associated with diseases (for example, through trio analysis which includes the affected offspring and its unaffected parents). Both WES and TS have similar limitations, such as difficulties in achieving uniform coverage due to some sequences being difficult to capture. On the other hand, whole-genome sequencing (WGS) has an advantage over WES and TS as the coverage is the most uniform over the whole genome. WGS can also detect sequence variants outside of exons. In addition, there are other benefits, such as improving detection of CNVs (Caspar, S.M. et al., 2018).

1.1.3.3 Methods of high-throughput sequencing

There are at least two modes of HTS sequencing: short- and long-read sequencing. Short-read sequencing (SRS) generally produces reads that are up to 400 bp in length and this length has been increasing over time. Illumina, Thermo Fisher Scientific and Beijing Genomics Institute (Ion Torrent) are some of the well-known manufacturers offering SRS methods (Caspar, S.M. et al., 2018), however, the vast majority of SRS is now done with Illumina. Usually, SRS incorporates a library preparation method in which a DNA sample is being randomly fragmented, followed by adapter ligation and finally the formed fragments are sequenced (Goodwin et al., 2016). While SRS provides high-throughput the short read length itself is the main limitation of the platform and that leads to challenges during the alignment of reads, such as correctly mapping them on repetitive or paralogous regions, regions with large structural variation and low-complexity (Pollard et al., 2018).

Some limitations of read length can be overcome by using the long-read sequencing (LRS) methods. LRS offers reads that are multiple kilobases long, and that attribute improves alignment of reads on a reference genome as long reads can fully span complex and repetitive regions which are difficult to resolve when using SRS data (Caspar, S.M. et al., 2018). LRS can be classified into two types. Firstly, true-LRS (also known as third-generation sequencing) is a single-molecule real-time sequencing method (Mantere et al., 2019) which can generate reads up to 7–8 kb long when using methods developed by

Oxford Nanopore Technologies (Leggett & Clark, 2017) or even 10–15 kb when using Pacific Biosciences platform (Rhoads & Au, 2015).

Secondly, synthetic-LRS is a long-read library preparation method (such as Illumina TruSeq Synthetic Long-Reads or 10X Genomics linked read sequencing) where long sequences are first fragmented and indexed, then sequenced and finally short reads are merged bioinformatically back into long ones (Caspar, S.M. et al., 2018). These split fragments are linked together using molecular barcodes, a series of known bases that are ligated to the fragments before sequencing. After sequencing, these long fragments can be reassembled from shorter ones by matching the barcode they contain (Goodwin et al., 2016). This ultimately offers results similar to long-LRS, despite some disadvantages (such as potentially missing some regions that can be detected by true-LRS) (Caspar, S.M. et al., 2018). LRS has also its own disadvantages compared to SRS, such as lower throughput and higher cost per sample and at this time a substantially higher error rate (above 10%). Also, fewer bioinformatic tools are available for the analysis of LRS data (Maroilley & Tarailo-Graovac, 2019).

HTS is an important technology in research and in clinical contexts for identifying rare and new variants (Liu et al., 2013). Finding either SNVs, INDELs, CNVs or other types of variations from sequencing data is a process of many steps and requires a computational pipeline comprised of multiple tools and steps (Liu et al., 2013). A common workflow for such a pipeline is illustrated in Figure 1.1.

1.1.3.4 Variant calling from high-throughput sequencing data

In order to detect sequence variants such as SNVs and INDELs in the human genome that might be associated with diseases, a bioinformatics analysis pipeline is run. This pipeline can be divided into three parts. In the first part, the sequencing process takes place where raw DNA fragments are being converted to digital information (A, T, G and C strings) (Lebo et al., 2020).

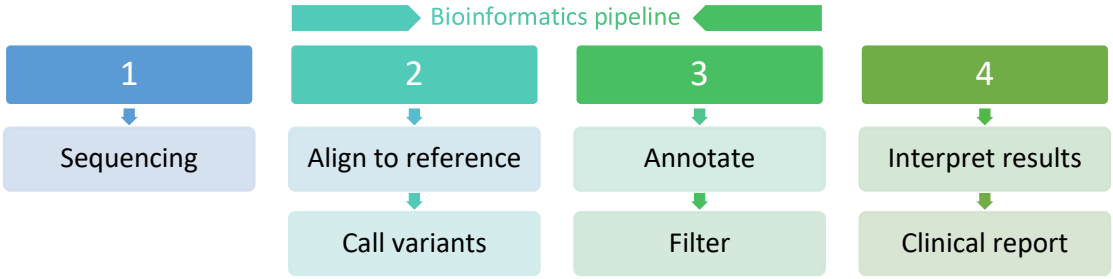


Figure 1.1. The general workflow from HTS data to results. First, specific regions or the whole genome are sequenced, followed by aligning HTS reads on a reference genome. Specific tools are used to call variants from the aligned data and then variants being annotated, filtered and analysed that may or may not lead to a molecular diagnosis.

The second part incorporates alignment and variant calling. Millions or even billions of sequenced HTS reads are aligned to a reference genome by using specialised computer software (Treangen & Salzberg, 2012). For example, Burrows-Wheeler Aligner (Li & Durbin, 2009) or Bowtie (Langmead et al., 2009) are common aligners. Their throughput is around 10-40 million reads per hour on one processor (Treangen & Salzberg, 2012). A human reference genome (where reads are aligned by an aligner) is a mosaic haploid genome composed of genomes obtained from several people. For instance, the two major reference genomes currently used, GRCh37/hg19 (released in 2009) and the newer GRCh38/hg38 (released in 2013) are assembled from genomes of 13 volunteers (Maroilley & Tarailo-Graovac, 2019).

The variant calling step is a key process to identify variants. There are different types of tools for this process. Some of them just use bases detected on the high confidence read alignments which are in contrast to the reference genome sequence, while most widely used methods do probabilistic calling which includes base and mapping quality scores (Pereira et al., 2020). Some of the most common variant callers are GATK (DePristo et al., 2011; McKenna et al., 2010) and SAMtools (Li et al., 2009). Commonly, variant calling tools use BAM/SAM files as input and output the results into a standardised variant calling format (VCF) file (Pereira et al., 2020). Variants can be either small (such as SNVs, small INDELs) or large (such as SVs) (Lebo et al., 2020). Often a pipeline for detecting SNVs and/or INDELs uses BWA and GATK best practises (Auwera et al., 2013; Caspar, S.M. et al., 2018). However, when looking at specific types of variants, such as CNVs and tandem repeats, then use of specialised algorithms/tools are required (Treangen & Salzberg, 2012).

In the third and final part, the identified variants are annotated, filtered and interpreted. Variants called in the previous step, which are in the VCF file, mainly contain information such as the genomic position, reference and alternate bases (Pereira et al., 2020). However, these variants need to be annotated, to provide biological context by incorporating any clinically and biologically relevant information on genomic location, transcript or protein changes to assess their potential effects on function (Lebo et al., 2020). For example, it is important to define the gene where the variant is found and its impact, such as whether there is a resulting missense, nonsense, synonymous or other type of mutation (Pereira et al., 2020). Proteins with different sequence are coded when the reading frame shifts by one or two nucleotides. Some genes can also overlap and therefore, a single DNA sequence can code proteins for both genes. In that case, mutations in overlapping genes can affect both proteins (Lèbre & Gascuel, 2017). A variant can affect either both or one coded protein. For instance, a SNV can result in a synonymous mutation to one gene and non-synonymous to the other one (Wei & Zhang, 2014). When comparing a sequenced genome to the reference one, on average about one variant is detected per 1000 bp of sequence (Lebo et al., 2020). The vast majority of these variants are common in the population and not related to disease and can be annotated as such from databases such as ExAC and gnomAD (Karczewski et al., 2020). Therefore, the vast numbers of called variants should go through annotation and appropriate filtering processes, as manual examination of each call and variant is not practical (Lebo et al., 2020).

Many methods and sources of information can be used for filtering and annotation. For example, literature searches can be used to determine reported disease-associated variants (for example, NCBI). Public online databases, such as OMIM, ClinVar, HGMD and LSDB are useful to find relationships between variations and phenotypes. SNVs and INDELs can also be prioritised based on population frequencies from datasets (for instance, ExAC/gnomAD), and determine the effects of variants on splicing (such as dbscSNV) or protein structure and function (for example, SIFT and PolyPhen2) (Mardis, 2013).

Following annotation, filtration and prioritisation should be performed to sort the unlikely functional variants out. The first filtration step could involve limiting variants to only genes or exons, or the medical exome which is a set of genes (currently 2000–4500) that

are known to have associations with diseases. This filtration could precede the annotation. Many other types of filtering could also be used, such as keeping only *de novo* variants when analysing trios (genetic-based filtration), filtration based on rarity in population, type and location of variant (variant-based filtration) (Lebo et al., 2020).

However, every result after filtering and prioritisation still requires manual review. The final identification of disease causing variants involves using all the evidence gained in the previous steps (Lebo et al., 2020). Using more than one independent analysis pipeline is suggested as that would decrease the chance of false negatives. Sequence variants found from HTS data could go through an additional sequencing step by using the gold-standard Sanger method to confirm the calls (Caspar, S.M. et al., 2018).

1.2 Short tandem repeats as a source of genetic variation

1.2.1 Overview

Tandem repeats are found in nearly all organisms, including higher eukaryotes and humans, where they are very common across the genome (Gemayel et al., 2010; Nithianantharajah & Hannan, 2007). An increasing number of studies have demonstrated a link between tandem repeats and altered gene expression and/or its function (Fotsing et al., 2019; Gymrek et al., 2016; Jansena et al., 2012; Quilez et al., 2016). A class of tandem repeats called short tandem repeats (STRs) were first classified in early 1980s (Ellegren, 2004). For example, the sequence "CAG|CAG|CAG|CAG|CAG" is a STR that comprises of CAG repeat units that has a size of three and repeat length of five repeat units (Figure 1.2).

There is no consensus in the shortest or longest motif length but they are considered in the range of 1–10 bp (Hannan, 2010), or 1–6 bp (Sawaya et al., 2013; Subramanian et al., 2003; Tang & Nzabarushimana, 2017) or 2–6 bp in length when discarding the homopolymers (Gymrek et al., 2016; Kristmundsdóttir et al., 2017; Press et al., 2014; Willems et al., 2014). Occasionally, motifs that are longer, such as in length of 9 bp (Duitama et al., 2014) or 10 bp (Press et al., 2014) have been classified as STRs as well. Motifs longer than the upper limit of STRs are termed as minisatellites, often the limit is 10 bp (Hannan, 2010). However, this is also arbitrary, as some studies have considered STRs with repeat unit length of 13 bp and then the range of minisatellites starts from 14 bp (Lander et al., 2001).

STRs have been linked to almost 50 hereditary disorders, with the vast majority of them degenerative without a treatment (Khristich & Mirkin, 2020). Some of these diseases are caused by long expansions, for example a newly discovered cerebellar ataxia, neuropathy and vestibular areflexia syndrome (CANVAS) with a pathogenic range between 400 and 2000 pentanucleotide repeats (Cortese et al., 2019). In contrast, other diseases can be a result of a few-repeat difference from the wild-type allele, such as in case of oculopharyngeal muscular dystrophy (Brais et al., 1998).

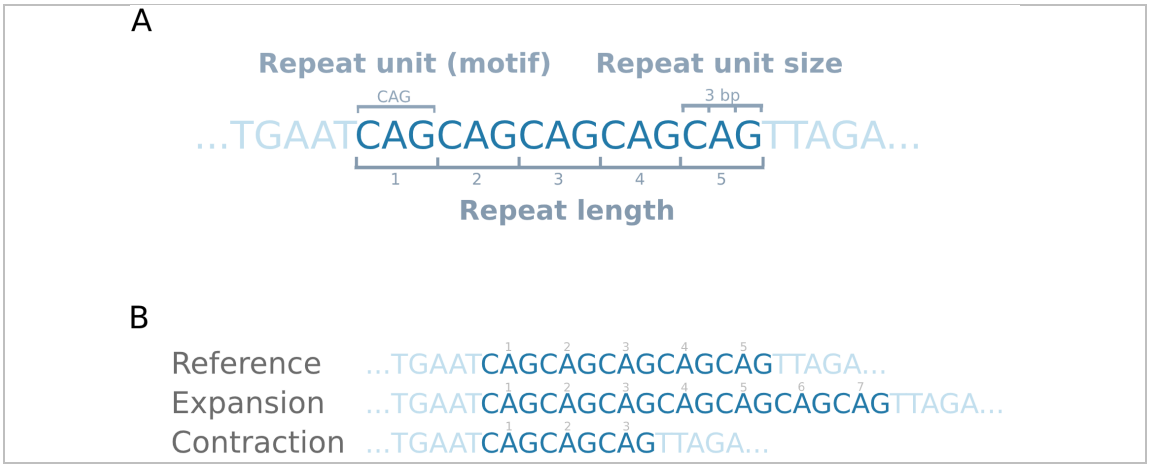


Figure 1.2. Example of STR structure and mutations. (A) STR composition. (B) Types of STR mutations: expansion gains repeats, whereas contraction loses repeats, compared to the reference repeat length.

1.2.2 Distribution

At least half of the human genome consists of different types of repeated regions, including STRs that makes up around 3% of the genome (Lander et al., 2001). Over a million STR loci are found in the human genome, however this number is not well defined as it is related to the parameters used on the search algorithm (for example, including more imperfect repeats can increase the count) (Ellegren, 2004).

Dinucleotide repeats are the most common STRs, comprising about 0.5% of the human genome (Lander et al., 2001). However, when including mononucleotide repeats then STRs become the most frequent repeat units in the human genome, which is mainly attributed to the polyA/polyT repeats (Katti et al., 2001). There is a significant variation between repeats types across different genomic regions. For example, tri- and hexanucleotide repeats have around two-fold higher frequency in exonic regions than in intronic and intergenic regions in all chromosomes except the chromosome Y (Subramanian et al., 2003). This can be explained by the fact that non-trimeric STRs in coding regions causes frameshift mutations that are more likely to be selected against

(Metzgar et al., 2000). Other repeats, such as di-, tetra- and pentanucelotide are more abundant in non-coding regions (Subramanian et al., 2003).

As STRs are highly variable in the population, they have been applied extensively in forensics to match suspects with DNA collected from crime scenes (Butler, 2007). For example, Federal Bureau of Investigation of the United States has defined 20 STR loci (CODIS Core Loci) which are being used to create the DNA fingerprint in forensics casework (Hares, 2015). STRs are also used to identify individuals, such as victims of mass disaster, missing people, as well as in parentage testing (Butler, 2007).

1.2.3 Mutations

1.2.3.1 Mutation models

Many models have been created to describe how STRs mutate, and most of them originate from the stepwise mutation model (SMM) which is based on the addition or removal of one motif at a fixed rate (Ellegren, 2004; Ohta & Kimura, 1973). It soon became clear that the simplistic SMM, does not result in stable length distributions of STRs and does not consider the fact that STRs seem to have a length limit. Improved versions of SMM were proposed that accounted for the upper limit of repeat sizes, as well as for mutation bias where large alleles show preference for contracting mutation or which were based on more complex mathematical models (Ellegren, 2004). One of the models that performed well describes the mutational process as two opposing forces – length of mutations, where the mutation rate is positively correlated with the repeat count, and point mutations, which break long repeats into smaller ones to slow the mutation rate (Bell & Jurka, 1997; Kruglyak et al., 1998).

1.2.3.2 Mutation mechanisms

A change in repeat length is believed to be caused by DNA replication slippage events (Levinson & Gutman, 1987). When the replication of a STR-containing DNA sequence has been initiated, then two strands of the DNA may separate and the growing strand could misalign, resulting in an increase in STR repeat length (Ellegren, 2004), see Figure 1.3. The DNA mismatch repair system repairs most of the erroneous mutations, but a small number of unrepaired ones will end up as a mutation in the DNA sequence (Strand et al., 1993). It has been shown that only DNA polymerase is required for the slippage event (Schlötterer & Tautz, 1992), and while it is pausing and separated from the strand,

the end of the newly synthesized strand dissociates from the template, followed by annealing to another repeat unit which causes misalignment (Hile & Eckert, 2004).

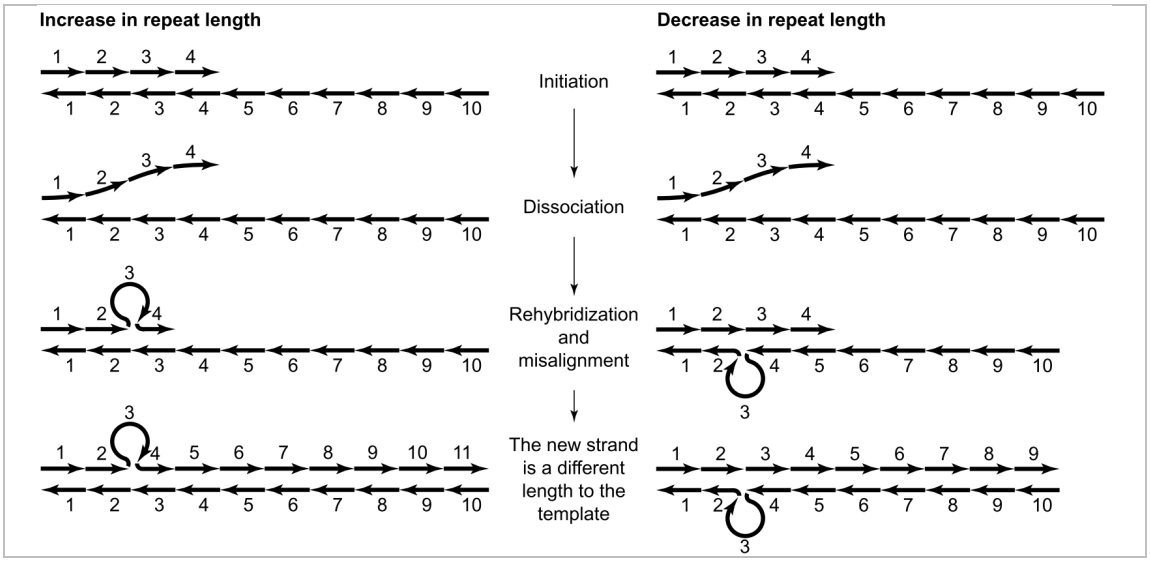


Figure 1.3. DNA replication slippage mechanism. The first step is the initiation where replication of a repeat tract starts. The two strands then separate and the nascent strand misaligns back. If the nascent strand forms a loop then it will lead to an increase in repeat length. If the loop forms in the template strand then a decrease in repeat length is observed. (Ellegren, 2000b)

Replication slippage can also happen during PCR amplification cycles causing artificial alleles with wrong size, also known as the "stutter" (Salipante et al., 2014). All STRs except the poly(A) tails, are believed to be a result of polymerase slippage events during DNA replication (Lander et al., 2001).

Kruglyak and colleagues analysed 2–4 bp repeats and determined differences between repeat unit lengths. They found that the highest mutation rate is present in dinucleotides, followed by tri- and tetranucleotide repeats (Kruglyak et al., 1998), which could be because more slippage events are possible for shorter motifs in the same length of DNA (Katti et al., 2001). Another possible factor could be the sequence composition. For example, there are studies which found differences in mutability between different sequence motifs in humans (Kelkar et al., 2008) as well as in macaques (Liu et al., 2017). It has also been suggested that mismatch repair system repairs more efficiently $(AT)_n$ mutations than $(GT)_n$ ones (Harr et al., 2002), leading to differences between motifs. In addition, there are also differences in mutation rates found between organisms, with the highest slippage rate found to be in mice, followed by humans, yeast and then fruit flies (Kruglyak et al., 1998).

1.2.3.3 Magnitude of mutation and mutation rates

The majority (89%) of mutation events result in one repeat difference (Ellegren, 2000b). There is no certain mutation rate that can be applied to all cases, instead, it is variable between loci, alleles and probably between different species as well (Ellegren, 2000b). It has been found that in eukaryotic cells there are more insertions of repeats than deletions (Twerdi et al., 1999; Wierdl et al., 1997).

One factor that has been repeatedly reported as affecting the mutation rates of STRs is the repeated sequence length, and the correlation between allele lengths and mutation rates has been observed in many studies (Brinkmann et al., 1998; Fondon et al., 2012; Huang et al., 2002; Kelkar et al., 2008; Lai & Sun, 2003). Also, some studies have found a linear relationship between allele lengths and mutation rates (Sibly et al., 2003), whereas others reported it to be exponential (Leopoldino & Pena, 2003). Studies have indicated that the long alleles tend to have contraction mutations more frequently (Xu et al., 2000) and therefore mutating to shorter ones instead of growing infinitely (Ellegren, 2000a).

1.3 Importance of short tandem repeats

Some STRs are clinically important as their mutations in the genome can have pathogenic results, which has been shown for at least 50 different loci (Depienne & Mandel, 2021). The majority of repeat expansion diseases are neurological (Hannan, 2018). STRs role in human diseases was discovered in early 1990s when an "unstable region of DNA" was described in affected individuals with fragile X syndrome. Sequencing of the region revealed later that it was composed of expanded CGG repeats (Monckton, 2021). There are many ways in which tandem repeats can cause disease, that depend on the genomic location where a repeat expands or contracts (Murmann et al., 2018). Expansions of tandem repeats are found in coding and non-coding areas of a genome (Figure 1.4), affecting processes such as gene transcription, translation to protein, protein structure and its functions, RNA splicing and RNA transport and stability – all of which can result in a rare genetic diseases (Nithianantharajah & Hannan, 2007; Xu et al., 2016).

Many STRs that are in gene regulatory regions are thought to act as "tuning knobs" for gene expression, such as regulating transcription and formation of binding sites for transcription factors (Hannan, 2018). Whereas it was originally assumed that STRs in coding regions cause diseases via translation of amino acids to toxic proteins, later

disease-causing STRs were discovered in untranslated regions (UTRs) as well, which can result in toxic RNAs (Murmann et al., 2018). It is now known that STR expansions can cause a disease through different mechanisms depending on the location, including protein gain-of-function, protein loss-of-function and RNA-mediated gain-of-function (Orr & Zoghbi, 2007).

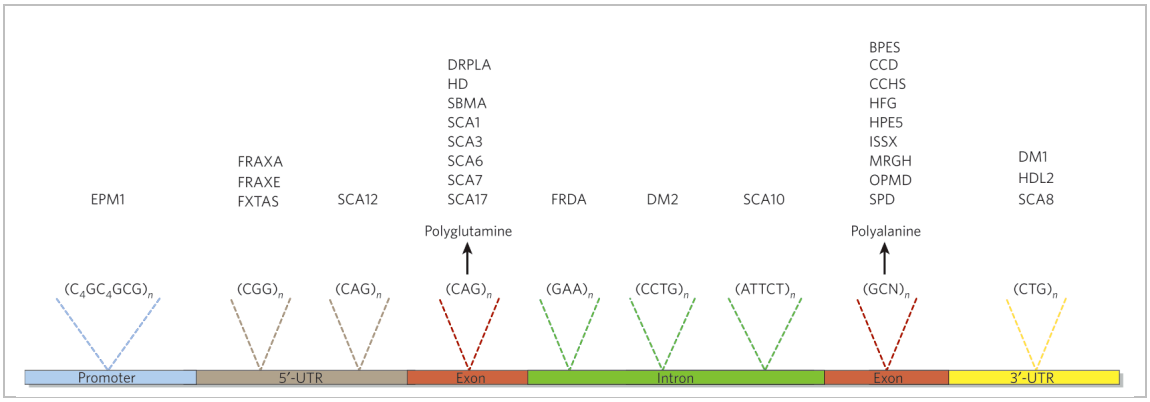


Figure 1.4. Genomic location of some disorders caused by STR expansions in humans. EPM1 - progressive myoclonic epilepsy 1, FRAXA - fragile X syndrome, FRAXE - fragile X mental retardation, FXTAS - fragile X-associated tremor ataxia syndrome, SCA - spinocerebellar ataxia, HD - Huntington's disease, DRPLA - dentatorubral-pallidoluysian atrophy, SBMA - spinal and bulbar muscular atrophy, FRDA - Friedreich's ataxia, DM - myotonic dystrophy, HDL2 - Huntington's disease-like 2, BPES - blepharophimosis-ptosis-epicanthus inversus syndactyly, CCD - Cleidocranial dysplasia, CCHS - congenital failure of autonomic control, HFG - hand-foot-genital syndrome, HPE5 - holoprosencephaly 5, ISSX - X-linked infantile spasm syndrome, MRGH - X-linked mental retardation with isolated growth hormone deficiency, OPMD - oculopharyngeal muscular dystrophy, SPD - synpolydactyly (Mirkin, 2007).

1.3.1 Repeats in coding regions

1.3.1.1 Polyglutamine disorders

All of the disorders caused by expansions of trinucleotide STRs in protein-coding region are either polyglutamine or polyalanine disorders (Hannan, 2018). The largest class, polyglutamine diseases, are caused by expansions of CAG repeats that are translated to glutamine tracts (Li et al., 2004). Huntington's disease (HD), dentatorubral-pallidoluysian atrophy (DPLA), spinobulbar muscular atrophy (SBMA) and spinocerebellar ataxias (SCA) 1, 2, 3, 6, 7 and 17 are the examples of polyglutamine disorders (Hannan, 2018). They are all progressive and cause neural dysfunction as well as neural loss as disease progresses (Li et al., 2004). These disorders are life-shortening (Jones et al., 2017) and there is no concrete therapy available at the moment to stop or slow down the progression of any of them (Rüb et al., 2013).

HD is a well-known example, and is a progressive neurodegenerative disorder characterised by CAG repeat expansions in the coding region of HTT gene that leads to

extended polyglutamine tract (increased portion of repetitive CAGs) in the huntingtin protein (Hannan, 2018). The normal number of CAG repeats in HTT gene is between 6 and 34, whereas more than 36 repeats leads to disease (Zoghbi & Orr, 2000). It has also been shown that the higher number of CAG repeats results in an earlier age of disease onset as well as increased severity (Orr & Zoghbi, 2007). Symptoms of HD patients usually start between the age of 30 and 50 years and the condition of patients worsens over a period of 10–25 years (Murmann et al., 2018). However, childhood-onset of HD has been described (Patra & Shirolkar, 2015), which occurs in approximately 5% of cases. Similar to HD, the spinal and bulbar muscular atrophy (SBMA) is another polyglutamine disease, and is caused by expanded CAG repeats in exon 1b of the androgen receptor gene (Murmann et al., 2018).

A large group of other polyglutamine diseases are spinocerebellar ataxias (SCAs), which are characterised by imbalance and incoordination in patients (Rüb et al., 2013). SCA1, SCA2, SCA3, SCA6, SCA7 and SCA17 are all progressive, mainly adult-onset nervous system disorders which have average disease duration between 15 and 30 years, followed by death (Rüb et al., 2013). Often, when CAG repeats exceed 35–45 repeats (20 in case of SCA6), a polyglutamine region becomes prone to taking an abnormal conformation that results in protein aggregation and abnormal interaction with other neuronal proteins (Rüb et al., 2013).

The reason why numerous diseases are caused by CAG repeat expansions could be due to a polyglutamine tolerance in proteins (Katti et al., 2001). That is, a small CAG expansion in coding regions are tolerated whereas hydrophobic or basic amino acids caused by a small change in a repeat length could be lethal and therefore carriers of these mutations would have been eliminated from the population (Katti et al., 2001). An expanded glutamine stretch, while not embryonic lethal, can affect protein function and in most cases cause a toxic gain-of-function in diseases (Galvão et al., 2001). In some cases, it can also result in a change-of-function protein (Orr, 2012).

1.3.1.2 Polyalanine disorders

While the most common trinucleotide repeats are caused by polyglutamine expansions, many other pathological expansions can be a result of expanded polyalanine repeats (Galvão et al., 2001). Polyalanine expansions are causing disorders that affects development of an organism and are biochemically analogous to polyglutamine disorders

(Hannan, 2018), but they have imperfect triplets of $(GCN)_n$ that usually have a small number of expanded repeats (Mirkin, 2007). There are at least nine diseases caused by expanded polyalanine repeats in humans (Shoubridge & Gecz, 2012). However, in contrast to polyglutamine expansions, polyalanine disorders do not appear to be progressive or neurodegenerative (Hannan, 2018).

1.3.2 Repeats in non-coding regions

Many disorders caused by STRs can also be located in non-coding regions, such as introns and 3′ and 5′ untranslated regions (Jansena et al., 2012). While it is clear that STRs can cause protein toxicity, the evidence also shows a role in RNA-mediated diseases that became apparent after STRs were found in non-coding regions (Murmann et al., 2018).

In untranslated regions, myotonic dystrophy type 1, Huntington's disease-like 2 and spinocerebellar ataxia 8 are caused by a CTG expansion in the 3′ UTR. Whereas spinocerebellar ataxia 12, fragile X mental retardation 2, glutaminase deficiency, desbuquois dysplasia-2, SCA37 and neuronal intranuclear inclusion disease are all caused by expansions in the 5′ UTR (LaCroix et al., 2019; Li et al., 2004; Ranum & Day, 2002; Seixas et al., 2017; Sone et al., 2019; van Kuilenburg et al., 2019). In addition, both fragile X mental retardation 1 (FMR1) and fragile X-associated tremor ataxia (FXTAS) syndrome are caused by expansions of CGG in the 5′ UTR of FMR1 gene (Jansena et al., 2012).

Myotonic dystrophy (DM) is the most studied disease that is caused by toxic RNA (Murmann et al., 2018). DM is the most common form of muscular dystrophy in adults (Liquori, 2001) that is caused by CTG repeat expansion on chromosome 19 in the 3′ UTR of the myotonic dystrophy protein kinase (DMPK) gene (Liquori, 2001). In myotonic dystrophy type 1 (DM1), DMPK encodes DM1 protein kinase where the expansion results in toxic RNA that affects RNA splicing (Liquori, 2001). Unaffected individuals have less than 50 CTG repeats in DMPK gene, while expansions of up to more than 1000 were reported as pathogenic (Murmann et al., 2018). Myotonic dystrophy type 2 (DM2), on the other hand, is caused by a CCTG repeat expansion that is present in intron 1 of the zinc finger protein 9 (ZNF9) gene and encodes CCHC-type zinc-finger nucleic-acid binding protein (CNBP) (Liquori, 2001).

Some other disorders are a result of expansions in introns, such as spinocerebellar ataxias 10 (Ranum & Day, 2002), 31 and 36 (Kobayashi et al., 2011). Friedreich's ataxia is caused by GAA expansions in the intron (Nageshwaran & Festenstein, 2015), which results in loss of function as the FXN gene is no longer expressed due to the formation of stable triple helix formation at the gene locus that blocks transcription of the gene (Kennedy & Shelbourne, 2000).

In addition, some examples of the most recent findings are the different types of familial adult myoclonic epilepsy caused by the TTTCA repeat that is not present in the reference genome, but was found to be expanded in introns of a number of genes in patients with the disorder (Corbett et al., 2019; Florian et al., 2019; Ishiura et al., 2018; Lagorio et al., 2019). Secondly, expansions of AAGGG repeats in an intron of RFC1 gene are linked with cerebellar ataxia, neuropathy, and vestibular areflexia syndrome (CANVAS) which is a different repeat motif to what is present in the reference genome (Cortese et al., 2019).

1.3.3 Somatic mosaicism of short tandem repeats

High mutation rates of STRs can result in somatic variation, also known as mosaicism, which is characterised by genetic variants that are different in somatic cells or tissues and are not inherited from the germline (Press et al., 2014). STRs can expand or contract through meiosis, but there is some evidence that it can happen through mitosis as well which can result in mosaicism (Nithianantharajah & Hannan, 2007). For example, a study found that CAG repeats in the HD gene had higher mutability and increased in length in CNS tissues (especially high in striatum), in contrast to non-CNS tissues, and these changes increased with age (Kennedy & Shelbourne, 2000). Another study in HD mice showed that the mRNA repeat length of HD gene in striatum was longer than in cortex and brainstem, indicating its toxicity properties (Gonitel et al., 2008). This mosaicism also developed with age (Gonitel et al., 2008), which is an important piece of information as the length of repeats in HD patents are known to be negatively correlated to the age of onset (Duyao et al., 1993). Another study observed somatic instability in different brain regions (except cerebellum) in transgenic mouse models of HD (R6/1, R6/2 and R6/5), but also in non-CNS tissues in contrast to the study by Kennedy and colleagues (Unseld et al., 1997). After the characterisation of mosaicism in HD, other studies followed that investigated somatic instability in SCA1 and SCA3 as well as in DRPLA that also showed patterns of mosaicism (Kraus-Perrotta & Lagalwar, 2016).

Such expansions that happen through mitosis can result in genetically different populations of cells and tissues within an individual (Nithianantharajah & Hannan, 2007). However, it is difficult to measure somatic mosaicism in relevant tissues such as the brain and therefore pathogenic variants can be missed completely as traditional genotyping methods involve using either blood, saliva or buccal cells (Hannan, 2010; Hannan, 2018).

In cancer research, STRs are better known under the microsatellite instability (MSI) term (Salipante et al., 2014). Somatic mutations of STRs in cancer have been shown to be caused by defective mismatch repairing genes which can be a due to a somatic or germline mutation . When a mismatch repair gene, such as hMLH1, hMSH2, hMSH6 or hPMS2, is inactivated, then expansions and contractions of STRs cannot be accurately repaired (Meyer et al., 2009; Zhu et al., 2015). This has been shown in patients with Lynch syndrome where mutations in one of these genes are observed and which in turn cause different size of STR repeats (Meyer et al., 2009). However, MSI can be also caused by only defective hMLH1 gene through epigenetic promoter methylation (Fang et al., 2012). Moreover, STR mutations can result in oncogenesis via inactivation of tumour suppressor genes (Woerner et al., 2009). Besides Lynch syndrome, MSI is associated with other types of cancers too, including sporadic colorectal, endometrial and gastric (Zhu et al., 2015). Moreover, a study conducted by Hause and colleagues analysed nearly six thousand exomes from 18 cancer types and found MSI positive tumours in 14 of them, indicating the importance of STRs in various types of cancer (Hause et al., 2016).

1.4 Methods of short tandem repeat analysis

Analysing STRs became accessible towards the end of 1980s when the PCR technology was introduced. This allowed the specific amplification of a repeated region using primers that matched with the unique flanking ends of the locus. PCR products were then sized on a sequencing gel by using Sanger's radiolabelled nucleotide method (Tautz, 1989). This made it possible to use STRs in genomic mapping and population genetic studies (Ellegren, 2004). Genome-wide analysis of STRs began from the late 2000s (Molla et al., 2009; Payseur et al., 2011). These studies were limited to few genomes as the throughput of the used Sanger's sequencing methods is low (Willems et al., 2014). The rise of HTS has transformed the field of genetics and increased the throughput of sequencing data orders of magnitude (Goodwin et al., 2016). However, Sanger sequencing still remains as the gold standard method to confirm sequencing variants (Willems et al., 2014).

HTS has opened new doors in the analysis of STRs and now enables researchers to profile STRs from a genome at a much bigger scale which was not possible beforehand (Caspar, S. M. et al., 2018). Nowadays it takes only hours or days to sequence whole-exome or whole-genome (Caspar, S. M. et al., 2018). However, genotyping STRs from HTS data has its own challenges (Figure 1.5). Library preparation methods that use PCR amplification will often cause stutter noise that produces alleles with the wrong size due to *in vitro* DNA polymerase slippage (Gymrek, 2017), however it is becoming more common to generate data with PCR free protocols. Many analysis algorithms require reads to be long enough to span the whole STR region and have an unique flanking sequence in both ends to accurately align reads on a reference genome (Press et al., 2014). Difficulties in sequencing GC rich regions, complex STRs or repeats being in large chunks that are difficult to align create additional challenges (Caspar, S. M. et al., 2018; Gymrek, 2017). And finally, aligning reads to a reference genome usually results in detecting a variant because the source DNA nearly never matches with the reference (Treangen & Salzberg, 2012).

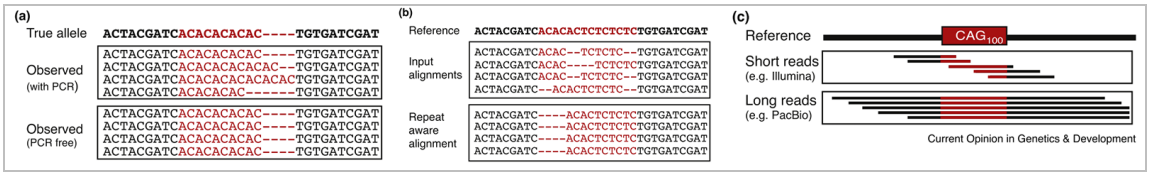


Figure 1.5. Technical difficulties in genotyping STRs from HTS data. (a) PCR amplification during a library preparation can result in stutter noise, that is, creating alleles with wrong size of repeats (top box), in contrast to PCR-free sequencing (bottom box). (b) Issues with alignment as it is difficult for complex repeats and results in inconsistent results (top box), whereas tools using repeat-aware technique aligns reads consistently (bottom box). (c) Short reads (for example, less than 100) cannot span over the allele that has at least 100 repeats (top box), whereas, long reads can cover the whole STR region (bottom box). STR repeats are illustrated in red and non-repetitive regions in black (Gymrek, 2017).

Sequencing data which consists of short reads (for example, 36 bp, 50 bp and 76 bp) and with low genome coverage (5–20x) has been found to have low accuracy for genotyping STRs as calls that are based only on few reads are not enough to make confident calls even for homozygous germline alleles (Figure 1.6) (Press et al., 2014).

Overcoming these challenges has, however, progressed enormously over the past 5 years. In particular sequencing technology is developing with cheaper sequencing and longer read length and Illumina has introduced amplification-free library preparation methods (Kozarewa et al., 2009). Technologies for long-read sequencing exists now as well, including capturing GC-rich regions in a genome (Caspar, S. M. et al., 2018).

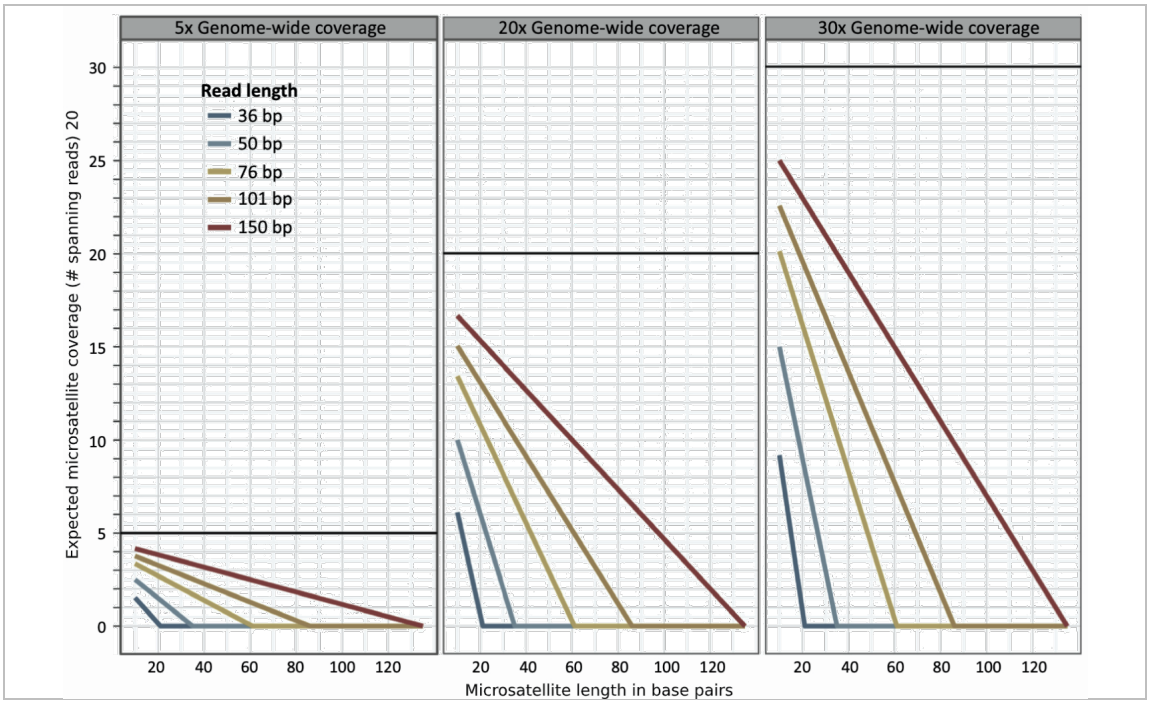


Figure 1.6. Expected STR coverage depending on the read length and sequencing depth (8 bp flanking end sequence on each side is assumed). The longer the read length and sequencing depth is, the higher is the expected coverage of STRs will be. Black bars show the sequencing coverage (Press et al., 2014).

With the rise of HTS, many analysis tools have been developed over the past decade for genotyping STRs from HTS data. Some of these tools are genome-wide genotypers, such as lobSTR (Gymrek et al., 2012), RepeatSeq (Highnam et al., 2013), HipSTR (Willems et al., 2017) and GangSTR (Mousavi et al., 2019). Others have a specific aim, such as to capture long expansions of STRs. For example, STRviper (Cao et al., 2014), ExpansionHunter (Dolzhenko et al., 2017), TREDPARSE (Tang et al., 2017), STRetch (Dashnow et al., 2018) and exSTRa (Tankard et al., 2018) are built for this purpose. A couple of tools are meant for targeted analysis and require a specified set of genomic loci to be described. For instance, ExpansionHunter, Dante (Budiš et al., 2019) and STRScan (Tang & Nzabarushimana, 2017). A tool named popSTR (Kristmundsdóttir et al., 2017) is specifically built for a population analysis.

Many of these tools also aim to overcome the challenges described here before. In particular, many of them have built-in stutter models to identify amplification artefacts from sequencing data. Some of them also use data features beyond genome alignment to try to overcome the limits of read lengths and identify long expansions from a short-read sequencing data.

1.5 Aims and outline of the thesis

STRs are challenging to analyse from HTS data due to multiple technical issues. Several methods that address these issues are now widely used to perform genotyping in various contexts. It has been shown by previous research that STRs are complex and different genotyping performance can be seen for different size of repeat units, different repeat unit composition, different length of repeated sequence and also their location in a genome. However, the accuracy of available tools for genotyping STRs is relatively unknown because tools have only been tested by authors of the publication and using different datasets and various evaluation methods (such as Mendelian inheritance, simulated data or capillary electrophoresis).

This thesis has two central aims. Firstly, to determine the accuracy of common STR genotyping tools and investigate repeats with different motif length and composition separately to provide better strategies for filtering STRs data and maximise the use of it. Secondly, validate, improve and create new methods to detect the highest number of pathogenic alleles in known diseases which are caused by STR expansions.

In Chapter 2, four of the most widely used STR genotyping tools are examined on the Simons Simplex Collection dataset of whole exome sequencing samples. A novel method of evaluation is used where calls in the X chromosome of male samples will be analysed for this purpose. The accuracy is examined for repeats in the size range of 1 to 6 bp and related to confidence scores of calls as well as to the coverage.

In Chapter 3, a single genotyping tool is used to observe accuracy across the entire genome in replicates of whole genome sequenced samples of NA12878. Accuracy of STR genotyping is looked at in more depth than in the previous chapter. This is determined for each repeat unit separately, as well as for each motif and for different repeat sequence length. Finally, mutation rate for each repeat unit is examined to see the magnitude of erroneous calls. As replicate samples had different pre-sequencing (such as number of PCR-cycles) as well as sequencing (such as insert size) protocols, these parameters were also investigated in relation to previously described accuracy of STR genotyping.

In Chapter 4, we focus on enabling the analysis of known pathogenic STR loci from a short-read sequencing data. To do this we perform an extensive search of the literature to find all diseases caused by STR mutations and compile a comprehensive catalogue. We use the data in the literature and upgrade a current method available for targeted analysis to be able to genotype all known disease-caused STR loci. The upgraded method is then validated by simulating samples with mutations in all pathogenic loci and finally applied on a real rare diseases' dataset. Moreover, we develop our own software application which enables a user in a quick and straightforward way to genotype nearly all known pathogenic STRs loci.

Finally, in Chapter 5, the findings of this thesis are drawn together and discussed.

Chapter 2: Accuracy of short tandem repeats genotyping tools in whole exome sequencing data

Published on 23 March 2020: Halman & Oshlack. (2020). Accuracy of short tandem repeats genotyping tools in whole exome sequencing data. F1000Research, 9(200).

Abstract

Background: Short tandem repeats are important source of genetic variation, they are highly mutable and repeat expansions are associated dozens of human disorders, such as Huntington's disease and spinocerebellar ataxias. Technical advantages in sequencing technology have made it possible to analyse these repeats at large scale; however, accurate genotyping is still a challenging task. We compared four different short tandem repeats genotyping tools on whole exome sequencing data to determine their genotyping performance and limits, which will aid other researchers to choose a suitable tool and parameters for analysis.

Methods: The analysis was performed on the Simons Simplex Collection dataset, where we used a novel method of evaluation with accuracy determined by the rate of homozygous calls on the X chromosome of male samples. In total we analysed 433 samples and around a million genotypes for evaluating tools on whole exome sequencing data.

Results: We determined a relatively good performance of all tools when genotyping repeats of 3-6 bp in length, which could be improved with coverage and quality score filtering. However, genotyping homopolymers was challenging for all tools and a high error rate was present across different thresholds of coverage and quality scores. Interestingly, dinucleotide repeats displayed a high error rate as well, which was found to be mainly caused by the AC/TG repeats. Overall, LobSTR was able to make the most calls and was also the fastest tool, while RepeatSeq and HipSTR exhibited the lowest heterozygous error rate at low coverage.

Conclusions: All tools have different strengths and weaknesses and the choice may depend on the type of analysis. In this analysis we demonstrated the effect of using different filtering parameters and offered recommendations based on the trade-off between the best accuracy of genotyping and the highest number of calls.

2.1 Introduction

2.1.1 Overview of short tandem repeats (STRs) and methods of analysis

STRs, also known as microsatellites, consist of repeated units of 1 to 6 base pairs (bp) in length and cover about 3% of the human genome (Gymrek, 2017). STRs are highly mutable and often vary in their number of repeat units across the population. They can be found in various regions of the genome, including in or near protein coding regions and introns (Hannan, 2018). Expanded variants contribute to several dozen human disorders, including Huntington's disease, fragile X syndrome, spinocerebellar ataxias and other diseases. In addition, variation in STR length has been shown to associate with quantitative traits such as gene expression (Gymrek, 2017). The standard method to genotype the length of STRs is to perform polymerase chain reaction (PCR) amplification on the region of interest and gel electrophoresis (Tang & Nzabarushimana, 2017). Sanger sequencing has high accuracy, but low throughput, limiting analysis to a few genes at a time (Caspar, S.M. et al., 2018).

Recent technology advances in high-throughput sequencing (HTS) have revolutionised the genomics field and brought us the opportunity to detect sequence variants at a scale that was impossible before (Caspar, S.M. et al., 2018). HTS of the whole genome provides the potential to profile over a million STRs in the human genome. Recent advances in bioinformatics have brought us tools to analyse STRs from sequencing data (HTS), but genotyping still remains challenging for many reasons, including issues with extreme GC content, short read lengths that do not span over the entire repeat, and issues with alignment due to variation in STRs appearing as large insertions or deletions relative to the reference. In addition, using PCR amplification during library preparation will often cause stutter noise and produce artificial variability in the sequence (Caspar, S.M. et al., 2018; Gymrek, 2017). Stutter noise is a result of *in vitro* slippage of DNA polymerase during PCR cycles that leads to erroneous reads of incorrect repeat length (Willems et al., 2014), which contributes to challenges in genotyping.

Illumina has developed a method for amplification-free (PCR−) library preparation (Kozarewa et al., 2009), which theoretically eliminates the STR stutter error during PCR amplification in sample preparation (PCR+) and therefore improves the accuracy of STR genotyping. The developers of STR-FM evaluated the new protocol by running their tool in both PCR− and PCR+ samples and found that the PCR− protocol compared to PCR+

has up to nine-fold fewer errors (Fungtammasan et al., 2015). However, huge amounts of sequencing data have already been generated by using the PCR+ protocol, where some data will not be resequenced due to time and/or cost (Fungtammasan et al., 2015). In addition, despite the advantages of whole genome sequencing (WGS), whole exome sequencing (WES) is still widely used in human genetics due to its lower cost and higher coverage and WES is a PCR+ process (Björn et al., 2018). Therefore, tools that can accurately genotype STRs not only from PCR− but also from PCR+ data are essential.

While there are a number of computational tools that have been developed to genotype STR alleles in HTS data, there have been few independent comparisons of their performance. Evaluation of methods for genotyping STRs is difficult. The gold standard measurement of STRs is by capillary electrophoresis (Willems et al., 2014), but these methods have low throughput. Further evaluations have used Mendelian inheritance as a measure of accuracy (Gymrek et al., 2012; Highnam et al., 2013; Mousavi et al., 2019). Other studies have used simulated data for the evaluation of genotyping accuracy (Fungtammasan et al., 2015; Highnam et al., 2013). While simulation can generate many loci with known alleles, it is difficult to simulate the true complexity of real data.

Here we propose to compare and evaluate STR genotyping methods on exome data using a different but complementary approach. We used the natural hemizygous state of the X chromosomes in males to look for incorrect calls revealed by a heterozygous call. With repeats on the X chromosome in males there is only one allele, so we expect all calls to be homozygous. While this approach does not evaluate the accuracy of the allele length, it has advantages in that (a) the data sets are large so we can test thousands of calls, and (b) the data comes from real patients with all the noise and biases found in real data.

In our study, we compared LobSTR (Gymrek et al., 2012), RepeatSeq (Highnam et al., 2013), HipSTR (Willems et al., 2017) and a recently published tool GangSTR (Mousavi et al., 2019). In addition, we included a common variant calling tool GATK HaplotypeCaller (McKenna et al., 2010) as a comparison of genotyping accuracy.

There are a number of tools that have been developed for STR analysis and which were excluded from this analysis. For example, popSTR (Kristmundsdóttir et al., 2017) is a population based STR genotyper and is optimised for whole genome sequencing (WGS) data. STRviper is another method for genotyping STRs that is able to pick up repeats

longer than the read length; however, it has no built-in stutter model and it is not suitable for diploid dataset as it assumes only one allele (Cao et al., 2014). Galaxy environment has an STR analysis tool called STR-FM which we were unable to run (Fungtammasan et al., 2015). Dante (Budiš et al., 2019) and STRScan (Tang & Nzabarushimana, 2017) are designed for targeted searches and require a user-defined list of STR loci.

Tools such as Expansion Hunter (Dolzhenko et al., 2017; Dolzhenko et al., 2019), TREDPARSE (Tang et al., 2017), STRetch (Dashnow et al., 2018) and exSTRa (Tankard et al., 2018) were excluded from our analysis as well because they are classified as tools specifically looking for expansions that might be disease causing and are often longer than the physical read length or expansion relative to a control set.

Our analysis focuses on comparing the performance of STR genotyping tools on the X chromosome of more than 400 males. Using this data set, we investigate the overall ability of tools to call genotypes, the accuracy as a function of coverage and repeat unit and also investigate quality scores of the tools. We find most tools are able to call a majority of homozygous alleles and different tools have different advantages in terms of repeat unit and coverage.

2.1.2 Computational tools to genotype STRs from HTS data

First, we will give a short overview of STR genotyping tools included in our analysis and their reported accuracy. The tools evaluated in our analysis are summarised in Table 2.1. All of the tools require a set of defined STR loci. Tandem Repeats Finder (TRF) is a tool that can be used to detect STRs that have two or more copies of the same repeat unit in a row in the reference genome (Benson, 1999). In addition, it can detect repeats for which the repeat unit size is up to several hundreds of bp long. Running TRF generates a report that includes all the loci detected in the genome, with genomic start and end location of the STR, repeat unit and its size, number of copies aligned with the consensus pattern and other relevant information. For this study, we limited the loci defined by TRF to repeat units up to 6 bp (see *Methods*).

Table 2.1. Feature comparison of short tandem repeats (STRs) specific genotyping tools used in our analysis.

	RepeatSeq	**LobSTR**	**HipSTR**	**GangSTR**
Latest version of the tool used in this study	0.8.2 (2014)	4.0.6 (2016)	0.6.2 (2018)	2.4 (2019)
Built-in stutter noise model	✓	✓	✓	✓
Ability to detect STRs that are longer than the read length	✗	✗	✗	✓
Input file types	BAM	BAM FASTA FASTQ	BAM CRAM	BAM CRAM
Sequencing read types	Single- and paired-end reads	Single- and paired-end reads	Single- and paired-end reads	Paired-end reads

2.1.2.1 LobSTR

LobSTR was one of the first successful STR genotyping tools for HTS data. It initially used its own inbuilt aligner but can also use data aligned with BWA-MEM (Li, 2013). LobSTR identifies reads that completely contain the STR and which also have flanking sequence with no repetitive sequence when aligned to a reference genome. As mentioned, PCR amplification during library preparation can create stutter noise at an STR locus, and LobSTR tackles this issue with an included stutter model that aims to detect and account for noise to improve genotyping accuracy. The stutter noise model used can be custom generated from the data or the standardised one supplied by the tool developers. As a result, LobSTR determines and reports the maximum likelihood estimates of the genotype in each locus (Gymrek et al., 2012).

LobSTR was validated using concordance of biological replicates (blood and saliva samples) from the same subject to measure the precision of the tool. At 21x coverage, the discordance rate for genotype was 3% and for allelotype was 2%. While for lower 5x coverage, the discordance rate for genotype was 11% and for allelotype was 5%. STR length differences were analysed in discordant calls that were heterozygous in both blood and saliva samples and found that at coverage 5x or higher, 90% of the errors were one repeat unit difference and 99% of errors were in 2 bp repeat unit size (Gymrek et al., 2012). However, it is important to note that LobSTR validated 2-6 bp repeat unit size

STRs and did not validate homopolymers (Gymrek et al., 2012), which are common in the genome and a known source of genetic variation (Highnam et al., 2013).

2.1.2.2 RepeatSeq

RepeatSeq (Highnam et al., 2013) uses data aligned by an external tool, such as BWA or Bowtie. It uses Bayesian model selection to determine the most probable genotype and requires all reads to fully contain STRs and at least two reads at a locus to make a call. The RepeatSeq noise model is based on genomes derived from over 100 inbred isolates of fly.

RepeatSeq's accuracy was evaluated by analysing a trio WGS data to test consistency with Mendelian inheritance. The authors reported that on minimum coverage of two, 92.1% of repeat calls were consistent with the Mendelian inheritance, while with a minimum coverage of nine it was 95.3% and on minimum coverage of 17 it was 98.0% (Highnam et al., 2013).

2.1.2.3 GangSTR

One of the major drawbacks of the first series of STR profiling methods was that they were limited to genotyping repeats within the read length in HTS data. GangSTR (Mousavi et al., 2019) is a more recent method that incorporates additional information besides repeat-enclosing reads to estimate the length of repeats. This includes available information such as fragment length, coverage and information about partially enclosing reads where only one end contains flanking sequence. More specifically, reads are divided into four classes: 1) enclosing read pairs that have at least one read that includes the whole STR and a flanking region in both ends; 2) spanning read pairs that have a mate pair where one read is aligned to one side of the STR and the second read of the pair on the other side; 3) flanking read pairs that include a read which partially extends into the STR region; and 4) fully repetitive read pairs that have one or two reads which are entirely made of STR (Mousavi et al., 2019). These four classes of reads are used to not only genotype repeats less than the read length but can also be used to genotype longer alleles such as repeat expansions.

The GangSTR method was evaluated by first simulating paired-end 150 bp reads (40x coverage) for 14 repeat expansions involved in STR disorders. Tool accuracy was measured by comparing true and observed alleles and also compared to TREDPARSE

and ExpansionHunter. In this evaluation, GangSTR showed a lower root mean square error (RMSE) rate between true and observed allele lengths for all tested repeats. The authors demonstrated that GangSTR had an advantage over ExpansionHunter and TREADPARSE, especially in alleles that were close to the read length or longer. In addition, GangSTR and ExpansionHunter improved significantly with higher coverage and longer read length.

GangSTR genotyping for disease causing alleles was also tested on validated 14 Huntington's Disease and 25 Fragile X Syndrome real PCR-free WGS data and they reported an RMSE (7.9 and 29.3, respectively) that was lower than for TREDPARSE (8.3 and 34.8, respectively) and ExpansionHunter (10.1 and 27.3, respectively). In evaluations of genotyping a WGS trio, GangSTR was found to have similar performance to HipSTR for shorter alleles (Mousavi et al., 2019).

2.1.2.4 HipSTR

HipSTR (Willems et al., 2017) is a haplotype-based method for genotyping, haplotyping and phasing STRs. While other STR tools are made for finding true length of repeats independently along the genome, HipSTR takes into account the whole repeat structure on the allele, which may also have missing data. HipSTR accuracy was tested by comparing calls from 118 PCR− WGS samples to capillary electrophoresis data, reporting about 98.8% consistency between the two datasets (Willems et al., 2017).

2.1.2.5 GATK HaplotypeCaller

GATK HaplotypeCaller (GATK-HC) (McKenna et al., 2010) can also be used for finding SNPs and indels in repeat regions, but it is not specifically made for STR analysis. It has been widely documented that indel calling is not as accurate as SNP calling and indel callers are not ideal for identifying STR mutation due to the lack of reporting repeat genotypes. Instead, indel callers report insertions or deletions of bases relative to the reference, which may or may not be a multiple of the repeat unit, as well as including SNP differences. Dedicated STR callers, however, use information about the repeat unit, composition and repeat length in order to make more accurate genotype calls (Highnam et al., 2013).

2.2 Results

In order to evaluate the accuracy and performance of STR genotyping methods, we used a novel evaluation approach applied to exome sequencing data of more than 430 individuals. Several previous comparisons determined accuracy by comparing the estimated lengths of repeats to a known truth, determined from either simulations or alternative assays such as PCR. Here, we took only male individuals and looked at the heterozygosity of the calls only on the X chromosome. As there is only one X chromosome in males, a method that reported only homozygous calls was defined to be more accurate than those that reported heterozygous calls.

2.2.1 Dataset

We began with a dataset of 472 males from the Simons Simplex Collection. We had to remove 39 samples for a variety of reasons: six samples were not sequenced with a paired end approach, three samples had no coverage on the Y chromosome so were assumed to be females mislabelled as males, 28 samples produced an error in GangSTR and two samples could not be aligned to the reference genome by BWA due to a software error. The remaining 433 samples (Supplementary Table 1, see *Underlying data*) (Halman, 2020b) were analysed with LobSTR (Gymrek et al., 2012), RepeatSeq (Highnam et al., 2013), HipSTR (Willems et al., 2017) and GangSTR (Mousavi et al., 2019) (see *Methods*). In addition, variant calling was performed using the GATK best practices pipeline.

In brief, the FASTQ files were mapped using BWA-MEM (Li, 2013) and the same BAM files were used as the starting point for running each STR calling method. Each method requires a set of intervals that define repeats to be genotyped. To generate this, we used Tandem Repeats Finder (Benson, 1999) to locate tandem repeats in the hg19 reference genome and detected 224774 STR loci in the X chromosome. Because we are using exome data, we only analysed calls in the 6860 capture regions on the X chromosome. In total, we found 2322 STR loci overlapping the capture regions (Figure 2.1A), where almost 60% of loci consist of 6 bp repeat units (Figure 2.1B). In our full data set, across the 433 individuals, we have over a million STR loci for analysis.

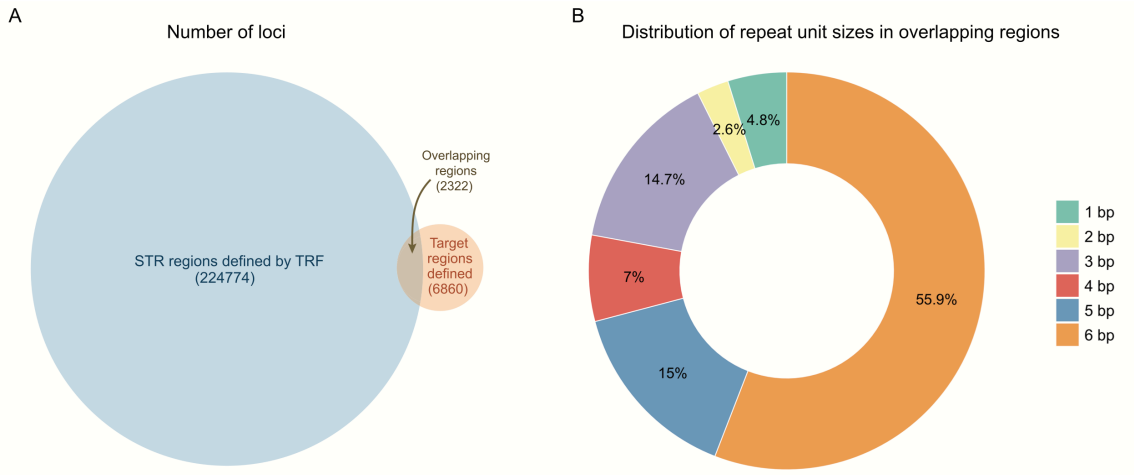


Figure 2.1. STR loci and repeat unit distribution in the X chromosome. (A) Number of STR loci defined by TRF and number of regions in the capture regions of the X chromosome. Overlapping regions include all STR loci that are completely or partially overlapping the target region. Total number of STRs found in target regions is 2322. (B) Distribution of all repeat unit sizes in overlapping regions (2322).

First, we looked the ability of a method to make a call at any given locus in the capture regions. By looking at the total number of calls on the X chromosome for each method, we found that LobSTR reported the highest number of loci (Figure 2.2A). However, the number reported for each individual was variable. Figure 2.2B shows the distribution of the number of reported loci per individual, with the highest median number of calls by LobSTR (2015), outperforming GangSTR (1967), RepeatSeq (1847) and HipSTR (1834). GATK-HC reported a median of 11 loci per individual but, rather than genotyping all loci, GATK only makes a report when it deems there is a difference from the reference genome. Out of the 2322 loci we investigated, there were 23 loci for which the reference STR length was longer than the read length (this does not necessarily mean that the allele length is longer). Out of these 23 loci, genotypes for six loci were not reported by any tool in any individual. For the remaining 17 loci, all of them were reported by GangSTR in some individuals and four of them were genotyped by other tools as well. We also found another 13 loci were not detected by any of the tools in any of the individuals.

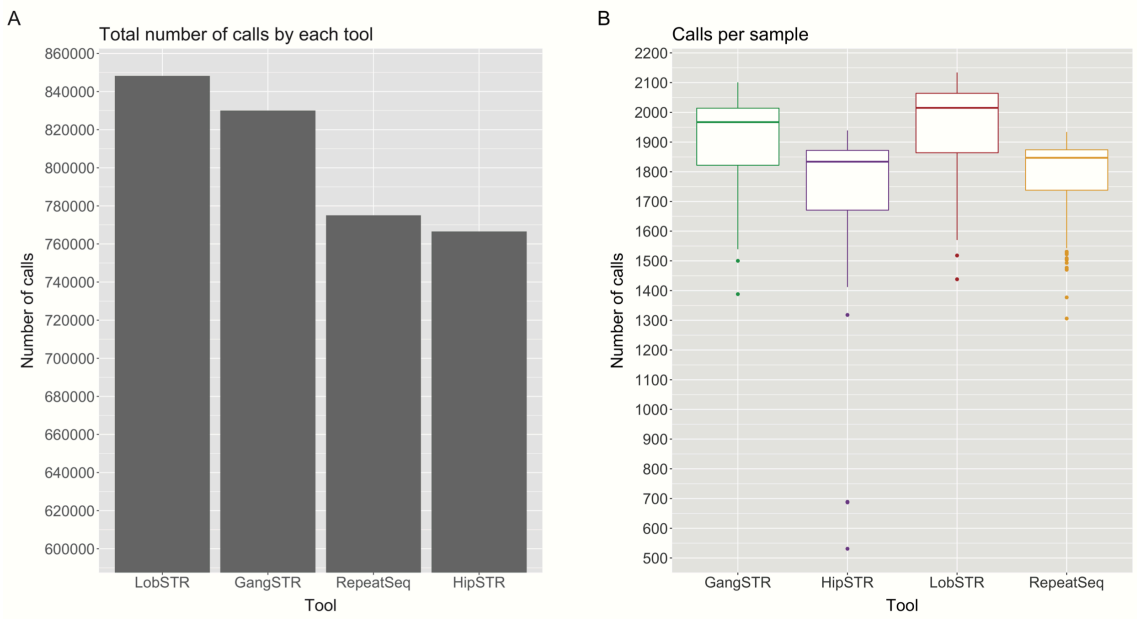


Figure 2.2. Comparison of different short tandem repeat genotyping tools. (A) Total number of calls on the X chromosome over all 433 samples made by each tool. (B) Number of calls made per sample by each tool out of a possible 2322.

GATK only makes calls at positions where there is evidence that the allele is different from the reference. In this dataset, GATK made calls in a total of 346 (14.9%) different loci and 21.4% of these were heterozygous, giving a minimum overall heterozygous rate of 3.1%, assuming all uncalled positions are homozygous. No call either means the allele is a homozygous reference or there is not enough data to make a call. This is one reason why specialised STR callers are better suited for genotyping STR loci.

To determine the genotyping accuracy of the four specialised STR callers, we first looked at the overall percentage of heterozygous calls to estimate the error rate for each method. Overall, RepeatSeq had the lowest median error rate, with 8675 (1.09%) of its calls being heterozygous. Next was HipSTR with 19459 (2.23%), LobSTR with 27410 (2.96%) and GangSTR with 33204 (3.29%). Again, error rates were variable across individual samples, ranging between 0% and 47.3%. RepeatSeq, HipSTR and LobSTR are generally consistent, with three sample outliers with respect to error rate, while GangSTR has higher variability in error rates across samples (Supplementary Figure 2, *Extended data*) (Halman, 2020a). Interestingly, for these three samples, the heterozygous percentage increased for LobSTR after the strict filtering. All tools except RepeatSeq recommend filtering the outputs based on quality metrics of the calls (see *Methods* for details on filtering parameters that were used). Once the recommended filters are applied, we found that the performance of GangSTR improved by 2.76% to 0.53%, HipSTR by 2.14% to

0.09% and LobSTR by 1.32% to 1.64% (Figure 2.3). However, the median number of calls per sample dropped for LobSTR by 201, HipSTR by 1462 loci and GangSTR by 1512.

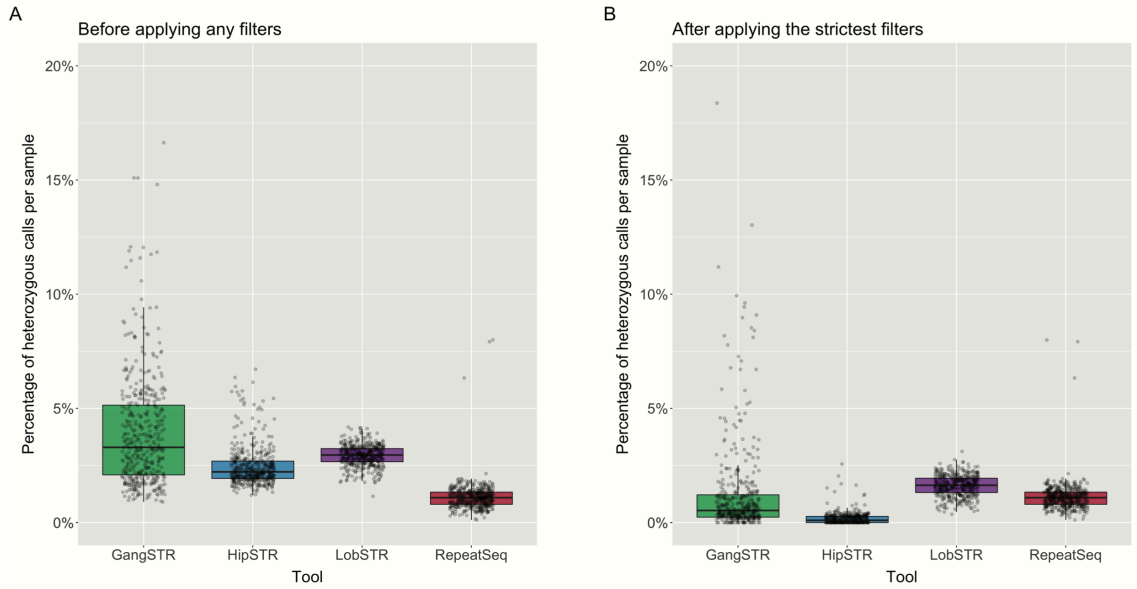


Figure 2.3. Comparison of different short tandem repeat genotyping tools on making heterozygous calls with and without filters. (A) Percentage of heterozygous calls over all samples (each dot is a sample) - no filters applied. (B) Percentage of heterozygous calls over all samples after applied the strictest recommended filters for the tools. Since no filters were recommended for RepeatSeq then it has the same values on both plots.

The recommended filters for each tool were different (for instance, minimum coverage of 100 for HipSTR and 50 for GangSTR) and therefore we next decided to analyse the effect of these filtering parameters separately.

2.2.2 Effects of repeat unit and coverage on accuracy

We investigated the number of heterozygous calls as a function of the repeat unit length, ranging from 1 to 6 bp. We found that all tools exhibit high error rates for 1 bp repeats, which is not surprising as it is difficult to genotype homopolymers due to higher rates of polymerase slippage events. More surprisingly, 2 bp repeats were poorly genotyped by HipSTR, LobSTR and GangSTR and the best results were obtained with RepeatSeq. All other repeat unit lengths produced much more accurate genotypes.

To investigate the effect of coverage and quality scores on results, we applied call-level filters to our data according to developers' recommendations to get two different datasets: low filtering, where we included all suggested filters except coverage and quality scores, and strict filters, where we also included filters for coverage and quality scores. Then, we

looked at the effects of coverage and plotted the error rate as well as percentage of the remaining number of calls as a function of the minimum number of reads supporting the call. We expected the error rate to drop as coverage increased and this was the case for 3, 4, 5 and 6 bp repeat units. However, mono- and dinucleotide repeats did not follow a consistent pattern and the pattern was different between tools. While dinucleotide repeats showed a trend towards a lower error rate with increasing coverage for RepeatSeq and LobSTR, the trend was reversed for GangSTR (Figure 2.4).

Two call-level filters are recommended for GangSTR: level 1 filters, which require filtering out calls that have less than 20 reads, and level 2 filters, which require at least 50 reads to support a call. Filtering out all calls with coverage of less than 20 reads brings the heterozygous error rate for the 3-6 bp repeats to between 1.6 and 3.1% at this minimum coverage. Filtering out calls with less than 50 reads improves the error rate even further to 0.83-1.97%. However, by filtering out calls with coverage of less than 20, we lose on average 39.4% of 2-6 bp repeats data, with a median of 1120 loci reported per sample. By filtering out calls with coverage less than 50, we lose on average 71.7% of 2-6 bp repeats data and have a median of 464 loci reported. Interestingly, we see an increase in error rate as a function of coverage when genotyping 2 bp STRs (Figure 2.4A). Unlike the other tools, where heterozygous percentage decreases when coverage increases and remains relatively steady, GangSTR is not so consistent and fluctuates around 1%.

HipSTR excludes calls that have coverage less than 100 by default unless specified otherwise. This coverage gives high accuracy but also filters out 75.9% of data. We investigated reducing the minimum coverage and found HipSTR has excellent accuracy even with minimum coverage of 25: 0.02-0.04% heterozygous rate for 3, 5 and 6 bp and 0.16% for 4 bp repeats (Figure 2.4B). In addition, a minimum coverage of 50 improves results even more and the maximum error rate for these repeats is 0.01% at this minimum coverage (median of 899 calls per sample). HipSTR also struggles to genotype 2 bp repeats, having a heterozygous error rate around 5.2% for coverage of both 25 and 50. Therefore, by decreasing the filtering parameters for coverage to only exclude the calls less than 25x coverage, we retain on average 76.6% of data for 2-6 bp repeat units, with a moderate call rate (median of 1369 calls per sample).

LobSTR recommends filtering out calls with coverage less than five, which gives us a heterozygous percentage for 3-6 bp repeat units at this minimum coverage of 0.44-0.90%

per repeat unit length. This filters out on average 10.0% of calls for 2-6 bp repeats, giving a median of 1840 reported alleles per sample. The accuracy for these repeats increases as a function of coverage and considering the amount of calls that are filtered out, a minimum coverage of 5-10 might be the best compromise. By filtering out calls with coverage less than 10, the heterozygous error rate at this minimum coverage is 0.55% for 4 bp and 5 bp repeats, 0.48% for 3 bp repeats and 0.36% for 6 bp repeats, while still retaining 74.5% of calls on average for 2-6 bp repeats (median of 1555 calls per sample). Increasing the minimum coverage to 20 reads would improve our results by a further 0.2% but also filter out significantly more calls. For dinucleotide repeats, accuracy seems to get better with higher coverage and reaches 1% error rate when the minimum coverage is 33; however, the error rate increases again when coverage increases over 42 (Figure 2.4C).

Authors of RepeatSeq do not recommend any additional filtering and we found high accuracy even at low coverage. Filtering out all calls less than coverage of five will result in an error rate for 3-6 bp repeat units of 0.04-0.08% and only an average 7.1% of 2-6 bp repeat calls filtered out, leaving an average of 1728 reported calls per sample. Among all tools, RepeatSeq shows the best accuracy for dinucleotides, having an error rate no more than 1.81% (Figure 2.4D).

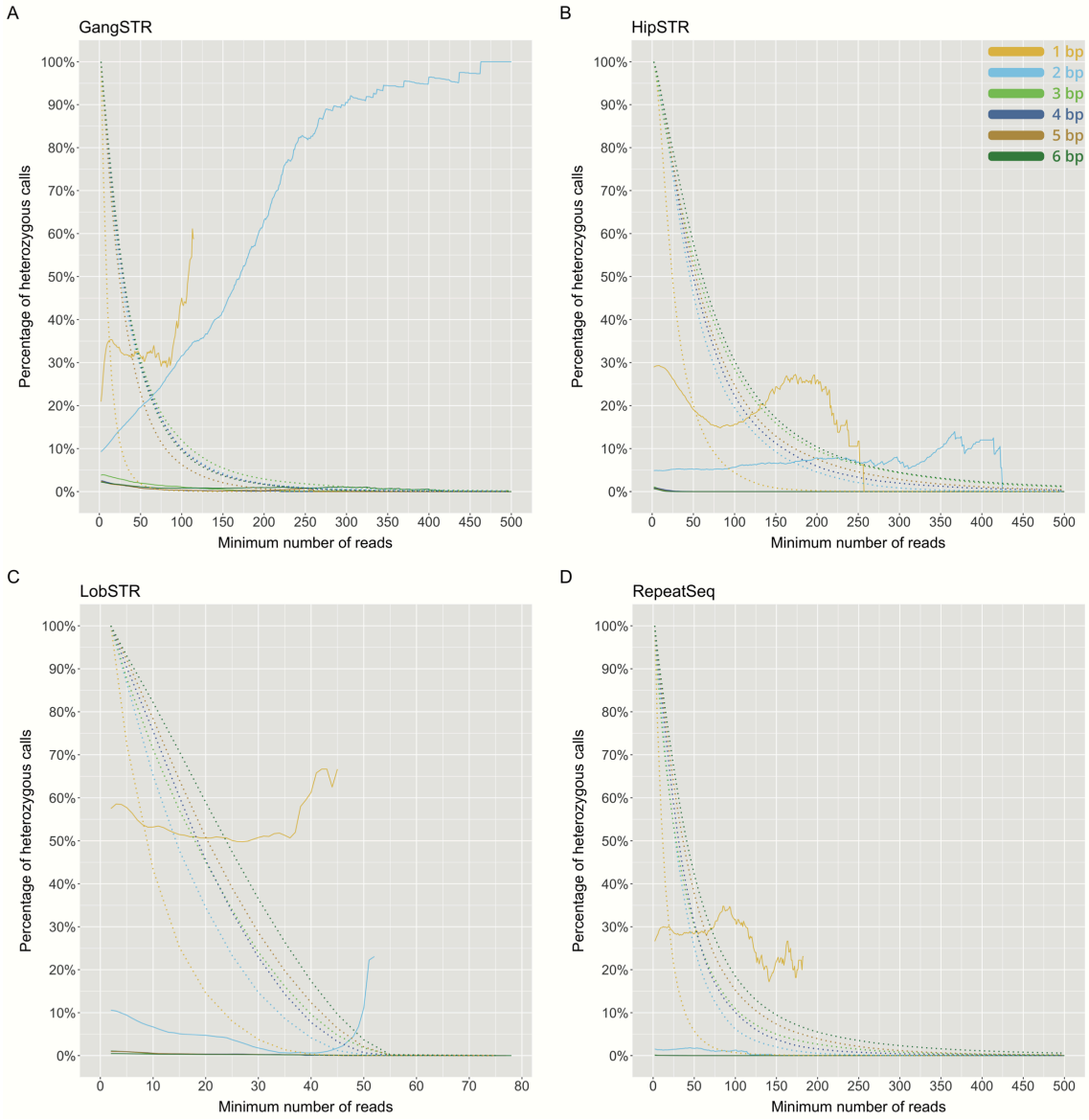


Figure 2.4. Percentage of heterozygous calls per minimum number of reads for each short tandem repeat genotyping tool. (A) GangSTR, (B) HipSTR, (C) LobSTR, (D) RepeatSeq. Solid line shows the percentage of heterozygous calls as a function of minimum number of reads. Dotted line represents the percentage of remaining calls as a function of minimum number of reads. Y-axis limited to 500 reads. Heterozygous calls are represented in percentages, but the total number of calls is different for each tool, where 100% is 796775 calls for GangSTR, 757432 calls for HipSTR, 848252 calls for LobSTR and 775030 calls for RepeatSeq.

As mentioned, we saw unusually high error rates for dinucleotide repeats in nearly all tools, so we examined these repeats in more detail. Curiously, we found that for GangSTR and LobSTR, the unusually high error rate of 2 bp repeats were due to AC/TG repeats, while other repeat units do not exhibit these characteristics (Figure 2.5). The same pattern was observed for RepeatSeq but at lower error rates. This information is not easily available for HipSTR as it does not report the repeat unit.

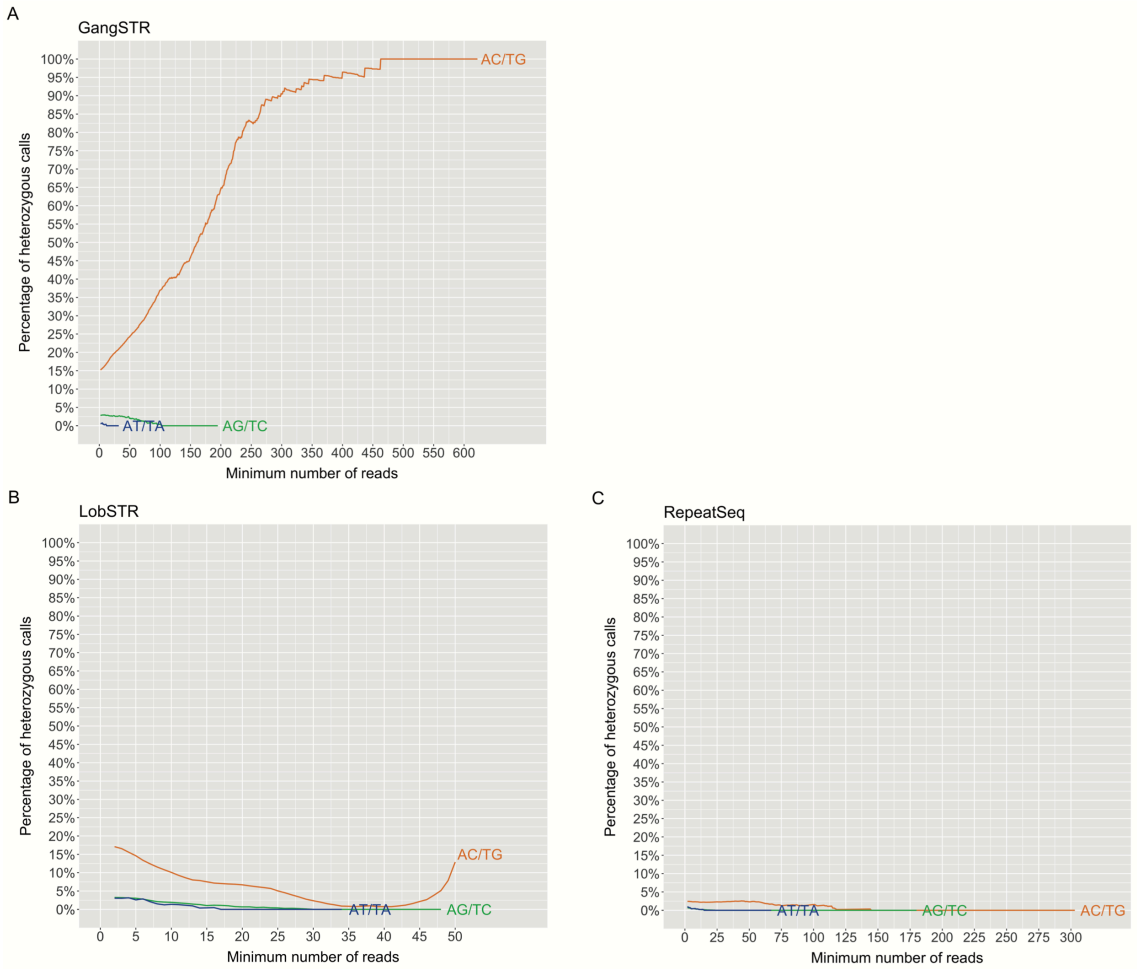


Figure 2.5. Percentage of heterozygous calls per minimum number of reads for dinucleotide repeats. (A) GangSTR, (B) LobSTR and (C) RepeatSeq, divided by the three possible repeat unit sequences.

2.2.3 Effects of repeat unit and quality scores on accuracy

Besides coverage, the second parameter which is commonly used for filtering in all tools is the call quality or quality score produced by each algorithm. We next investigated the effects of quality scores on accuracy by relaxing the coverage filtering and looking at the quality scores in different bins across the score range (Figure 2.6). GangSTR's level 2 filters recommend filtering out calls that have quality scores below 0.9. We see in Figure 2.6A that the heterozygous error rate fluctuates at lower scores and starts to decrease from quality scores above 0.6 for all repeats except for homopolymers. We reach a 1% heterozygous error rate when excluding calls that have quality scores below 0.81 for 3 bp repeat units, 0.66 for 4 bp units, 0.77 for 5 bp units and 0.93 for 6 bp repeat units. We see the lowest heterozygous error rate at the quality score of 1.0, but we also determined a sharp drop in the number of reported genotypes after excluding calls with quality scores below 1.0, and we also lose 61.7% of 3-6 bp repeats data on average. The recommended 0.9 seems a reasonable suggestion for balancing the accuracy with how much data we

will have left after the filtering. We can also see that the accuracy of 1 bp repeat calls improves with the highest quality score (1.0) and a stronger filter for this repeat unit may be appropriate.

HipSTR, similarly to GangSTR, recommends filtering out calls with a quality score less than 0.9. In Figure 2.6B, we can see that on average the calls with a score below 1.0 have a high heterozygous error rate, and keeping only the ones with the highest quality score will improve overall accuracy. Indeed, 98.5% of calls have a quality score of 1.0 and therefore, we only lose a fraction of data while filtering out calls with a quality score below 1.0, which may be a good trade-off.

LobSTR recommends filtering out calls with a call quality score less than 0.8. As seen in Figure 2.6C, the heterozygous error rate fluctuates but generally shows higher accuracy with higher quality scores for 1-3 bp repeats, while for 4-6 bp there is no clear trend below a quality score of 0.9. There is no significant improvement in overall accuracy when we remove calls with quality scores less than 0.8, or even 0.9, which might be due to the fact that, similar to HipSTR, the majority of calls (94.3%) have a quality score of 1.0. We do see improvement when only leaving calls with quality of 1.0, particularly in 1-2 bp repeat units, which shows utility in filtering out the least accurate calls and since the majority of data has a quality score of 1.0, this filtering could be a good choice, as also suggested for HipSTR (see Supplementary Figure 3, *Extended data*) (Halman, 2020b).

RepeatSeq does not recommend any filtering and reports quality scores on a Phred scale (Figure 2.6D). We determined that filtering out calls with Phred quality score of 10 or less improves the accuracy of all repeat units. Accuracy of genotyping mono- and dinucleotide repeats continues to improve as a function of Phred scores, while the best accuracy is observed at the highest quality score. On the flip side, the number of calls also decreases, and at the highest Phred scores, we are left with 31.6% of homopolymers and 87.8% of dinucleotide repeats. However, it only filters out 1.2% of 3-6 bp repeats data. Overall, filtering data based on the quality scores may be reasonable if looking at mono- or dinucleotide repeats and accuracy is an important factor.

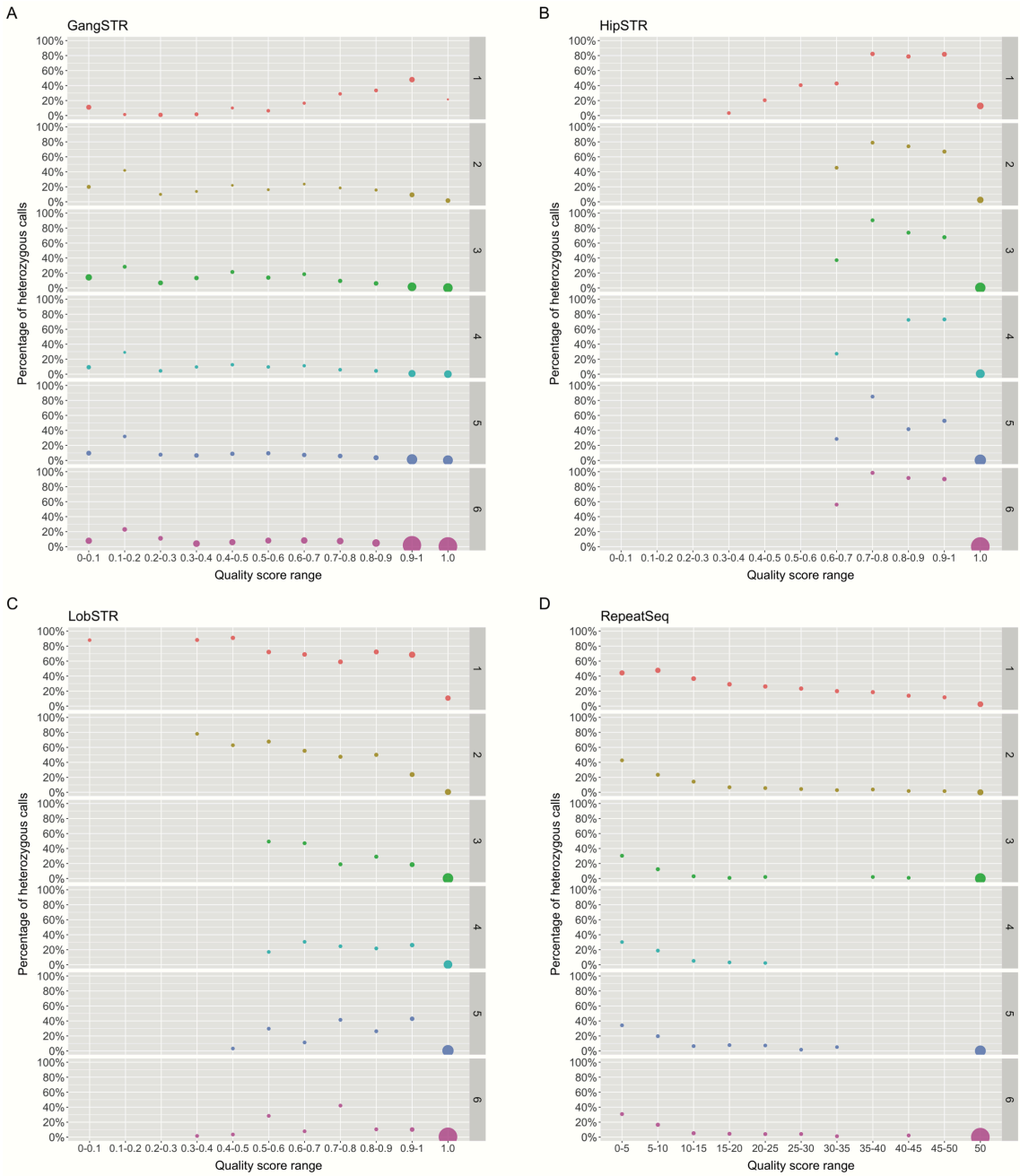


Figure 2.6. Percentage of heterozygous calls in different bins of quality score separated by each repeat unit length. (A) GangSTR, (B) HipSTR, (C) LobSTR, (D) RepeatSeq. Each bin is from marked range (inclusive) to the end value (exclusive), except the last one. Dot size represents the number of calls in that range of the repeat unit length (rows).

2.2.4 Accuracy of GangSTR by looking at only the enclosing class of reads

LobSTR, HipSTR and RepeatSeq use types of reads where the STR region has to be completely in the read. However, GangSTR uses more classes of reads that may give rise to false positives. Therefore, to make a more direct comparison, we decided to look at GangSTR results where we filtered out all other classes of reads besides the enclosing ones, marked here as GangSTR (enc.). Compared to the previous GangSTR results, we now see lower error rates for all repeat units as well as no substantial fluctuation in higher

coverage that was apparent previously in Figure 2.4A. At a coverage of 20, GangSTR (enc.) has a heterozygous error rate of 0.53-0.99% for 3-6 bp repeats, while 58.6% of 2-6 bp repeats data is filtered out, with a median of 699 calls per sample (Figure 2.7A). When we increase the minimum number of reads to 50, we can see even further improvement. This results in an error rate of 0.01-0.27% for 3-6 bp repeats; however, this also filters out 86.3% of 2-6 bp repeats data (median of 152 calls per sample).

We also looked at the dinucleotides separately (see Supplementary Figure 4, *Extended data*) (Halman, 2020b) and found that it is considerably lower compared to the results of GangSTR. However, we determined that AC/TG repeats still have a higher error rate compared to other repeat units (Figure 2.4A and Figure 2.7A). Results of quality scores are quite similar to GangSTR's, where we see improvements at high quality scores (Figure 2.7B).

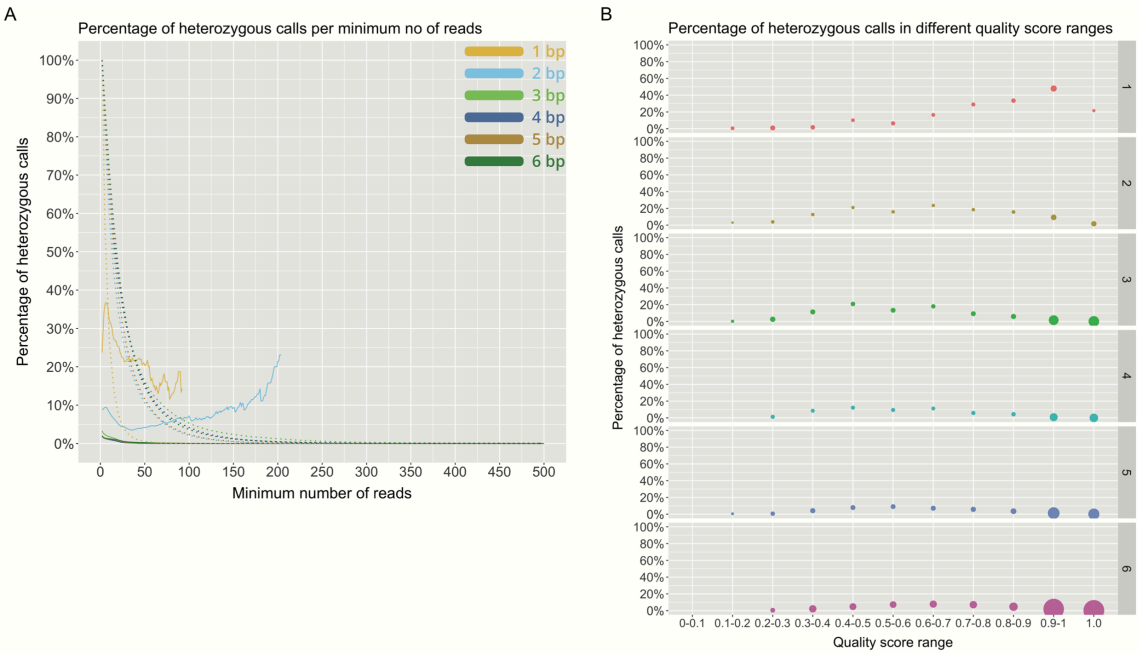


Figure 2.7. Percentage of (A) heterozygous calls per minimum number of reads and (B) in different quality score ranges for GangSTR (enc.). Solid line on (A) represents the percentage of heterozygous calls per minimum number of reads and dotted line represents the percentage of remaining calls per minimum number of reads. Dot size on (B) represents the number of calls in that range of the repeat unit length (rows) and each bin is from marked range (inclusive) to the end value (exclusive), except the last one.

2.2.5 Running time of tools

Finally, we benchmarked all tools to determine their average running time. We ran each tool genome wide by using one, two, four and eight cores and determined that only RepeatSeq supports multithreading, which allows the tool to run faster when utilising more processor cores. In particular, we saw that for samples that have coverage on target regions of around 90x, the average running time on one core for LobSTR was 9 min 31 sec, giving it a clear advantage compared with HipSTR at 1 h 22 min, GangSTR at 3 h 4 min and RepeatSeq 2 h 58 min. By utilizing more cores, LobSTR's, GangSTR's and HipSTR's running time remained the same, while RepeatSeq's time improved.

2.3 Discussion

We used a novel way of evaluating the error in short tandem repeats genotyping methods where we analysed STR calls on the X chromosome of male samples. Because this is a hemizygous chromosome, we determined a relative error rate as the rate of heterozygous genotypes. We performed this evaluation on a human exome dataset of 433 samples, resulting in the evaluation of more than a million STR loci. Exome sequencing is widely used, but the PCR step in library preparation causes a challenge for STR genotyping tools due to the interference of stutter noise. This is the first independent evaluation of these STR genotyping tools that we are aware of.

Many of the tools do not claim to be able to accurately genotype homopolymers and we found that indeed all tools had difficulty with these repeats, resulting in a high error rate. There was also no clear correlation between minimum coverage and accuracy of genotyping homopolymers, but using the highest quality scores did improve the accuracy. Interestingly, most tools produced high error rates for genotyping dinucleotides as well, which we later found to be mainly caused by AC/TG repeat units. One who analyses dinucleotide repeats with these tools should be aware of the differences in accuracy of genotyping different repeat units and carefully interpret the results of AC/TG repeat units. Repeat units with a length of 3-6 bp were all relatively accurate and similar across tools, with only minor differences. However, genotyping was slightly less accurate for 3 bp length repeats in low coverage and low quality scores, but differences were reduced with proper filtering. We found that LobSTR was able to report the highest number of genotypes at a heterozygous error rate of less than 1%.

There are certain filtering parameters suggested for each tool and we examined the effects of coverage and quality scores across all tools. However, some tools have further parameters that could be explored that were not part of our investigation. In general, we found that higher quality scores increased the accuracy of results at the cost of losing some potentially accurate calls. The relationship with coverage was more complex but some coverage filtering improved results for all tools. Which parameters to use depends on the aim of the analysis. For example, more calls may be desirable to begin a screen, or more accuracy may be desirable if selecting potential disease associated loci. When one does an exploratory analysis to find potential loci of interest that can be followed up with alternative methods, then lowering filtering parameters for coverage and quality scores

for certain tools could be a good approach as it leaves us with larger portion of data. We found that even in exome data, we can use these tools to genotype tens of thousands of loci.

Unlike the other tools we used in this analysis, GangSTR utilises four different types of reads, which can help to pick up the locus other tools cannot (such as those longer than the read length). However, these can also produce genotyping errors. In our analysis, we first looked at GangSTR results that included all four classes of reads and then we excluded all calls where only spanning or bounding class reads were present, as suggested by the tool authors. This filtering increased the genotyping accuracy of the tool (we also looked at the results where we skipped this filtering parameter but this did not improve results). Still, compared to other tools, GangSTR showed a higher error rate. Finally, we decided to look only at the enclosing class of reads as the other tools do and determined an error rate around three times lower at 20x and bigger gains at higher coverage. On the other hand, that change will result in losing the ability to genotype alleles longer than the read length, which is GangSTR's important addition. We also found that HipSTR has a very high accuracy for 3-6 bp repeats when coverage is at least 50x. Excellent accuracy was also found for RepeatSeq at very low coverage and this was the most accurate among the tools for genotyping dinucleotides. In addition, RepeatSeq is the only tool that supports multithreading and therefore can run faster by allocating more cores.

Here, we have presented one way of performing an evaluation and this approach does not look at accuracy of the estimated allele length, which is a limitation of the study. In addition, it is difficult to rule out a bias towards tools that default to genotyping an allele as a homozygous reference by the software. Our comparisons were specifically analysing an exome dataset that was PCR amplified, where a tools' noise model may play an important role. Therefore, tools may perform differently when we analyse PCR-free WGS datasets.

In conclusion, all these tools are built to genotype STRs but have different strengths and weaknesses. Based on our analysis there is no clear overall winner. RepeatSeq and HipSTR are the best when considering genotyping error rate even with low coverage. On the other hand, GangSTR has an advantage because it is the only tool among them that can call alleles longer than the read length but shows a higher error rate, unless looking at only the enclosed class of reads, which in turn would lose the GangSTR's advantage of

picking up long genotypes. In addition, GangSTR is the newest tool and so comes with reference files for different reference builds that are periodically updated according to the tool's webpage. The correct choice of a tool and the subsequent filtering depends on the aim of the analysis, and might be influenced by available hardware resources and time limit for running tools.

2.4 Methods

2.4.1 Dataset

In order to compare all STR tools, we ran each one of them on the same dataset. We used the data from the publicly available Simons Simplex Collection (SSC) for our analysis.

Samples' genomic DNA was extracted from whole blood. Exomes were captured with NimbleGen EZ Exome v2.0 (Roche Nimblegen, Inc., Madison, WI) reagents and sequenced using Illumina (San Diego, CA) GAIIx (N = 271) or HiSeq 2000 (N = 162) at the Yale University School of Medicine.

In total there were 238 families where only males were selected for our analysis to avoid heterozygous sites in the X chromosome, assuming that any multiallelic STR calls should be a result of PCR and/or sequencing errors. Male samples were determined by using metadata of samples (472 samples) and quality controlled by looking at the coverage on X and Y chromosomes. Results of the analysis led to the exclusion of three samples as they had no coverage on the Y chromosome. Out of the remaining 469 samples, we excluded six single-end read sequenced files as well as 28 paired-end read sequenced samples that did not work on GangSTR, and two additional samples that had issues with mapping, which left us with 433 samples in total (Supplementary Table 1, see *Underlying data*) (Halman, 2020b).

All computational steps (tools and parameters used) are described in this section.

2.4.2 Computational steps

2.4.2.1 SRA to FASTQ conversion

All whole-exome sequencing files in Simons Simplex families were downloaded from NCBI Sequence Read Archive (SRA) and converted to FASTQ files by using fastq-dump (https://ncbi.github.io/sra-tools/fastq-dump.html):

```
fastq-dump \
 --gzip \
 --skip-technical \
 --readids \
 --read-filter pass \
 --split-3 \
 --dumpbase \
 --clip \
 FILE.SRA
```

2.4.2.2 Defining STR regions

The human reference sequence hg19 (February 2009 assembly) was downloaded from UCSC (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.gz) and the "hg19.fa" file was created by and indexed using Samtools v1.10 (Li et al., 2009):

```
cat *.fa > hg19.fa
samtools faidx hg19.fa
```

Creating FASTA sequence dictionary file for GATK analysis:

```
gatk CreateSequenceDictionary
 -R hg19.fa
 -O hg19.dict
```

Tandem Repeats Finder v4.09 (Benson, 1999) was used to find STRs (1-6 bp repeat unit length) in the hg19 reference genome using the following command and parameters:

```
./trf409.linux64 hg19.fa 2 7 7 80 10 24 6 -h
```

A custom-made Python script named trf2bed.py (Halman, 2020a) was used to extract data from the TRF output file to generate a BED regions file for LobSTR, GangSTR, HipSTR, RepeatSeq and GATK.

```
python3 trf2bed.py \
 --dat hg19.fa.2.7.7.80.10.24.6.dat \
 --bed hg19.fa.2.7.7.80.10.24.6_$TOOL.bed \
 --tool $TOOL
```

2.4.2.3 FASTQ alignment and calculating BAM coverage

Reads from FASTQ files were aligned to the hg19 reference genome using BWA-MEM v0.7.17 and aligned BAM files were merged and indexed using Samtools v1.10 (Li et al., 2009).

```
bwa mem -M -t 8 -R
"@RG\\tID:$id\\tPL:$PLATFORM\\tPU:$BARCODE\\tSM:$SAMPLE\\tLB:$LIBRARY" hg19.fa
$INPUT_FILE1.fastq $INPUT_FILE2.fastq | samtools sort -@hg19.fa -o $OUTPUT_FILE.bam -

samtools merge $OUTPUT_FILE.merge.bam $INPUT_FILE1.bam $INPUT_FILE2.bam
samtools index $INPUT_FILE.merge.bam
```

To follow the best practices of GATK duplicate reads were removed:

```
gatk MarkDuplicatesSpark \
 INPUT=$INPUT_FILE.merge.bam \
 OUTPUT=$OUTPUT_FILE.bam \
 --remove-sequencing-duplicates \
 --create-output-bam-index
```

Coverage of BAM files on target regions was found with the MosDepth v0.2.4 tool (https://github.com/brentp/mosdepth/), followed by calculating the median and average coverage:

```
mosdepth \
 -n \
 --fast-mode \
 --by $TARGET_REGIONS.bed \
 $OUPUT_FILE.coverage \
 $INPUT_FILE.merge.bam

gunzip -c $INPUT_FILE.regions.bed.gz | sort -n -k 5 | awk
'{a[NR]=\$5}END{print(NR%2==1)?a[int(NR/2)+1]:(a[NR/2]+a[NR/2+1])/2}' >
$OUTPUT_FILE.avgcov

gunzip -c $INPUT_FILE.regions.bed.gz | sort -n -k 5 | awk '{ sum += \$5; n++ } END { if
(n > 0) print sum / n; }' > $OUTPUT_FILE.medcov
```

2.4.2.4 Calling STRs and genotyping

GangSTR v2.4 (Mousavi et al., 2019) was executed with the following parameters:

```
GangSTR \
 --bam $INPUT_FILE.merge.bam \
 --ref hg19.fa \
 --regions hg19.fa.2.7.7.80.10.24.6_gangstr.bed \
 --out $OUTPUT.vcf \
 --nonuniform \
 --coverage X*

* where X = mean coverage for the particular sample that was calculated by MosDepth tool
as described previously.
```

Strict filtering was done as recommended by the developer using dumpSTR, which is part of STRTools package (https:/github.com/gymreklab/STRTools):

```
dumpSTR \
 --vcf $INPUT_FILE.vcf \
 --out $OUTPUT_FILE \
 --filter-spanbound-only \
 --filter-badCI \
 --max-call-DP 1000 \
 --min-call-DP 50 \
 --min-call-Q 0.9
```

Since we were looking the relationship between coverage and quality scores, and genotyping accuracy separately, we did additional filtering (partial filtering), where we discarded the filtering on calls with low coverage or low-quality scores:

```
dumpSTR \
 --vcf $INPUT_FILE.vcf \
 --out $OUTPUT_FILE \
 --filter-spanbound-only \
 --filter-badCI \
 --max-call-DP 1000
```

RepeatSeq v0.8.2 (Highnam et al., 2013) was executed with the following parameters:

```
repeatseq \
 $INPUT_FILE.bam \
 hg19.fa \
 hg19.fa.2.7.7.80.10.24.6_repeatseq.bed
```

LobSTR v4.0.6 (Gymrek et al., 2012) was downloaded and a custom lobSTR reference was made using lobstr_index.py and GetSTRInfo.py scripts as follows:

```
python ./lobstr/scripts/lobstr_index.py
 --str hg19.fa.2.7.7.80.10.24.6_lobstr.bed \
 --ref hg19.fa \
 --out ./lobstr/hg19_custom/

python ./lobstr/scripts/GetSTRInfo.py \
 hg19.fa.2.7.7.80.10.24.6_lobstr.bed hg19.fa >
 ./lobstr/hg19_custom/lobstr_hg19_custom_strinfo.tab
```

LobSTR's allelotype was used to call STRs and it was run with default parameters, with and without the the *--no-rmdup* flag:

```
./lobstr/bin/allelotype \
 --command classify \
 --bam $INPUT_FILE.merge.bam \
 --index-prefix ./lobstr/hg19_custom/lobstr_hg19_custom_ref/lobSTR_ \
 --strinfo ./lobstr/hg19_custom/lobstr_hg19_custom_strinfo.tab \
 --noise_model ./lobstr/share/lobSTR/models/illumina_v3.pcrfree \
 --out $OUTPUT_FILE.vcf \
 --no-rmdup
```

Willems and colleagues explored the effects of recommended allelotype options for lobSTR (--filter-mapq0, --filter-clipped, --max-repeats-in-ends and --min-read-end-match), but found the optimal settings for lobSTR does not include these parameters and best results are obtained with default ones, which was also reported for RepeatSeq (Willems et al., 2017), and therefore we decided to run both tools with the default parameters.

We did the strict filtering with the lobSTR's filtering tool, based on the author's recommendations for whole genome data:

```
python ./lobstr/share/lobSTR/scripts/lobSTR_filter_vcf.py \
 --vcf $INPUT_FILE.vcf > $OUTPUT_FILE.vcf \
 -loc-cov 5 \
 --loc-log-score 0.8 \
 --loc-call-rate 0.8 \
 --loc-max-ref-length 80 \
 --call-cov 5 \
 --call-log-score 0.8 \
 --call-dist-end 20
```

And the partial filtering:

```
python ./lobstr/share/lobSTR/scripts/lobSTR_filter_vcf.py \
 --vcf $INPUT_FILE.vcf > $OUTPUT_FILE.vcf \
 --loc-call-rate 0.8 \
 --loc-max-ref-length 80 \
 --call-dist-end 20
```

HipSTR v0.6.2 (Willems et al., 2017) was executed with the following parameters:

```
HipSTR \
 --min-reads 2 \
 --def-stutter-model \
 --fasta hg19.fa \
 --regions hg19_2.7.7.80.10.24.6_hipstr.bed \
 --str-vcf $OUTPUT_FILE.vcf.gz \
 --bams $INPUT_FILE.merge.bam
```

Strict filtering was done according to the developer's recommendations:

```
python ./HipSTR/scripts/filter_vcf.py \
 --vcf $INPUT_FILE.vcf \
 --min-call-qual 0.9 \
 --max-call-flank-indel 0.15 \
 --max-call-stutter 0.15 \
 --min-call-allele-bias -2 \
 --min-call-strand-bias -2 > $OUTPUT_FILE.vcf
```

Since we ran HipSTR with "--min-reads 2" parameter, we additionally filtered out all calls that had less than 100 reads, as this is the default parameter that HipSTR uses.

Partial filtering was done:

```
python ./HipSTR/scripts/filter_vcf.py \
 --vcf $INPUT_FILE.vcf \
 --max-call-flank-indel 0.15 \
 --max-call-stutter 0.15 \
 --min-call-allele-bias -2 \
 --min-call-strand-bias -2 > $OUTPUT_FILE.vcf
```

GATK v4.1.2 was executed with the following standard parameters:

```
gatk HaplotypeCaller \
 --reference hg19.fa \
 --intervals hg19_2.7.7.80.10.24.6_gatk.bed \
 --genotyping-mode DISCOVERY \
 --input $INPUT_FILE.merge.bam \
 --output $OUTPUT_FILE.vcf
```

Data were analysed using GATK best practice guidelines (DePristo et al., 2011) up to variant calling. Variant calling was performed with the HaplotypeCaller in GATK (Poplin et al., 2017).

2.4.2.5 Data extraction from variant calling files (VCFs) and analysis

A custom-made Python script named extract-data.py (Halman, 2020a) was created to extract data from outputted VCFs of all tools ran:

```
python3 extract-data.py --tool toolname --vcf ./vcf_folder --chr chrX --out disk

* Where toolname was either gangstr, lobstr, hipstr, repeatseq or gatk
```

In particularly, only calls in the X chromosome were extracted out. In case of filtering, only calls that passed the filter were extracted out. The output file contained information about all STR loci found in the VCF file, having the following fields: sample name, locus, chromosome, start and end coordinates of the STR region, motif (repeat unit), length of

motif, length of the reference, length of alleles, genotype, number of total reads, number of reads supporting the call in each class and the quality score.

The data was then analysed in R by using str-analyse.R together with str-analyse.functions.R script (Halman, 2020a). Bioconductor's GenomicRanges package for R (Lawrence et al., 2013) was used to find and then filter out all calls that fell outside of the target regions. All STR regions that were entirely or partially inside of the target region were included in the analysis, however, all duplicate loci were removed. When calculating heterozygous percentage per minimum number of reads or quality score bins, we only included the results when there were minimum of ten results (samples) to use and calculated the percentage again for each repeat unit length after each read or after 1/10 quality score bin.

2.4.2.6 Running time and multithreading

To see the performance of the STR specific tools we selected five WES samples that had a median coverage on target regions closest to 90x (between 88.6x and 91.6x) and calculated the time each tool ran on each sample individually by using either one, two, four or eight processor cores on the same server that has Intel(R) Xeon(R) 2.60 GHz processors and maximum of 16 GB RAM. Each test was repeated three times and the average time was calculated. Timing was performed with the UNIX time command.

2.4.3 Data availability

The Simons Simplex Collection dataset, Accession number SRP010920: https://identifiers.org/insdc.sra:SRP010920

Harvard Dataverse: Supplementary information for the "Accuracy of short tandem repeats genotyping tools in whole exome sequencing data" article. https://doi.org/10.7910/DVN/RWTGWK (Halman, 2020b)

2.4.4 Software availability

Source code available from: GitLab: https://gitlab.com/andreassh/research-str-wes

Archived source code at time of publication: Zenodo: Source code for the "Accuracy of short tandem repeats genotyping tools in whole exome sequencing data" article. https://doi.org/10.5281/zenodo.3695971 (Halman, 2020a)

Chapter 3: Analysis of short tandem repeats in whole genome sequencing data

3.1 Introduction

In the previous chapter, accuracy of STRs was determined from analysing an exome dataset and specifically looking at the chromosome X. However, the regions captured in whole exome sequencing data account only for 1.5% of the genome (Björn et al., 2018). Expansions of STRs that occur in coding regions are responsible for around half of the diseases discovered so far, but STR expansions in non-coding regions such as introns and untranslated regions (UTRs) are still responsible for the other half of known STR diseases (Khristich & Mirkin, 2020). It is clear that STRs, which are ubiquitous all over the whole genome, can be clinically important despite their location and therefore are important to analyse. While numerous diseases are linked to repeat expansions in exons, some disorders in UTRs and introns have been found as well, for example, myotonic dystrophy type 1 is caused by expansions in the 3′ UTR and the type 2 is a result of expansions in intron (La Spada & Taylor, 2010; Liquori, 2001). Without a doubt, it is important to accurately analyse STRs from both exons, UTRs and introns. In this chapter, a whole genome analysis of STRs is carried out. We survey the detection of STRs in all genomic regions and look at the data in more depth than in the previous chapter. That includes characterising accuracy of STRs in different lengths of alleles, different motif groups and determining the frequency and direction of mutations for each repeat unit size separately.

In this study we used the genomic DNA of the gold-standard NA12878 (genome in the bottle) sample and sequenced it in seven replicates (Table 3.1) on the same flow cell while varying several of the library preparation steps including the number of PCR-cycles beforehand. The NA12878 sample is a gold standard reference often used for benchmarking, including in the area of short tandem repeats (Doi et al., 2014; Gymrek et al., 2012; Highnam et al., 2013; Mousavi et al., 2019; Willems et al., 2017). In our data, a new transposome-based library preparation technology that achieves a consistent fragment sizes across a range of DNA inputs, was used (Bruinsma et al., 2018). However, the library was prepared by using a PCR amplification and it is known that PCR can cause mutations *in vitro* and create alleles with an artificial repeat length, also known as the stutter noise (Caspar, S.M. et al., 2018; Gymrek, 2017). This adds an additional challenge on determining the accuracy of our results and thus, these results were compared with the

publicly available PCR-free, high-quality Illumina Platinum PCR-free NA12878 sample to determine the accuracy of calls. The software of our choice was the most recent whole genome STR genotyping tool GangSTR, as it is also able to detect repeats longer than the read length (Mousavi et al., 2019).

With this data set we also investigated the effects of changing library preparation parameters such as number of PCR-cycles and sequencing-related parameters such as length of an insert size and coverage on recall and accuracy of STRs.

3.2 Results

3.2.1 Dataset

The dataset used in this analysis consists of seven replicate samples of NA12878 which were run on the same flow cell to sequence their whole genome. These samples could be divided into two groups based on the pre-sequencing manipulation: (1) four samples had different DNA input amount used (15.6, 27.2, 54.5 and 108.9 ng) and (2) three samples had different DNA fragmentations prior to library preparation (500, 1000 and 1500 bp). Samples names are related to the library preparation.

Different numbers of PCR cycles were used due to the variations in DNA input amount, which also affects the library yield. Table 3.1 shows how the sequencing coverage for all samples, except one, follows the same trend as the library yield. The exception is the sample frag_500 that had half the library yield but higher sequencing coverage than the sample with the same input DNA and PCR cycles (frag_1000).

Table 3.1. All seven samples and their amount of DNA input, fragmentation size, PCR cycles, library yield and calculated sequencing coverage. Table is sorted by sequencing coverage in descending order. PCR-free sample used in the 3.2.4 section is added in the end (Eberle et al., 2017).

Sample	**Input DNA (ng)**	**Initial fragmentation size**	**PCR cycles**	**Library yield (nM)**	**Coverage (X)**
dna_16	15.6	–	8	35.7	56.1
dna_109	108.9	–	5	27.6	53.4
dna_27	27.2	–	6	24.2	36.2
frag_1500	33.0	1500 bp	6	23.0	32.1
frag_500	30.0	500 bp	6	7.9	30.8
dna_55	54.5	–	5	19.2	29.9
frag_1000	30.0	1000 bp	6	17.7	27.0
PCR-free	–	–	0	–	50

In our initial investigation we chose to eliminate effects of sequencing coverage from the results by downsampling files to have a consistent coverage. However, we first examined STR recall and accuracy over the downsampling process. The sample dna_109 was chosen for this analysis as it has high coverage and low number of PCR cycles. The sample was downsampled ten times, starting from 100% (coverage of 53.4x) and removing further 10% fragments at each step. Downsampling was completed when 10% fragments of the original file were left (coverage of 5.3x). STR recall and accuracy was

determined by comparing the downsampled calls to the original one that contains all fragments.

As expected, STR recall decreases as fragments are being taken out (where recall is defined as the percentage of positions where calls were made in the new data set relative to the calls in the initial data set). STR accuracy is defined as the proportion of the calls with identical allele estimates as the truth data. Predictably accuracy has a larger drop off when the coverage is reduced. Filtering out half of the fragments decreases the proportion of accurate calls by 9.3%, whereas a sharper drop is seen over the lower coverage steps (Figure 3.1). At the lowest coverage (5.3x) STR recall is 75.1% and accuracy only 56.5%, but increases above 90% at the coverage of 27x. STR recall increases faster and plateaus early whereas accuracy has a more gradual increase. As the lowest coverage among our samples is also 27x (frag_1000), we chose this value to downsample all samples to have uniform (27x) coverage for the following analysis. Downsampling was performed using Picard's DownsampleSam function (see Methods).

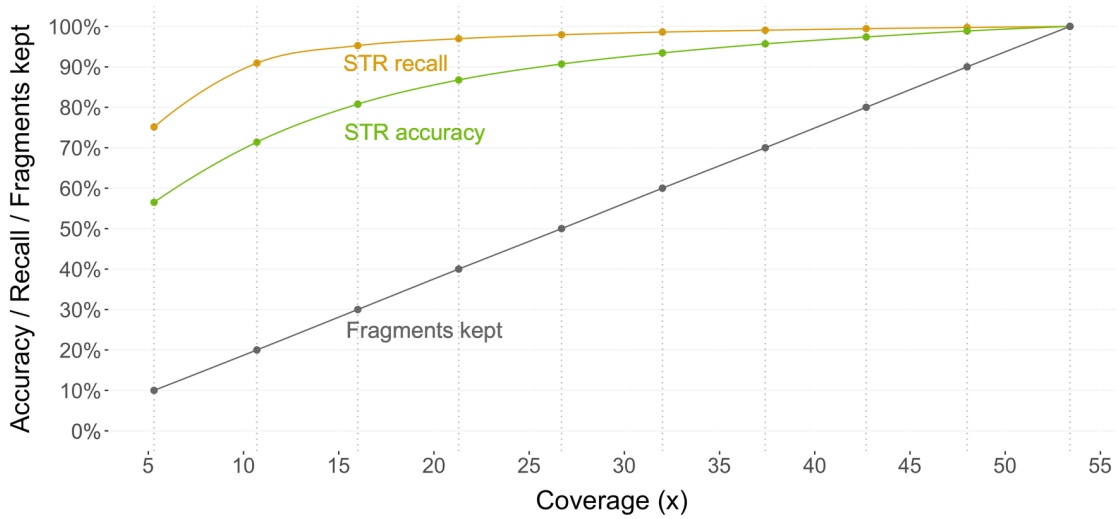


Figure 3.1. STR recall and accuracy of the dna_109 sample downsampled in 10 steps. The orange line shows percentage of the full coverage STR calls that were also called in the dwonsampled data (recall). The green line is the proportion of calls with the same genotype as the original sample (STR accuracy) and black line is the percentage of fragments kept after downsampling. Sequencing coverage of BAM file is on the X-axis

After downsampling, insert size of fragments and the resulting coverage was calculated for all samples (see Methods). Two samples had the median insert size at least 300 bp (Table 3.2) whereas the other five were between 205 bp and 294 bp. Considering that the reads are 150 bp in length, having an insert size less than 300 bp will result in overlapping read pairs. Overlapping reads contain less information than read pairs with longer insert

sizes as some of the bases are redundant. The amount of usable sequence information will be decreased when such reads occur at significant levels.

Table 3.2. General parameters of samples after downsampling. Table is sorted by the median insert size in descending order.

Sample	PCR cycles	Coverage (X)	Median insert size (bp)
dna_109	5	26.97	301
dna_55	5	26.92	300
frag_1500	6	26.94	294
dna_16	8	26.98	272
dna_27	6	26.92	255
frag_500	6	26.93	232
frag_1000	6	26.90	205

Here we use this data set to investigate the effects of different library preparation procedures including PCR cycles and fragment size on STRs recall and accuracy. We use GangSTR as one of the best performing tools in our previous comparison from Chapter 2 and compare the results from applying this tool across the data set.

3.2.2 Initial analysis of seven WGS samples

Total number of loci containing STRs, which was found by the Tandem Repeats Finder (TRF) was 236,489 (See Methods). Using our downsampled dataset, the number of unique loci called by GangSTR over all samples was 220,884, which is 93.4% of the total number of STRs given. Most of the repeat units (73.5%) were either 2 or 4 bp long (Figure 3.2).

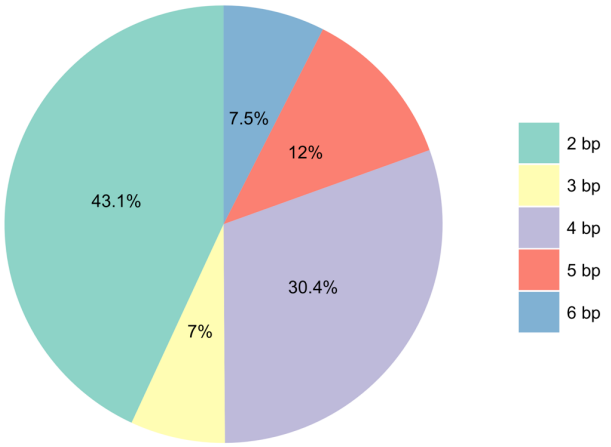


Figure 3.2. Frequency of repeat units in the genome as reported by TRF.

Overlapping read pairs could potentially be beneficial in analysis of STRs as merging reads might result in detecting STRs that are longer than the read length. On the other hand, the STR genotyping tool GangSTR used in this analysis incorporates different information acquired from the read pairs to make a call, where fragments with long insert size could be advantageous. Thus, all calls where at least one of the alleles was longer than the read length (150 bp) was included in an analysis to determine whether longer insert size will result in more calls, or *vice versa*.

No correlation between the number of calls made for long alleles and insert size was detected (Figure 3.3a). However, a strong correlation was determined between the insert size of the sequencing and coverage of loci (Figure 3.3b) potentially resulting in better accuracy of long alleles with longer insert sizes. Pearson's correlation test between the average coverage of samples loci and its median insert size revealed very strong correlation of $r = 0.99$ with a p-value of < 0.001 demonstrating the effect of insert size on locus coverage.

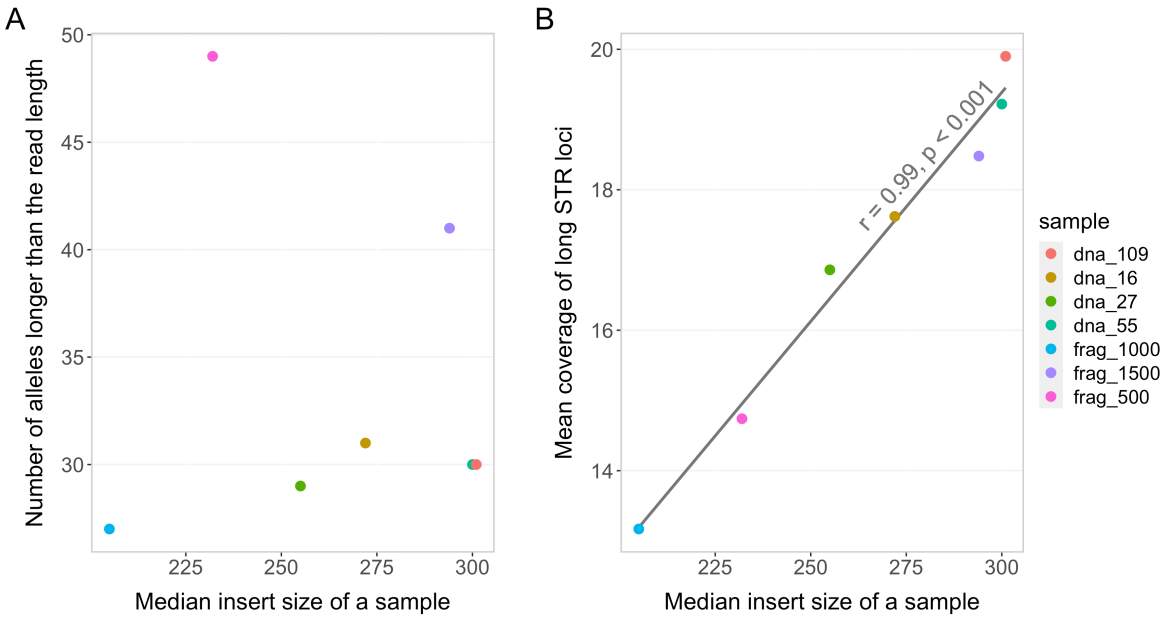


Figure 3.3. Median insert size relative to (A) total number of genotyped alleles longer than the read length and (B) mean coverage of STR loci that are longer than the read length. Number of calls is the sum of calls where at least one of alleles is longer than the read length.

3.2.3 Recall of STRs

The effects of insert size and PCR cycles on STR recall were investigated. STR recall is defined here as the percentage of STR loci detected in a sample out of all total loci defined by Tandem Repeats Finder (236,489). The number of unique loci per sample was

calculated and divided by the number of total loci. In addition, the median number of reads per locus was calculated for each sample.

Results in Table 3.3 show very strong positive correlation between STR recall and the insert size (Pearson's $r = 0.98$, $p < 0.001$). However, the recall was not correlated with the number of PCR cycles (Pearson's $r = -0.23$, $p = 0.6$). The insert size was also found to be strongly correlated with the median coverage of the STRs, so that the higher STR recall may be a result of having more reads per STR loci. One reason for this result may be that reads containing STR sequence are harder to map to the genome and rely more on the location of the read-pair. With longer insert size there is more independent information (fewer overlapping bases) to indicate the location of the reads. This result, together with the analysis of long STRs indicates that sequencing with longer insert sizes and less overlapping bases, results in greater detection of STRs from WGS.

Table 3.3. STR recall compared to median insert size and coverage of all loci. Table is sorted by the STR recall column in descending order.

Sample	PCR cycles	Median insert size (bp)	Median coverage (of reported repeats)	STR recall (%)
dna_109	5	301	20	91.7
dna_55	5	300	20	91.5
frag_1500	6	294	19	91.0
dna_16	8	272	18	89.8
dna_27	6	255	17	89.5
frag_500	6	232	15	87.7
frag_1000	6	205	13	84.5

3.2.4 Accuracy of STRs

Having a high STR recall is important, however, the quality of these results are equally crucial. Therefore, we next conducted an analysis to examine accuracy of STR allele calls and determine the effect of varying the insert size or the number of PCR cycles. All seven samples were compared with the Illumina Platinum pedigree PCR-free sample of NA12878 that is a common sample used for benchmarking. Taking into account that a PCR-free library preparation should not cause stutter noise in sequencing data, then we assumed the repeats detected in this sample were more accurate and we considered them as the true length of an allele.

We generated a high-quality "truth" data set from the set of calls from the PCR-free sample. For each call a quality score between 0 and 1 was reported by GangSTR. These quality scores were used to filter PCR-free sample calls as well as examining how filtering at different quality scores would affect results. The aforementioned quality score is a measure of confidence in allele calls and for GangSTR lower quality scores are given for alleles longer than the read length as their estimation is based on less confident read classes. Increasing the quality score threshold will increase the confidence of all calls.

To create the truth data, we aimed to determine a quality score parameter to use with the PCR-free sample which would give a good accuracy while still retaining a sufficient amount of calls to work on. The sum of two alleles was calculated for the PCR-free sample and for the sample that had the highest STR recall (dna_109). All calls which had a quality score less than 1.0 were filtered out from the data of dna_109 sample to provide the most accurate results against which the PCR-free sample was compared. The difference in the sum of alleles of each locus between samples was calculated. In addition, the percentage and the number of remaining calls in the whole dataset (220,884) at a particular minimum quality score was computed to assess the loss of data while filtering out calls.

The percentage of matching length of alleles between samples is shown on Figure 3.4. A gradual increase in matching number of calls between samples can be seen with the increase of minimum quality score. Without any filtering (at the minimum quality score of 0.0) 91.2% of calls had the same combined allele length, which increased to 99.4% when the minimum quality score was set to 1.0. However, when looking at the number of remaining calls in the dataset, loss of 18% of calls can be seen at a quality score of 0.1, followed by a gradual reduction until quality score of 0.9 when 42% of calls in total have been filtered out. Finally, when keeping only calls with quality score of 1.0, which gives the highest number of matching calls, result in 83% of calls being filtered out from the dataset.

As a result, the quality score of 0.9 was selected as a filtering parameter for the PCR-free dataset as this will provide the highest accuracy while keeping an adequate amount of the remaining data (58% or 128,420 calls). In addition, this is also a recommended filtering parameter from the author of GangSTR according to the documentation of the tool.

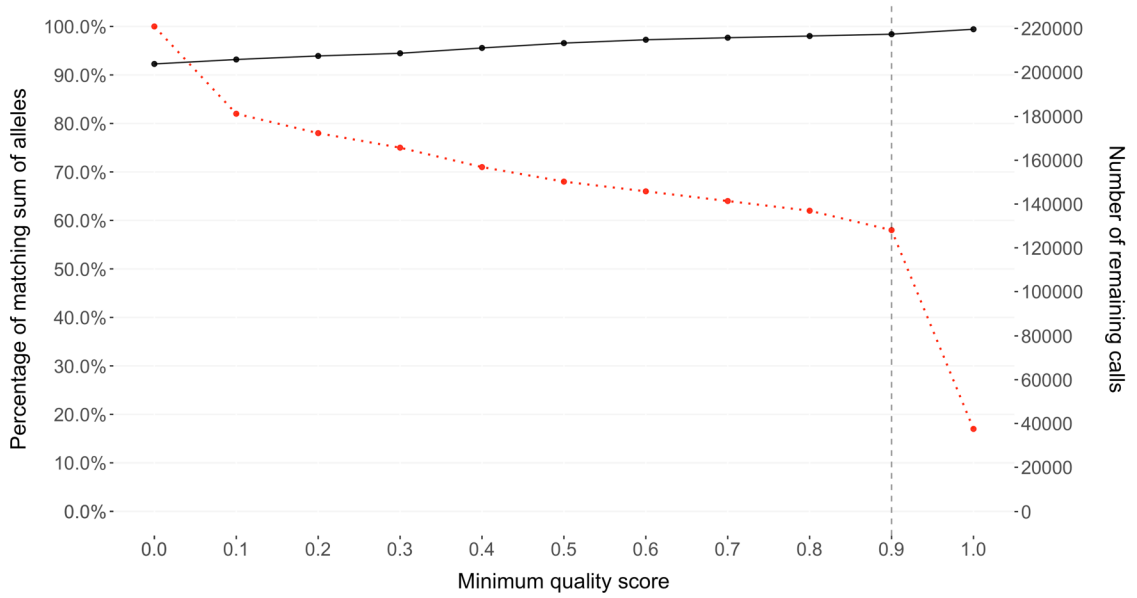


Figure 3.4. PCR-free sample compared with the sample that had the highest STR recall (dna_109). Minimum quality score of the PCR-free sample is on the X-axis. Solid black line represents the percentage of combined length of two alleles that are identical between two samples. Red dotted line represents the percentage (left Y-axis) as well as the total number (right Y-axis) of remaining calls in the whole dataset. Dashed grey line is the chosen quality score parameter for further analysis.

We took these high-quality (quality score >0.9) results of the gold standard PCR-free sample and then compared all samples against these calls to determine the STR accuracy and its relation to the number of PCR cycles and insert size. The sum of both alleles in repeat units were compared at three different filtering stage. First, where no filtering was done to the test samples (minimum quality score of 0.0), followed by the minimum quality score of 0.9 and finally at the maximum quality score (1.0). The percentage of identical sum of both alleles is defined as the STR accuracy in this section.

In every case the accuracy of the calls is at least 85%. Unsurprisingly, accuracy is the lowest when no quality filtering was done and increases with increasing the quality score filtering (Figure 3.5). Differences between samples are within 3.55% for no quality filtering and within 1.26% at a minimum quality score of 0.9. At the strictest filtering, accuracy of all samples is very similar (within 0.22%) and any difference between these samples is negligible.

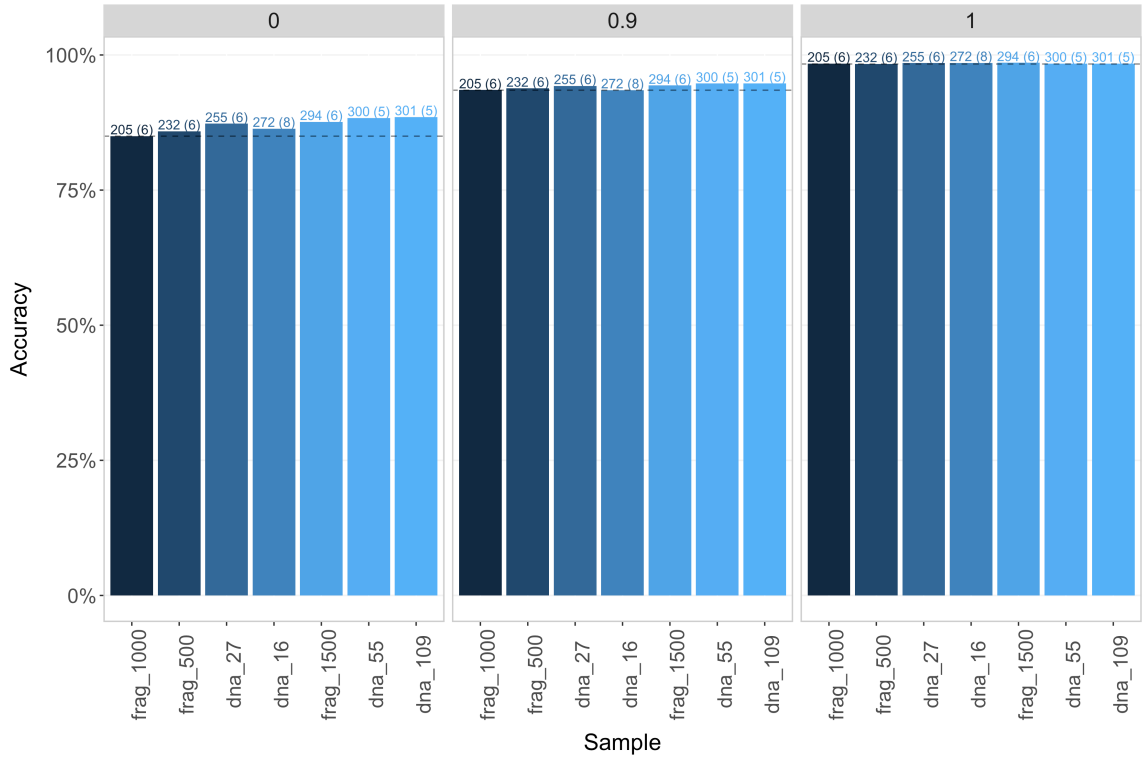


Figure 3.5. All samples with no filtering and minimum quality score at least 0.9 or 1.0 compared to the PCR-free sample. X-axis is ordered and coloured by the median insert size which is also displayed on top of each bar in conjunction with the number of PCR cycles in brackets. Black dashed line represents the sample with the lowest accuracy.

A Pearson's correlation coefficient was computed to assess the relationship between accuracy and insert size. Very strong positive correlation ($r = 0.93$, $p < 0.01$) was detected where no filtering was done. Moderate correlation ($r = 0.77$, $p < 0.05$) was seen between variables when filtered at minimum quality score of 0.9. However, as seen on the Figure 3.5 then this trend has an exception – a drop in accuracy is observed for the dna_16 sample when no filtering was done. Moreover, the dna_16 sample has the lowest accuracy after filtering at 0.9, whereas it was third lowest before filtering. This sample has the highest number of PCR amplification cycles and could indicate that the sample has a greater proportion of calls with the wrong allele size but higher quality scores. It is known that PCR creates stutter noise and the probability for stutter noise present in reads potentially increases with the number of cycles. In addition, it has been demonstrated how slippage events at the early stage of a PCR when using a small amount of initial template, such as in the case of the dna_16 sample, creates higher stutter peaks and can lead to calling a wrong size allele (Duffy et al., 2017).

Results which have high confidence but wrong allele length can be explained by higher numbers of PCR as follows. When an allele with a mutated length is created during an

early stage of PCR then more copies of this will be created during the following cycles. The proportion of wrong size alleles can also increase when correct length amplicons fails to amplify due to the stochasticity (Kebschull & Zador, 2015). When fragments with wrong size alleles are being selected for sequencing then they could be in a high proportion enough in sequencing data to make wrong calls with higher confidence (for example, detect a homozygous locus to be a heterozygous one due to reads in two different sizes). On the other hand, when a lower number of PCR cycles are used then there is less replication events and lower chance for stochastic and polymerase amplification errors.

At the strictest filtering no significant correlation ($r = 0.34$, $p = 0.48$) was detected and the sample dna_16 does not have lower accuracy compared to others. Hypothetically this could be due to filtering out results where there is even a small probability of making an erroneous call by the tool (for example, potentially requiring reads to have allele fraction close to 50% to be classified as heterozygous at the quality score of 1.0, whereas the requirements for allele fraction were more relaxed at lower quality scores). That in turn may improve accuracy which would then be comparable to other samples.

3.2.5 Effect of repeat unit size on accuracy of STRs

STRs are known to have a higher mutation rate for certain repeat unit sizes. For example, it has been demonstrated experimentally that the mutation rate for dinucleotide repeats is 1.5–2 times higher than for tetranucleotides and the mutation rate for non-disease-causing trinucleotides is between these two (Chakraborty et al., 1997). It has been also shown on theoretical models that higher mutation rates are observed for 2 bp repeats, followed by 3 bp and 4 bp repeats (Kruglyak et al., 1998). Considering that the dataset under analysis contains 43.1% of 2 bp repeats (Figure 3.2) and that are highly mutable during replication, it can potentially have a significant impact during PCR amplification. Therefore, accuracy of STRs was investigated for each repeat unit size individually at three filtering stages.

Results in Figure 3.6 show how accuracy of STRs is improving with increasing quality scores for each repeat unit size. At the base level where no filtering was done a range in accuracy across different samples can be seen for the "all" (all repeat unit sizes combined) and separately for dinucleotides. As shown above, samples with the lowest insert size (frag_1000 and frag_500) also have the lowest accuracy among the samples, whereas the sample with the highest insert size had the best accuracy in both di- and trinucleotides.

The difference between samples is then decreasing with the motif length and the differences between samples in the penta- and hexanucleotides are minor. Also, as we have observed samples become closer to each other in terms of accuracy as the minimum quality score improves to 0.9 and the difference is especially small when at the strictest quality filtering stage. At the minimum quality score of 0.9 a clear separation between samples is still present in dinucleotides, whereas the results are more uniform at other motif lengths. The difference in accuracy between samples in dinucleotides can be up to 7.5% when no filtering has been done, but around two times lower at the minimum score of 0.9.

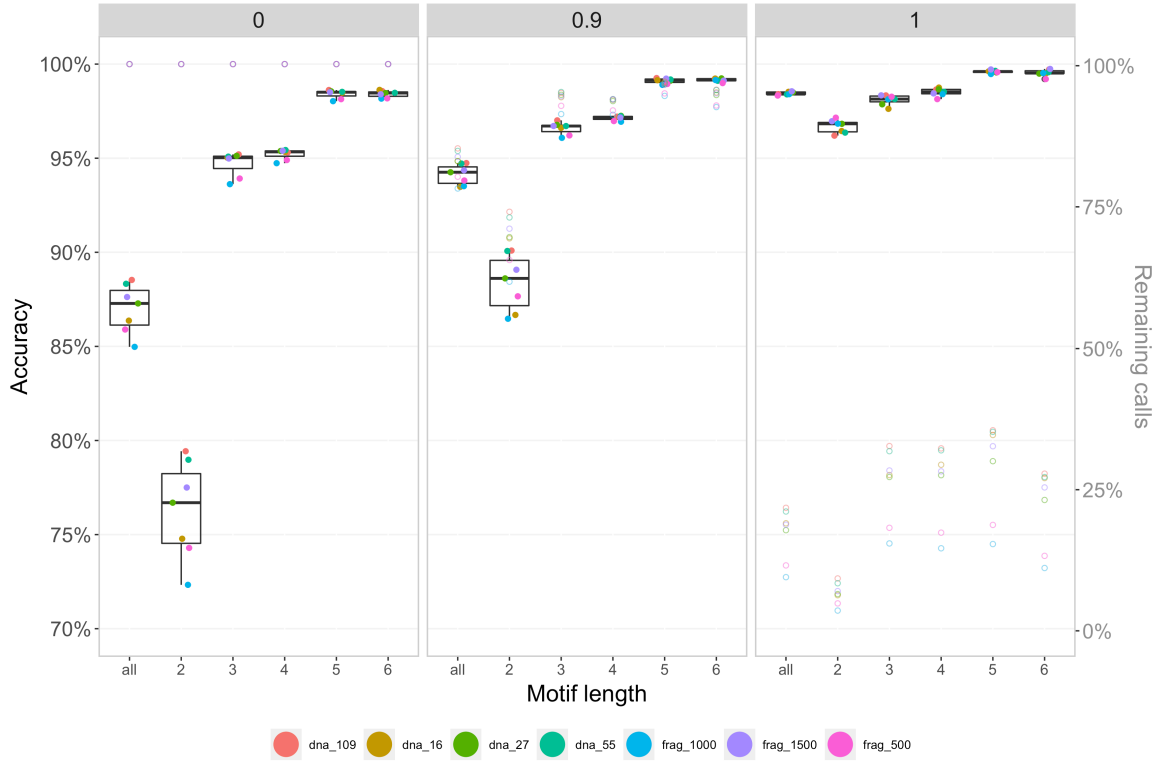


Figure 3.6. Differences in accuracy in different repeat unit sizes at three filtering strategies. Each filled dot represents the accuracy of a sample in the specific motif length ("all" includes all motif lengths). Empty dots represent the percentage of remaining calls in the sample. For each sample 100% on the right Y-axis means a different number of calls: 127,974 for dna_16; 127,908 for dna_27; 128,126 for dna_55; 128,112 for dna_109; 127,576 for frag_500; 126,759 for frag_1000; 128,055 for frag_1500.

Far better results are observed for 3 bp and 4 bp repeats where the overall accuracy is close to 95% even without filtering. Moreover, 5 bp and 6 bp repeats shows the best results among all repeat unit sizes, with the largest difference between samples and the PCR-free one only about 3% without filtering and reaching to even 99% accuracy at the quality score of 1.0. This may indicate that quality score filtering should only be applied to calls on 2, 3 or 4 bp repeats and all calls in 5 and 6 bp repeats can be retained.

A drop in the amount of remaining calls in the dataset is also seen with stricter filtering, which is the sharpest for dinucleotides, pointing out the large proportion of low-quality calls which could be responsible for the low accuracy at lower quality score values (Figure 3.6). The best accuracy is observed at the strictest filtering, but as a result a significant proportion of the dataset has been filtered out. Differences in the remaining proportion of calls between samples can be seen at the strictest filtering where two samples with the lowest insert size (frag_500 and frag_1000) are losing more calls than others. This means that these samples have the highest proportion of calls with lower quality scores (below 1.0) compared to other samples.

There are also differences between repeat units regarding the proportion of data being filtered out when increasing the quality score. When filtering data by using the highest quality score, the proportion of data that passes the filtering threshold is on average between 21.7 and 28.5% for 3–6 bp motifs, but for dinucleotides it is only 6.1% which corresponds to 3,665 calls per sample on average. Dinucleotides as the most abundant repeat units have also the highest proportion of low-quality calls. However, as a result, the accuracy of dinucleotides improves by 8.3% from the previous filtering step (minimum quality score of 0.9) and reaches to the highest, 96.7% accuracy, which is very close to the results of other repeat unit sizes.

These results together show that once again the lowest accuracy is in dinucleotides. An improvement in accuracy is seen with the increase of minimum quality score as the results are getting more confident. The best accuracy is observed when using the strictest filtering (minimum quality score of 1), which in turn results a significant proportion of calls being filtered out. Therefore, the following analysis was done by using minimum quality score of 0.9 as the filtering parameter. At that filtering parameter, a variability between these samples is visible mainly at dinucleotides, where the sample with the highest number of PCR-cycles and the sample with the shortest insert size have the lowest accuracy. For tri- and tetranucleotides the lowest accuracy is observed for both samples that have the shortest insert size. Penta- and hexanucleotides have higher accuracy and differences between samples are smaller than for other motif lengths. Similarly to 3 and 4 bp motif lengths, also 5 bp have the lowest accuracy determined for the shortest insert size samples, whereas for 6 bp the results are mixed.

3.2.6 Effects of repeat length on accuracy of STRs

It has been previously reported that a higher rate of STR mutations are seen in longer alleles (Brinkmann et al., 1998; Fondon et al., 2012; Huang et al., 2002; Kelkar et al., 2008; Lai & Sun, 2003). To test this, all samples were compared to the PCR-free one to determine accuracy of STRs over all repeat lengths for each repeat unit length. The number of total calls for each repeat length was recorded and accuracy of the total length of both alleles (the proportion of correct calls) for each length calculated.

Dinucleotides show a constant decrease in accuracy from repeat length of 12 repeats to the longer repeat lengths, excluding the ones where low number of calls were found which results can be unreliable (Figure 3.7). Interestingly, there is an increase in accuracy observed between 6–11 repeats and lengths of 12–16 had the highest accuracy before gradually decreasing again. These results are similar to previous research which analysed mutability for different repeat lengths and showed a similar trend where a higher mutability in dinucleotides was seen between below 12 repeats and then above 16 repeats with 13–15 repeat lengths being the most accurate (Kelkar et al., 2008). The authors of the paper speculated the repeats with lower mutability may been influenced by another process besides the replication slippage, such as more efficient repair mechanism (Almeida & Penha-Gonçalves, 2004) or due to the dynamics of different rates of expansions and contractions (Xu et al., 2000).

The samples with the lowest median insert size (frag_500 and frag_1000) show consistently lowest accuracy and samples with the highest median insert size (dna_55 and dna_109) the highest accuracy. The order of samples in terms of accuracy is mixed in repeats with smaller number of calls, which could refer to these results being rather unreliable due to low number of calls, instead of showing a difference between samples.

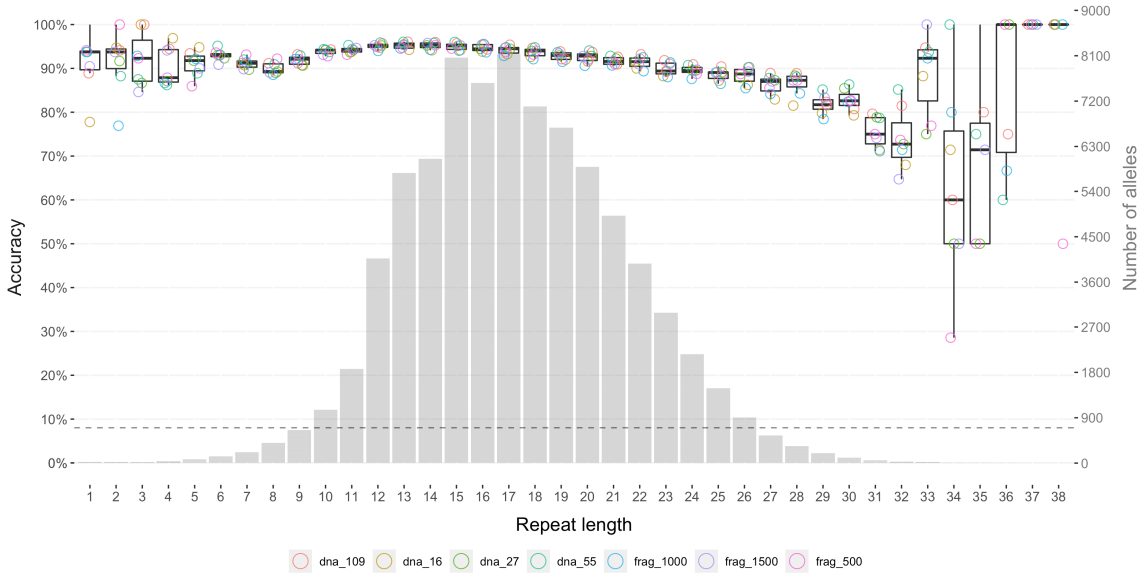


Figure 3.7. Accuracy of dinucleotide repeat alleles of different repeat length. On the left Y-axis is the percentage of matching alleles (shown by the box plot) and on the right (coloured in grey) it is the number of calls in this repeat length (shown by the histogram). Bars over the dashed grey lines means results where each sample has at least 100 calls to calculate accuracy.

Results of tri- and tetranucleotide repeats are similar to each other considering repeat length and its range. However, tetranucleotide repeats show slightly higher accuracy in repeat lengths 6–10, but decreasing more sharply and therefore had lower accuracy in shorter and longer repeats, whereas trinucleotide decrease at slower rate and show better accuracy (Figure 3.8). Similarly, to dinucleotides, a lower accuracy is detected for shorter repeat length.

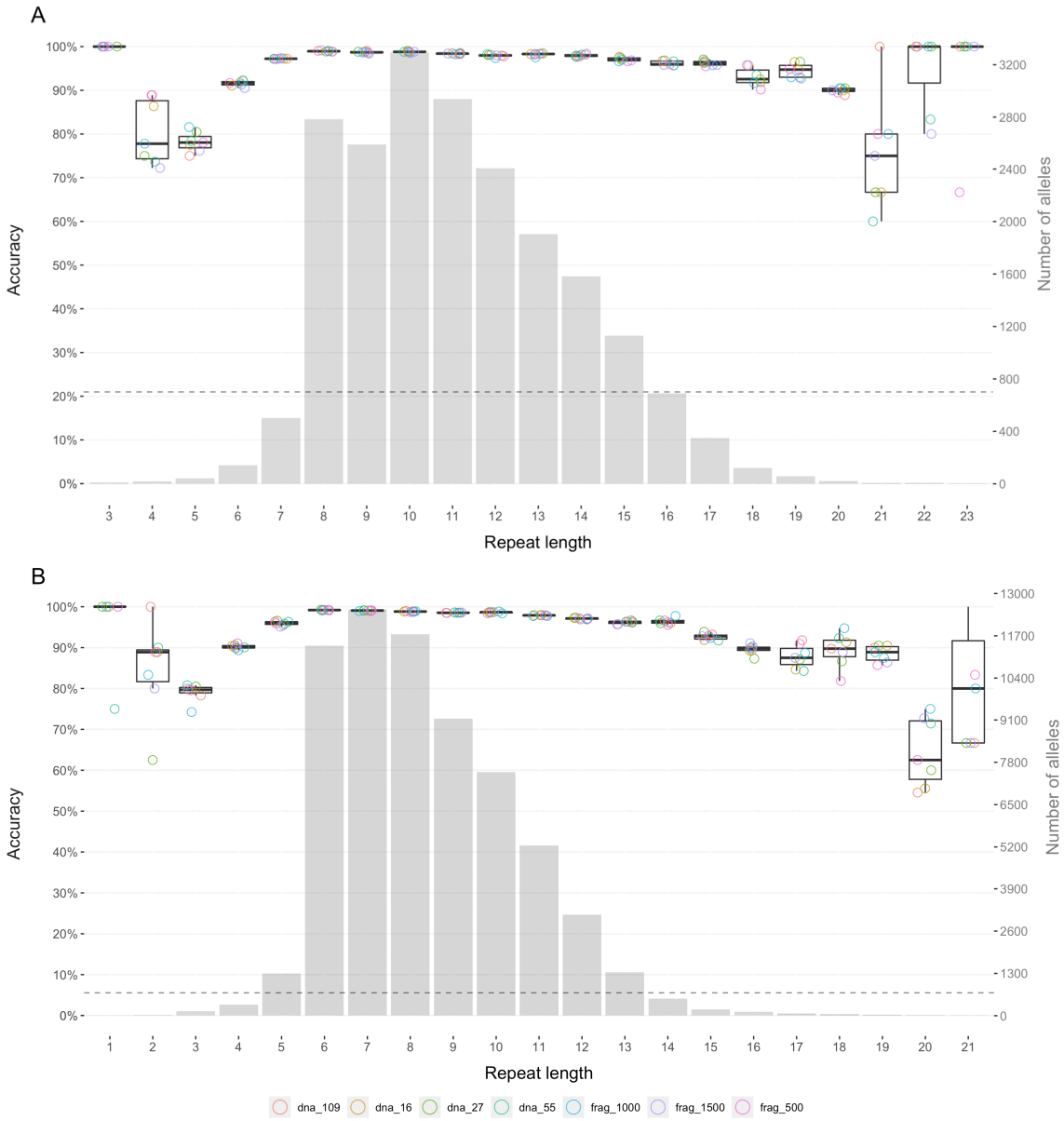


Figure 3.8. Accuracy of (A) trinucleotide and (B) tetranucleotide repeat alleles of different repeat length. On the left Y-axis is the percentage of matching alleles (shown by the box plot) and on the right (coloured in grey) it is the number of calls in this repeat length (shown by the histogram). Bars over the dashed grey lines means results where each sample has at least 100 calls to calculate accuracy.

Penta- and hexanucleotide repeats have majority of results in a smaller range of lengths as expected. Accuracy for both repeat units is higher than for any other repeat units and remains stable over repeat lengths (Figure 3.9). As accuracy is high for all samples then no individual differences between samples can be observed.

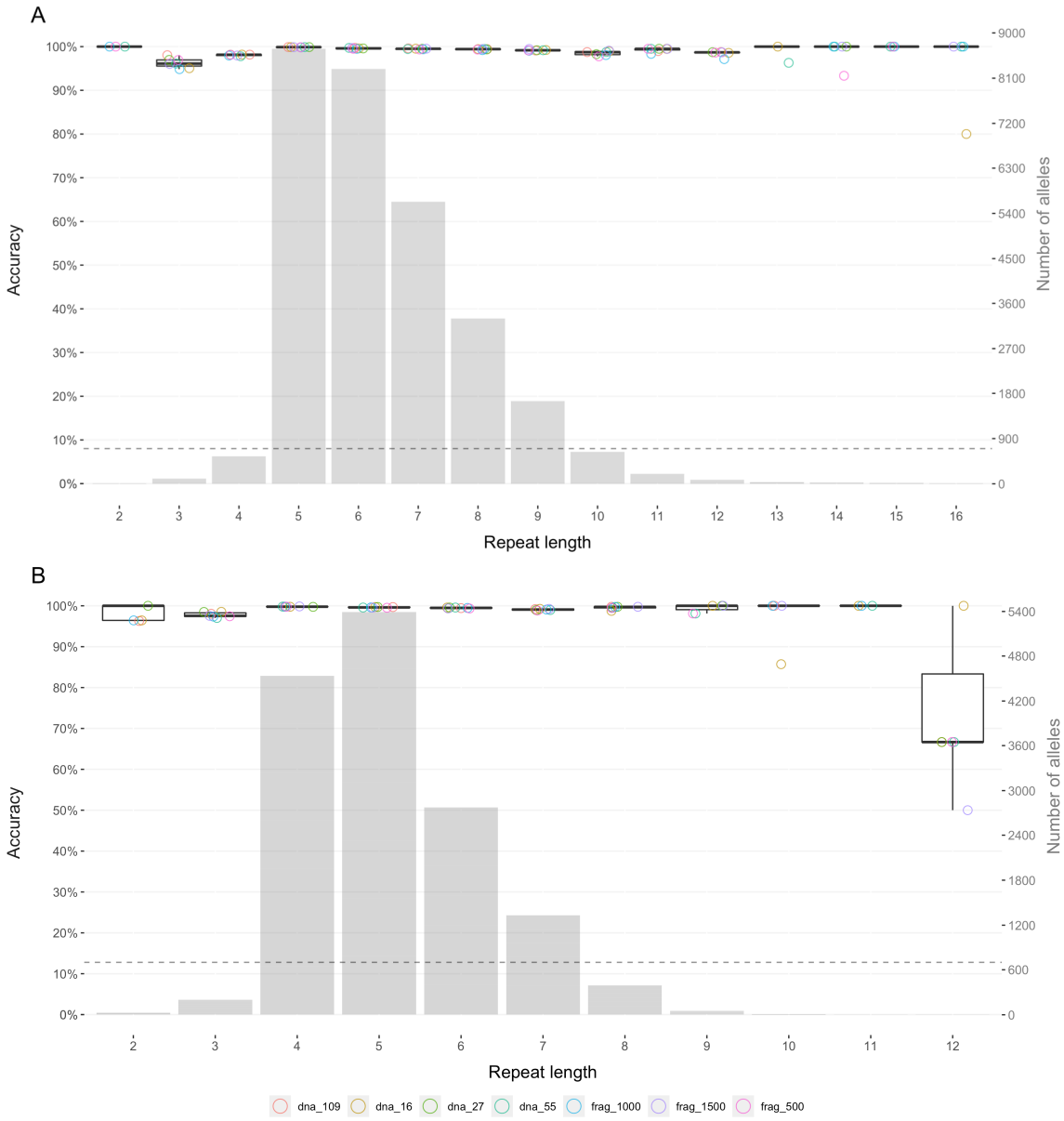


Figure 3.9. Accuracy of (A) pentanucleotide and (B) hexanucleotide repeat alleles of different repeat length. On the left Y-axis is the percentage of matching alleles (shown by the box plot) and on the right (coloured in grey) it is the number of calls in this repeat length (shown by the histogram). Bars over the dashed grey lines means results where each sample has at least 100 calls to calculate accuracy.

These results together show how in all repeat unit sizes the repeat length has an impact on accuracy. The longer the repeat length, the less accurate the results are. This could be due to mutations that arise during a DNA replication in PCR. For the same length of DNA there would be more slippage events possible for shorter motifs than for longer ones. This could also explain why overall accuracy improves with the repeat unit size. Because the read length in sequencing data is fixed, the maximum number of repeat units is limited. For example, when a decrease in accuracy is apparent to happen above 10–12 repeats (as seen for 2–4 bp repeats), then this cannot be observed for 6 bp repeats as there are not

enough calls above 8 repeats for reliable results. However, longer repeat units show higher accuracy for smaller repeat lengths, which suggests the involvement of other repeat unit characteristics influencing accuracy, such as biochemical properties and/or composition of a repeat unit.

In terms of differences between samples with different preparation procedures, no significant differences were seen for samples with a higher number of PCR-cycles or for samples with longer insert size. In the whole range of allele lengths where we had a reasonable number of calls for accuracy calculation, we saw a similar pattern across all samples. This indicates that the lower accuracy seen in certain samples is a result of having a higher number of erroneous calls across all repeat lengths and not just longer repeats.

3.2.7 Magnitude of STR genotype differences

We next wanted to examine, in more depth, the frequency and direction of mutations in the genotyping data as this was not explored in our previous exome analysis. We aimed to assess to what extent allele lengths of samples are deviating from the truth and secondly, to compare the samples with respect to PCR cycles and/or insert size.

For this analysis, both alleles were compared between each sample and the PCR-free one to find the difference in repeat lengths, which is represented in a negative or positive number (for example, –1 or +2) or 0 in case of no differences. This shows how many repeats a genotype of a sample diverge from the truth and gives the direction of mutations, such as whether repeats in STRs loci of a sample are commonly gained or lost. The frequencies of repeat differences was calculated for each repeat unit length separately. To calculate the differences in the number of repeats we first took the estimated allele sizes for each locus that was ordered from smallest to largest. Then, we took the same locus in the PCR-free sample and subtracted the allele sizes. Finally, we evaluated all the STR sites across the genome to plot the distribution of differences between the sample alleles and the PCR free alleles.

Reassuringly, a big proportion of allele length difference between our samples and the PCR free sample are one repeat unit longer or shorter. One repeat difference comprises 60.2% of all mutations found in dinucleotides and it is lower for tri-, tetra-, penta- and hexanucleotides (49.6%, 45.1%, 44.1% and 45.9%, respectively). Larger differences

were increasingly less likely, however there were cases in all repeats except hexanucleotides where the allele lengths differed by more than 10 repeat units (4.0% or 1657 calls on average per sample for 2 bp, 3.3% or 344 calls for 3 bp, 3.9% or 1256 calls for 4 bp and 0.6% or 94 calls 5 bp repeat units).

For the dinucleotide repeats where allele lengths differ the most, we also found the largest variability between samples with different DNA preparation steps (Figure 3.10). The mean number of calls per sample was 40,950 and the overall frequency of any mutation (error rate) for dinucleotides was 6.72% which is significantly higher than for other repeat units. The sample with the highest number of PCR cycles (dna_16) stands out from the rest as it has the highest frequency (3.42%) of –1 repeat unit mutations (that is, where observed alleles are 1 repeat shorter than the truth). The two samples with the lowest median insert size have the next highest error rates (2.90% for frag_1000 and 2.70% for frag_500). The lowest frequency of –1 mutation are seen in samples with the highest median insert size (1.92% for dna_55 and 1.87% for dna_109). Mutations +1 are the second most common deviation in length, however, the difference between samples is much smaller.

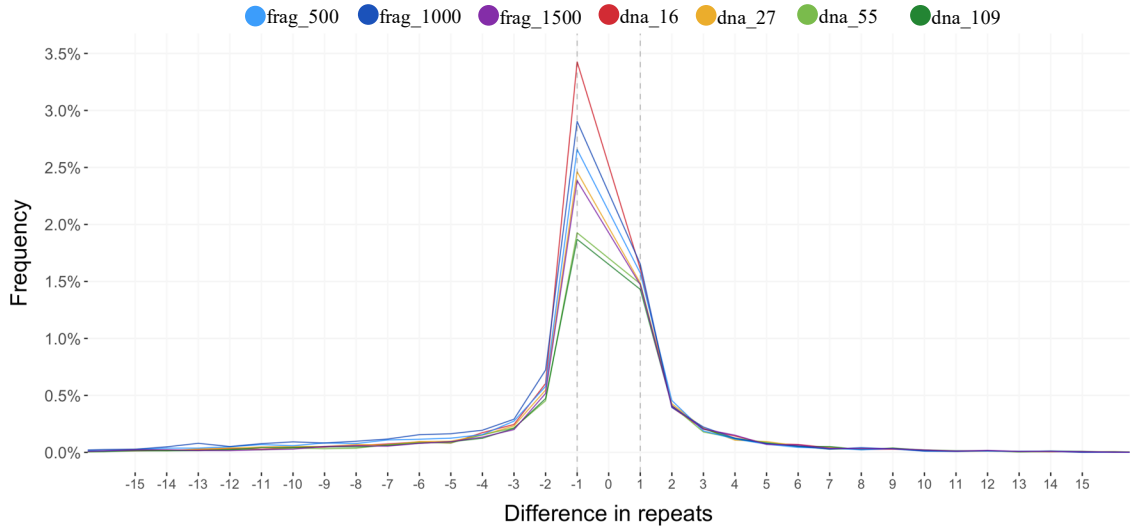


Figure 3.10. Frequencies of 2 bp repeat unit's differences between samples and the PCR-free one. On the X-axis the results of zero (no difference) is omitted from the figure and the axis is limited to –15 to 15 (leaving out results until –26 and 43 values). Dashed grey line indicates top two repeat units with the highest frequencies.

Similarly, to the 2 bp repeats, 3–4 bp repeats have –1 and +1 repeat unit mutations as the most frequent, however in a much smaller magnitude and in an opposite direction (Figure 3.11). Samples with the lowest median insert size (frag_500 and frag_1000) showed the highest frequency of allele differences, whereas the sample with the highest number of

PCR cycles has slightly fewer one-repeat mutations (in contrast to dinucleotides). On the other hand, the frequency of length differences between samples in tri- and tetranucleotides is around ten times smaller than it is in dinucleotides and therefore questionable whether any conclusions can be drawn. The overall error rate for trinucleotide repeats is 1.90% with 10,334 calls on average, and for tetranucleotides it is 1.59% with considerably more, 32,586 calls on average per sample. Nearly half (49.6% for trinucleotides and 45.1% for tetranucleotides) of these are different by one repeat unit.

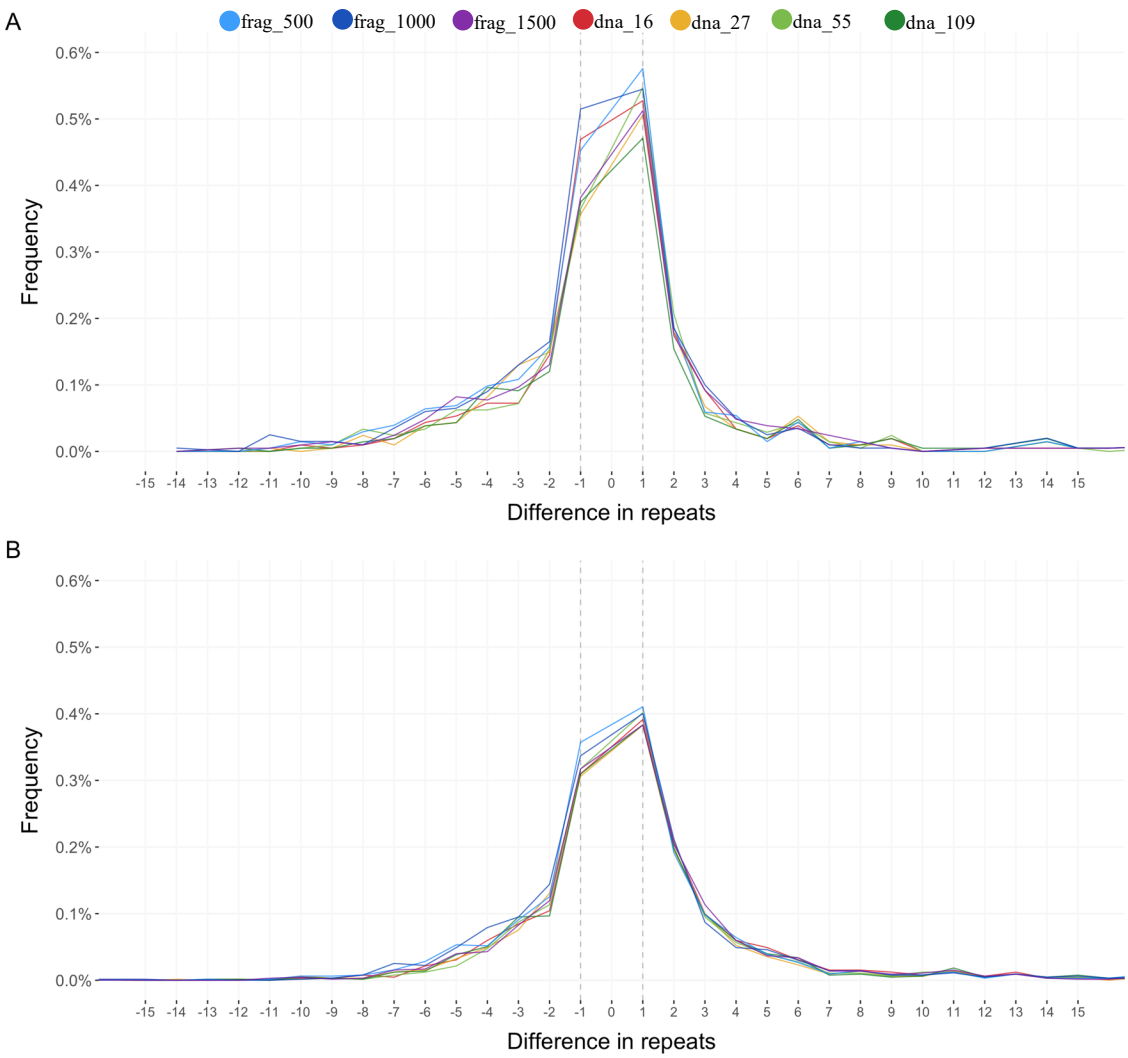


Figure 3.11. Frequencies of (A) 3 bp and (B) 4 bp repeat unit's differences between samples and the PCR-free one. On the X-axis the results of zero (no difference) is omitted from the figure and the axis is limited to –15 to 15 (leaving out results until 32 for figure A, and until –24 and 32 for figure B). Dashed grey line indicates top two repeat units with the highest frequencies.

Penta- and hexanucleotides have around four times lower frequency of mutations than 3 and 4 bp ones and 16-times lower than the dinucleotide repeats. The same samples (frag_500 and frag_1000) have, again, higher frequency of mutations in pentanucleotide repeats (Figure 3.12a). Still, the results are mixed in hexanucleotide repeats and do not

form a clear pattern (Figure 3.12b). The overall error rate for pentanucleotides is 0.46% with 14,714 calls and 0.44% with 7,412 calls for hexanucleotides, both being the lowest among all repeat units.

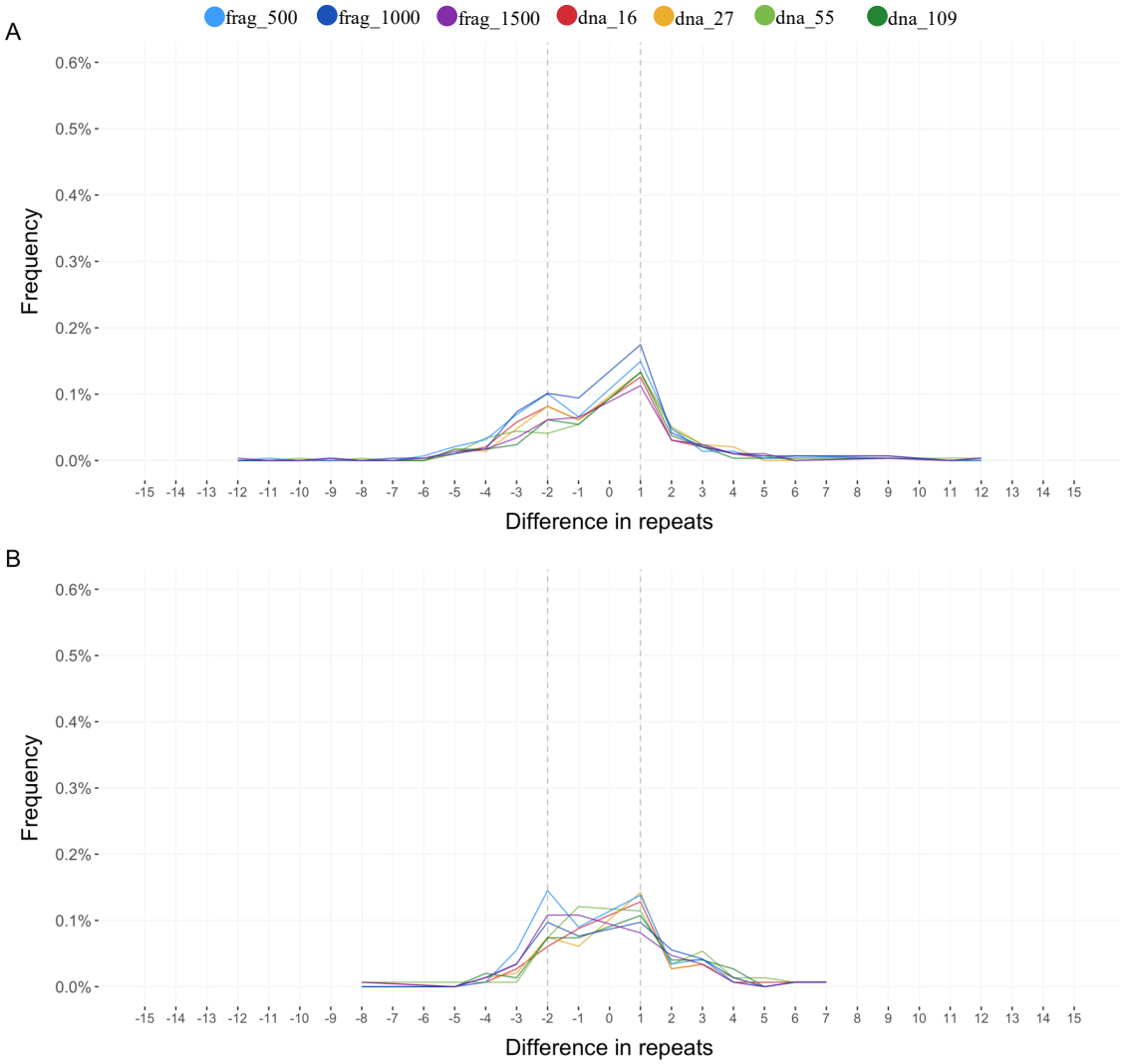


Figure 3.12. Frequencies of (A) 5 bp and (B) 6 bp repeat unit's differences between samples and the PCR-free one. On the X-axis the results of zero (no difference) is omitted from the figure. Dashed grey line indicates top two repeat units with the highest frequencies.

When looking at all the calls that are different between samples and the PCR-free one in more detail, it is seen that most of the mutations are only one repeat or a few different and thus have a minor effect on the overall length of an allele. It has been previously reported that the majority of STR mutations have only one repeat difference which would align with our results (Ellegren, 2000a). However, it is not known whether these observed mutations are a result of stutter noise from the PCR cycles or wrongly genotyped alleles by the genotyping tool. Results from dinucleotides showed higher number of inaccurate calls in samples that had more PCR-cycles, suggesting an effect from the PCR. On the

other hand, this trend was not seen when looking at the other repeat unit sizes. However, differences in longer motif sizes were also minor compared to differences in dinucleotides.

3.2.8 Effects of repeat units on accuracy of STRs

The composition of STR motifs was investigated to determine whether accuracy can be associated to any specific type of motif and whether any differences between samples can be seen.

Results show that among dinucleotides, AC was the most abundant motif, followed by AG, AT and finally CG motif (Figure 3.13a). Tandem Repeats Finder only defined nine loci with the CG/GC motif. Therefore, the average accuracy for this motif are unreliable. The highest accuracy was observed for AC and lower for AG and AT motifs. Samples with the highest number of PCR-cycles had lower accuracy in AG and AT motifs, whereas it was second lowest in AC motif, indicating that the higher number of erroneous calls in dinucleotides could be attributed to the AG and AT motifs.

The predominant motifs in trinucleotides were AAT and AAC, which were also the only ones to have at least hundred calls per sample for accuracy calculation. Motif AAC had a slightly higher accuracy over AAT, whereas there were not enough calls for calculating accuracy for the rest of the motifs in a confident manner (Figure 3.13b). Therefore, it is difficult to confirm lower accuracy for AAG or AGG.

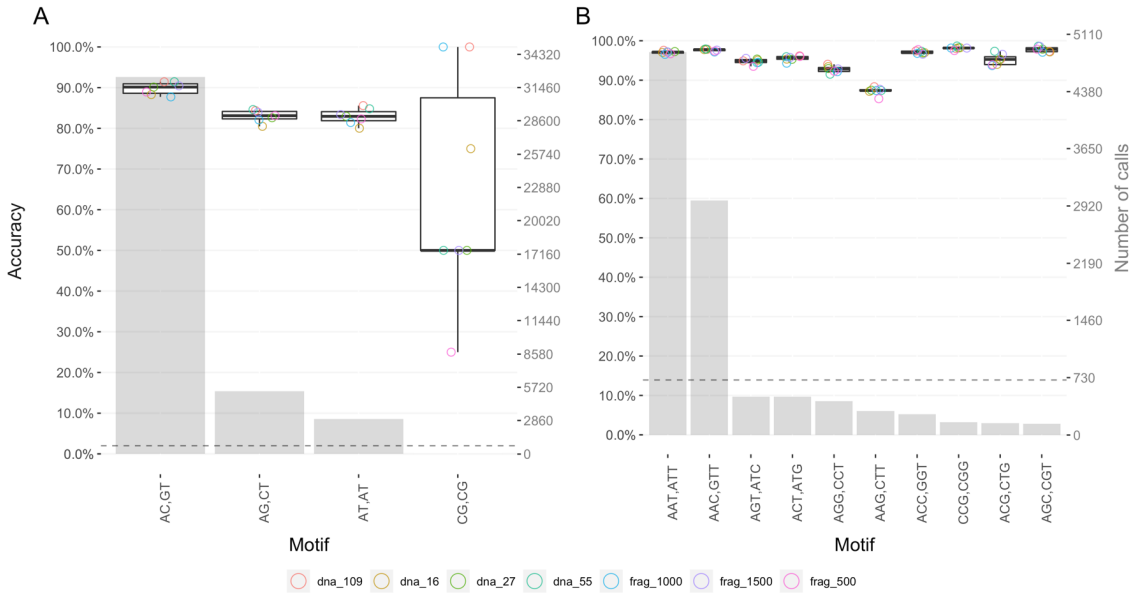


Figure 3.13. Accuracy and amount of (A) dinucleotide and (B) trinucleotide motifs. Grey bars show the total number of calls per each motif (scale on the right Y-axis) and boxes represent the proportion of identical sum of alleles, where each dot is the mean accuracy of a sample. Bars over the dashed grey lines means results where each sample has at least 100 calls to calculate accuracy.

The most common tetranucleotide motifs were AAAT, AAAC, AGAT, AAAG, AAGG, AATG and AAGT, in that order of abundance. Among these, a range of 98.2–98.9% accuracy was observed for AAAT, AAAC, AATG and AAGT. Lower accuracy, between 92.6% and 94.2% was recorded for AGAT, AAAG and AATG motifs (Figure 3.14). All samples have relatively similar accuracy, with the exception of AGAT, ACAT, AGGG, AGGC and AGAG where the sample with the lowest insert size (frag_1000) had clearly lower accuracy than others, and ACAG where frag_1500 was the lowest.

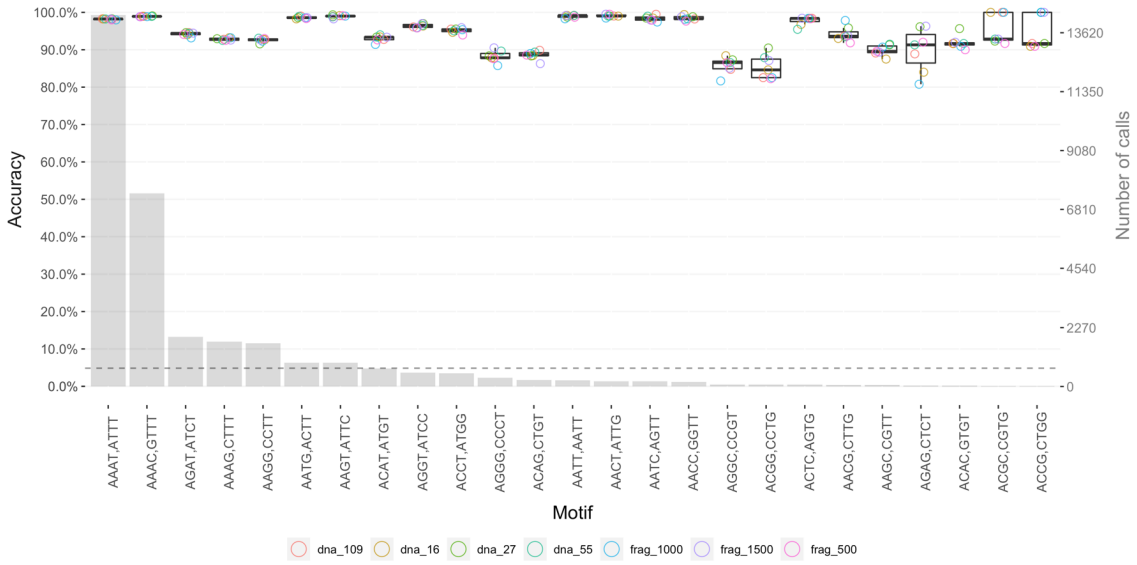


Figure 3.14. Accuracy and amount of tetranucleotide motifs. Grey bars show the total number of calls per each motif (scale on the right Y-axis) and boxes represent the proportion of identical sum of alleles, where each dot is the mean accuracy of a sample. Bars over the dashed grey lines means results where each sample has at least 100 calls to calculate accuracy. Motifs where the mean number of total calls were less than 10 are excluded from the figure.

Motifs AAAAC, AAAAT and AAAAG were the dominant ones in pentanucleotides, whereas others were present in significantly lower frequency (Figure 3.15). Both, AAAAC and AAAAT has 99.3% of accuracy, followed by AAAAG with slightly lower, 98.7% of accuracy. Whereas the results in the highest representative motifs are very similar, some differences in others can be seen. For example, frag_1000 stands out in AGGGG, ACCCG, AAATG, AACAC motifs with lower accuracy, as well as frag_500 in AATAT, AACAT, ACGGG and CCCGG motifs and dna_27 in AATAC, AGGGC, AGTAT, AAACT, ACCCC. However, only a small number of calls are present in all of these motifs.

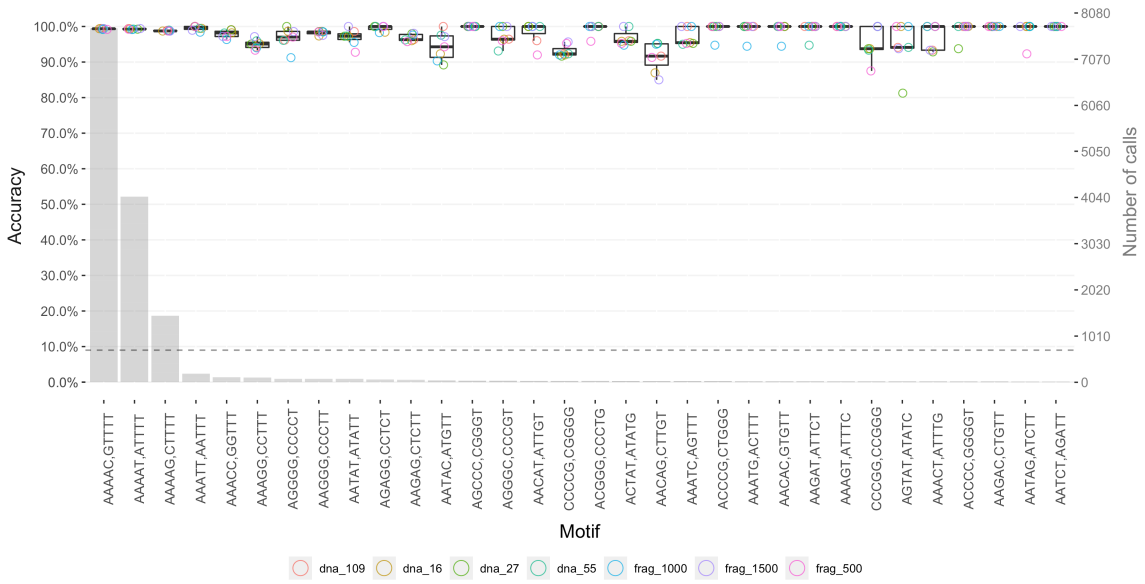


Figure 3.15. Accuracy and amount of pentanucleotide motifs. Grey bars show the total number of calls per each motif (scale on the right Y-axis) and boxes represent the proportion of identical sum of alleles, where each dot is the mean accuracy of a sample. Bars over the dashed grey lines means results where each sample has at least 100 calls to calculate accuracy. Motifs where the mean number of total calls were less than 10 are excluded from the figure.

Finally, similarly to 5 bp repeats, the motifs with the largest frequency in hexanucleotides were AAAAAC, AAAAAT and AAAAAG, all having accuracy of 99.5% (Figure 3.16). The difference between samples is minimal in motifs with the highest calls. Despite that there are some differences in other motifs, there are only handful of calls used for calculation the accuracy.

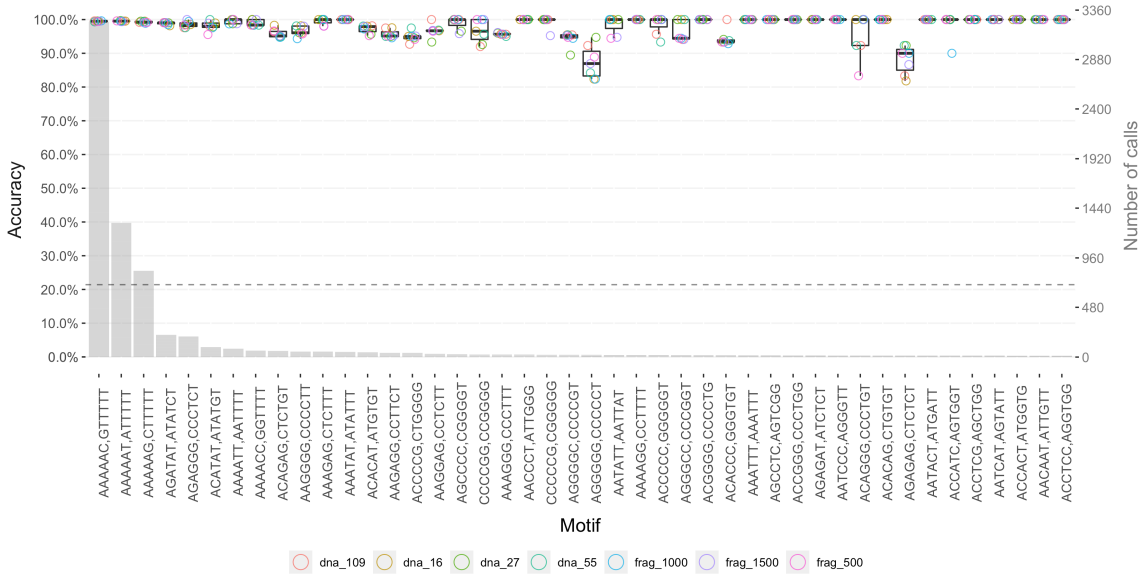


Figure 3.16. Accuracy and amount of hexanucleotide motifs. Grey bars show the total number of calls per each motif (scale on the right Y-axis) and boxes represent the proportion of identical sum of alleles, where each dot is the mean accuracy of a sample. Bars over the dashed grey lines means results where each sample has at least 100 calls to calculate accuracy. Motifs where the mean number of total calls were less than 10 are excluded from the figure.

Three most common motifs of dinucleotide (AC/GT, AG/CT, AT) and tetranucleotide repeats (AAAT/ATTT, AAAC/GTTT, AGAT/ATCT) as well as two most common trinucleotide repeat units (AAT/ATT, AAC/GTT) are consistent with the previous research where the same motifs were observed as the most common ones in humans (Subramanian et al., 2003), however, the rest of the order is different from the previous research. Similarly, the three most dominant motifs of 5 bp and 6 bp repeats are consistent with the most common repeat motifs found in macaques (Liu et al., 2017). Considering the results presented in this section, a difference in accuracy was detected in some of the motifs, but there was only enough data to assess accuracy for small number of motifs for each repeat unit size.

The difference in accuracy between samples of different motifs was minor. Motifs with the lowest accuracy of 2 bp (AT), 3 bp (AAG) and 4 bp (AAAG and AAGG) were also found to have the highest mutability in another study that lines up with the results presented here (Kelkar et al., 2008). In general, all samples have minor differences in motifs where there is sufficient amount of calls. Deviation can be seen mostly in motifs with low number of calls.

3.3 Discussion

Library preparation is a crucial step prior to HTS and experimental factors such as DNA input amount, quality, fragmentation and errors in PCR can have downstream effects on sequencing and variant calling (Bruinsma et al., 2018; Kebschull & Zador, 2015; McNulty et al., 2020). In this study we investigated how the differences in library preparation, particularly PCR-cycles and fragment/insert size affect genotyping STRs from WGS data. In addition, we examined how the genotyping accuracy across different STRs (such as repeat unit size, repeat length and motif) varies. We used replicates of the widely used NA12878 sample that had alterations in library preparation and called STRs by using the recent genome-wide genotyping tool GangSTR that is able to also call STRs that are longer than the read length. We determined dozens of long alleles per sample but did not detect any correlation between the number of long alleles called and the median insert size of a sample. However, insert size was strongly correlated with coverage of STR loci and we found that the STR recall has very strong correlation with the median insert size. Therefore, longer insert size results in higher number of alleles shorter than the read length being detected, but has no impact on capturing long alleles. That is interesting considering that longer insert size of paired-end sequencing potentially increases the likelihood that the fragment would span a long STR region, which however was not observed here. On the other hand, it is less accurate to estimate repeat length from insert size data signals (Kojima et al., 2016). Therefore, when quality score filtering is applied, then calls where an allele was determined to be over the read length, were filtered out and hence, losing the main advantage of the tool.

We examined the accuracy of STRs by using the gold standard Illumina Platinum PCR-free sample of the NA12878 as the truth. When analysing quality scores reported by the genotyping tool, we saw a gradual increase in accuracy and simultaneously a decrease in remaining number of calls in the dataset, with a significant drop in number of calls at the highest minimum quality score. When comparing samples to each other at different filtering strategies we saw a correlation between the overall accuracy and median insert size, except at the strictest filtering where all samples had the highest accuracy and differences were negligible. The potential contribution of additional PCR-cycles on accuracy became apparent when a sample, despite its high median insert size, had a lower accuracy than samples with lower insert size but fewer number of PCR-cycles. This could

indicate the presence of stutter noise in the dataset even at the quality scores of 0.9 and therefore resulting in relatively confident calls with wrong allele length. The strictest filtering seems to eliminate such calls, providing very high accuracy. However, a significant loss of calls is a major downside of the strictest filtering and therefore the quality score of 0.9 was also chosen for further analysis.

We determined the most common repeat units are di- and tetranucleotides that comprise over 70% of all repeats. Due to observations that repeat units with different length have different characteristics (Kruglyak et al., 1998) we analysed each repeat unit size individually. Overall, the lowest accuracy and deviation between samples was seen for dinucleotide repeats. The accuracy for other repeat unit sizes were at least 9% higher and improved as the repeat unit size increased while the differences between samples became trivial. Similarly, the higher proportion of low-quality calls were in the dinucleotides and was observed less in others. Considering that the increase of accuracy in different filtering strictness varies between repeat unit sizes, applying different filtering strictness based on the repeat unit size could be recommended. For instance, the overall accuracy of dinucleotides increased by 8.3% when increasing the minimum quality score from 0.9 to 1.0, which is a significant improvement. However, the increase in other repeat unit sizes with the same filtering is only 1.5% or less. Taking into account that this is also filtering additionally out at least 67% of all calls then applying the same filtering strictness to repeats other than 2 bp may not be favourable.

We analysed each repeat unit size individually to find out the distribution of mutation size errors. In general, roughly half of the mutations are only one repeat difference, followed by two repeat difference mutations. On one hand this is good as we can estimate the correct allele size in majority of cases with only ±1–2 repeats error which is probably acceptable when looking at repeats in intronic loci. On the other hand, repeat mutations in the coding region where the sum of nucleotides is not divisible by three will change the coding frame and potentially causing significant changes in the amino acid sequence. Therefore, mutations found in coding region have to be carefully examined. The source of these mutations could be either from stutter noise that arises during PCR or wrong estimation by the tool itself. In the latter case, the profile of mutations could be different when using different STR genotypers.

Finally, we examined different repeat lengths and each motif separately. When examining repeat lengths then lower accuracy was seen for shorter as well as for longer repeats, with the best accuracy roughly in the middle where also a significant proportion of calls were present. These findings are in line with previous research (Fondon et al., 2012; Huang et al., 2002; Kelkar et al., 2008; Lai & Sun, 2003). In terms of motifs, lower accuracy was detected for certain motifs which partially aligns with the previous research as well (Kelkar et al., 2008), while we also add new information regarding penta- and hexanucleotides. However, many motifs were in low number and the results for them could not be considered as confident enough for making conclusions.

One of the caveats of this study was the data was not generated with these comparisons in mind. Therefore, there was only a small number of samples with the highest number of PCR-cycles. Also, as the extent of amplification is linked to the amount of input DNA, we cannot eliminate the other possible effects that could arise from using low amount of DNA. Secondly, it would be also more reliable to analyse the impact of insert size while keeping the other factors (such as the DNA input and number of PCR-cycles) constant. Similarly, it would have been preferable to have paired-end reads with the same profile of insert size while looking at the effects of PCR-cycles, which, however, is more difficult to control. Also, some of the samples were fragmented prior to library preparation which could have some effects on the results, but it is expected to be minor as the fragments were longer than the fragments generated on beads. Finally, we did not perform additional filtering based on the coverage that is expected to improve the accuracy and probably decreasing the differences between samples even further.

In conclusion, it was seen that the samples with the smaller median insert size have lower accuracy, and the higher number of PCR-cycles can have contributing effect as well. The lower accuracy can be overcome by using strict filtering and preferably based on the repeat unit size. However, looking at the big picture then all samples are quite similar to each other when doing appropriate filtering, especially in longer repeat units where the accuracy is very high for all samples.

3.4 Methods

Sequencing data of seven replicate samples were provided by Victorian Clinical Genetics Services. Three of the samples first underwent a fragmentation step where they were targeted to produce fragments of 500 bp, 1000 bp and 1500 bp to simulate degraded DNA. Then, the library of all samples was prepared by using Nextera DNA Flex Library Prep kit (Illumina) following the manufacturer's protocol. To analyse the impact of PCR on STRs genotyping accuracy different number of PCR cycles were used regarding the amount of sample's DNA as recommended by the protocol: 8 cycles of PCR for 15.6 ng of DNA, 6 cycles for 27.2–33.0 of DNA and 5 cycles for 54.5–108.9 ng of DNA. All samples were sequenced on Illumina NovaSeq 5000 platform.

Following the sequencing, FASTQ files were aligned using BWA-MEM v0.7.13 (Li, 2013), duplicates removed with Sambamba markdup v0.7.1 (Tarasov et al., 2015) and indexed by using Samtools v1.10 (Li et al., 2009).

All bioinformatic scripts and commands used can be found from the project's repository on GitLab (https://gitlab.com/andreassh/research-str-wgs). In short, after the alignment all samples were downsampled in order to achieve a uniform coverage. First, Samtools v1.10 (Li et al., 2009) *depth* function was used to find the number of reads at each position or region of all samples and then finding the average coverage by calculating the mean of all positions. Downsampling was performed with the Picard's (http://broadinstitute.github.io/picard/) *DownsampleSam* function by using the default parameters, and the proportion of reads to keep was calculated by using the mean coverage found in the previous step. The quality of sequences were determined with Qualimap v2.2.1 (Okonechnikov et al., 2015), as well as providing information about the insert size and the coverage after downsampling.

Tandem Repeats Finder v4.09 (TRF) (Benson, 1999) was used to find and generate a custom set of STR loci found in the human reference sequence hs37d5 (1000 Genomes Project Phase II) with the hs37d5 decoy sequence. TRF was run with the following parameters: match score: 2, mismatch penalty: 7, indel penalty: 7, match probability: 80, indel probability: 10, minimum alignment score: 24 and maximum period size: 6.

GangSTR v2.4 (Mousavi et al., 2019) was used to call STRs from the sequencing data. The regions file required by GangSTR was generated with a Python script *trf2bed*

(https://gitlab.com/andreassh/trf2bed) by using the set of STR loci found by TRF. GangSTR was run with the default parameters on each sample separately. A custom-made Python script named *extract_data.py* (available at the project's repository) was used to create analysis ready files from GangSTR's variant calling files. These files included information about each STR loci called, such as: sample name, locus, chromosome, start and end coordinates of the STR region, length of repeat unit in base pairs, repeat unit (motif) sequence, number of repeats in the reference genome, length of both alleles in repeat units, genotype, number of reads per each read class separately, confidence interval for both of alleles, quality score and whether the call passed the filter. Calls did not pass the filter on two conditions: (1) when at least one of the alleles were outside of the confidence interval of that particular allele, and (2) when only spanning or bounding class of reads were present.

The Illumina Platinum pedigree PCR-free high coverage genome of NA12878 was downloaded from The International Genome Sample Resource (IGSR) portal (https://www.internationalgenome.org/data-portal/sample/NA12878). FASTQ files were aligned on the human reference sequence hs37d5 by using BWA-MEM v0.7.17 (Li, 2013), and aligned BAM files were sorted and indexed with Samtools v1.10 (Li et al., 2009). The process for calling STRs from the PCR-free sample is identical to the one used for samples as previously described. The data was then analysed and visualised in R using the scripts found in the project's repository.

Chapter 4: Detecting disease-causing short tandem repeat mutations in known pathogenic loci

4.1 Introduction

Today, over 50 monogenic human diseases are known to be caused by short tandem repeat (STR) expansions and many new ones have been discovered over the past three years (Khristich & Mirkin, 2020). Recent findings have shown that some diseases, such as different types of familial adult myoclonic epilepsy, are caused by STR motifs that are not present in the human reference genome nor healthy populations, but only found in affected individuals and considered to be pathogenic (Corbett et al., 2019; Florian et al., 2019; Ishiura et al., 2018; Lagorio et al., 2019).

Utilization of high-throughput sequencing (HTS) data is an established method in clinical contexts to uncover rare diseases (Frésard & Montgomery, 2018). To detect pathogenic variants, such as SNVs and small INDELs, public databases such as ClinVar that contain data of known pathogenic variants are being used (Rančelis et al., 2017). However, there is currently no standard method for genotyping pathogenic STRs. First of all, while it has been challenging to analyse STRs from HTS data, methods have been improving significantly over the past decade. In previous chapters, tools for genome-wide analysis were used, some of which were only able to genotype repeats within a read. In this chapter, we are focusing on small set of loci and two tools, ExpansionHunter (Dolzhenko et al., 2017) and GangSTR (Mousavi et al., 2019). These tools have been developed to be suitable for targeted analysis and are also able to estimate repeat length for alleles longer than the fragment length. Two other similar tools are also available, exSTRa (Tankard et al., 2018) and TREDPARSE (Tang et al., 2017). However, the former cannot estimate repeat size and the latter cannot genotype long alleles that are over a fragment length (Mousavi et al., 2019). In addition, potentially other, genome-wide tools could be used to perform targeted analysis when doing some modifications. For example, STRetch (Dashnow et al., 2018) is able to detect long expansions and could be used for a focused analysis by replacing the genome-wide list of STR loci with the disease-associated one.

ExpansionHunter is a tool created specifically for targeted genotyping analysis using a variant catalogue containing 22 known pathogenic STR disorders/loci (at the time of our research). On the other hand, GangSTR is a genome-wide genotyping tool by default,

with an option to do targeted analysis. The authors of GangSTR provide a reference catalogue created by Tandem Repeats Finder (TRF) including 829,933 genome-wide loci in which 23 of them overlap with known pathogenic loci, as counted by us. However, only 14 of them were validated in the research article by authors (Mousavi et al., 2019) and it is not known whether the other 10 have been validated. Moreover, ExpansionHunter and GangSTR both provide off-target regions for only 2 and 12 disease loci, respectively, which allows for more accurate genotyping of long alleles. Taken together, both available tools offer genotyping of less than half of all known diseases caused by STR mutations and for many diseases it is not possible to detect long alleles without specified off-target regions.

A second reason why STR genotyping is not routinely performed in a clinical setting is that information regarding known pathogenic STRs loci is scattered across multiple research articles and it is also challenging to determine the correct genomic regions for many of these loci due to the lack of annotation. In addition, genomic coordinates in research papers are often reported in different genome assemblies and do not specify state whether it is in 0-based or 1-based coordinates. The length of a locus is also variable between articles, depending on whether only intact repeated region and/or interruptions are included. Therefore, these coordinates may not be suitable for direct usage with a genotyping tool without verifying a region, and likely some modifications are needed. As a consequence, it requires considerable effort to confidently analyse new or rarer STR loci that are not included into the variant catalogues of STR genotyping tools. Moreover, manually adding a new locus into the catalogue should be validated before using the method on real samples. Without any help from bioinformaticians it is especially difficult to do this by clinicians or researches from other fields.

Finally, there is not much known about population wide variation in STRs loci. It would be useful to have repeat length distributions for different populations when working on affected individuals to determine whether an expansion might be disease causing in general, or for an individual in a specific population. For example, repeat expansions in C9orf72 that causes amyotrophic lateral sclerosis-frontotemporal dementia have reported to be pathogenic from 30 repeats (Renton et al., 2011), whereas another study using different population shows individuals with repeats of this length without

neurodegenerative or psychiatric disease (Kaivola et al., 2019), which could indicate some population differences.

The result of the aforementioned issues has been that while the currently available methods for genotyping STRs from HTS data have good accuracy, less than half of the known diseases can be genotyped. Furthermore, a straightforward workflow/pipeline for genotyping is not available for non-bioinformaticians. Thus, the work in this chapter aims to solve both these issues by providing accessible and easy-to-use tool for non-bioinformaticians that can accurately genotype STR loci of all known diseases.

We will use ExpansionHunter as the choice of our genotyping tool as it a recognised STR genotyping tool and it is relatively easy to use and fast as it does not require an additional alignment step (Bahlo et al., 2018). We aim to increase the number of loci genotyped in the provided catalogue by over two-fold to include all known disease-causing loci, including the rarest and also the latest discovered. The new catalogue will be made publicly available to enhance the use of HTS data for the analysis of STR expansion. We use simulations of PCR-free samples to validate all heterozygous and very rare homozygous expanded loci and determine genotyping accuracy over different lengths of repeat expansions (including the ones already found in the catalogue and newly added). Additionally, we also simulate stutter noise to validate the performance of the tool on PCR+ samples. As one of our goal is also to improve genotyping estimation for long alleles, we determine the most likely regions for each loci over the genome where reads full of repeats could misalign and use these as off-target regions to improve accuracy of genotyping. Finally, we develop a software application for a straightforward genotyping of pathogenic STR loci that will be released for public use to simplify and increase the accessibility of analysing STRs in clinical or research context.

4.2 Results

4.2.1 Compiling the list of diseases caused by mutations of short tandem repeats

A search in literature was conducted to find all diseases caused by monogenic STR mutations. We found exactly 58 diseases in total that are caused by STR mutations in 52 genes, as four genes were associated with two different diseases and one with three syndromes. For example, expansions between 55-200 repeats in the FMR1 gene causes fragile X tremor/ataxia syndrome whereas long, over 200 repeat expansions result in fragile X syndrome. In these 52 genes, there were total of 55 different pathogenic loci due to the fact that in two genes (ARX and HOXA13) there were multiple tracts associated with different diseases.

The list of diseases was divided into three sections based on the type of repeats which we classified into either standard, imperfect GCN or replaced/nested types of repeats (Table 4.1). For each locus we noted its location in the reference genome, name of the gene, repeat unit, associated disease(s) with inheritance, range of repeat lengths and MIM numbers. All genomic coordinates were converted to be in hg38 assembly (unless already given). The start and end positions of each locus were assessed manually and adjusted to be in 0-based coordinate system. For specifying reference regions, only pathogenic motifs with at least two consecutive repeats at the start or end of the sequence were included.

Table 4.1. Types of short tandem repeats.

Type	**Description**	**Example**	**Example disease**
Standard repeats	One specific motif is repeated in a row and becomes pathogenic when expanded above a certain threshold.	Normal: CAG CAG CAG Pathogenic: CAG CAG CAG CAG CAG CAG ...	Huntington's disease
Imperfect GCN repeats	As standard repeats, with the exception that repeated motifs can have different sequence but they are all coding alanines.	Normal: GCA GCA GCG Pathogenic: GCA GCA GCG GCG GCC GCT ...	Developmental and epileptic encephalopathy
Replaced repeats	An endogenous motif is replaced with a novel one which expansion over a certain threshold is shown to be pathogenic.	Normal: AAAAG AAAAG AAAAG Pathogenic: AAGGG AAGGG AAGGG AAGGG ...	Cerebellar ataxia, neuropathy, and vestibular areflexia syndrome
Nested repeats	Expanded stretch of novel repeats is inserted between or next to the endogenous repeat sequence.	Normal: AAAAT AAAAT AAAAT AAAAT Pathogenic: *AAAAT* TGAAA TGAAA ... *AAAAT*	Familial adult myoclonic epilepsy 1

We have made this curated set of disease causing STR expansions available as a public resource on https://stripy.org/database which will be updated as new disease-causing loci are discovered.

4.2.1.1 Standard type of repeats

Standard repeats are considered to be the ones where one specific repeat unit (for example, CAG or CCG) is in healthy individuals and its expansion is pathogenic. The 34 standard repeats are presented in the Table 4.2. Additional information about healthy and pathogenic repeat lengths in affected and unaffected individuals are provided in the following subsections.

Table 4.2. All known disease-causing STR loci made of standard type of repeats. The coordinates are in hg38 assembly and 0-based. Asterisk (*) denotes data where the pathogenic range vary in literature. [$]Denotes a range specified by us based on the findings in literature from multiple sources.

Gene	Reference and location	Motif	Disease	Inheritance	Normal range	Intermediate range	Pathogenic cut-off	References
AFF2 (FMR2)	chrX:148500638-148500683 (5′ UTR)	CCG	FRAXE mental retardation (FRAXE)	XLR	7-25	—	⩾200	(Knight et al., 1993)
AR	chrX:67545316-67545385 (Coding)	CAG	Spinal and bulbar muscular atrophy (SBMA)	XLR	17-26	—	⩾40	(Grunseich et al., 2014; Spada et al., 1991)
ATN1	chr12:6936728-6936773 (Coding)	CAG	Dentatorubral-pallidoluysian atrophy (DRPLA)	AD	7-23	—	⩾49	(Komure et al., 1995; Nagafuchi et al., 1994)
ATXN1	chr6:16327635-16327722 (Coding)	CAG	Spinocerebellar ataxia 1 (SCA1)	AD	⩽32	33-38	⩾39	(Orr et al., 1993; Rosas et al., 2020)
ATXN2	chr12:111598950-111599019 (Coding)	CAG	Spinocerebellar ataxia 2 (SCA2)	AD	⩽26*	27-33*	⩾35	(Gispert et al., 2012; van Blitterswijk et al., 2014)
ATXN3	chr14:92071010-92071040 (Coding)	CAG	Spinocerebellar ataxia 3 (SCA3)	AD	12-44	45-55	⩾56	(Ashizawa et al., 2018)
ATXN7	chr3:63912685-63912715 (Coding)	CAG	Spinocerebellar ataxia 7 (SCA7)	AD	7-27 (7-34)	28-35	⩾37	(Garden & La Spada, 2008; Gousse et al., 2018)
ATXN8OS	chr13:70139383-70139428 (3′ UTR)	CTG	Spinocerebellar ataxia 8 (SCA8)	AD	2-37	38-79[$]	⩾80	(Gupta & Jankovic, 2009; Schöls et al., 2003)
ATXN10	chr22:45795354-45795424 (Intron)	ATTCT	Spinocerebellar ataxia 10 (SCA10)	AD	10-29	30-799[$]	⩾800	(Teive et al., 2011)
C9orf72	chr9:27573528-27573546 (Intron)	GGGGCC	Amyotrophic lateral sclerosis and/or frontotemporal dementia (FTDALS1)	AD	2-6*	7-30*	⩾31*	(Kaivola et al., 2019; Van Mossevelde et al., 2017; Woollacott & Mead, 2014)

CACNA1A	chr19:13207858-13207897 (Coding)	CAG	Spinocerebellar ataxia 6 (SCA6)	AD	⩽18	19	⩾20	(Ashizawa et al., 2018)
CBL	chr11:119206289-119206322 (5′ UTR)	CCG	Jacobsen syndrome (JBS)	NA	11	80-99	⩾100	(Mattina et al., 2009)
CNBP (ZNF9)	chr3:129172576-129172656 (Intron)	CCTG	Myotonic dystrophy 2 (DM2)	AD	⩽29	30-54	⩾55	(Bachinski et al., 2009; Radvanszky et al., 2013)
COMP	chr19:18786034-18786049 (Coding)	GAC	Multiple epiphyseal dysplasia (MED)	AD	5	—	6	(Délot et al., 1999)
			Pseudoachondroplasia (PSACH)	AD	5	—	4 & 7	(Délot et al., 1999; Hecht et al., 1995)
DIP2B	chr12:50505003-50505024 (5′ UTR)	CGG	FRA12A mental retardation (MRFRA12A)	AD	6-23	—	⩾270	(Winnepenninckx et al., 2007)
DMD	chrX:31284557-31284605 (Intron)	GAA	Duchenne muscular dystrophy (DMD)	XLR	11-33	—	⩾59	(Kekou et al., 2016)
DMPK	chr19:45770204-45770264 (3′ UTR)	CTG	Myotonic dystrophy 1 (DM1)	AD	5-34	35-49	⩾50	(Brook et al., 1992; Gutiérrez Gutiérrez et al., 2020; Jaspert et al., 1995)
FMR1	chrX:147912050-147912110 (5′ UTR)	CGG	Fragile X Syndrome (FXS)	XLD	6-40	—	⩾201	(Mishra et al., 2019)
			Fragile X tremor/ataxia syndrome (FXTAS)	XLD	6-40	—	⩾55 ⩽200	(Mishra et al., 2019)
FXN	chr9:69037286-69037304 (Intron)	GAA	Friedreich ataxia (FRDA)	AR	5-30	—	⩾70	(Pandolfo, 2001)
GIPC1	chr19:14496041-14496074 (5′ UTR)	GGC	Oculopharyngodistal myopathy 1 (OPDM2)	AD	12-32	—	⩾73 (⩽500 ?)	(Deng et al., 2020)
GLS	chr2:190880872-190880920 (5′ UTR)	GCA	Glutaminase deficiency GD)	AR	5-26	—	⩾680	(van Kuilenburg et al., 2019)
HTT	chr4:3074876-3074933 (Coding)	CAG	Huntington's disease (HD)	AD	⩽26	27-35	⩾36	(Testa & Jankovic, 2019)
JPH3	chr16:87604287-87604329 (3′ UTR)	CTG	Huntington disease-like 2 (HDL2)	AD	6-27	—	⩾41	(Holmes et al., 2001a)
LRP12	chr1:149390805-149390831 (5′ UTR)	CGG	Oculopharyngodistal myopathy 1 (OPDM1)	AD	13-45	—	⩾90	(Ishiura et al., 2019)
NOP56	chr20:2652733-2652757 (Intron)	GGCCTG	Spinocerebellar ataxia 36 (SCA36)	AD	3-14	—	⩾650	(García-Murias et al., 2012; Kobayashi et al., 2011)

NOTCH2NLC (NBPF19)	chr1:149390802-149390829 (5′ UTR)	GGC	Neuronal intranuclear inclusion disease (NIID)	AD	7-39	—	⩾93	(Ishiura et al., 2019)
			Hereditary essential tremor type 6 (ETM6)	AD	4-41	—	⩾60	(Sun et al., 2020)
NUTM2B-AS1	chr10:79826377-79826403 (Intron)	CGG	Oculopharyngeal myopathy with leukoencephalopathy 1 (OPML1)	AD	3-16	—	⩾700	(Ishiura et al., 2019)
PABPN1	chr14:23321472-23321490 (Coding)	GCG	Oculopharyngeal muscular dystrophy (OPMD)	AD or AR	6	—	⩾7	(Ishiura et al., 2019)
PPP2R2B	chr5:146878728-146878758 (5′ UTR)	CAG	Spinocerebellar ataxia 12 (SCA12)	AD	7-31	—	⩾51	(Holmes et al., 2001b; Srivastava et al., 2001)
PRDM12	chr9:130681606-130681642 (Coding)	CCG	Neuropathy, hereditary sensory and autonomic, type VIII (HSAN8)	AR	⩽14	—	⩾18	(Chen et al., 2015)
TBP	chr6:170561907-170562015 (Coding)	CAG	Spinocerebellar ataxia 17 (SCA17)	AD	25-40	—	⩾41	(Koide et al., 1999; Maltecca et al., 2003; Nakamura, 2001)
TCF4	chr18:55586155-55586227 (Intron)	CTG	Fuchs endothelial corneal dystrophy-3 (FECD3)	AD	⩽31	32-49	⩾50	(Wieben et al., 2012; Zarouchlioti et al., 2018)
XYLT1	chr16:17470907-17470922 (5′ UTR)	GGC	Desbuquois dysplasia-2 (DBQD2)	AR	9-20	—	⩾110	(LaCroix et al., 2019)
ZIC3	chrX:137566826-137566856 (Coding)	GCC	X-linked VACTERL syndrome (VACTERLX)	XLR	10	—	⩾12	(Wessels et al., 2010)

STR mutations in AFF2

FRAXE mental retardation (FRAXE) is caused by CCG repeats in the AFF2 gene (also known as the FMR2 gene; MIM: 300806). Knight et al. (1993) found that in all male patients at least 200 repeats were present. Also, carrier females with over 200 repeats could have mental impairment. Repeat range in the control group was between 6 and 30 repeats (Knight et al., 1993).

STR mutations in AR

X-linked spinal and bulbar muscular atrophy (SBMA), also known as the Kennedy disease is caused by CAG repeat expansions in the AR gene (MIM: 313700). The pathogenic repeat range in multiple studies is reported to be at least 40 (Atsuta et al., 2006; Doyu et al., 1993; Fratta et al., 2014; La Spada et al., 1992; Spada et al., 1991). A patient with early onset of disease was found to have 68 repeat expansion, which is the known largest so far (Grunseich et al., 2014). In 75 control samples the range was determined to be between 17 to 26 repeats (Spada et al., 1991).

STR mutations in ATN1

CAG repeat expansions in ATN1 gene (MIM: 607462) are causing the dentatorubral-pallidoluysian atrophy (DRPLA) disease. Two studies have showed that in healthy individuals the repeat size varies from 7–23 repeats (Komure et al., 1995; Nagafuchi et al., 1994). In patients, repeat size has been observed to be between 49–90 (Komure et al., 1995; Nagafuchi et al., 1994; Sugiyama et al., 2020).

STR mutations in ATXN1

Normal range of repeats in ATXN1 gene (MIM: 601556) has been considered to be in the range of 6–39 repeats and where CAG expansions beyond 39 repeats are causing the spinocerebellar ataxia type 1 (SCA1) (Orr et al., 1993). Recently there are reports showing the potential risk factor of intermediate length alleles (at least 33 CAG repeats) on the ALS (Lattante et al., 2018) as well as on progressive non-fluent aphasia (PNFA) (Rosas et al., 2020).

STR mutations in ATXN2

The normal range of repeats in ATXN2 gene (MIM: 601517) in healthy population is less than 35 repeats, usually between 22 and 23 repeats and expansions over 34 repeats are known to cause spinocerebellar ataxia type 2 (SCA2) (Gispert et al., 2012).

There is no consensus regarding the intermediate length of repeats and it ranges from 24 to 32 repeats in many studies, depending on the studied population. It was found by van Blitterswijk et al (2014) that intermediate length expansions (27–33 repeats) could act as a disease modifier and influence phenotype of carriers of the C9orf72 expansions, making them more susceptible to motor neuron diseases (MND), such as ALS. However, no association was found between ATXN2 repeat length and age of onset of MND or the survival rates after onset of the disease. In contrast to MND, repeat expansions in ATXN2 gene are not associated with frontotemporal dementia (FTD) (van Blitterswijk et al., 2014).

STR mutations in ATXN3

Expansions in the ATXN3 gene (MIM: 607047) cause spinocerebellar ataxia type 3 (SCA3), also known as the Machado-Joseph disease (MJD) which is a neurodegenerative disease (Kawaguchi et al., 1994). The normal range of CAG repeats is up to 44, whereas in affected individuals it has been found to be in 56–87 repeats range (McLoughlin et al.,

2020). Repeats in the range of 45–55 have shown to present incomplete penetrance of disease's symptoms and can be considered as the intermediate range (Ashizawa et al., 2018).

STR mutations in ATXN7

In a healthy population, CAG repeats in the ATXN7 gene (MIM: 607640) ranges from 7–34, whereas expansions of 37 or more repeats cause spinocerebellar ataxia type 7 as reported by Garden et al. (2008). They also showed a strong negative correlation between the repeat length and age of onset. In addition, individuals with 59 or more repeats often experience visual impairments as the first symptom (Garden & La Spada, 2008). Repeats between 28 and 35 increases risk of pathogenic alleles during meiosis which would result in SCA7 in offspring (Gousse et al., 2018), but repeats in this range are extremely rare (Stevanin et al., 1998).

STR mutations in ATXN8OS

ATXN8OS (MIM: 603680) is the opposite strand of the ATXN8 gene that transcribes an mRNA that is associated with the spinocerebellar ataxia type 8 (SCA8) (Daughters et al., 2009). CTG repeat expansions in the 3′-UTR of ATXN8OS are leading to a toxic gain-of-function protein and causing the disease (Daughters et al., 2009).

The cut-off values for pathogenic length of repeats is debatable as many studies have shown the pathogenic range to be above 80 repeats, whereas one of the biggest ones so far has reported that majority of patients with symptoms had repeat length between 45 and 109 CTGs (in the range of 2–37 repeats in controls) (Schöls et al., 2003). However, there are also reports of many healthy individuals who had a repeat length over 91 CTGs (Zeman et al., 2004). This shows the overlap in repeat length of individuals with and without SCA8 and it is recommended to interpret genotyping results in clinical context (Gupta & Jankovic, 2009). In general, the majority of studies show a repeat length at least 80 for affected individuals but at least 50 should be expected for a SCA8 patient (Gupta & Jankovic, 2009).

STR mutations in ATXN10

Expansions of pentanucleotide ATTCT repeats in an intron of the ATXN10 gene (MIM: 611150) is responsible for the development of spinocerebellar ataxia type 10 (SCA10) as reported by Matsuura et al. (2000, 2006). They reported that the number of repeats in a

healthy population was between 10 and 29, whereas pathological repeats in affected individuals were in the range of 800–4500 (Matsuura et al., 2000; Matsuura et al., 2006). Apart from that, there is also evidence of repeat lengths with reduced penetrance. For example, an affected individual was identified with a 400 repeat expansion, whereas her siblings with 360 repeats and the father with 370 repeats were unaffected (Alonso et al., 2006). Also, a patient with early-onset ataxia was described, having 280 ATTCT repeats, while her mother was unaffected with the same repeat length (Matsuura et al., 2006).

STR mutations in C9orf72

Expansions in C9orf72 gene (MIM: 614260) are linked to the ALS and FTD and is a common cause for ALS, FTD and FTD-ALS diseases, accounting for 11%, 4–29% and 17–28% of cases, respectively (Van Mossevelde et al., 2018). In addition, the frequency is higher for familial cases of the diseases (38% in ALS, 29% in FTD and up to 88% in FTD-ALS) (Van Mossevelde et al., 2018). Repeat expansions have been also detected in patients diagnosed with other neurodegenerative diseases, such as Alzheimer's, Parkinson, and Huntington disease-like disease (Beck et al., 2013; Cacace et al., 2013; Lesage et al., 2013; Woollacott & Mead, 2014).

In a healthy population repeat length is considered to be in the normal range for those with fewer than 30 repeats in an allele and alleles less than 10 repeats are significantly more common ones than the ones between 10–30 (Beck et al., 2013; Kaivola et al., 2019; Rutherford et al., 2012). However, it is also population dependent. For example, Finnish population have alleles in range of 20–30 repeats, twice more compared to the North American or British population (Kaivola et al., 2019). The intermediate range has not been clearly established, but often is defined as either 7–30 or 20–30 repeats and thought to contribute to neurodegenerative diseases (Kaivola et al., 2019). Pathogenic range is generally considered above 30 repeats and likely reaching up to thousands of repeats, but the minimum length is not defined (Woollacott & Mead, 2014). It is also suggested that alleles between 30–45 should not be considered as pathogenic (Kaivola et al., 2019). However, most expansions in patients are at least several hundred repeats long (Van Mossevelde et al., 2017).

STR mutations in CACNA1A

The normal length of repeats for CACNA1A gene (MIM: 601011) is up to 18 repeats, and repeat length of 19 is considered to be intermediate and individuals who have at least

20 repeats of CAG are affected by the spinocerebellar ataxia type 6 (Ashizawa et al., 2018).

STR mutations in CBL

Jacobsen syndrome (JBS) is usually caused by gene deletion of the 11qter region as described by Mattina et al. (2009). However, they found that CCG repeat expansion in the 5′ UTR of the CBL gene (MIM: 165360) could be responsible for the pathogenesis of the disease through hypermethylated expanded repeats, which would have a negative effect on DNA replication. In 70% of analysed individuals had 11 copies of CCG repeats, but this can expand to 80–100 repeats (premutation) and the full mutation can be up to several hundreds of repeats (Mattina et al., 2009).

STR mutations in CNBP

Myotonic dystrophy type 2 (DM2) is disease similar to DM1 but it is caused by tetranucleotide CCTG expansions in the intron 1 of CNBP gene (also known as ZNF9; MIM: 116955) which can even reach to 11000 copies of CCTG (Liquori, 2001). In a healthy population the tract length has shown to be between 94 to 217 bp (Bachinski et al., 2009). Radvanszky et al. (2013) identified stretches of CCTG repeat in the tract that were often interrupted by GCTG, TCTG or ACTG motifs in healthy individuals. Instability was correlated with the length of uninterrupted tracts and unstable pre-mutations may be between 30–55 CCTG repeats (Radvanszky et al., 2013). The smallest disease-causing allele has found to have 55 repeats of CCTG (Bachinski et al., 2009).

STR mutations in COMP

Mutations in the COMP gene can lead to two diseases: pseudoachondroplasia (PSACH) and a type of multiple epiphyseal dysplasia (MED), both of which are skeletal disorders (Lin et al., 2018). Repeat mutations in COMP gene (MIM: 600310) can cause both diseases, as shown in two patients described by Délot et al. (1999). One affected individual who was diagnosed with a typical PSACH phenotype had 7 repeats of GAC in the allele. The other, patient with a mild MED was carrying an allele with 6 repeats in the same locus (Délot et al., 1999).

Moreover, PSACH patients with a deletion of one GAC repeat has been described, demonstrating the pathogenic effect of contraction of a repeat (Hecht et al., 1995).

STR mutations in DIP2B

CGG repeat expansions in the 5′ UTR of DIP2B gene (MIM: 611379) have been associated with FRA12A mental retardation (MRFRA12A). Research by Winnepenninckx et al. (2007) determined the repeat length in the control group to be between 6 and 23 repeats. In the individuals, who cytogenetically express fragile site (FRA12A) were found to have repeat expansions. In unaffected individuals, one allele was found to be 650–850 bp, that was longer in affected – over 1050 bp (Winnepenninckx et al., 2007). Therefore, the pathogenic length for the affected samples could be over 270 repeats.

STR mutations in DMD

Duchenne muscular dystrophy (DMD) is the most common and lethal type of muscular dystrophies as described by Kekou et al. (2016). Typically, DMD is known to be caused by large deletions and duplications in the DMD gene (MIM: 300377), but they found that repeat expansions can be disease-causing as well. In an intron of the DMD gene, expansions between 59 and 82 GAA repeats was discovered in affected individuals. In the healthy controls, GAA was in the range of 11–33 repeats (Kekou et al., 2016).

STR mutations in DMPK

Myotonic dystrophy type 1 (DM1) is a progressive neuromuscular disease due to CTG trinucleotide expansions in the 3′ UTR of DMPK gene (MIM: 605377), which is caused by at least 50 repeats (Brook et al., 1992). Salehi et al. (2007) research demonstrated that disease severity is correlated with repeat length and more muscular disability is seen at longer repeats. Affected individuals with expansions up to 100 repeats usually develop mild symptoms or asymptomatic (Salehi et al., 2007). The most severe phenotype of this disease can be observed in individuals with long repeats, which can even reach up to thousands of repeats (Jaspert et al., 1995). In a healthy population, allele length ranges 5–34 CTG repeats, while premutation alleles are defined to be between 35 and 49 repeats (Gutiérrez Gutiérrez et al., 2020). These individual has no symptoms but whose offspring can acquire the allele and may reach to the pathogenic range (Gutiérrez Gutiérrez et al., 2020).

STR mutations in FMR1

Both, Fragile X syndrome (FXS) and fragile X-associated tremor/ataxia syndrome (FXTAS) are related but clinically distinct diseases that are caused by CGG repeat

expansion in the FMR1 gene (MIM: 309550) (Mishra et al., 2019). In a healthy population, the repeat length in FMR1 is between 6 and 40 repeats (Aziz et al., 2003). Repeat expansions in the range 55–200 (pre-mutations) causes FXTAS (gene is overexpressed) whereas repeats over 200 repeats causes FXS (gene is silenced) (Mishra et al., 2019).

STR mutations in FXN

Friedreich ataxia (FRDA) disease is caused by GAA repeat expansions in intron 1 of the FXN gene (MIM: 606829), with pathogenic repeats ranges from 70 repeats to more than 1000 (most often between 600–900) (Pandolfo, 2001). In healthy individuals, the range is found to be 5–30 GAA repeats (Pandolfo, 2001).

STR mutations in GIPC1 and LRP12

Oculopharyngodistal myopathy (OPDM) is a rare neuromuscular disease and around 200 affected individuals have been found to date since 1977, although the genetic cause still remains elusive (Deng et al., 2020). Type 1 of the disease (OPDM1) is thought to be caused by CGG repeat expansions in the LRP12 gene (MIM: 618299) which was recently described by Ishiura et al. (2019). They discovered repeats in the length of 280 or 380 bp (corresponding to ~93 or ~127 repeats, respectively) in the affected individuals. However, this was only observed in a subset of the samples. In 1000 controls, CGG repeats were reported to be between 13–45 repeats, except for two of them who also had expanded alleles (Ishiura et al., 2019).

The article by Ishiura and colleagues investigating repeats in the LRP12 gene was shortly followed by another study conducted by Deng et al. (2020) that examined Chinese and Japanese individuals. In these patients, repeat expansions in the 5′ UTR of the GIPC1 gene (MIM: 605072) was discovered and this was associated with OPDM as well, followed by naming the disease as type 2 (OPDM2). In 12 out of 24 affected individuals from Chinese cohort and 7 out of 194 patients from Japanese cohort had expansions in the GIPC1 gene, ranging from 88–164 for the Chinese and 73–161 for the Japanese cohort. No expansions were detected in any of the controls. The range of GGC repeats in 550 control subjects was determined to be between 12–32 (Deng et al., 2020).

In the same research, it was suggested that expansions in two different genes could cause OPDM through shared pathogenesis. Interestingly, one unaffected individual with over

500 GGC repeats in 5′ UTR of the GIPC1 gene was discovered, whose both offspring, with 117 and 113 repeats, were symptomatic. That raises a possibility where a certain range of repeats can cause a disease, similarly to CGG repeats in the FMR1 gene whereby 55–199 repeats causes FXTAS while 200 and more repeats results in FXS. More research is needed to confirm this (Deng et al., 2020).

STR mutations in GLS

Recently, the role of tandem repeats in the GLS gene (MIM: 138280) was discovered in three unrelated patients of glutaminase deficiency (GD) disease, carrying either 680, 900 or 1500 CAG repeats in the GLS gene (van Kuilenburg et al., 2019). In the majority of controls, the repeat length varied from 5 to 26 repeats (van Kuilenburg et al., 2019).

STR mutations in HTT

The normal length of CAG repeats in the HTT gene (MIM: 613004) is below 27 repeats (Testa & Jankovic, 2019). Allele lengths 40 or more repeats causes Huntington's disease (HD) and is fully penetrant, however, the disease may also occur in individuals with 36–39 repeats in which case they often encounter later age of disease onset and slower progression (Testa & Jankovic, 2019). The number of CAG repeats is negatively correlated with the age of motor onset (Gusella et al., 2014).

Intermediate alleles in the range of 27–35 have been associated with normal phenotype, however, there are also reports showing repeats in this range can increase the risk for Alzheimer's disease (Menéndez-González et al., 2019).

STR mutations in JPH3

Expansions of CTG repeats in 3′ UTR of JPH3 gene (MIM: 605268) have shown to cause Huntington's disease-like 2 (HDL2) syndrome and affected individuals have been found to carry alleles with 41–58 repeats, while in normal individuals, it varies from 6 to 27 repeats (Holmes et al., 2001a; Margolis & Holmes, 2003).

STR mutations in NOP56

Expansions of a hexanucleotide repeat GGCCTG in the intron 1 region of NOP56 gene (MIM: 614154) causes spinocerebellar ataxia type 36 (SCA36) (Kobayashi et al., 2011). The healthy population typically has 3–14 repeats whereas in affected individuals, the pathogenic allele contains repeats ranging from 650 to 2500 (García-Murias et al., 2012; Kobayashi et al., 2011).

STR mutations in NOTCH2NLC

Neuronal intranuclear inclusion disease (NIID) is a neurological disease inherited in an autosomal dominant manner, although around two thirds of cases have been found to be sporadic (Ishiura et al., 2019; Sone et al., 2016). It was recently described by two research groups that GGC repeats in the 5′ UTR of NOTCH2NLC gene (MIM: 618025) are causing the disease (Ishiura et al., 2019; Nakamura et al., 2020). Ishiura et al. (2019) found that in 28 patients, the expansion was in the range of 90–180 repeats. In the control group they did not see any expansions and in 182 unaffected individuals, the repeat range was between 7 and 39 repeats (Ishiura et al., 2019). This was confirmed by another study a year later showing 330–520 bp repeat length (corresponding to 110–173 trinucleotide repeats) in seven NIID patients (Nakamura et al., 2020).

In addition, expansions in NOTCH2NLC gene are also associated with the hereditary essential tremor type 6 (ETM6), as recently reported (Sun et al., 2020). The repeat length in affected individuals from mainland China was determined to be between 60 and 250 repeats, whereas in the control group it was between 4 to 41 repeats (Sun et al., 2020).

STR mutations in NUTM2B-AS1

A new disease named oculopharyngeal myopathy with leukoencephalopathy 1 (OPML1) has been lately described by Ishiura et al. (2019) that was found in one family. Expansion of CCG repeats was observed in the NUTM2B-AS1 gene (MIM: 618639) that is transcribed bidirectionally (in CGG direction) in the noncoding RNA LOC642361 gene. In one out of four patients, the expanded allele was determined to be 2.1 kb which corresponds to 700 trinucleotide repeats. In controls the repeat range was between 3 to 16 repeats (Ishiura et al., 2019).

STR mutations in PABPN1

Oculopharyngeal muscular dystrophy (OPMD) is a muscle disease which is caused by short trinucleotide (GCG) expansions in the coding region of PABPN1 gene (MIM: 602279) as described by Brais et al. (1998). They showed that normal individuals carry 6 repeats of GCG in the allele, whereas affected subjects have 7–13 repeats. The pathogenic repeat length usually varies from 8 to 13 repeats which leads to autosomal dominant disease, but more severe disease was caused by homozygous alleles with 7 repeats (Brais et al., 1998).

STR mutations in PPP2R2B

Pathological CAG repeat expansions in the PPP2R2B gene (MIM: 604325) leads to spinocerebellar ataxia type 12 (SCA12) (Srivastava et al., 2001). The range of repeats in healthy individuals ranges from 7–31 and in affected individuals, 51 to 69 repeats (Holmes et al., 2001b; Srivastava et al., 2001).

STR mutations in PRDM12

Repeat mutations in the PRDM12 gene (MIM: 616458) have been shown to be one of the causes of the hereditary sensory and autonomic neuropathy type VIII (HSAN8) (Chen et al., 2015). Patients with homozygous alleles of 18 and 19 alanine repeats have been defined, while in the general population the length of alleles does not exceed 14 alanine repeats (Chen et al., 2015).

In some cases, the type of repeats causing HSAN8 has been classified as imperfect GCN (Khristich & Mirkin, 2020; Lavoie et al., 2003). However, we consider repeats in this locus as the standard type due to the fact that in both affected and unaffected individuals, the allele consists of alanines coded by GCC repeat units only, except the last residue that is coded by GCG (Chen et al., 2015). This means that the repeat expansion in patients resulted in extended stretch of only GCC repeats, and therefore do not have the characteristics of imperfect GCN types of repeats.

STR mutations in TBP

Research has shown that individuals who have had 41–66 repeats of CAG in the coding region of the TBP gene (MIM: 600075) are developing the spinocerebellar ataxia type 17 (SCA17) disease, while in a healthy population repeats range from 25 to 42 (Koide et al., 1999; Maltecca et al., 2003; Nakamura, 2001).

STR mutations in TCF4

Fuchs endothelial corneal dystrophy type 3 (FECD3) is an eye disease which is caused by CTG repeat expansions in the TCF4 gene (MIM: 602272) (Zarouchlioti et al., 2018). The conservative approach has defined 50 or more CTG repeats as the cut-off for disease-causing (Zarouchlioti et al., 2018). In one research conducted by Wieben et al. (2012), it was reported that 79% of patients had more than 50 repeats (reaching to over 1000 repeats in 4 cases) and 20% of the cases had less than 40 repeats. In the control group of 63 individuals, 95% of the subjects had below 40 repeats and 5% over 50 repeats (Wieben

et al., 2012). However, it was recently recommended that the length of at least 32 repeats should be considered as risk associated (Zarouchlioti et al., 2018).

STR mutations in XYLT1

Desbuquois dysplasia-2 (DBQD2), also known as the Baratela-Scott syndrome is caused by mutations in the XYLT1 gene (MIM: 608124) and LaCroix et al. (2019) recently discovered that GGC repeat expansions in the XYLT1 can be disease-causing as well. By investigating 12 individuals from 10 families with DBQD2, a 238 bp sequence missing from the reference genome (hg19 and hg38) was discovered. In that sequence, expansions ranging from ~110 to ~850 repeats were found in all patients. The 238 bp sequence was seen to be present in the control group (100 individuals) as well, where GGC repeats were in the range of 9–20 repeats (LaCroix et al., 2019).

STR mutations in ZIC3

It has been discovered only in one patient by Wessels et al. (2010) that two repeat expansion in the exon 1 of the ZIC3 gene (MIM: 300265) was responsible for developing X-linked VACTERL syndrome. The patient with the mutation died an hour after birth. It was later determined that the repeat length expanded from 10 to 12 repeats of alanine coded by GCC repeat. None of the 336 controls nor the mother had an expansion in the locus (Wessels et al., 2010).

4.2.2 Imperfect GCN type of repeats

Imperfect GCN type refers to repeats which encodes for the amino acid alanine, but the sequence can be composed of different repeat units (either GCA, GCT, GCC or GCG). Compared to the vast majority of standard repeats, the diseases of this type are usually caused by short expansions in conservative loci and all pathogenic cut-off values are below 30 residues of alanine (Table 4.3). Currently we have curated nine disease causing repeat expansions of this type.

Table 4.3. All known disease-causing STR loci made of imperfect GCN type of repeats. Given coordinates are in hg38 assembly and 0-based. $^{+}$, † and ‡ are used to distinguish values for different tracts in a locus.

Gene	Reference and location	Motif	Disease	Inheritance	Normal range	Intermediate range	Pathogenic cut-off	References
ARX	†chrX:25013529-25013565 ‡chrX:25013649-25013697 (Coding)	GCN	Developmental and epileptic encephalopathy-1 (DEE1)	XLR	$12^{\dagger}$ $16^{\ddagger}$	—	$\geqslant 20^{\dagger}$ $\geqslant 23^{\ddagger}$	(Strømme et al., 2002)
			Partington syndrome (PRTS)	XLR	$12^{\dagger}$	—	$\geqslant 20^{\dagger}$	(Strømme et al., 2002)
			X-linked mental retardation with or without seizures (MRXARX)	XLR	$12^{\dagger}$ $16^{\ddagger}$	—	$\geqslant 20^{\dagger}$ $\geqslant 18^{\ddagger}$	(Bienvenu et al., 2002)
FOXL2	chr3:138946020-138946062 (Coding)	GCN	Blepharophimosis, ptosis, and epicanthus inversus syndrome (BPES)	AD	14	—	$\geqslant 19$	(De Baere et al., 2003; Nallathambi et al., 2007)
HOXA13	$^{+}$chr7:27199924-27199966 †chr7:27199825-27199861 ‡chr7:27199678-27199732 (Coding)	GCN	Hand-foot-genital syndrome (HFG)	AD	14^{+} $12^{\dagger}$ $18^{\ddagger}$	—	$\geqslant 22^{+}$ $\geqslant 18^{\dagger}$ $\geqslant 24^{\ddagger}$	(Goodman et al., 2000; Innis et al., 2004; Utsch et al., 2002)
HOXD13	chr2:176093058-176093097 (Coding)	GCN	Synpolydactyly (SPD)	AD	15	—	$\geqslant 22$	(Albrecht et al., 2004)
PHOX2B	chr4:41745971-41746031 (Coding)	GCN	Central hypoventilation syndrome (CCHS)	AD	20	—	$\geqslant 24$	(Sivan et al., 2019)
RUNX2	chr6:45422750-45422801 (Coding)	GCN	Cleidocranial dysplasia (CCD)	AD	17	—	$\geqslant 27$	(Mundlos et al., 1997)
SOX3	chrX:140504316-140504361 (Coding)	GCN	X-linked mental retardation (XLMR)	XLR	15	—	$\geqslant 26$	(Laumonnier et al., 2002)
			X-linked panhypopituitarism (PHPX)	XLR	15	—	$\geqslant 22$	(Woods et al., 2005)
TBX1	chr22:19766762-19766807 (Coding)	GCN	Tetralogy of Fallot (TOF)	AD	15	—	$\geqslant 25$	(Rauch et al., 2010)
ZIC2	chr13:99985448-99985493 (Coding)	GCN	Holoprosencephaly 5 (HPE5)	AD	15	—	$\geqslant 25$	(Brown et al., 2005)

STR mutations in ARX

There are four polyalanine tracts in the ARX gene (MIM: 300382), two of which are in exon 2 and are associated with multiple syndromes. Healthy individuals have 12 repeats of alanine in one of these tracts and the most frequent mutation found there, an eight polyalanine expansion, is linked to Partington syndrome (PRTS), X-linked mental retardation with or without seizures (MRXARX) and the developmental and epileptic

encephalopathy-1 (DEE1) (Bienvenu et al., 2002; Strømme et al., 2002). In all of these patients, a 24 bp duplication of alanine, resulting in 20 repeat allele was observed (Bienvenu et al., 2002; Strømme et al., 2002). Moreover, a longer, 33 bp insertion in the same tract has been also described in two brothers, stretching an allele to 23 repeats (Demos et al., 2009).

The longer tract has a length of 16 alanines that causes MRXARX and DEE1. Patients in one family with the MRXARX syndrome had 2 repeat expansion, resulting in 18 residues of alanine (Bienvenu et al., 2002). In addition, alleles with the length of 23 and 25 repeats have been reported as well (Kato et al., 2007; Strømme et al., 2002).

STR mutations in FOXL2

A polyalanine expansion in the FOXL2 gene (MIM: 605597) was discovered by Nallathambi et al. (2007) in one Indian family diagnosed with the blepharophimosis, ptosis, and epicanthus inversus syndrome (BPES). The allele is highly conserved and has 14 repeats of alanine. They found a homozygous 5 repeat expansion in two individuals of the family that resulted in a stretch of 19 residues of alanine long, demonstrating recessive allele behaviour (Nallathambi et al., 2007). However, other research has shown patients with BPES who have heterozygous alanine expansions and also longer, such as alleles with 24 alanine repeats present (De Baere et al., 2001; De Baere et al., 2003).

STR mutations in HOXA13

A polyalanine repeat expansions in exon 1 of the HOXA13 gene (MIM: 142959) has been described, causing the hand-foot-genital syndrome (HFG) (Mortlock et al., 2000). In exon 1 there are three alanine tracts, with a length in healthy humans corresponding to either 14, 12 or 18 trinucleotide repeats (Mortlock et al., 2000).

In the first tract, an eight repeat expansion has been detected in affected individuals of one family who carry an allele with 22 polyalanines, but not in the unaffected daughter (Innis et al., 2004). Pathogenic phenotype has been shown in patients who had six additional alanines inserted into the second tract, resulting in an 18 repeat long allele (Frisen, 2003). Finally, allele expansions in the third, largest tract has been found to be expanded from 18 repeats to 24 (Utsch et al., 2002), 26 (Goodman et al., 2000) or even to 29 and 30 repeats of alanines (Innis et al., 2004).

STR mutations in HOXD13

It has been described by Albrecht et al. (2004) that expansions of seven repeats (an increase from 15 to 22 alanines) in the exon 1 of HOXD13 gene (MIM: 142989) resulted in synpolydactyly (SPD). They found that the expressed protein in cells which had small expansion (22 repeats) was localised in the nucleus. However, at 24 repeats, the protein was in the nucleus in only 50% of cases. At longer expansions (29–36 repeats), the protein was found in the cytoplasm in almost all cases. Therefore, they suggested that expansions of alanine above a threshold of 29 repeats resulted the expressed protein to aggregate and that prevents its entering to nucleus (Albrecht et al., 2004). In another study on Chinese cohort, an eight-repeat expansion was observed in all affected individuals, which was not present in any of 50 controls (Xin et al., 2012).

STR mutations in PHOX2B

Polyalanine repeat expansions in the exon 3 of the PHOXB2 (MIM: 603851) has been shown to be the main cause of the central hypoventilation syndrome (CCHS) and pathogenic alleles range from 24–33 alanine repeats, whereas normal length of the allele is 20 repeats (Sivan et al., 2019).

STR mutations in RUNX2

Polyalanine expansions in the exon 3 of the RUNX2 gene (also known as CBFA1; MIM: 600211) causes cleidocranial dysplasia (CCD). In one study conducted by Mundlos et al. (1997) discovered a family where six members carried an expanded allele. The repeat size in these patients was 27, whereas the normal length is 17 alanine repeats (Mundlos et al., 1997).

STR mutations in SOX3

Polyalanine expansions in the SOX3 gene (MIM: 313430) have been found in one patient who was diagnosed with X-linked mental retardation with isolated growth hormone (XLMR) (Laumonnier et al., 2002). That patient had 11 additional alanines in the tract, resulting in 26 repeats long allele (Laumonnier et al., 2002).

In addition, insertion of seven alanines (resulting a 22 alanine allele) have been found in three brothers who were diagnosed with X-linked hypopituitarism, but did not have mental retardation (Woods et al., 2005).

STR mutations in TBX1

When Rauch et al. (2010) were analysing patients with tetralogy of fallot (TOF) heart disease, they discovered one individual with repeat expansion in the TBX gene (MIM: 602054). The allele was expanded from 15 to 25 repeats of alanine and none of the 185 controls nor patient's mother had the expansion (Rauch et al., 2010).

STR mutations in ZIC2

There has been only one report which demonstrates the pathogenic role of polyalanine repeat expansion in the exon 3 of the ZIC2 gene (MIM: 603073) (Brown et al., 2005). An expansion of 10 alanines was detected in one patient of ZIC2, resulting in an allele of 25 alanine residues and causing brain malformation, known as holoprosencephaly 5 (HPE), while the normal length was found to be 10 repeats (Brown et al., 2005).

4.2.3 Replaced and nested types of repeats

Replaced and/or nested type of repeats are those for which the pathogenic motif is not present in the reference genome and usually not found in healthy individuals. Out of the nine diseases discovered so far, only one (CANVAS) is caused by replaced type repeats, where a stretch made of repeat units with one sequence (such as AAAAG) is replaced with a sequence composed of another motif (such as AAGGG). All the other diseases are caused by a nested type of repeats where a stretch made of the pathogenic motif (such as TTTCA) is inserted between or next to the non-pathogenic stretch of standard repeats (such as TTTTA). All diseases with relevant information are presented in the Table 4.4.

Table 4.4. All known disease-causing STR loci made of nested and replaced types of repeats. The coordinates are in hg38 assembly and 0-based. Coordinates for the reference region denotes the STR location of endogenous repeats. The motif indicates the pathogenic motif with the non-pathogenic reference motif in parentheses. [§]The normal range is set to 0 because the pathogenic motif is usually not in the healthy population, however, heterozygous alleles with long expansions have been found in healthy individuals. [#]Size of the whole allele (including non-pathogenic repeats). Number of repeats in the pathogenic cut-off column for many diseases has been calculated from the allele size assuming they are composed solely of pentanucleotide repeats.

Gene	**Reference and location**	**Motif**	**Disease**	**Inheritance**	**Normal range**	**Intermediate range**	**Pathogenic cut-off**	**References**
BEAN1	chr16:66490398-66490466 (Intron)	TGGAA *(TAAAA)*	Spinocerebellar ataxia 31 (SCA31)	AD	0	—	⩾500[#]	(Sato et al., 2009)
DAB1	chr1:57367043-57367118 (Intron)	ATTTC *(ATTTT)*	Spinocerebellar ataxia 37 (SCA37)	AD	0	—	⩾31	(Corral-Juan et al., 2018; Seixas et al., 2017)
RFC1	chr4:39348424-39348485 (Intron)	AAGGG *(AAAAG)*	Cerebellar ataxia, neuropathy, and vestibular areflexia syndrome (CANVAS)	AR	0[§]	—	⩾400	(Cortese et al., 2020; Cortese et al., 2019)
SAMD12	chr8:118366816-118366914 (Intron)	TTTCA *(TTTTA)*	Familial adult myoclonic epilepsy 1 (FAME1)	AD	0	—	⩾100	(Cen et al., 2018; Ishiura et al., 2018; Lei et al., 2019)
STARD7	chr2:96197066-96197121 (Intron)	TTTCA *(TTTTA)*	Familial adult myoclonic epilepsy 2 (FAME2)	AD	0	—	⩾274	(Corbett et al., 2019)
MARCH6	chr5:10356347-10356407 (Intron)	TTTCA *(TTTTA)*	Familial adult myoclonic epilepsy 3 (FAME3)	AD	0	—	⩾668[#]	(Florian et al., 2019)
YEATS2	chr3:183712187-183712222 (Intron)	TTTCA *(TTTTA)*	Familial adult myoclonic epilepsy 4 (FAME4)	AD	0	—	⩾1000[#]	(Yeetong et al., 2019)
TNRC6A	chr16:24613439-24613529 (Intron)	TTTCA *(TTTTA)*	Familial adult myoclonic epilepsy 6 (FAME6)	AD	0	—	⩾1100[#]	(Ishiura et al., 2018)
RAPGEF2	chr4:159342526-159342616 (Intron)	TTTCA *(TTTTA)*	Familial adult myoclonic epilepsy 7 (FAME7)	AD	0	—	⩾60	(Ishiura et al., 2018; Lei et al., 2019)

STR mutations in BEAN1

Spinocerebellar ataxia type 31 (SCA31) is caused by an insertion of a long stretch of TGGAA repeats in intron of BEAN1 gene (MIM: 612051), which was inserted as a complex before a stretch of TAAAA as found by Sato et al. (2009). In homozygous patients, the inserted sequence was determine to have the following confirmation: [TCAC-$(TGGAA)_{exp}(TAGAA)_{exp}$(TAAAA TAGAA)$_{exp}$]. In healthy controls, TGGAA repeats were not found. However, in two cases an expansion in the region was found, but which also did not include the pathogenic motif, having the following confirmation: [TCAC-$(TAGAA)_{exp}$(TAAAA TAGAA)$_{exp}$]. The size of the inserted sequence complex was found to be between 2.5-3.8 kb, corresponding to approximately 500 and 760 pentanucleotide repeats (Sato et al., 2009).

STR mutations in DAB1

Several studies have shown the pathogenic role of inserted ATTTC repeats into the intron of DAB1 gene (MIM: 603448), causing spinocerebellar ataxia type 37 (SCA37) (Corral-Juan et al., 2018; Seixas et al., 2017). Affected individuals have shown to have insertions of 31–75 ATTTC repeats in the DAB1 gene (Corral-Juan et al., 2018; Seixas et al., 2017). Seixas et al. (2017) determined that the pathogenic repeats were always flanked on both ends by at least 58 of ATTTT repeats. The ATTTT/AAAAT repeats that are present in normal alleles were found to be in the range of 7–400 repeats (mostly under 30) in the healthy controls, but without the presence of any ATTTC repeats, which only associates the latter repeat unit with the disease. The pathogenic confirmation of repeats was determined as the following: $[(ATTTT)_{60–79}(ATTTC)_{31–75}(ATTTT)_{58–90}]$ (Seixas et al., 2017).

STR mutations in RFC1

It has been recently found by Cortese and colleagues (2019, 2020) that patients with cerebellar ataxia, neuropathy, and vestibular areflexia syndrome (CANVAS) also carry a different repeat unit in their RFC1 gene (MIM: 102579). The reference region with 11 AAAAG repeats was replaced with biallelic variable number of AAGGG repeats. The number of repeats ranged from around 400 repeats to 2750 (Cortese et al., 2020; Cortese et al., 2019).

In the first article (Cortese et al., 2019), the locus in healthy controls was determined to be highly polymorphic. After analysing 608 controls, they determined that 75% of them

have the reference $(AAAAG)_{11}$ allele and 13% carried an expanded (15–200 repeats) version of the same repeat. In 8% of cases there was an expansion (400–1000 repeats) of another (AAAGG) motif. Finally, four individuals had expansions in the range of 40–1000 repeats of the pathogenic (AAGGG) motif, which, however, was in the heterozygous state and therefore not pathogenic. In addition, some healthy individuals also had biallelic expansions of different repeat unit combinations, either $(AAAAG)_{exp}/(AAAGG)_{exp}$, $(AAAAG)_{exp}/(AAGGG)_{exp}$ or $(AAAGG)_{exp}/(AAGGG)_{exp}$, demonstrating that only biallelic expansions of AAGGG are disease-causing (Cortese et al., 2019).

STR mutations in SAMD12

Insertions of pentanucleotide TTTCA repeats into the intron 4 of the SAMD12 gene (MIM: 618073) has been shown to be responsible for familial adult myoclonic epilepsy type 1 (FAME1) disease that has been recently described by a few different research groups. The first published study that reported pathogenic TTTCA expansions in FAME 1, 6 and 7 was conducted by Ishiura and colleagues (2018). They showed that the insertion of expanded TTTCA repeats were present in all patients in conjunction with TTTTA repeats. The range of expansions for TTTCA and TTTTA combined was estimated to be in the range of 440 to 3680 repeats. For example, one patient had 598 repeats of TTTTA and 458 repeats of TTTCA. In 82 patients the configuration of expansion was determined to be: $(TTTTA)_{exp}(TTTCA)_{exp}$ and in three patients from one family it was as follows: $(TTTTA)_{exp}(TTTCA)_{exp}(TTTTA)_{exp}$. They did not see any expansions of TTTCA repeats in controls, but in 5.9% of healthy individuals they observed expansions of TTTTA, suggesting that TTTTA does not contribute to the disease (Ishiura et al., 2018). Two other studies investigating Chinese cohorts has been published as well, which reported that the disease-causing repeat length can be as low as 100 of TTTCA (Cen et al., 2018; Lei et al., 2019).

STR mutations in STARD7

The expansion of inserted repeats of ATTTC in the intron 1 in the STARD7 gene (MIM: 616712) causes familial adult myoclonic epilepsy type 2 (FAME2) disease, as reported by Corbett et al. (2019). In affected individuals, a long stretch of AAATG in conjunction with AAAAT repeats was found in all cases. The pathogenic configuration was defined as: $(AAATG)_{exp}(AAAAT)_{exp}$. By using two different types of long-read sequencing, they

analysed the region in two patients. In one of them, 274 repeats of AAATG and 387 repeats of AAAAT was determined using one method whereas with the other method, they determined 345 repeats of AAATG and 390 repeats of AAAAT. The second patient had 558 AAATG and 340 AAAAT repeats. None of the control samples had any repeats of the pathogenic AAATG motif (Corbett et al., 2019).

STR mutations in MARCH6

Familial adult myoclonic epilepsy type 3 (FAME3) has been recently described by Florian et al. (2019) caused by TTTCA repeat expansions that has been inserted into the intron 1 of MARCH6 gene (MIM: 613297). By investigating four unrelated families, they determined the expansions size to be variable among patients and ranges between 3 and 14 kb, corresponding to approximately 668–2814 pentanucleotide repeats. The repeat configuration was determined to be as follows: $(TTTTA)_{exp}(TTTCA)_{exp}$. In their control group of 83 individuals, the endogenous repeat unit TTTTA was found to be in the range of 9–20 repeats and never expanded. No repeats of the pathogenic motif TTTCA was found in controls (Florian et al., 2019).

STR mutations in YEATS2

Expanded stretches of inserted TTTCA repeats in the intron 1 of YEATS2 gene (MIM: 613373) has been described by Yeetong et al. (2019), causing familial adult myoclonic epilepsy type 4 (FAME4) disease. They found this in one Thai family with 21 members (13 affected and 8 unaffected). In one affected member, 819 repeats of TTTTA in conjunction with 221 repeats of TTTCA was found. The allele size for all affected members ranged from 5 to 8 kb which corresponds to approximately 1000–1600 pentanucleotide repeats (including expansions of both repeat unit). Configuration of the mutant sequence was determined as follows: $(TTTTA)_{exp}(TTTCA)_{exp}$. No TTTCA repeats were found in the unaffected family members nor any of the 1116 control subjects. Moreover, four of the healthy control subjects had TTTTA expansions (in the range of 719–1219 repeats), collectively suggesting the pathogenic role of TTTCA repeats (Yeetong et al., 2019).

STR mutations in TNRC6A

Familial adult myoclonic epilepsy type 6 (FAME6) is was described by Ishiura and colleagues (2018), caused by insertions of expanded TTTCA repeats into the intron of the TNRC6A gene (MIM: 610739). They were able to find 5 affected individuals from

one family who carried the mutation. The expanded allele size was determined to be around 5.6 kb longer than the normal one, corresponding to approximately 1100 pentanucleotide repeats. The configuration of the repeat was determined to be the following: $(TTTTA)_{22}(TTTCA)_{exp}(TTTTA)_{exp}$. TTTCA expansions were not found in any of the 1000 control samples. The endogenous TTTTA repeat was expanded in nine healthy subjects, where in seven of them it was in the range of 40–120 repeats and in two individuals potentially even longer (Ishiura et al., 2018).

STR mutations in RAPGEF2

Familial adult myoclonic epilepsy type 7 (FAME7) is thought to be caused by expanded TTTCA repeats in the intron 14 of the RAPGEF2 gene (MIM: 609530) as reported by Ishiura et al. (2018) and Lei et al. (2019). The first study was conducted in one family with one member having the identified expansions. The affected individual had 11.1 kb longer allele compared to the normal one, of the unaffected sibling. The expanded allele, in the case of composing only pentanucleotide repeats, would correspond to over 2200 repeats. The configuration of the sequence in the affected individuals is as follows: $(TTTTA)_{exp}(TTTCA)_{exp}(TTTTA)_{exp}$. None of the 1000 control samples had TTTCA repeats in the locus. Five controls was found to have an expanded stretch of TTTTA repeats, at least in three controls it was in the range of around 60 to 600 repeats (Ishiura et al., 2018). The second published study found just 60 repeats of TTTCA in the RAPGEF2 gene in three affected individuals (Lei et al., 2019).

4.2.4 Summary of known disease-causing short tandem repeats

This data collectively shows that pathogenic short tandem repeats can be divided into three different types based on their sequence structure. The first type – "standard repeats", refers to the expansion of motifs with specific sequences. The second type – "imperfect GCN repeats", are motifs with varying sequence but all of which encode for alanine residues. The third type will be named here as "replaced/nested repeats", which includes the recently found FAMEs, CANVAS and also SCA type 31 and 37. These repeats have a pathogenic motif with a specific sequence. However, unlike the first two types, these repeats are not present in the reference genome and, according to the research, are not present in the healthy population except CANVAS where heterozygous expansions of pathogenic motifs are found to be non-disease causing.

4.2.5 Validation of genotyping of the new variant catalogue

4.2.5.1 Overview of the method

In this section we aimed to validate our created new variant catalogue. For this we chose the tool ExpansionHunter 4.0.1 (Dolzhenko et al., 2017), which was the latest version at the time of the analysis. The tool comes with a variant catalogue including genomic coordinates for numerous known pathogenic STRs genes. Comparison of the catalogue with our list of loci (Table 4.2–Table 4.4) illustrated that although it includes 23 disease associated loci, the other 32 loci in 29 genes are not present in the catalogue and hence, would not be genotyped using the provided catalogue. Therefore, to be able to genotype all known monogenic disease-causing pathogenic STR loci, we created a new catalogue based on our compiled lists (Table 4.2–Table 4.4). The new catalogue would allow genotyping of STRs in total of 55 loci in the 52 genes that are causing 58 diseases.

To validate the newly created variant catalogue and assess ExpansionHunter's accuracy of genotyping these different loci and repeat types across samples, we conducted a series of simulations. In our simulations, we validated all loci except one in ARX and two in HOXA13 genes. In these genes, expansions in multiple tracts are shown to be pathogenic, but they are very similar to each other (all imperfect GCN types of repeats and the length of a healthy allele and pathogenic cut-off length is similar). Therefore, we selected one tract in each of these genes (that had the longest pathogenic cut-off value) which we used in our validation.

Details of our extensive simulations can be found in the Methods section and are summarised here. First of all, we simulated heterozygous samples whereby the shorter allele was set to be 60 bp in length and varied the length of the longer allele in the range of 60–2100 bp by adding one motif length for each simulation. As all the diseases in our list are caused by either tri-, tetra-, penta- or hexanucleotide motifs, the number of repeats in the longer alleles ranged from 20, 15, 12 and 10 repeats to 700, 525, 420 and 350 repeats, respectively. We included 2 kb sequence from the reference genome upstream and downstream of the locus for the simulated sequence, resulting in samples of 4060–6100 bp long. We simulated 150 bp paired-end PCR-free reads with 450 bp fragment length and total average coverage of 50x by using ART simulation tool with Illumina HiSeqX profile.

Secondly, we simulated rarer, homozygous genotypes of the pathogenic motif with expansions in both of alleles ranging from 60 to 2100 bp. The simulation process was similar to the heterozygous samples. These simulations were conducted for each of the 52 genes/loci to examine genotyping accuracy with increasing length of repeats.

In the case of imperfect GCN repeats, we created alleles whereby the "N" in "GCN" sequence was replaced randomly with another nucleotide, resulting in stretches of alanines coded by different triplets. This simulates an extreme case where an allele is composed of different repeat units mixed together, whereas in the majority of real-life cases, the allele sequence is expected to have longer stretches of one repeat unit, mixed with interruptions and/or expansions of other motifs that could be easier to align and genotype by the tool.

For the replaced and nested type of repeats we simulated alleles based on the definitions in the literature to simulate real-life findings. For example, SCA31 pathogenic alleles were simulated with the following formula: $(TCAC)_1[TGGAA]_{exp}(TAGAA)_{exp}(TAAAATAGAA)_{exp}$ which creates one TCAC sequence before the pathogenic (marked in square brackets) TGGAA repeats that increases in length, followed by TAGAA and TAAAATAGAA repeats which expansion was randomly chosen for each sample separately. Other formulas for the expanded sequences are specified in the Table 4.9. See sections 4.4.1 to 4.4.3 in Methods for more details about the simulations.

Once all simulations were performed (73,100 samples), we aligned the simulated reads to the hg38 (GRCh38 patch release 12) reference genome and genotyped all heterozygous and homozygous samples with the latest release of ExpansionHunter (v4.0.1) at the time and using our compiled variant catalogue. Estimated genotypes were extracted out from the VCF files and compared to the true, simulated length of repeats. We calculated root mean square error (RMSE) for each locus across the different lengths of simulations to determine the accuracy and plotted combined results of simulations in Figure 4.1 (see the section 4.4.4 in Methods).

Results were observed at three different ranges considering whether the simulated repeat length is less than the read length, is between the read and the fragment length or is longer than the fragment length. As different types of reads (e.g. spanning reads, fully repeat

reads or reads with flanking sequence) are used in the likelihood model to determine a genotype, we expect to see differences in accuracy between these length ranges. We first look at the standard and imperfect GCN samples shorter than the read length and determine the performance of the tool on these samples before moving on to longer and replaced/nested types of repeats that potentially have lower accuracy.

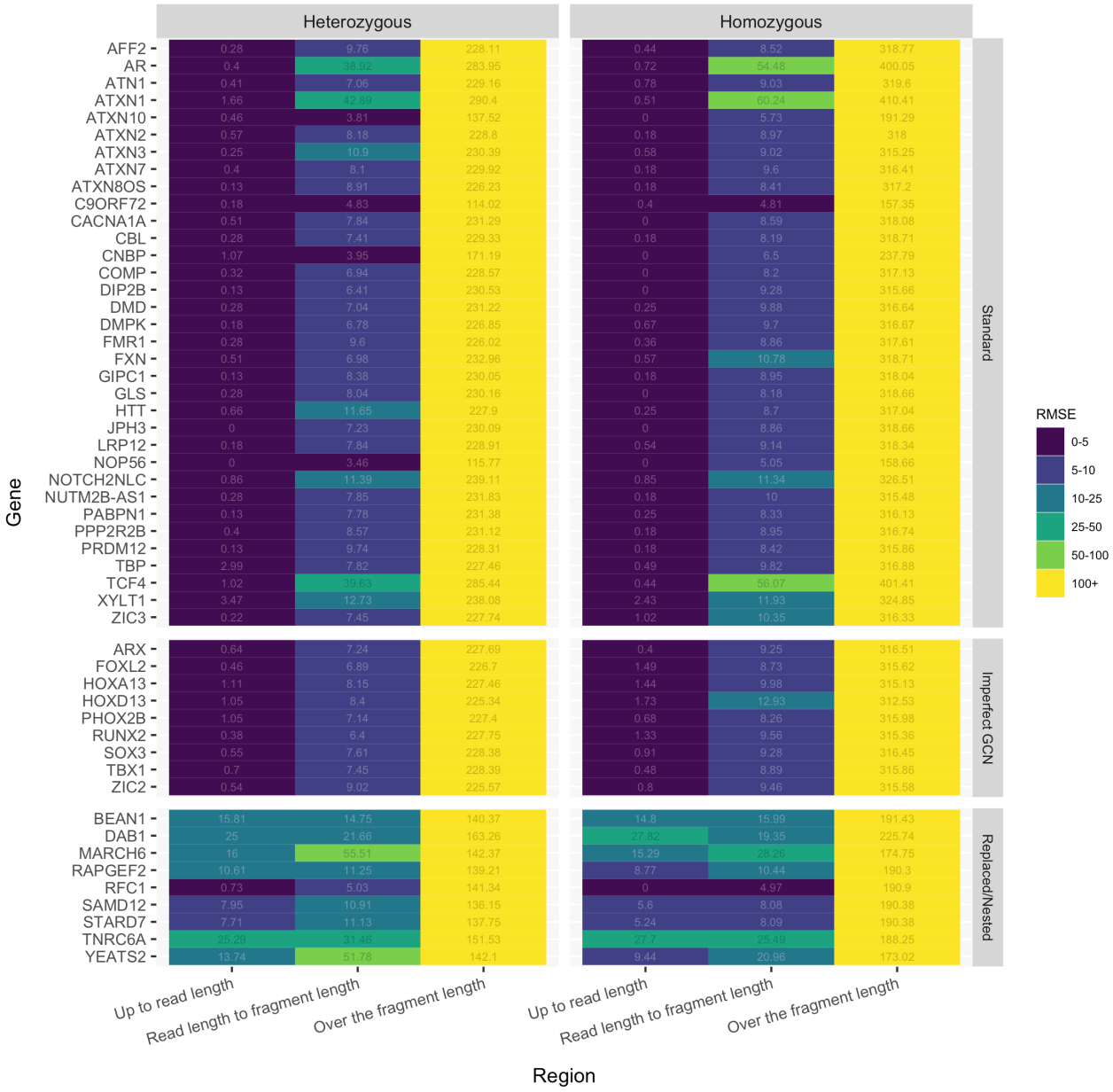


Figure 4.1. Genotyping accuracy for simulated heterozygous and homozygous samples of each gene. "Up to read length" represents samples containing repeats of 60 bp to 150 bp, "Read length to fragment length" are samples with repeats 151 bp to 450 bp in length, and "Over the fragment length" are all samples with repeat length over 450 bp. RMSE is divided into ranges that is used to colour each cell.

4.2.5.2 Genotyping samples up to the fragment length

Up to the read length (alleles in the range of 60–150 bp), both standard and imperfect GCN repeats have excellent accuracy where the RMSE in majority of cases is below 1, and are below 5 in all cases. Genotyping a longer region, from the read length to the mean size of the fragment length (151–450 bp) shows lower accuracy with RMSE below or close to 10 in most cases.

The slight drop in accuracy is expected as it is more difficult to estimate correct repeat length from reads that do not span the STR region. A closer examination revealed that the average estimated length of an allele is quite accurate up to the fragment size but is often limited to this. An example is seen in Figure 4.2 showing the genotyping results for DMPK gene where STR expansion causes the myotonic dystrophy type 1 disease.

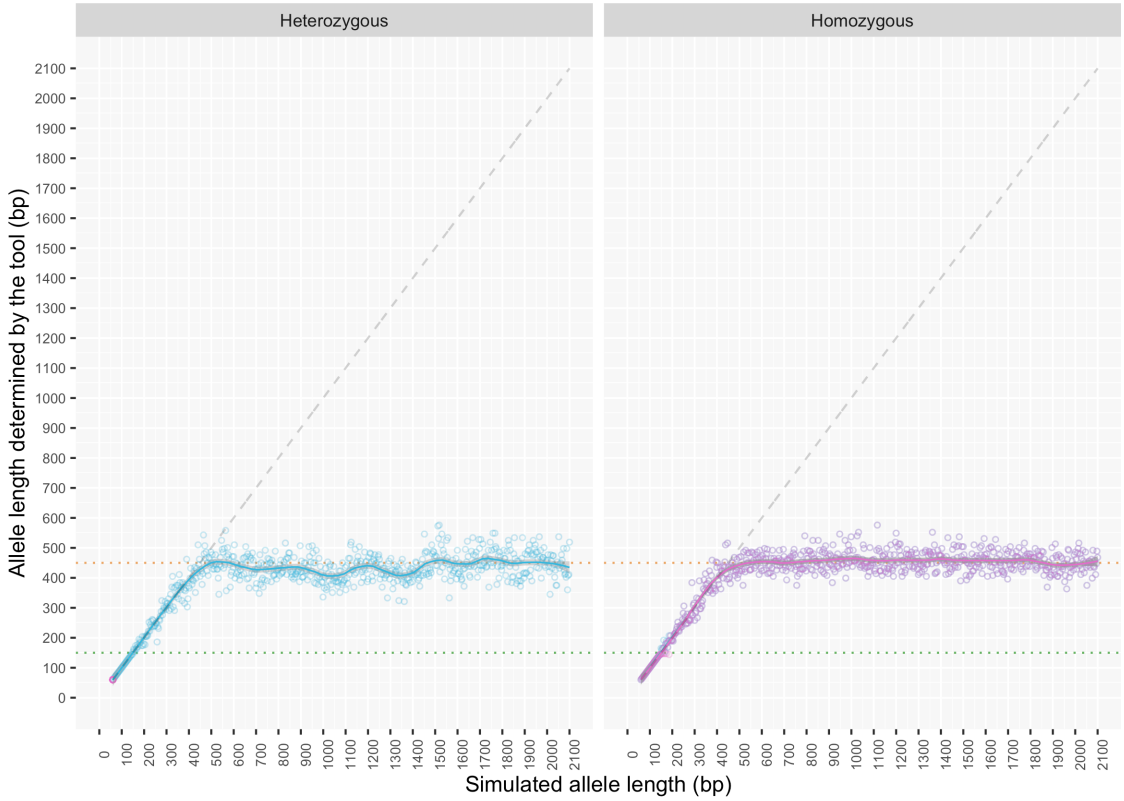


Figure 4.2. Genotyping results of increasing number of repeats in the locus in DMPK gene that is causing myotonic dystrophy type 1 disease which has good genotyping accuracy up to the fragment length (450 bp). Pink dots represent one allele, which for heterozygous samples is fixed to 60 bp or 20 repeats (bottom left dot in the heterozygous plot). Blue dots represent the other allele that has increasing number of repeats. Purple dots are overlapping pink and blue ones (as seen in the homozygous plot). The dashed line is true length of the longer allele. The dotted green line represents the read length and orange one, the average fragment length.

There are a few loci which have almost four times higher RMSE than others for lengths up to 450 bp (Figure 4.1). These loci are in the AR, ATXN1 and TCF4 genes, coding CAG/CTG repeats where expansions are known to cause SBMA, SCA1 and FECD3 disease, respectively. On closer inspection we discovered that the estimated genotype for all ATXN1 simulated samples was limited to the read length (150 bp). For AR and TCF4, the estimated length was slightly above the read length (Figure 4.3), but generally not over 200 bp for heterozygous or 250 bp for homozygous samples. In addition, homozygous samples of AR and TCF4 were genotyped as heterozygous when the allele length exceeded the read length.

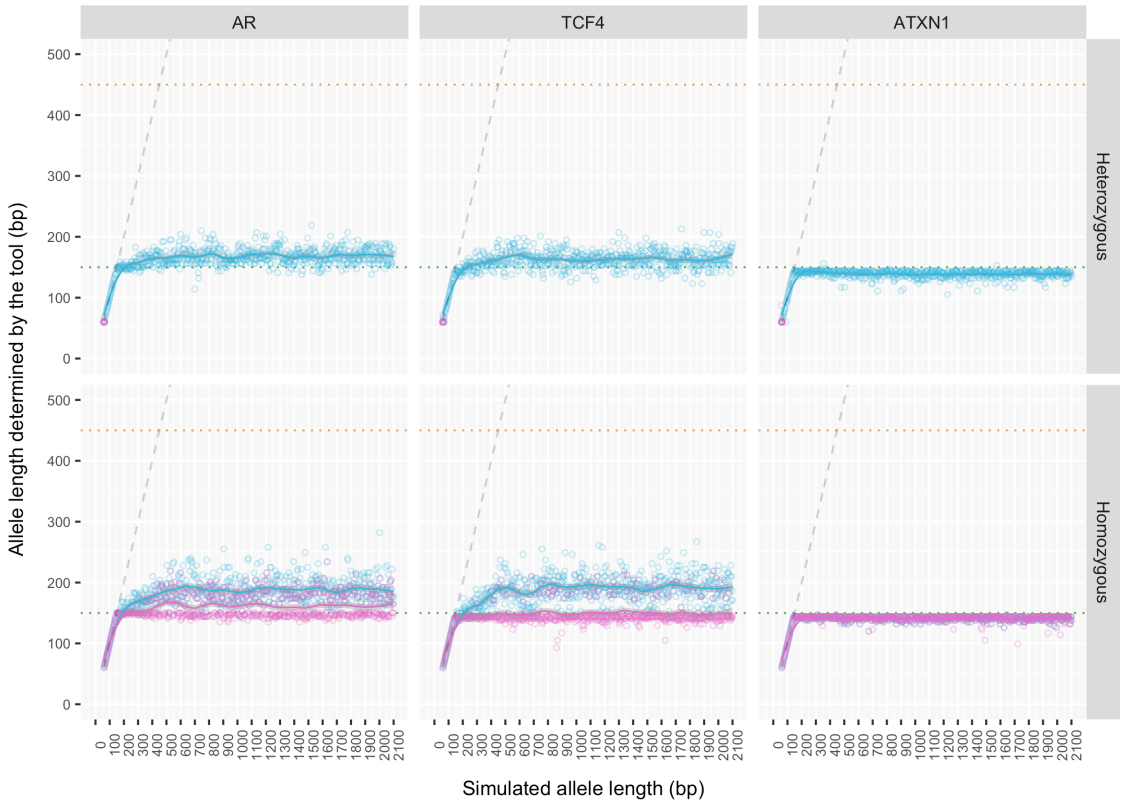


Figure 4.3. Genotyping results for three problematic loci (AR, TCF4 and ATXN1) that are made of standard type of repeats. These loci seem to have a maximum length coinciding with the read length (150 bp). Pink dots represent one allele, which for heterozygous samples is fixed to 60 bp or 20 repeats, whereas blue dots represent the other allele that has increasing number of repeats. Purple dots are overlapping pink and blue ones. The dashed line is the expected length of the longer allele. The dotted green line represents the read length and orange one the average fragment length.

This is an interesting finding and also a concern. To confirm that the issue does not lie in our variant catalogue, we used the catalogue which comes with ExpansionHunter that also includes all of these three loci. Identical results were obtained for AR and TCF4, whereas the problem with underestimating repeat lengths in both, homozygous and heterozygous samples in ATXN1 was not present anymore. After examining the

differences in the catalogues, we determined that the specified repeat unit in the ExpansionHunter's catalogue was "TGC" with reference region "chr6:16327633-16327723". In our catalogue, we used "CTG" repeat unit and "chr6:16327635-16327722" as the reference region. The difference is the order of motif's sequence and its coordinates. ATXN1 encodes for ataxin-1 protein on the minus strand that contains glutamine coded by CAG repeats where expansion of 39 or more repeats are pathogenic (Orr et al., 1993). As the motif and coordinates has to be adjusted for the plus strand, our specified CTG repeat unit is reverse complementary to the protein coding CAG and therefore it would be correct to use with the specified reference coordinates. This issue with the genotyping is further addressed below.

The problem with underestimating the genotype when it exceeds the read length for AR and TCF4 is concerning given that this tool with the provided catalogue is utilized by other researchers as well as clinicians which could potentially lead to false negative results. In particular, the pathogenic cut-off for repeated allele to cause SBMA (AR gene) is 40 CAG repeats and 50 repeats for the FECD3 caused by an expansion in the TCF4 gene (Table 4.2). This corresponds to 120 and 150 bp, respectively. By using sequencing data containing shorter reads, such as 100 bp (or even shorter for some older datasets), we may not be able to genotype these samples to detect repeats in the pathogenic range of FECD3 or SBMA.

To determine whether this could be possible, we simulated a set of hetero- and homozygous samples containing repeat expansions in AR and TCF4 genes similarly as previously described. However, this time we used Illumina HiSeq 2000 profile and simulated 100 bp long reads with 250 bp fragment size. We found that only 55.5% of heterozygous samples simulated to contain 40 or more repeats in the locus of the AR gene, were genotyped to be in the pathogenic range. Since SBMA is a recessive disorder, we determined that in the case of homozygous samples, only 48.4% of the time both of the alleles were genotyped to have at least 40 repeats (Figure 4.4).

Considering FECD3 in which pathogenic range starts from 50 repeats (150 bp), we found that none of the heterozygous samples simulated to be in the pathogenic range was determined to have at least 50 repeats. In case of homozygous samples, only 18.3% of the time one allele was determined to be in the pathogenic range and only 0.5% of the time both alleles exceeded 50 repeat length.

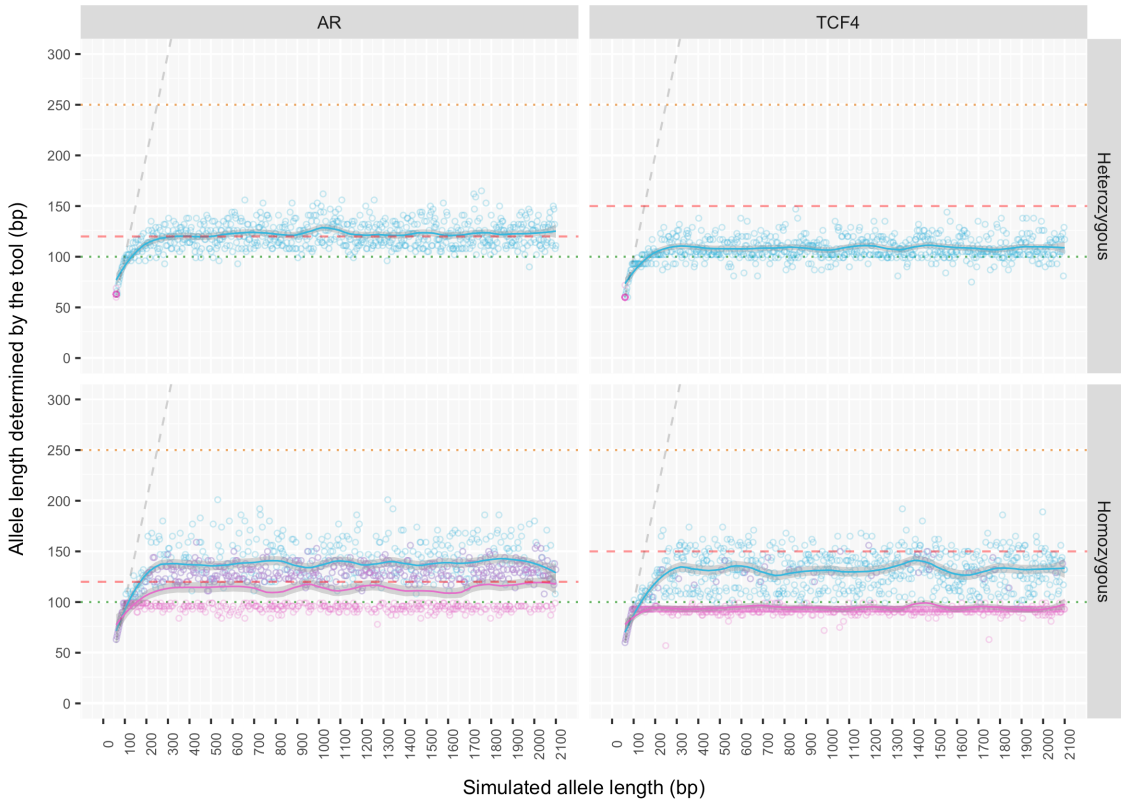


Figure 4.4. Results of genotyping AR and TCF4 loci with smaller length of reads and fragments. Pink dots represent the fixed, 60 bp or 20 repeats allele, whereas blue dots represent the other allele that has increasing number of repeats. Purple dots are overlapping pink and blue ones. The solid line is a linear regression line for each of the allele. The dotted green line represents the read length and orange one the average fragment length. Dashed red line is aligned to 120 bp for AR and 150 bp for TCF4 bp showing the minimum pathogenic range for these diseases.

This demonstrates the possibility of not being able to detect any pathogenic heterozygous FECD3 samples and roughly half of the SBMA ones by using sequencing data matching with our simulated parameters. As a result, we contacted the author of the tool to report our findings. As the author of ExpansionHunter explained (E. Dolzhenko, personal communication, November 23, 2020), the sizes of expanded AR, ATXN1 and TCF4 repeats were underestimated because the flanking sequence of each of these repeats contains another repeat with similar sequence composition. This has caused the sequence graph alignment algorithm to misalign some of the reads originating within expanded AR/TCF4 repeats. To address this issue, they have generalized the alignment algorithm to explore all non-unique kmer matches between reads and the corresponding graph. The updated algorithm generates alignments of reads against the target repeat and the flanking repeat and then reports the higher-scoring alignment. This led to improvements in ExpansionHunter and issues regarding underestimating alleles above read length in AR,

ATXN1 and TCF4 loci were fixed and the changes in the code were incorporated into the newer (4.0.2) version of the tool.

We confirmed the fix by running the updated version of ExpansionHunter on our simulated samples again. Both loci can now be genotyped until close to the fragment length (Figure 4.5), which is consistent with results of other loci (Figure 4.1). All samples which had repeats in the pathogenic range were now correctly genotyped. In addition, homozygous samples of AR are not determined as heterozygous anymore.

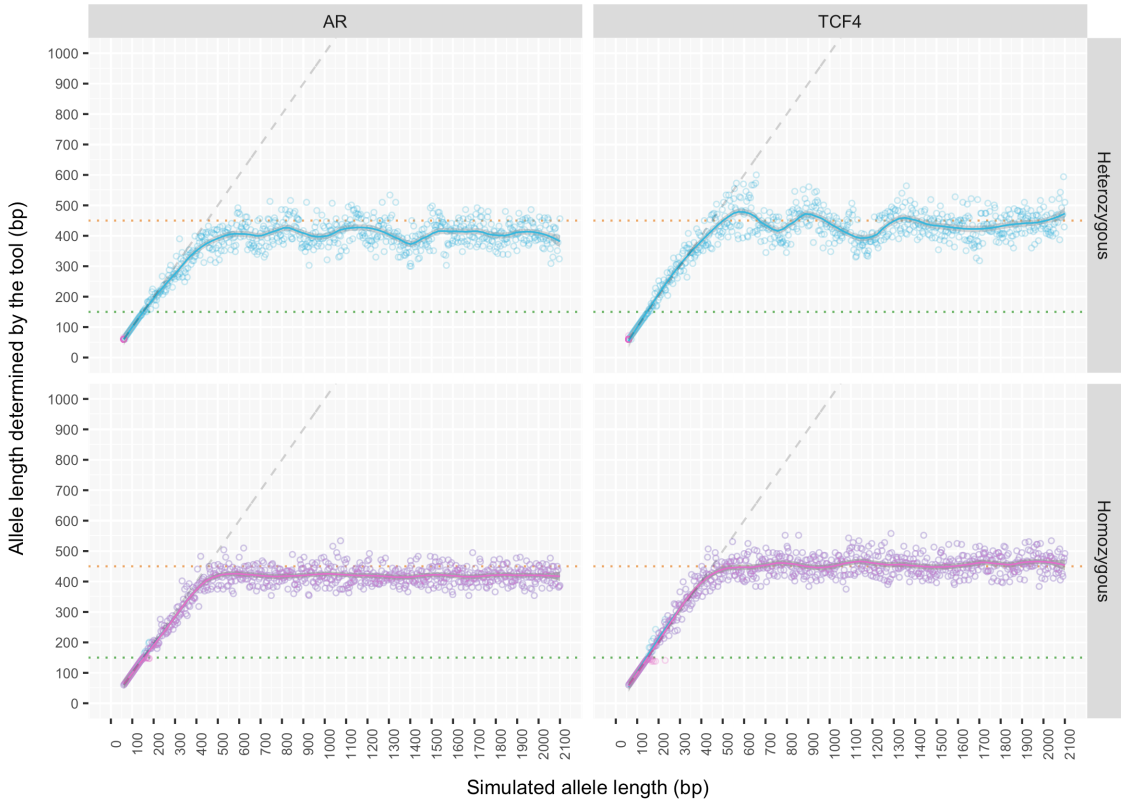


Figure 4.5. Results of genotyping SBMA (AR locus) and FECD3 (TCF4 locus) loci by fixed version (4.0.2) of the tool. Pink dots represent one allele, which for heterozygous samples is fixed to 60 bp or 20 repeats, whereas blue dots represent the other allele that has increasing number of repeats. Purple dots are overlapping pink and blue ones. The solid line is a linear regression line for each of the allele. The dotted green line represents the read length and the orange one the average fragment size.

4.2.5.3 Genotyping replaced and nested STR expansions

The third type of repeats, named as replaced and nested types, are difficult to genotype even in the range of read length. The best results are obtained for the RFC1 locus which is the only locus belonging to the replaced type of repeats. Expansions in the gene causes the CANVAS disease where the healthy repeat unit (AAAAG) is completely replaced by the pathogenic AAGGG motif, which is in contrast to the nested type of repeats where both, the endogenous and the pathogenic motif is present. Therefore, it is easier to

estimate a genotype for these samples when no other repeats in the locus are interfering. The RMSE up to read length for RFC1 is 0.73 and from read length to fragment size 5.03. In contrast, the average RMSE for nested types of repeats is significantly higher at 15.26 up to read length and 26.06 from read length to fragment length. Therefore, diseases caused by nested type of repeats, such as FAMEs, SCA 31 and 37 are the hardest to genotype and have the lowest accuracy. None of these loci are in the catalogue provided by ExpansionHunter and therefore genotyping this type of repeats has not been validated by the tool's authors. As we simulated variable number of repeats of the healthy motif before and, for some diseases, after the pathogenic stretch (depending on the repeat formula; see Table 4.9), we show an example on Figure 4.6 comparing one nested (FAME6, TNRC6A gene/locus) and the only replaced type (CANVAS, RFC1 gene/locus) repeats side-by-side.

For the TNRC6A locus, a repeat of the healthy (TTTTA) motif is attached before as well as after the pathogenic stretch of repeats, posing a difficult challenge for the tool in estimating the correct allele length. Despite the fact that we specified only the pathogenic repeat unit TTTCA in the variant catalogue, the tool is including the healthy repeats into the reported allele size and therefore the estimated repeat length diverges from the actual one even at short alleles. As seen on Figure 4.6, the lowest number of repeats detected in a pathogenic allele is over the read length (175 bp or 35 repeats), whereas it was simulated to be 12 repeats of TTTCA. In addition, as one allele in heterozygous samples has only the healthy motif and does not include the pathogenic one at all, we still observed expansions of the pathogenic motifs detected in the healthy allele as well.

That being said, we believe the estimation for this type of repeats can be improved. For example, as we know the sequence of the pathogenic and non-pathogenic motif(s), we can specifically look for the pathogenic motif and exclude the non-pathogenic ones from the genotyping likelihood models. A new feature for ExpansionHunter could be built that would allow to specify the non-pathogenic sequences which the tool can take into account. Therefore, these findings including our recommendations were also forwarded to the author of the tool. The feature to accurately genotype nested type of repeats is yet to be developed.

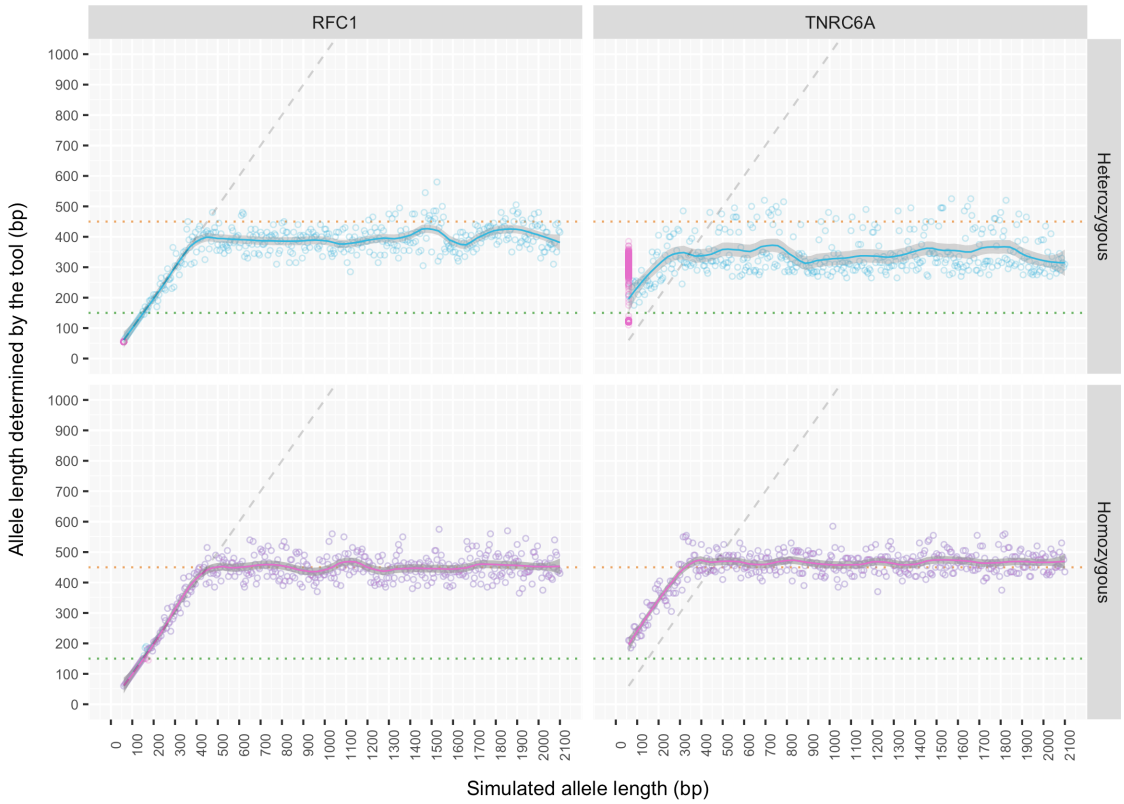


Figure 4.6. Results of genotyping CANVAS (RFC1 locus) and FAME6 (TNRC6A locus) loci representing replaced and nested types of repeats. Pink dots represent one allele, which for heterozygous samples is fixed to 60 bp or 12 repeats of non-pathogenic motif, whereas blue dots represent the other allele that has increasing number of pathogenic repeats. Purple dots are overlapping pink and blue ones. The solid line is a linear regression line for each of the allele. The dotted green line represents the read length and orange one the average fragment length.

In summary, the highest genotyping accuracy for nearly all loci (with the exception of some replaced/nested type ones) is seen in repeats with length lower than the read length. In such cases, the spanning class of reads are used to estimate the genotype. These reads contain the whole STR region as well as flanking sequence which helps align reads to the correct position and provide confident genotyping results. As the repeat size increases over the read length, then flanking and potentially in-repeat reads are used by the algorithm to determine the likelihood of a genotype. Flanking reads are type of reads that contain the STR region and also a flanking end of the non-repeated sequence, whereas in-repeat reads are fully made of repeated sequence. In-repeat reads are also being used when the repeat size exceeds the fragment length. We found samples with repeats longer than the fragment length had very high RMSE (Figure 4.1) and the estimated allele length was usually limited to the fragment length (a typical example seen in Figure 4.2).

4.2.5.4 Genotyping alleles longer than the fragment length

The inability to genotype samples over the fragment length could be due to the loss of in-repeat reads assigned to the locus. We hypothesised that a significant portion of in-repeat reads are being mis-mapped to other genomic loci where a longer stretch of these repeats are present. Taking this into account, we decided to find all similar regions to the pathogenic one in the genome where in-repeat reads could mis-align. These regions can then be specified as off-target regions for the STR locus in the variant catalogue which would allow the tool to use misaligned reads.

To define off-target regions for all standard and replaced/nested types of repeats, we used a data driven approach on a new set of simulated samples. In particular, only a long stretch (2100 bp) of repeated sequence for each locus for the aforementioned type of repeats was created. We simulated the samples having a high coverage (100x) and slightly increased the sequencing error rate, aiming to produce a large number of reads with minor differences in their sequence. We next aligned these reads back to the genome and determined all regions where these reads were aligned, merging the ones that were close to each other. Then, we filtered out the regions with low coverage and used the remaining region to specify off-target regions for each locus in our variant catalogue. We were able to do this for all STRs except the imperfect GCNs type of repeats due to the fact that there are a massive number of loci in the genome where these sequences with variable motif composition align.

We used the updated catalogue with off-target regions specified and genotyped all our previously generated samples again with the updated version of ExpansionHunter (4.0.2). Since we were using the fixed version of a tool, we also genotyped all samples by using a catalogue without off-target regions in order to accurately compare results in the AR, ATXN1 and TCF4 genes that was not genotyped correctly with the previous (4.0.1) version. The genotyping results after specifying off-target regions are shown in Figure 4.7, compared side-by-side with the results obtained using the previous catalogue where off-target regions were not included.

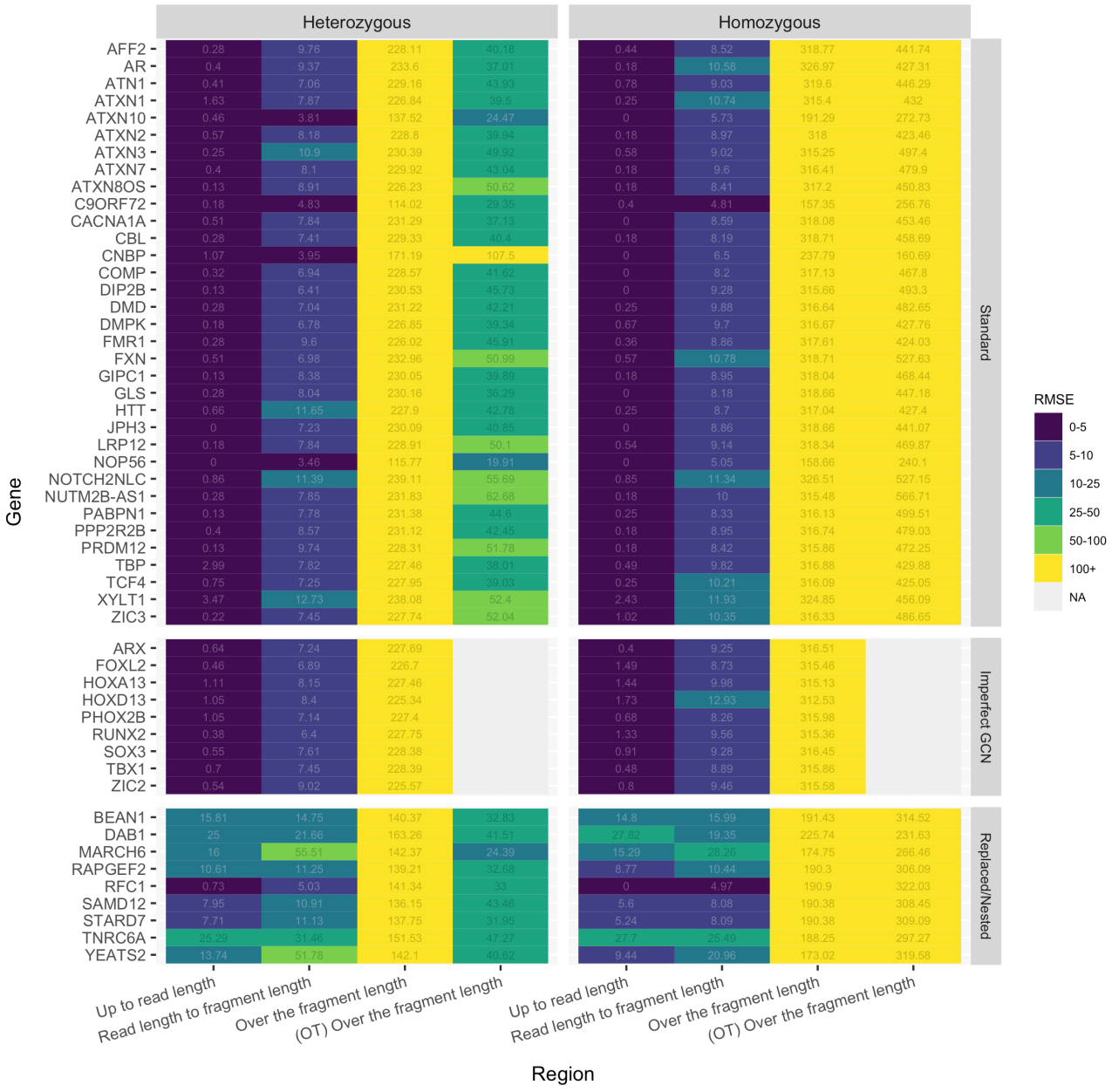


Figure 4.7. Genotyping accuracy for simulated heterozygous and homozygous samples of each gene including results after specifying off-target regions. "Up to read length" represents samples which the length of repeats is from 60 bp to 150 bp, "Read length to fragment length" is from 151 bp to 450 bp and "Over the fragment length" are all samples where the repeat were over 450 bp. RMSE is divided into ranges that is used to colour each cell. (OT) represents the results obtained with the catalogue containing off-target regions. Off-target regions were not specified for imperfect GCN repeats which are therefore missing from the last column.

It is evident that specifying off-target regions has a significant positive effect on genotyping accuracy for all heterozygous samples, bringing the RMSE down from over 200 to less than 50 in numerous cases (Figure 4.7). The lowest effect was observed for the CNBP locus (DM2 disease) whose allele size is underestimated above the fragment length (Figure 4.8) and it remains difficult to genotype accurately even after specifying

off-target regions. Figure 4.8 also shows results for another (DMPK) locus which represents the typical results obtained for the majority of other loci. For comparison, results of both locus when no off-target regions were used were plotted on the same figure in grey.

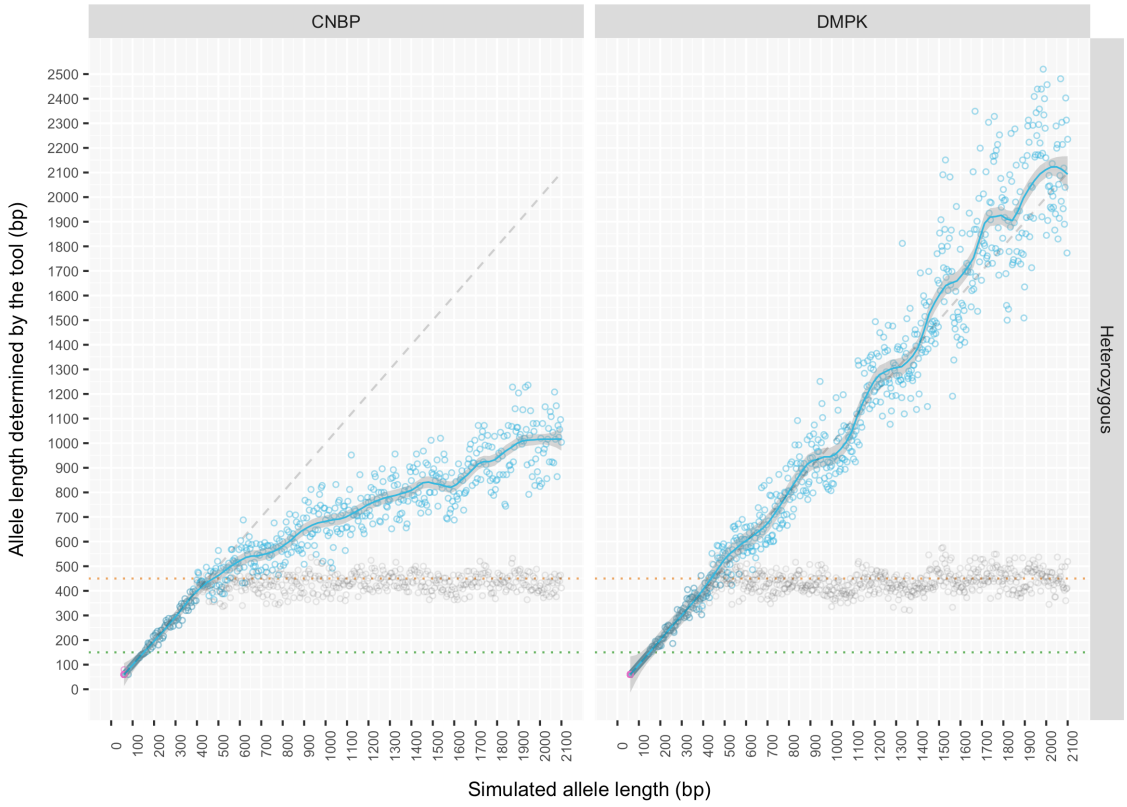


Figure 4.8. Genotyping results of heterozygous samples when using off-target regions in the CNBP and DMPK locus. Pink dots represent one allele, which for heterozygous samples is fixed to 60 bp, whereas blue dots represent the other allele that has increasing number of repeats. Grey dots show the results without using off-target regions. The solid line is a linear regression line for each of the allele. The dotted green line represents the read length and orange one the average fragment length.

Surprisingly, the effect of using off-target regions had a negative effect on the accuracy of genotyping all homozygous locus and increased RMSE by roughly a hundred, except for the CNBP. As this was unexpected, we investigated homozygous samples in more details and found that they all follow the same pattern. Specifically, genotyping estimation over the fragment length is good until to a certain point, until around 500–800 bp for the majority of cases. Above that, the tool starts to estimate homozygous alleles as heterozygous ones which diverge from each other in opposite direction as the repeat length increases. Ultimately that results in genotypes that can be even thousands of base pairs off from the actual length. As an example, results of genotyping homozygous samples having expansions in the CNBP and DMPK loci are shown in Figure 4.9. The

results for the DMPK locus are similar to the majority of other standard and replaced/nested type of homozygous loci genotyped by using off-target regions. The increase in accuracy for samples of CNBP was due to the fact that the genotyping error is not so significant for this locus and alleles diverge less from the truth.

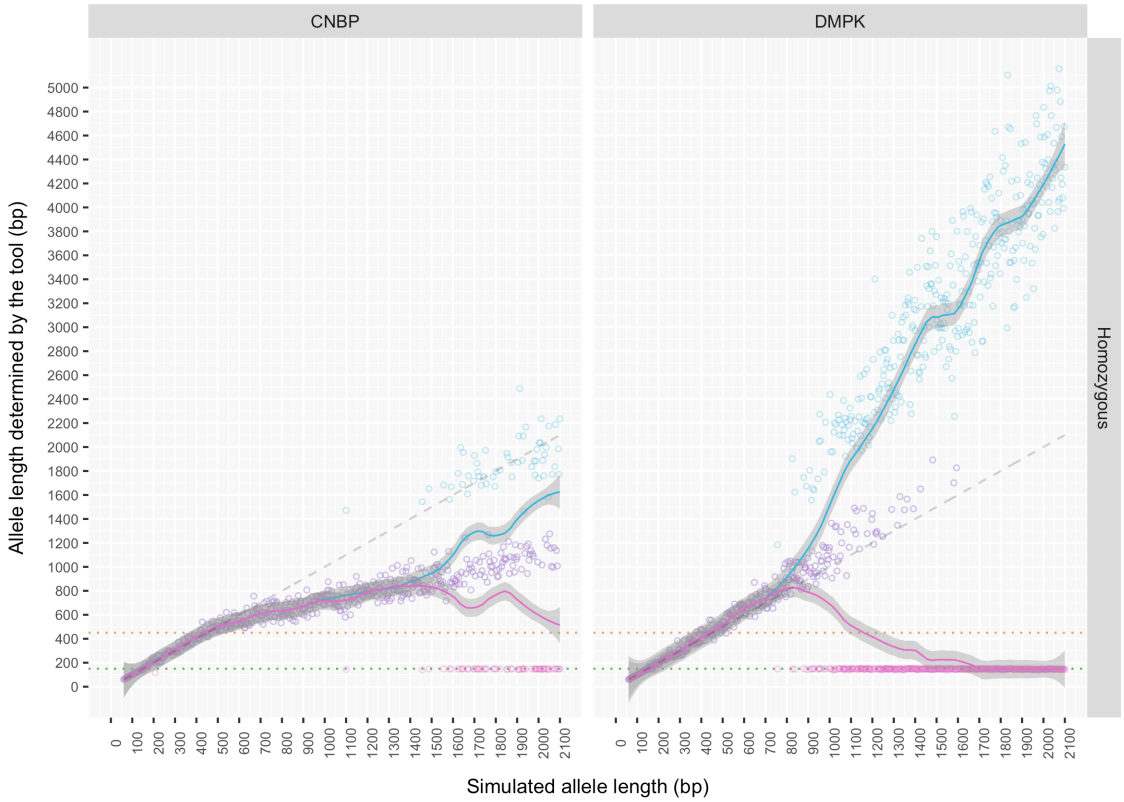


Figure 4.9. Genotyping results of homozygous samples when using off-target regions in the CNBP and DMPK locus. Pink dots represent one allele, whereas blue dots represent the other allele that has increasing number of repeats. Purple dots are overlapping pink and blue ones. The solid line is a linear regression line for each of the allele. The dotted green line represents the read length and orange one the average fragment length.

These findings were reported to the author of the tool as well, which led to the development of an improved algorithm that would enable to genotype homozygous alleles in better accuracy. The updated algorithm is yet to be released and therefore the validation of a new version of ExpansionHunter could not be done at this point.

4.2.5.5 Ability to genotype STRs into all pathogenic ranges

The pathogenic range for some diseases can be very high (such as 800 pentanucleotide repeats in SCA10, 650 hexanucleotide repeats in SCA36 and so on) that is in the range of thousands of base pairs and exceeding the maximum length of our previously simulated samples. To confirm whether we can genotype all samples into the pathogenic range of any disease, we simulated seven samples for each disease where the pathogenic allele was

simulated to have repeats in seven lengths. The first sample was simulated to have repeat length matching the pathogenic minimum of a disease. The following five samples were all in 60 bp increasing length and the last sample had +300 bp in length. For example, the pathogenic samples of Huntington's disease had the following length: 35, 55, 75, 95, 115, 135 and 235 repeats, corresponding to a range from 105 to 705 bp. We simulated heterozygous samples for both dominant and recessive diseases due to the inability of ExpansionHunter to genotype long homozygous alleles as described previously. The results are shown in Figure 4.10 confirming that we can reach to the pathogenic range in all diseases, including the ones with pathogenic cut-offs of thousands of base pairs. Genotyping some loci that have high pathogenic cut-off could be inaccurate for several reasons discussed before and therefore may result in false negatives for these samples. On the other hand, the results will still be close to the pathogenic cut off and therefore will be likely to be followed up with a more precise assay along with indications of the clinical features.



Figure 4.10. Genotyping results of pathogenic samples of all diseases. Green colour shows healthy range of repeats, yellow intermediate and red pathogenic range associated with a disease. Green and orange dotted lines represent the read and fragment length, respectively. Dark red dotted line marks the upper limit of previously simulated samples (2100 bp). Each dot is a one simulated sample. X-axis is limited to 7500 bp.

4.2.5.6 Summary of improvements made to the pathogenic STR catalogue

This work has now enabled us to genotype samples with standard and replaced/nested type of repeats with alleles over the fragment length by using the new variant catalogue that includes off-target regions for each of these loci. This allows the determination of pathogenic alleles that could have very long STRs, such as FXS where the pathogenic range starts from ~200 repeats/600 bp and others with even longer alleles, including GD, OPML1, SCA type 10 and 36, which would not be possible to do without specifying off-target regions. We can also estimate the whole allele length for replaced and nested type of repeats with the caveat that currently we cannot be sure the expansion is of the pathogenic motif, but we expect this to be solved soon as well. Regarding the inability to genotype imperfect GCN repeats longer than the fragment length, it is not of practical

concern in known real life situations as the pathogenic range for all GCN type of repeats found so far has not exceeded 40 repeats or 120 bp. This is below modern read lengths of 150 bp and therefore these repeats can be genotyped with a very high accuracy.

Currently, the variant catalogue which comes with ExpansionHunter includes off-target regions for two loci (C9orf72 and FMR1). Our catalogue specifies off-target regions for 44 loci (including all standard and replaced/nested type of repeats), and considering the low pathogenic cut-off for imperfect GCN repeats, our work allows accurate genotyping of all heterozygous pathogenic alleles. Once the tool receives a fix to genotype long biallelic expansions, all 58 diseases can be genotyped up to the known pathogenic range.

In order to standardise STRs genotyping in the scientific community and increase flexibility, we created an online tool for creating variant catalogues for ExpansionHunter. By using the locus specific information (such as the correct reference regions and repeat units) and also off-target regions determined in this chapter a user can create custom variant catalogues by selecting the targeted loci by hand or in groups (for example, selecting all loci in coding region, all loci causing neurological disorders, and so on). The catalogue creator is available at https://stripy.org/expansionhunter-catalog-creator and works with hg38/GRCh38 and hg19/GRCh37 genomes.

4.2.6 Effects of expansions in off-target regions on genotype estimation

One potential factor that might influence genotype estimation is an additional expansion in one or even many of the off-target regions specified. To examine effects of such expansions in off-target regions or other disease loci on a genotype estimation, we simulated a set of samples composed of two whole chromosomes in which where we created heterozygous expansions to simulate different scenarios. We choose ATN1, DMPK, ATXN2 and CACNA1A genes where mutations can cause DRPLA, DM1, SCA2 or SCA6 disease, respectively. ATN1 and DMPK are both in different chromosomes, but the STR locus is composed of the same motif (CAG). ATXN2 and CACNA1A are also located on different chromosomes and have the CTG repeat unit, which is reverse complementary to the former two loci. Potentially, reads from any of these loci can misalign to each of the other loci and affect genotype estimation. To test this, we firstly created a sample without any expansions in any loci as a control (Table 4.5). We used detected genotypes in the control sample as the true genotype for the following simulated samples.

Table 4.5. Information about control samples with their true and detected genotype by ExpansionHunter. True genotype was determined manually by counting repeats in the reference file. Detected genotype was determined by ExpansionHunter. CI – ExpansionHunter's confidence interval. OT – off-target region (not associated with any disease).

Sample	Disease	Gene	Motif	Reference Coordinates	True genotype	Detected genotype	Confidence interval
Control	DRPLA	ATN1	CAG	chr12:6936716-6936773	17/17	19/19	19-19/19-19
	DM1	DMPK	CAG	chr19:45770204-45770264	20/20	20/20	20-20/20-20
	SCA2	ATXN2	CTG	chr12:111598950-111599019	21/21	23/23	23/23-23/23
	SCA6	CACNA1A	CTG	chr19:13207858-13207897	13/13	13/13	13/13-13/13
	–	*OT-1*	CAG	chr12:113634296-113634326	10/10	–	–
	–	*OT-2*	CTG	chr19:46898776-46898812	12/12	–	–

Next, we simulated seven samples with different allele length configurations in ATN1 and DMPK gene and in one off-target region (Table 4.6). We simulated Sample 1 to determine whether an expansion in one of the off-target regions can affect the estimated genotype of a disease locus which is not expanded. We did not observe any effect of that. This is also expected as genotyping short alleles does not require the use of in-repeat reads and therefore off-target regions are not being used by the tool.

Secondly, we simulated a short expansion (65 repeats/195 bp) in one pathogenic locus which is between the read and fragment length. To determine the effect of expansion in one of the off-target regions we also created a 110 CAG repeat (330 bp) expansion in one sample. By comparing these two samples, we saw that when there was an expansion in the off-target region the genotype was detected to be 5 repeats longer than in a sample without the expansion in the off-target region. However, the correct length was within the confidence interval for both samples.

Thirdly, we repeated the previous simulation, but instead of a short expansion, a long one (215 repeats/645 bp) was created, which is well over the fragment length and where off-target regions will be used to search for in-repeat reads. We observed the same results as previously, whereby expansion in the off-target locus increases the estimation of a genotype. In this experiment, the estimated genotype was off from the actual by 24 repeats but the correct length was within the confidence intervals for both samples.

Finally, we created expansions in two pathogenic loci, which were also included in the off-target regions for both of the loci. We saw that an additional expansion can further affect the accuracy of genotype estimation, as seen on the Sample 7 in DMPK locus.

Table 4.6. True and detected genotype by ExpansionHunter for set of simulated samples with expansions in up to three loci. True genotype is the genotype detected in control group or the simulated one (in case of expansion). The red arrows are illustrating a change in the estimated genotype.

Sample	Gene	Expansion	True genotype	Detected genotype	Conf. interval
1	ATN1	–	19/19	19/19	19-19/19-19
	DMPK	–	20/20	20/20	20-20/20-20
	OT-1	Yes (long)	110/110	–	–
2	ATN1	Yes (short)	19/65	19/72	19-19/62-93
	DMPK	–	20/20	20/20	20-20/20-20
	OT-1	–	10/10	–	–
3	ATN1	Yes (short)	19/65	19/77↑	19-19/62-93
	DMPK	–	20/20	20/20	20-20/20-20
	OT-1	Yes (long)	110/110	–	–
4	ATN1	Yes (long)	19/215	19/219	19-19/175-263
	DMPK	–	20/20	20/20	20-20/20-20
	OT-1	–	10/10	–	–
5	ATN1	Yes (long)	19/215	19/243↑	19-19/203-319
	DMPK	–	20/20	20/20	20-20/20-20
	OT-1	Yes (long)	110/110	–	–
6	ATN1	Yes (long)	19/215	19/243↑	19-19/175-263
	DMPK	Yes (long)	20/117	20/179	20-20/152-230
	OT-1	–	10/10	–	–
7	ATN1	Yes (long)	19/215	19/243↑	19-19/203-319
	DMPK	Yes (long)	20/117	20/185↑↑	20-20/157-236
	OT-1	Yes (long)	110/110	–	–

To confirm whether using off-target regions containing expanded sequence which is reverse complementary to the pathogenic one, can also affect the estimated genotype, we generated two more samples. These samples had three more loci in addition to the ATN1, DMPK and OT-1 ones that had CAG expansions. The other three loci were ATXN2, CACNA1A and OT-2, which repeat unit is CTG. We are assuming that reverse complementary reads composed fully of repeated sequence in the first three loci can misalign to the latter three loci and *vice versa*. From Table 4.7, we can verify that expansions in off-target regions where the sequence is reverse complementary to the

reference one can further impact the estimated genotype. For example, Sample 8 had three expanded CAG loci and two expanded CTG loci, whereas Sample 9 had one expanded CTG loci in addition, which affected genotype estimation even further. However, this is an extreme case and in real life situation we will probably never see so many long expansions and in most cases even a double-expansion is likely a rarity.

Table 4.7. True and detected genotype by ExpansionHunter for set of simulated samples with expansions in up to six loci. True genotype is the genotype detected in control group or the simulated one (in case of expansion). The red arrows are illustrating a change in the estimated genotype.

Sample	Gene	Motif	Expansion	True genotype	Detected genotype	Confidence interval
8	ATN1	CAG	Yes (long)	19/215	19/252↑↑	19-19/190-291
	DMPK	CAG	Yes (long)	20/117	20/220↑↑↑	20-20/193-306
	OT-1	CAG	Yes (long)	110/110	–	–
	ATXN2	CTG	Yes (short)	23/73	23/175↑↑	23-23/146-227
	CACNA1A	CTG	–	13/13	13/13	13-13/13-13
	OT-2	CTG	Yes (long)	162/162	–	–
9	ATN1	CAG	Yes (long)	19/215	19/429↑↑↑↑	19-19/367-570
	DMPK	CAG	Yes (long)	20/117	20/381↑↑↑↑	20-20/332-503
	OT-1	CAG	Yes (long)	110/110	–	–
	ATXN2	CTG	Yes (short)	23/73	23/328↑↑↑↑	23-23/286-436
	CACNA1A	CTG	Yes (long)	13/250	13/456↑↑↑↑	13-13/352-527
	OT-2	CTG	Yes (long)	162/162	–	–

Taken together these results demonstrate that expansions in other regions can affect genotype estimation by ExpansionHunter. However, the expansions in other regions do not affect genotyping of normal length alleles and will not cause false positives according to this experiment if the allele is not expanded over the read length. When there are expansions in other loci besides the one being targeted, then the estimation of genotype can be affected, but not significantly. We also did experiments where we excluded other coordinates of other pathogenic loci from the off-target list (results not presented here), but we saw a significant decrease in genotyping accuracy above the fragment length, which indicates that many reads do misalign to other regions. Therefore, including all loci in the off-target regions will greatly improve genotyping accuracy for longer alleles. The potential caveat is that the detected genotype could deviate from the actual in the case of an additional long expansion in some of the off-target regions. However, that is probably a very rare event.

Interestingly, we also observed that estimated genotypes were affected when we removed the loci with expansions from the list of off-targets. This indicates that reads fully made of repeats can misalign from one region (such as from a random part of a genome) to another (such as the other off-target locus) and affect the genotype when the repeat length in the pathogenic locus exceeds the read length. In that case, removing these from off-target regions will not completely eliminate the occurrence of the event.

4.2.7 Genotyping PCR samples with ExpansionHunter

ExpansionHunter is a tool meant to work on PCR-free (PCR–) sequenced samples (Dolzhenko et al., 2017). However, as massive amounts of sequencing data have been generated using PCR (Fungtammasan et al., 2015) there is a demand to use the tool for genotyping such data as well. We aimed to examine ExpansionHunter's performance on a PCR+ dataset as well which has not been characterised well yet. To do this, we simulated another set of samples, similar to the previous ones, but now introducing PCR stutter noise into each of the samples (see section 4.4.2 in Methods describing creating stutter noise).

In this experiment, the maximum length of repeats was limited to half of the previous experiments (1050 bp) as simulating samples with stutter noise requires much more computational resources and as genotyping samples above the fragment size had a similar error rate from 1050 up to 2100. Instead of RMSE, we calculated the median absolute difference in repeats from the truth (simulated allele length) to provide additional information regarding how accurately we can estimate the size of STRs. As seen in Figure 4.11, there are only minor differences between PCR– and PCR+ samples in all sections of expansion size (up to read length, from read length to fragment length and over the fragment length) for standard types of repeats (long CNBP repeats being the only exception).

The median absolute difference from truth, for all PCR– and PCR+ samples up to read length having standard and imperfect GCN types of repeats, is 0 (no difference between samples) except XYLT1 which has 1 for PCR+ and 2.5 for PCR– samples. The detected genotype is the same as the simulated one for 94.5% of PCR+ standard and imperfect GCN types of repeats (including XYLT1) and 96.1% of these repeats in PCR– samples. For the replaced/nested type of repeats, the median difference is between 5 and 26 repeats with PCR-free and PCR-samples being very close to each other.

The median difference is higher for alleles that are longer than the read length and shorter than the average fragment length. For standard type of repeats, it is between 2 and 10 repeats (mean 6.2 repeats for both PCR– and PCR+), for imperfect GCN it is between 5 and 8 repeats (mean 6.1 for PCR– and 6.2 for PCR+) and for nested/replaced it is higher, between 3 and 15 repeats (mean 9.2 for PCR– and 8.9 for PCR+).

Repeat lengths over the fragment length have higher median difference which ranges from 9 to 48 (mean 20.5 for both PCR– and 21.1 for PCR+) in standard type of repeats, 90-109 in imperfect GCNs (mean 100.3 for both PCR– and 103.1 for PCR+), and finally, 12 to 27 in replaced/nested ones (mean 16.8 for PCR– and 17.5 for PCR+). Imperfect GCN repeats are difficult to genotype above the fragment limit because off-target regions cannot be specified (as discussed earlier). The largest differences between PCR– and PCR+ samples are above the fragment length, but few repeat differences at this length is also small enough not to have any concerns.

In conclusion, we demonstrated that the differences between PCR-free and PCR+ samples are mostly negligible. These results confirmed that ExpansionHunter will mostly estimate alleles correctly when reads with different length of repeats are present due to the PCR. However, it is important to note that in real-life samples PCR is also a source of other errors besides the stutter noise, such as the GC bias that can lead to low coverage regions and therefore affect the accuracy of genotyping these loci.

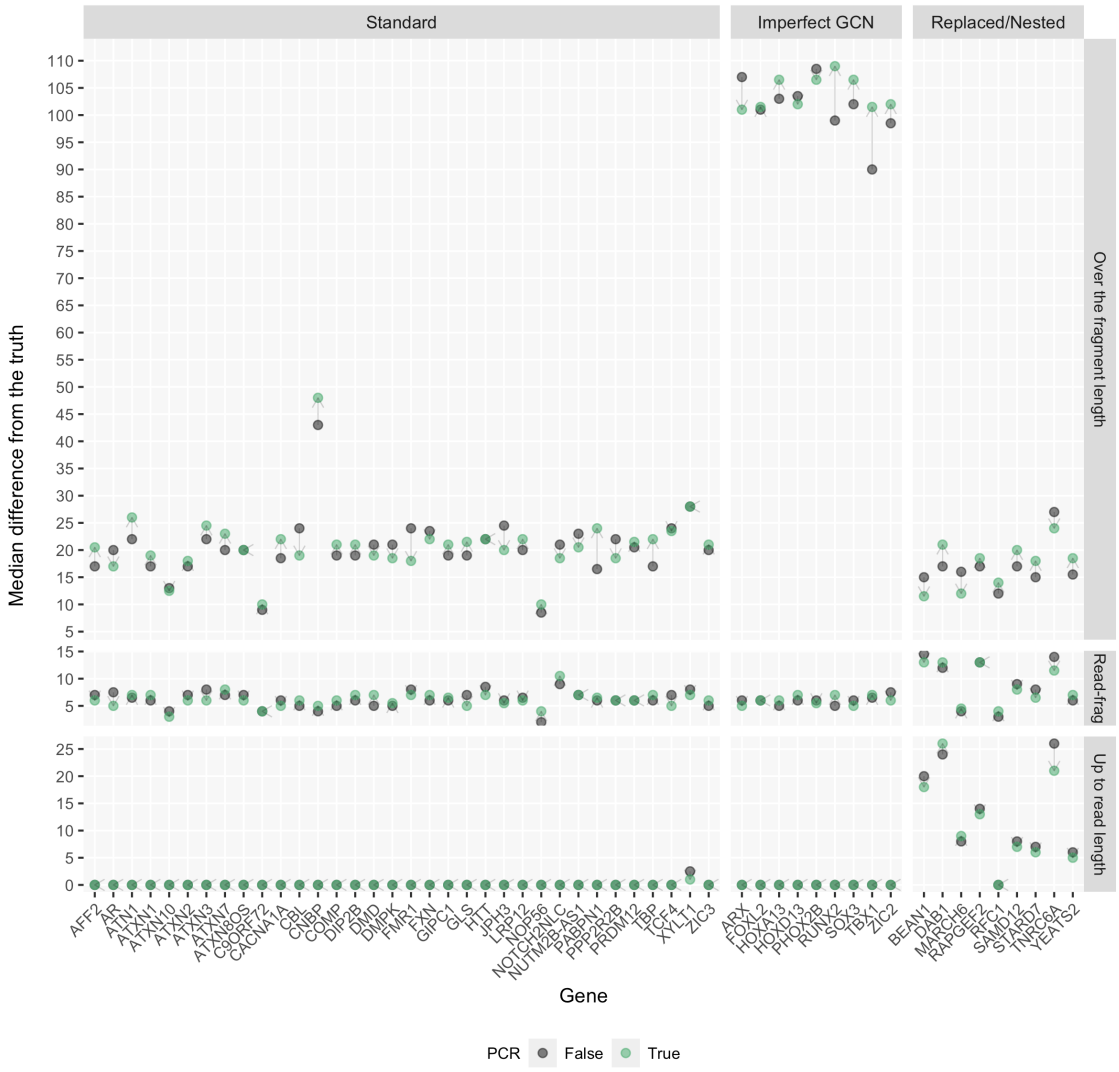


Figure 4.11. Comparison of simulated heterozygous PCR-free and PCR+ samples with stutter noise. "Up to read length" represents samples from 20 bp to 150 bp repeats, "Read-frag" samples from 151 to 450 bp and "Over the fragment length" samples from 451 to 1050 bp. Black colour represents simulated PCR-free samples and green colour simulated PCR+ samples with stutter noise.

4.2.8 Analysis of a real set of samples by using the new catalogue

We next applied our new STR catalogue to genotype a set of real rare disease cohort of undiagnosed patients provided by our collaborator. The dataset contains whole genome sequencing (WGS) samples from 39 families, including the affected individual and, in the majority of cases, the patient's parents as well. As STR diseases have not yet been analysed in this cohort, we analysed the dataset by running ExpansionHunter v4.0.2 with our catalogue that allows the detection of up to 58 diseases in WGS data (see section 4.4.6 in Methods).

We estimated several expansions longer than the normal range, including eight in intermediate ranges and nine with an estimated pathogenic length (Figure 4.12). Out of those nine, three of them were detected as heterozygous while the disease is recessive. We detected two alleles to be greater than the fragment length, one in the DMD gene and the other one in the GIPC1 gene, known to cause Duchenne muscular dystrophy and oculopharyngodistal myopathy type 2, respectively. As one patient's clinical features matched with a disease caused by an expansion in one of these genes, we reported our findings for further investigation. The other expansions were not a good fit for different reasons (such as phenotypic differences, age of onset, and so on). Interestingly, both DMD and GIPC1 were two of the new loci added to our catalogue which would not have been detected by using ExpansionHunter with the original provided catalogue.

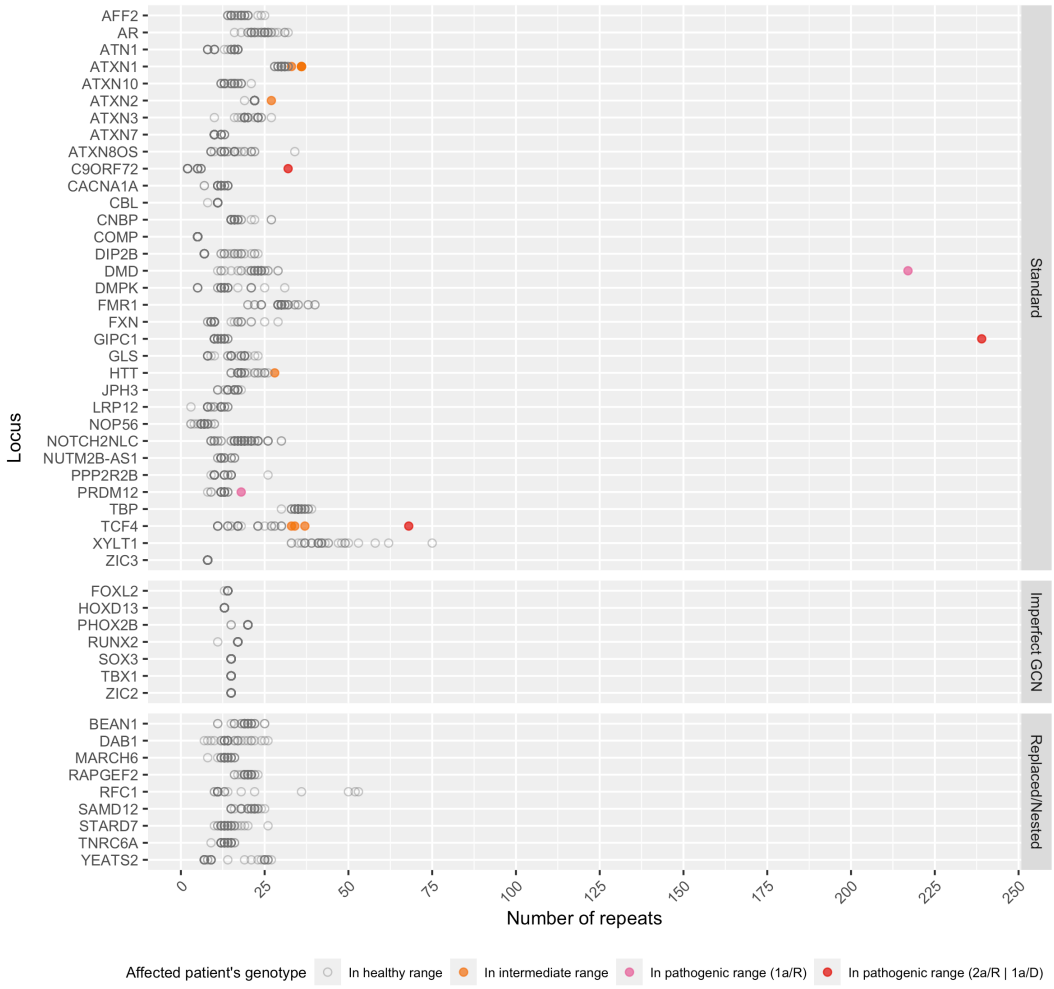


Figure 4.12. Genotyping results of WGS samples of 37 undiagnosed patients. Red dots represent samples where either two alleles were in the pathogenic range in the case of recessive disease or one allele in the case of a dominant disease. Pink dots represent samples where one allele was expanded, whereas the disease is thought to be recessive. Orange dots represent findings where the repeat length was in the intermediate range and grey ones are in the healthy range. In case of replaced/nested repeats, the results are for the whole allele, including the healthy motif as genotyping only the pathogenic repeats were not possible.

4.2.9 Development a new computational tool for STR analysis – STRipy

In this section, we introduce a new computational tool named STRipy that aims to make genotyping STRs an easy process and provide informative visual outputs. STRipy is designed for researchers or clinicians with no bioinformatics skills as STRipy requires no installation and has a graphical interface. Moreover, as genotyping can take place using cloud computing, no downloading of reference genomes is required or even knowledge of which genome assembly has been used for aligning a sample since this can be autodetected. STRipy will be a useful tool for STR researchers as well because it improves genotyping accuracy for some types of repeats and contains valuable data about populations and information acquired from the literature as described above. We have shown, in this chapter, ExpansionHunter is a robust and accurate method for STR analysis over different length of repeats. Therefore, we used ExpansionHunter as our genotyping method integrated into STRipy and developed extra functionality to improve genotyping and provide additional information regarding repeats (Figure 4.13).

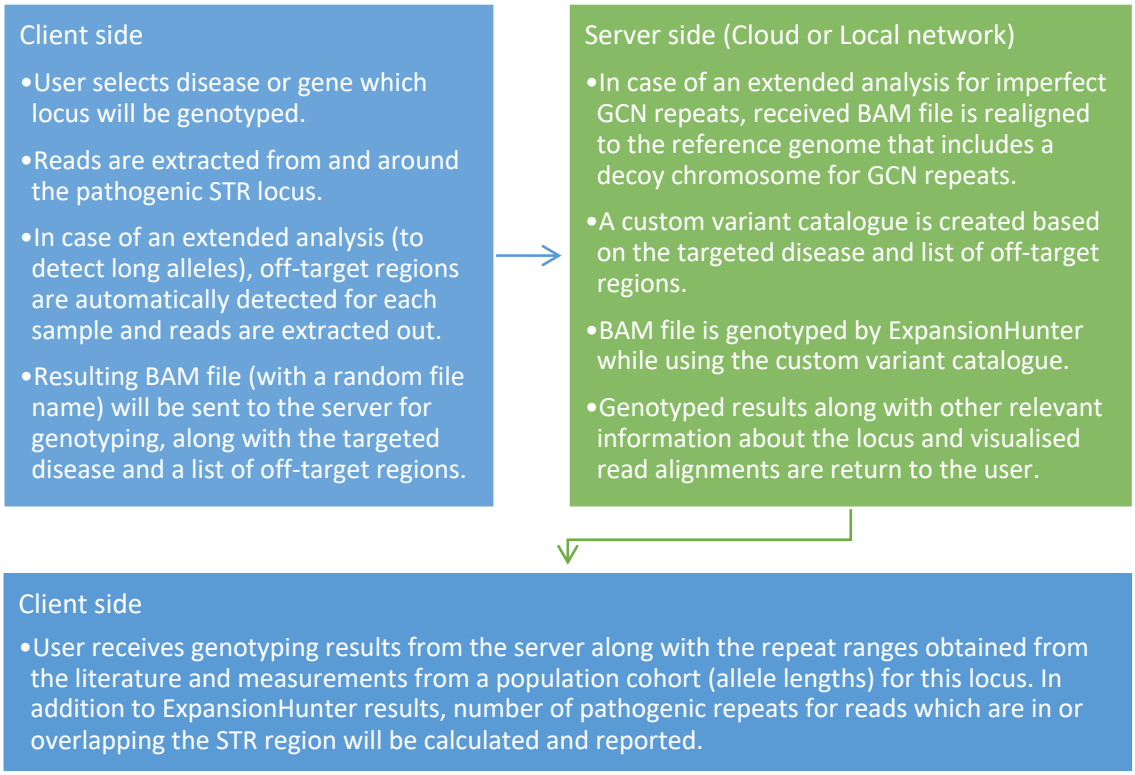


Figure 4.13. Overview of STRipy method describing the main steps. Server side can be either Cloud (external server set up by us to analyse samples) or Local (can be installed by the user into a local computer or network for internal analysis of samples.

One important feature of STRipy is the generation of a personalised, data driven definition of off-target regions that augments the variant catalogue on the fly for each sample individually, instead of using a general pre-defined catalogue. This allows the use of reads in regions where they are actually misaligned, as a pre-defined catalogue may not include all of them. Specifically, when an aligned paired-end file of a sample is loaded, all reads in the within or overlapping 2 kb upstream or downstream of the reference STR region are extracted from a BAM file. The fragment length and read lengths are calculated and used to determine whether the second read of a paired-end reads could be fully positioned in the STR locus. In that case, coordinates for all alterative mapping regions for these reads will be acquired and are used to define first level off-target regions. Secondly, we also determine regions where mates of reads that are aligned on the first level off-target regions are aligned, and extract reads from these regions as well (second level off-target regions). Based on our observations we believe that in the case of a long expansion, when the whole fragment is in the repeat region, both reads align to the genomic locus where there are the most repeats of that motif in the reference (that is, have better alignment). In that case, we found that capturing reads from the second-level off-target regions is beneficial to genotype long alleles. However, in case of detecting more than 50 off-target regions, the tool makes a compromise and does not include second level regions to avoid very long processing time. Finally, all regions overlapping or close to each other on the genome are merged (Figure 4.14) and a list of off-target regions is generated which is then incorporated into the personalised variant catalogue along with the coordinates for the targeted locus and repeat unit for a sample. The newly created BAM file composed of the extracted reads is genotyped by ExpansionHunter using the personalised variant catalogue and results are returned to the user via a graphical interface. The interface can compare the genotyping results with the population data of different superpopulations, visualise read alignments and download a PDF report. Read alignments are created with a separate tool named REViewer created by the authors of ExpansionHunter (https://github.com/Illumina/REViewer) and colourised by STRipy depending whether the allele is in normal, intermediate, unknown or pathogenic range. STRipy allows the genotyping of one locus at a time to avoid incidental findings.

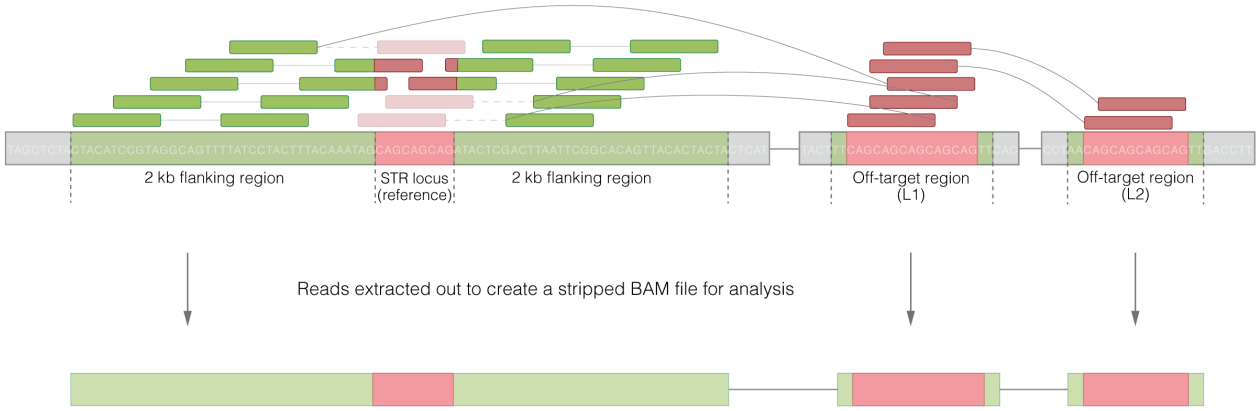


Figure 4.14. STRipy's method for defining off target regions and extracting reads. From the sequencing file, reads within or overlapping a 2 kb region of the STR locus will be extracted. In addition, the coordinates of misaligned mates of paired-end reads that are mapped to a reference STR locus will be determined and all reads from the first off-target region will be extracted (L1 regions). Similarly, coordinates for reads that are paired with reads mapping to the first level off-target regions are determined and extracted as well (L2 off-target regions). The final BAM file includes reads from the region surrounding the reference STR locus and all reads found from L1 and L2 off-target regions.

The second functionality developed for STRipy is improving the genotyping accuracy for imperfect GCN type repeats that are longer than the fragment length. As we demonstrated before, all imperfect GCN type of repeats can be genotyped into the pathogenic range as this is usually less than the read length. However, we incorporated a method in STRipy which can potentially genotype them over the fragment length in case of rare long expansions that could be present in the known loci or when new similar ones are being discovered that have longer repeats. To enable this, we created a 2 kb long decoy chromosome made of GCN repeats, which was added to the reference genome. During an analysis, in addition to the reads that are aligned on and near the reference STR locus and on other off-target regions, all unmapped reads will be examined as well. Unaligned reads made of GCN repeats will be extracted out and incorporated into a temporary BAM file which will be then realigned to the genome with the decoy sequence by using more relaxed parameters for alignment. Finally, the file will be genotyped with ExpansionHunter as described previously, with the addition of specifying the decoy chromosome as one of the off-target regions which would allow the genotyping tool to use long stretches of GCN repeats.

In STRipy a user has the option of performing either quick or extended analysis, meaning either excluding or including off-target regions (Figure 4.15). It is recommended to do the quick analysis first as this will significantly reduce the processing time. If the genotyping indicates an allele length close to or above the read length, a recommendation is given to run the sample again by using the extended option, which increases the

accuracy of estimating long alleles by incorporating reads from off-target regions, but also time of analysis. A genome assembly can be chosen between hg38/GRCh38, hg19/GRCh37 and autodetect. In the latter case, STRipy is using 20 bp pre-determined flanking sequence of a STR locus and looking for matches at specified coordinates for both genome assemblies to determine the one with a high match.

STRipy's output of genotyping results include information about the length of detected alleles and number of reads consistent with genotypes. In the case of an extended analysis and long alleles, STRipy also calculates independently the number of reads made solely of the pathogenic repeat unit (similar to ExpansionHunter's in-repeat reads, but only counting whole repeat units and reporting reads which are at least 90% made of repeats). This is done for reads that are mates of reads which are aligned next to the STR locus. In addition, the highest number of the pathogenic repeat units are calculated for flanking reads which are reported as well.

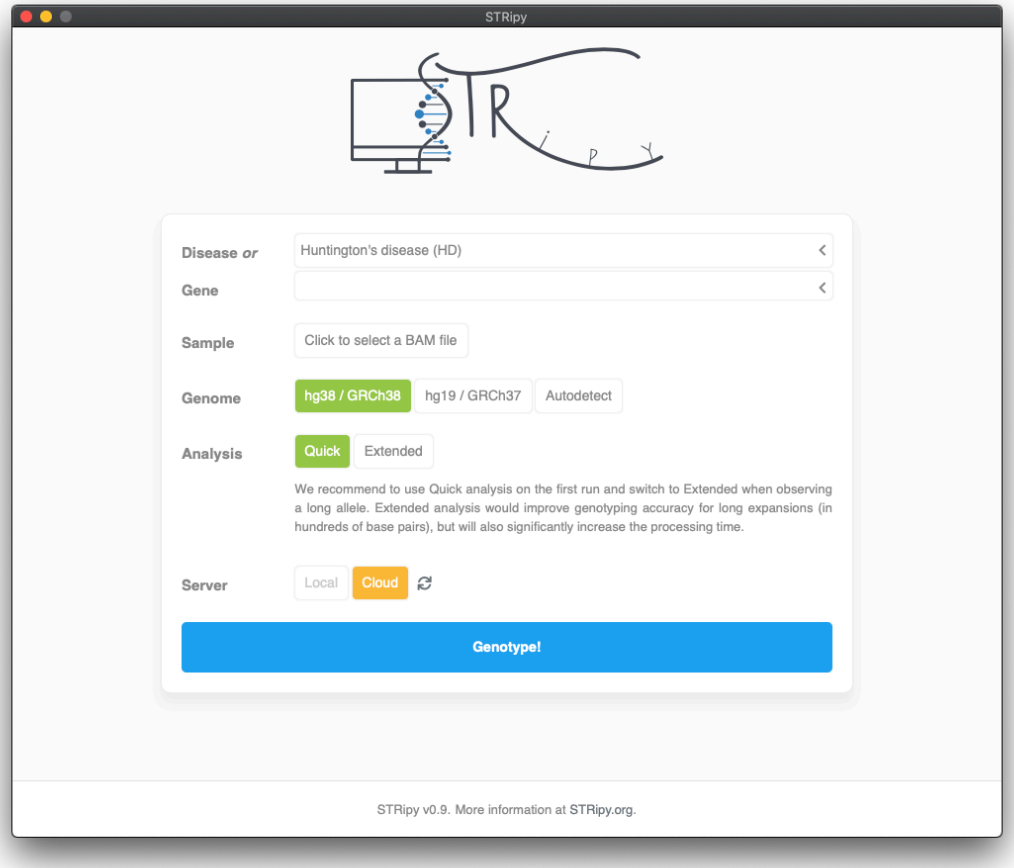


Figure 4.15. Screen capture of the front screen in STRipy v0.9. A user will have a choice to either select a disease or gene to genotype. A BAM file can be selected, followed by the genome assembly, the type of analysis and whether to use local or remote server (Cloud).

STRipy's feedback (Figure 4.16) includes information about the disease, such as coordinates for the STR locus, pathogenic repeat unit, gene and inheritance, as well as the normal, intermediate and pathogenic ranges of repeats based on the research articles published so far (section 4.2.1).

One important addition to complement this data is the population data we have generated for each pathogenic locus. Only a small number of research articles that reported findings of pathogenic STRs provided exact information regarding the distribution of allele lengths in healthy populations. Hence, we used ExpansionHunter to genotype the previously generated STR catalogue on 2504 samples from the DRAGEN reanalysis of 1000 Genomes project samples (see Methods section 4.4.8). This dataset includes samples in five super-population: African, European, South-Asian, East-Asian and Admixed-American. The allele distribution of genotypes estimated in this population data are displayed in the STRipy's output and can be viewed in combined form or separately for each super-population. Utilization of the relevant population data for user's sample helps to determine whether the genotyping results could be clinically important as the healthy range of repeats in disease loci could be population dependent.

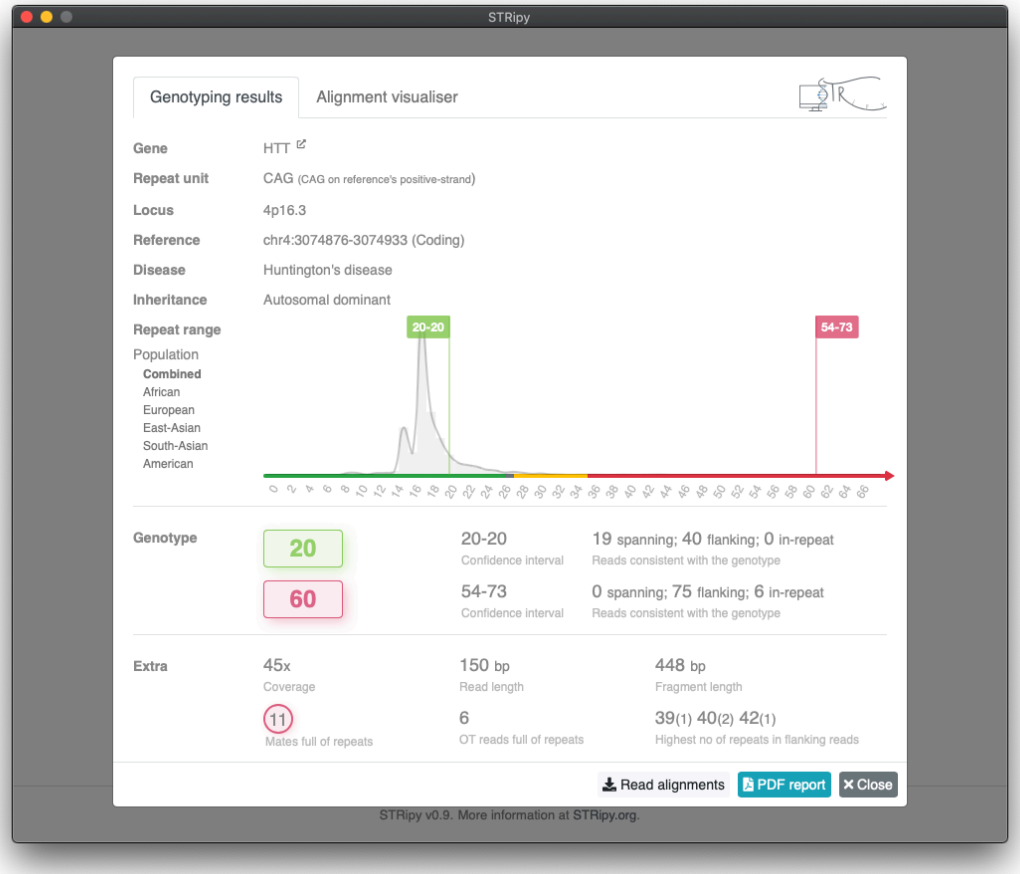


Figure 4.16. Screen capture of the results screen in STRipy. A user will get feedback regarding the gene and allele (disorder, type of inheritance, affected gene, pathogenic repeat unit and the reference coordinates for the STR locus). The repeat length ranges for healthy, intermediate and pathogenic repeats acquired from the literature is displayed graphically on the x-axis and the population data distribution as the combined line and bar chart on top of that. Genotyping results include the estimated repeat length for each allele including the confidence interval and reads consistent with the genotype. In addition, median coverage, read and fragment length of the analysed file will be calculated and displayed. The number of fully repeat reads found in mates and in off-target (OT) regions along with the highest number of repeats in flanking reads. Generated PDF report and read alignments can be downloaded as vector graphics.

The presence of reads that are fully comprised of repeats and are paired with the reads aligned next to the STR locus indicates the existence of a long allele. The information regarding pathogenic repeat units in mates and flanking reads is especially valuable when analysing replaced or nested types of repeats as ExpansionHunter itself cannot work properly with these repeats and does not distinguish normal and pathogenic repeats. STRipy looks only at the pathogenic repeat units which are present in mates and flanking reads in the potentially pathogenic locus. Moreover, the presence of pathogenic repeats can be also seen under the Alignment visualiser section when looking at the visualisation of read alignments created by a tool named REViewer (Figure 4.17).



Figure 4.17. Screen capture of the alignment visualiser screen in STRipy. In this example, a simulated heterozygous sample with 12 pathogenic repeats (TTTCA) inserted into one allele was analysed and reads visualised with REViewer, which are shown separately for both alleles. The reference sequence displayed on top of each allele's reads contain the pathogenic repeat unit. It can be seen for the allele on top that all spanning reads contain four TTTTA repeat units, which is the endogenous one. For the second allele, six repeats of TTTTA can be seen in the beginning of the STR locus, followed by twelve repeats of TTTCA, confirming the presence of pathogenic repeat units in one of the alleles.

STRipy is meant to serve as the first screening tool to quickly assess whether an STR expansion could be clinically important. As STRipy is using only a fraction of the genomic sequence (the areas most important for analysis), some information (reads) could be missed. Thus, running ExpansionHunter on a whole genome file may provide more accurate results as all reads in the file could be used. In case of an expansion, it is recommended to confirm the results by running ExpansionHunter with the catalogue made by using our online catalogue creator on the whole sample. With that being said, we used a subset of the simulated heterozygous samples in the 4.2.5 section to assess the accuracy of STRipy. These results were compared against the results obtained in section 4.2.5.4 which were genotyped using the full read set by ExpansionHunter v4.0.2 with our catalogue including pre-defined off-target regions. The same version of ExpansionHunter was also used by STRipy (without the pre-defined catalogue).

We determined the RMSE for samples in each locus (between 60 bp and 2100 bp simulated lengths for each locus, see section 4.4.7 in Methods) and plotted the results on Figure 4.18. In the standard types of repeats up to the read length, the median RMSE for ExpansionHunter + catalogue (EHC) was 0.28, whereas for STRipy, it improved to 0.25. In particular, RMSE for the two most inaccurate loci (XYLT1 and TBP) dropped from 3.47 to 0.32 in STRipy for XYLT1 and from 2.99 to 0.32 for TBP. The accuracy of genotyping imperfect GCN type of repeats in up to the read length section was higher for EHC with a median RMSE of 0.64, compared to 1.28 in STRipy. There were no differences in genotyping replaced/nested type of repeats up to the read length as both had median RMSE of 7.95.

The genotyping accuracy for alleles from the read length to the fragment length is very close in standard type of repeats, where the median RMSE for EHC is 7.76 and for STRipy it is 7.80. Again, EHC is performing better at genotyping imperfect GCN repeats with the RMSE 7.45, compared to the STRipy with median RMSE of 12.83. No differences in replaced/nested type of repeats was observed between methods as both had a median RMSE of 11.24.

The median RMSE in the standard type of repeats longer than the fragment length was also slightly better for STRipy, being 40.27, compared to the 42.33 for EHC. Despite having a minor difference in RMSE, there were two loci with noticeable differences. STRipy genotyped the samples having an expansion in the CNBP with significantly greater accuracy, having a RMSE only 24.42, compared to the EHC which had RMSE of 107.50. On the other hand, STRipy struggled to genotype XYLT1 locus and the RMSE was 138.28 for STRipy, while it was 52.40 for EHC. Many other loci were more accurately genotyped by STRipy (for example, FXN, LRP12, PRDM12 and ZIC3) and only a few were worse (for instance, COMP).

Regarding the imperfect GCN type of repeats, we demonstrate here the effectiveness of our method which results in average RMSE of 74.99, down from 227.46 that was obtained with EHC (where off-target regions were not be specified). The results for the replaced/nested type of repeats are nearly identical again (EHC_{RMSE}=33.00 and $STRipy_{RMSE}$=32.88).

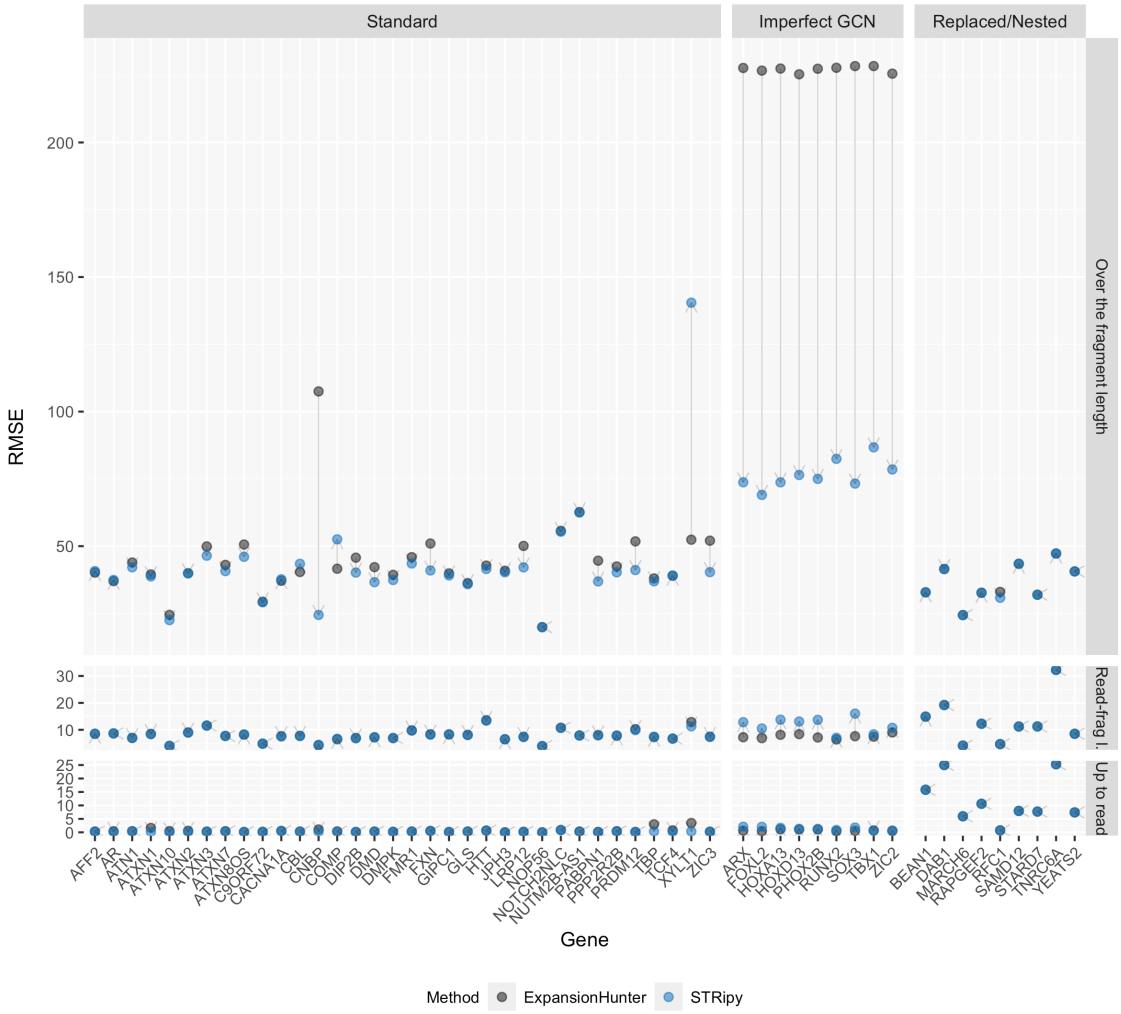


Figure 4.18. Comparison of accuracy by using STRipy and ExpansionHunter + catalogue method. Black dots represent the results obtained by using ExpansionHunter with the pre-defined catalogue and blue ones results from STRipy. The lines are connecting the results and arrows pointing to the change of the RMSE.

Finally, in addition to the accuracy, we also determined the sensitivity of detecting a pathogenic allele. We linked each disease with an STR and took the pathogenic allele cut-off length associated with that disease. In cases where one gene was responsible for multiple diseases (such as FMR1) we used the largest value among these diseases. Due to ExpansionHunter's issue with accurately genotyping long homozygous samples, we only used heterozygous simulations for all diseases. We calculated the number of samples which had the simulated allele length at least as long as the pathogenic value and compared it to the number of samples where the allele was genotyped into the pathogenic range. In other words, sensitivity is defined as the percentage of pathogenic simulations that were determined as pathogenic by the tool.

We included all STRs whose minimum pathogenic length cut-off was less than 1800 bp since the repeat length simulated in our samples was limited to 2100 bp and that would

allow us to use a minimum of 50 samples of hexanucleotide repeats or 100 trinucleotide ones for calculating sensitivity. As seen on Figure 4.19, the sensitivity for all these samples is over 97% regardless of the method. In four cases EHC method has a higher sensitivity. These results demonstrate that both EHC and STRipy can be used to detect samples that have pathogenic ranges of repeats at least until 1800 bp.

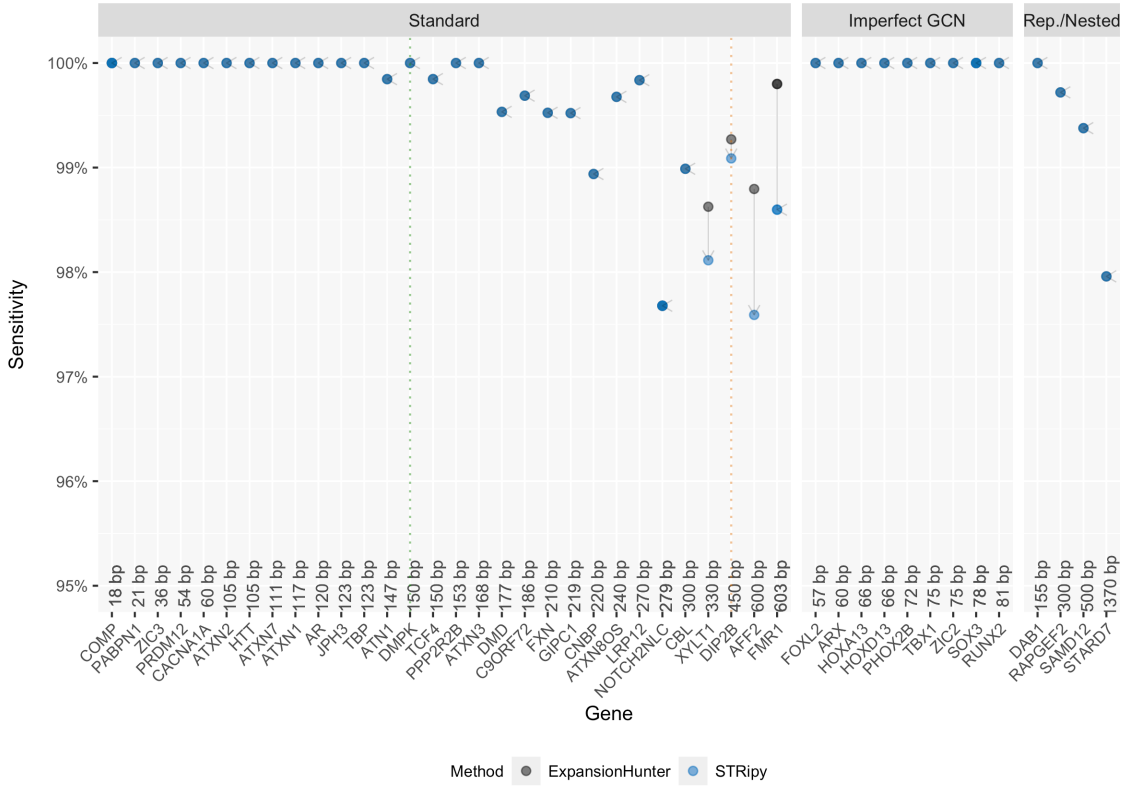


Figure 4.19. Comparison of sensitivity for pathogenic alleles using STRipy and ExpansionHunter + catalogue method. Sensitivity is the percentage of pathogenic length genotypes detected by the methods out of all pathogenic samples. Black dots represent the results obtained by using ExpansionHunter with the pre-defined catalogue and blue ones results from STRipy. The lines are connecting the results and arrows pointing to the change of the RMSE. The X-axis is ordered by the pathogenic cut-off value in base pairs which is also displayed on top of the gene name. The green dotted line represents the read length and orange dotted line the mean fragment length.

In conclusion, we described here a new computational tool, with a graphical interface, designed for an easy and fast STR analysis of pathogenic loci that utilizes the recognised STR genotyping tool, ExpansionHunter. Our method STRipy gives a fast, targeted analysis by only selecting reads in relevant regions of the disease locus and includes additional features to improve the genotyping estimation. We demonstrated the accuracy and sensitivity with our simulated samples and show that results obtained through STRipy are comparable with the conventional method by using ExpansionHunter which includes the whole sequencing file with a pre-defined catalogue (Table 4.8). In addition,

information regarding repeat lengths from the literature and population data help a user to assess the genotyping results.

Table 4.8. Comparison between ExpansionHunter and STRipy. ExpansionHunter with the original catalogue included the lowest number of loci in the catalogue as was limited to genotype 22 diseases and genes as well. ExpansionHunter with the new catalogue created by us and STRipy had similar performance over different types of reads. RMSE is the median RMSE as calculated before in this section.

Genotyping ability	**ExpansionHunter + original catalogue**	**ExpansionHunter + new catalogue (EHC)**	**STRipy**
Diseases	22	58	58
Genes	22	52	52
Loci	22	55	55
Loci with long alleles	2	43	55
Accuracy up to read length	$Standard_{RMSE}$	0.28	0.25
	Imperfect GCN_{RMSE}	0.64	1.28
	$Replaced/Nested_{RMSE}$	7.95	7.95
Accuracy from read length to fragment length	$Standard_{RMSE}$	7.76	7.80
	Imperfect GCN_{RMSE}	7.45	12.83
	$Replaced/Nested_{RMSE}$	11.24	11.24
Accuracy longer than the fragment length	$Standard_{RMSE}$	42.33	40.27
	Imperfect GCN_{RMSE}	227.46	74.99
	$Replaced/Nested_{RMSE}$	33.00	32.88
Average sensitivity		99.69%	99.62%

4.3 Discussion

There are already nearly 60 diseases known today that are linked with monogenic STR expansions. Several of these diseases have been discovered in recent years and many new findings may emerge in the following years given that STRs are abundant in the genome. As the list of diseases caused by STR expansion grows, so does their importance in research and clinical settings. Current STR genotyping tools have the potential to detect and genotype various loci and repeats, but they are focused on the most common ones. One of the best targeted STR genotyping tools, ExpansionHunter, includes 22 disease-causing loci in their variant catalogue and users are either limited to this or can add their own loci of interest to the catalogue. However, in the latter case, the accuracy of the tool at this locus is unknown and the correct specification of region can be tricky. In addition, using these tools and interpreting results requires skills and knowledge in bioinformatics.

We decided to tackle these issues in this chapter and after an extensive research in the literature, we created a comprehensive catalogue of 58 diseases caused by expansions in 55 pathogenic loci in 52 genes. Genomic coordinates, including the pathogenic motif and ranges of normal, intermediate and pathogenic repeats were defined. We have made this catalogue available as a resource for STR researchers at https://stripy.org/database. In putting this catalogue together, we divided the STR loci into three different types, based on the characteristics of the repeats they include: (1) standard, (2) imperfect GCN or (3) replaced/nested. These three types were analysed separately in a large simulation study where we assessed the ability to accurately genotype each STR region. This was done for both heterozygous and homozygous genotypes. Our results showed that all loci can be genotyped with the highest accuracy seen in standard, then imperfect GCN and finally in replaced/nested type of repeats. We found the low accuracy in the last type of repeats is due to the lack of support by ExpansionHunter. The majority of diseases that fall into this category are FAMEs which were discovered after the launch of ExpansionHunter, explaining the missing feature to accurately genotype these repeats.

We showed that ExpansionHunter works well on genotyping alleles with lengths up to the fragment length, but not being able to do this for the longer ones without utilising in-repeat reads. Thus, for each locus we empirically determined the potential locations of in-repeat reads in the genome and specified these as off-target regions in ExpansionHunter's variant catalogue. That enabled us to genotype all heterozygous

samples longer than the fragment length with a relatively good accuracy. Genotyping long homozygous alleles is yet to be solved and we reported this to the author of tool. In addition, we also found three standard type loci that ExpansionHunter had issues to genotype at lengths longer than the read length, both in homozygous and heterozygous samples. We demonstrated, by additional simulations, that due to this issue there is a good chance a pathogenic allele is genotyped with fewer repeats (limited by the read length) when using HTS data with smaller (such as 100 bp) reads. These findings were reported to the author of the tool as well, which were all fixed in the next release of the software.

Additionally, we showed that ExpansionHunter is able to genotype samples with stutter noise caused by PCR nearly as well as the PCR-free data, which increases confidence of using PCR+ samples with the tool. However, we focused here only on stutter noise and did not incorporate other PCR induced errors, such as the GC bias which potentially could result in genotyping errors due to the low coverage of GC-rich regions (Benjamini & Speed, 2012).

Our results confirmed that expansions in off-target regions are not affecting nor causing false positives in small alleles. However, they can affect the estimation of allele length for alleles that are longer than the read length. The size of this effect is probably dependent on the number of regions where there is an expansion and also the length of the repeats. We believe the reads from other genomic loci can misalign and interfere the estimation of a genotype, even when we exclude these regions from the off-target list. That being said, long expansions in multiple loci are thought to be rare and in real life situations it is probably not an issue in vast majority of cases.

As a result of our simulations, we were able to increase the variant catalogue to over twice its size, and increase the number of loci being able to genotype past the fragment length from 2 to 44, as well as providing information to fix important errors present in the tool. Moreover, we applied the new catalogue on a real disease dataset and found an undiagnosed individual who may have a STR expansion in one of the disease-causing loci matching with the phenotype. This would not have been possible to detect with the original catalogue, or even with a catalogue where this was added manually without specifying the correct off-target regions.

Our final aim for this chapter was to provide an easy to use method for analysing STRs to increase analysis of STRs from clinical samples, so we developed a free open source tool called STRipy. This uses ExpansionHunter as the genotyping method but provides additional functionality to improve accuracy for certain type of repeats, simplify the whole analysis process and provide an informative feedback for the user. We created this application for a quick and easy use and therefore built it with a graphical user interface which is straightforward even without any bioinformatic skills. By using a personalised catalogue for each sample, the application can find reads that are aligned to other regions which are not present in a pre-defined variant catalogue. In addition, unmapped reads full of GCN repeats can be recovered, aligned on a decoy chromosome and used during genotyping, which would allow genotyping GCN type of repeats even above the fragment length. Future improvements for the tool include adding additional functionality, such as the option to specify a custom STR locus on your own and target multiple loci of user's choice.

All our validation in this chapter was done by using simulated samples which is the main drawback of this study. Diseases caused by STR expansions are rare and some of them have only been found in single families or even individuals. It would be impossible to obtain sequencing data to validate all these diseases and loci and therefore this validation could realistically only be done by using simulations. Real life HTS samples can be noisy and more complex to analyse than simulated samples that tend to have fewer artefacts. However, we did introduce variability in the fragment size, simulated reads with natural occurring sequencing errors, generated stutter noise to simulate PCR samples and even created repeat sequences to match with real life findings for some type of repeats, all in order to make simulated samples resemble real HTS data. In addition, we simulated repeats in the size of 4–6 kb which included the pathogenic locus and 2 kb flanking ends. This is sufficient considering that ExpansionHunter's algorithm seeks reads in the range of 1 kb of the pathogenic locus by default. On the other hand, as we saw in our work previously, many in-repeat reads tend to misalign to other genomic regions. But by specifying them as off-target regions, ExpansionHunter can utilise reads that are misaligned there. Since we did not simulate reads in these off-target regions, then only the reads originated from the pathogenic locus will be aligned there and we are unsure whether other reads found there without expansions could affect the estimate of a genotype. We do not believe this has an impact on the pathogenic genotype unless the

allele in off-target regions is expanded as well, which is supported by our experiment where we simulated samples with expansions in multiple loci. As some of the off-target regions are another diseases' pathogenic locus, we also genotyped the samples by targeting the other disease but did not get any false positives.

In conclusion, we simulated over 90,000 STRs to characterise ExpansionHunter's performance on different types of repeats and lengths of alleles. We confirmed that the latest version of the tool is robust and performing well on all heterozygous samples. Our developed software application STRipy is using ExpansionHunter with additional functionality and allows the genotyping of 55 loci using only a fraction of the reads in a sample with the accuracy comparable to running ExpansionHunter on a whole genome sample. Our online tool to create a variant catalogue is useful to standardise the process of genotyping STRs with ExpansionHunter as all loci are being validated and makes results between different datasets confident and comparable.

Our curated online STRs database comes with information assisting analysis of expansions with ExpansionHunter. This database is a valuable resource for all researchers as it describes all reported loci and includes validated genomic coordinates for each of the locus along with the population data. For comparison, a genome-wide database at http://webstr.ucsd.edu includes population data from only a few hundred individuals from three super-populations and is not specifically disease-causing loci. Therefore, the reported repeat units and allele lengths are in some cases misleading (for instance, PABPN1 locus) and sometimes do not contain the pathogenic locus at all. One other database includes information about genes related to repeat expansions (DRED, http://omicslab.genetics.ac.cn/dred) however it is only based on the reports in literature and does not contain reference genomic regions.

All our developed tools are freely available at stripy.org for the whole community.

4.4 Methods

Based on the information found in literature we compiled a list of disorders that are known to be caused by STR expansion with information about the pathogenic range of repeats and motif, healthy range and genomic coordinates. All genomic coordinates that were not reported in coordinates of hg38 assembly were converted to hg38 with UCSC Lift Genome Annotations tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver). We used UCSC GRCh38.p12 reference genome for all simulations which was downloaded from https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/latest/hg38.fa.gz (20.08.2020).

Source code of all scripts used in this chapter to validate STR loci and referenced in this section can be found on GitLab repository: https://gitlab.com/andreassh/research-str-targeted. STRipy's client-side source code is accessible at: https://gitlab.com/andreassh/stripy-client and the server-side one at: https://gitlab.com/andreassh/stripy-server. STRipy's application, STRs database and online tools for ExpansionHunter can all be accessed at stripy.org website.

4.4.1 Simulating samples with variable length of STRs

Creation of samples with artificial length of repeats was done in two main steps. Firstly, a template sequence was created based on the reference genome, where the known STR sequence was swapped with a newly created one. Secondly, this template was used to simulate reads that were then aligned on the reference genome.

4.4.2 Template generation

We wrote a custom-made program that takes the genomic coordinates of a known pathogenic STR locus and extracts out 2 kb flanking sequence before and after the repeated sequence. Then, a new stretch of repeats was created with specific length, which was then merged with the flanking sequences to form over a 4 kb sequence and saved into a new FASTA file. For all the disorders that are caused by imperfect GCN types of repeats (stretches of the same amino acid but with an inconsistent repeat sequence coding alanine), we generated stretches of motifs that had one base randomly chosen. For example, to simulate polyalanine stretches we used GCN as the motif and replaced "N" randomly with either "A", "T", "G" or "C" in every repeat present in the sequence.

To determine genotyping accuracy across increasing length of repeats, we first simulated homozygous alleles that included the pathogenic motif from 60 bp until the 2100 bp limit was reached where each simulation increased the length of the allele by one repeat unit (in the range of 20–700 trinucleotide-, 15–525 tetranucleotide-, 12–420 pentanucleotide and 10–350 hexanucleotide repeats). In the case of the nested type of repeats, we created short alleles without the pathogenic motif. Secondly, for heterozygous alleles we designed the shorter allele to be fixed length (60 bp) and increased the length of the second allele from 60 bp until 2100 bp, similarly to the process of simulating homozygous repeats. For alleles of the nested type of repeats, we simulated alleles based on the formulas in Table 4.9, where *exp* for the non-pathogenic repeat unit was a random number between 2 and 10 and for pathogenic motif the repeat was in increasing length. For example, to simulate FAME4 disorder, a normal AAAAT repeats was created which length was randomly chosen to be between 2 and 10 repeats, followed by 12 to 240 repeats of pathogenic AAATG motif to create a 60 to 2100 bp pathogenic locus.

Table 4.9. Formulas used to simulate alleles. Motif in parentheses is the non-disease causing and motif in square brackets is the pathogenic one. The number after parentheses or square brackets shows the number of repeats of that motif, *exp* denotes an expansion which length is not specified and will be randomly assigned.

Gene	**Disease**	**Repeat type**	**Formula for allele simulation**
RFC1	CANVAS	Replaced	$(AAAAG)_{11}[AAGGG]_{exp}$
BEAN1	SCA31	Nested	$(TCAC)_{1}[TGGAA]_{exp}(TAGAA)_{exp}(TAAAATAGAA)_{exp}$
DAB1	SCA37	Nested	$(AAAAT)_{60\text{-}79}[GAAAT]_{31\text{-}75}(AAAAT)_{58\text{-}90}$
MARCH6	FAME1	Nested	$(TTTTA)_{12}[TTTCA]_{exp}$
RAPGEF2	FAME2	Nested	$(TTTTA)_{exp}[TTTCA]_{exp}(TTTTA)_{exp}$
SAMD12	FAME3	Nested	$(AAAAT)_{exp}[TGAAA]_{exp}$
STARD7	FAME4	Nested	$(AAAAT)_{exp}[AAATG]_{exp}$
TNRC6A	FAME6	Nested	$(TTTTA)_{22}[TTTCA]_{exp}(TTTTA)_{exp}$
YEATS2	FAME7	Nested	$(TTTTA)_{exp}[TTTCA]_{exp}$

For the experiment to find off-target regions (4.2.5.4), we simulated 2100 bp long repeats and created one FASTA file for each disorder to represent homozygous alleles, except for the eight disorders that can be caused by motifs with different compositions.

For the experiment to simulate PCR samples (4.2.7), we limited the maximum repeat size to 1050 bp and only heterozygous samples were simulated where the normal allele

was simulated to be 60 bp long. To simulate PCR stutter noise, we used the model described by Willems et al. (2017). In particular, for each STR locus x, we constructed a stutter model θ_x based on a geometric distribution with ρ_s denoting the parameter to control the size of stutter-induced changes, whereas (u) and (d) denotes the probability that stutter adds or removes repeats from the true allele in an observed read respectively. The probability of a stutter artefact δ repeat units in length is as follows (Willems et al., 2017):

$$P(stutter = \delta|\theta_x) = \begin{cases} 1-u-d & \delta = 0 \\ u\rho_s(1-\rho_s)^{\delta-1} & \delta > 0 \\ d\rho_s(1-\rho_s)^{-\delta-1} & \delta < 0 \end{cases}$$

4.4.3 Simulating reads and alignment

We generated paired-end reads in the FASTQ format from the template (FASTA) files with the ART (version MountRainier 2016-06-05) next-generation sequencing read simulator tool (Huang et al., 2012). Illumina HiSeqX PCR-free profile was used to simulate 150 bp reads with the mean fragment size of 450 bp and standard deviation 50 bp. Reads were simulated to have total coverage of 50x, both for homozygous and heterozygous samples.

For the experiment to genotype AR and TCF4 locus from lower read length data (4.2.5.2), we used Illumina HiSeq 2000 profile to simulate 100 bp reads with 250 bp fragment size (SD 10 bp) and total coverage of 50x.

For the experiment to find off-target regions (4.2.5.4), we simulated reads as described previously, with two exceptions. Firstly, we increased the total coverage to be 100x and secondly, we shifted down the quality score of every read by 10 to increase sequencing error rate to 1% of the default profile.

For the experiment to simulate PCR samples with stutter noise (4.2.7), we used Illumina HiSeq 2500 profile to simulate reads for each of the FASTA file that was created for this sample and which had different length of allele determined by the PCR stutter model. The other parameters were kept the same as previously. The total coverage was simulated to be 50x and a fixed seed for ART was used for all samples.

For the experiment to simulate multiple expansions (4.2.6), reads were simulated for merged chromosome 12 and 19 that included manually created expansions in specific loci.

For the experiment to simulate heterozygous samples into the pathogenic range (4.2.5.5), we simulated seven samples. First sample had the length of repeated sequence corresponding to the minimum of the pathogenic range, followed by increasing the allele length by 60 bp for the next five samples and ending with a sample where 300 bp of repeats were added.

All FASTQ files were aligned on the hg38 reference genome with BWA-MEM v0.7.17 (Li, 2013) and indexed with Samtools v1.10 (Li et al., 2009).

4.4.4 Genotyping samples with ExpansionHunter

We first used ExpansionHunter v4.0.1 (Dolzhenko et al., 2017) to genotype STRs, followed by the use of version 4.0.2 that included fixes for AR, TCF4 and ATXN1 loci (specified in the results section). A custom variant catalogue for the tool was generated with a script (*create_variant_catalog.py*) that was based on the previously composed table of STR disorders and where off-target regions were not specified. ExpansionHunter was performed on each of the aligned sample (both homozygous and heterozygous) by using the default parameters.

We made a custom-made script (*extract_eh_results.py*) to extract information regarding the determined genotypes for each locus out from the VCF files, merged this with the true length of alleles and analysed the data in R (*analyse_eh_results.R*).

We used root mean square error (RMSE) to determine genotyping accuracy for both, heterozygous and homozygous samples by comparing predicted and true repeat lengths as described by Mousavi et al. (2019). The $\langle x_1^i, x_2^i \rangle$ denotes a diploid genotype of sample i and is ordered by length such that $x_1^i \leq x_2^i$. To compare the actual $X = \{\langle x_1^1, x_2^1 \rangle, \langle x_1^2, x_2^2 \rangle \ldots \langle x_1^n, x_2^n \rangle\}$ and predicted $Y = \{\langle y_1^1, y_2^1 \rangle, \langle y_1^2, y_2^2 \rangle \ldots \langle y_1^n, y_2^n \rangle\}$ genotypes, then RMSE is defined by the following formula (Mousavi et al., 2019):

$$\sqrt{\sum_{i=1}^{n} \sum_{j=1}^{2} \frac{(y_j^i - x_j^i)^2}{2n}}$$

4.4.5 Determining off-target regions and creation of new variant catalogue

Following alignment and indexing, we used a custom-made script (*create_variant_catalog.py*) to extract out the start and end positions of all reads mapping to the genome. Overlapping and nearby regions within 100 bp were merged into one single entry. In addition, as we did not see any improvement in genotyping when using exact coordinates (or even adding 50 bp on both ends) then each region was expanded by 100 bp from each end of the coordinate that solved the issue. We set the threshold of at least five reads to include the region into the list of off-targets.

These coordinates were used to generate a new variant catalogue (if off-targets were specified). Off-target regions that included or overlapped with the reference region of the disease or regions on alternative contigs were removed from the list beforehand. The latter was due to ExpansionHunter crashing when a specified alternative contig was not found in a sample.

4.4.6 Analysis of the undiagnosed samples dataset

Clinical samples were sequenced on the Illumina NovaSeq 5000 platform, followed by alignment on the hg38 genome by using BWA-MEM v0.7.13 (Li, 2013). Duplicates were removed with Sambamba markdup v0.7.1 (Tarasov et al., 2015) and finally indexed by using Samtools v1.10 (Li et al., 2009).

ExpansionHunter v4.0.2 was performed on all samples using default parameters and the variant catalogue generated in the 4.4.5 with off-target regions specified. Results were then extracted from VCF files and the genotype of each locus in each sample was compared to the minimum value of intermediate- and pathogenic range. All results that did not meet the criteria were discarded.

4.4.7 Validation of STRipy

All heterozygous PCR-free samples from 60 to 2100 bp allele length were used to validate STRipy. All samples were genotyped by ExpansionHunter v4.0.2 using the custom-made variant catalogue with specified off-target regions for all loci except the ones composed of GCN repeats. Simultaneously, the same set of samples were genotyped by STRipy without the pre-defined variant catalogue and instead, the internal actual off-target loci specification system was used.

4.4.8 Population data analysis of STR loci for STRipy

We used over two and half thousand samples from the 1000 Genomes project to genotype all known pathogenic loci of standard and imperfect GCN type of repeats in each sample. We used samples from the DRAGEN reanalysis of the 1000 Genomes Dataset (1kGP-DRAGEN, https://registry.opendata.aws/ilmn-dragen-1kgp). This dataset includes 2504 unrelated samples from the phase 3 of 1000 Genomes project that were sequenced with the Illumina NovaSeq 6000 system to produce 150 bp paired-end reads at least 30x coverage. All BAM files in the dataset were realigned to hg38 assembly by using the Illumina DRAGEN v3.5.7b (https://www.illumina.com/products/by-type/informatics-products/dragen-bio-it-platform.html).

Amazon Web Services (AWS) cloud computing was used to genotype all samples from the following location in the AWS S3 instance: s3://1000genomes-dragen/data/dragen-3.5.7b/hg38_altaware_nohla-cnv-anchored/. ExpansionHunter v4.0.2 with our complete variant catalogue that included all genomic loci, plus different tracts in the same gene was used. Off-target regions were not included to speed up the process. Due to the ExpansionHunter's issues of genotyping loci that are very close to each other, we applied a filter to exclude results where there are less than 50 combined flanking and in-repeat reads in case of no spanning reads on all results to avoid incorrect results from genotyping different tracts in ARX and HOXA13 genes.

The resulting data in STRipy presented the five major super-populations (African, European, South-Asian, East-Asian and Admixed-American) and also combined version of the super-populations data.

Chapter 5: General discussion

This thesis is focused on short tandem repeats whose large-scale analysis was impossible until the arrival of high-throughput sequencing technology. STR genotyping tools emerged in the early 2010s with the start of the first successful genome-wide STR genotyper called lobSTR that was released in 2012. Many other tools followed and demonstrated that their performance was superior to the previous methods. As these tools had different data and techniques for validations and comparison to other tools, it has been difficult to independently assess their real performance and determine how well they compare to each other. Additionally, it is difficult for new researchers with an interest in STRs to select a tool that would be the best for his or her needs and type of analysis, considering both accuracy and sensitivity.

For many years, genotyping STRs were limited by read lengths and they were only able to determine alleles which were shorter than a read and could be completely sequenced within one read (called spanning reads). Later, a few tools, such as ExpansionHunter, STRetch and GangSTR were developed, each with novel approaches that incorporated other reads besides the spanning reads into the likelihood model for estimating a genotype. These tools can be powerful and help in facilitating clinical diagnoses through identifying expansions in known disease associated loci and discoveries of *de novo* disease causing STR expansions. In all cases, the software are complex research tools which may not be intuitive for anyone outside of the field and this limits their usage in clinical and research context.

In Chapter 2 of this thesis I investigated the main STRs genotyping tools built over the years aiming to find and characterise the best performance for genome wide calling of STRs. I assessed the performance of tools using a novel metric applied to exome sequencing data across a large dataset. More in-depth analysis of STR genotyping performance followed in Chapter 3 on whole genome sequencing data expanding our previous findings. Finally, in Chapter 4, I addressed the issues regarding genotyping known disease-causing STRs and improved ExpansionHunter, one of the most renowned STRs genotyping tool created for targeted analysis. In addition, I created a database of STR loci known to cause disease with information on the variant as well as population-wide genotypes for each of them. This database facilitated the development of a

computational tool which allows more accurate targeted genotyping of nearly all disease causing STR loci discovered today. The tool is designed to be straightforward so that it can be used by clinicians and researchers with limited knowledge in bioinformatics and STRs.

5.1 Scientific contributions

In Chapter 2, I compared four main STR genotyping tools, starting from the earliest one (lobSTR) to the most recently published GangSTR at the time. No such comparison has been performed previously and this addressed an important need considering that the correct choice of tools is essential for someone new to the field of STR research. I analysed a total of 433 whole exome sequencing samples and around a million genotypes in chromosome X of male individuals to determine the heterozygous call rate as a measure of accuracy. With this research, I demonstrated the differences in tools, but more importantly, I showed intra-tool variations and how accuracy and sensitivity is repeat unit dependant.

Overall, these results showed significant differences between repeat unit lengths. While we observed similar results for repeats between 3–6 bp, all tools were unable to genotype 1 bp repeats and also struggled with 2 bp repeats. This is important knowledge as it demonstrates the inability to accurately genotype homopolymers by any of the main STRs genotyping tools despite applying quality filtering and provides a valid reason to exclude mononucleotide results from STR analysis. Secondly, the inaccuracy of genotyping dinucleotide repeats was surprising, especially by the newest tool GangSTR. New efforts to more accurately genotype 2 bp repeats from exome data are needed in the future development of tools. Differences in accuracy between repeat unit sizes indicates the need to use filtering strategies based on the repeat characteristics (such as the repeat unit length) to improve accuracy. In fact, according to user manuals of lobSTR, HipSTR and GangSTR the recommended filtering of STRs is equally applied to all repeat units. A new tool dumpSTR (Mousavi et al., 2020) was specifically created for filtering VCF files with STR genotypes but applies the same filtering metric across different repeats. None of these methods are accounting for the differences between repeat units and as a result, they will either filter out too much data and thus, potentially exclude some true positives, or have more false positives in the dataset when filters are too relaxed.

In particular, when one is analysing the combined results of 2–6 bp repeats and uses filtering quality to reach certain accuracy then, due to the abundance of 2 bp repeats that have higher error rates, it will lead to increasing filtering strictness which would not be appropriate for the 3–6 bp repeats and therefore potentially filter out a significant amount of data. Therefore, we recommend using quality filtering based on the repeat unit size, such as having stricter options for 2 bp repeats and lower stringency for 3–6 bp repeats. We can then reach the same accuracy and retain more results in the dataset for 3–6 bp which may facilitate broader applications in these repeats. Collectively, our results propose performing more specialised filtering based on the characteristics of repeats which has not been suggested or done before.

This is important when trying to find novel disease-causing loci. Often disease datasets have a limited number of samples and including results as many results as possible could be critical for having enough power to determining a new repeat locus association. When we have more specialised methods to filter STRs data and as a result retain more data then it could help us to find novel loci. In order to gain more insights into this topic, I researched it in more depth in the following chapter.

In Chapter 3, I used GangSTR and analysed STR loci over the whole genome to include as many different STRs loci as possible. I confirmed the lower genotyping accuracy specifically for 2 bp repeat units, and showed additionally that accuracy drops at short and long repeat lengths, having the highest genotyping accuracy for repeats around 25–35 bp in length. In addition, I determined that the accuracy changes between different repeat sequences including all variations in 2 bp to 6 bp repeat units which has not been investigated at such a large scale before.

We assessed the STRs mutation profile by looking at repeat differences between PCR and PCR-free samples with regards to specific repeat motifs, repeat lengths and STR lengths. Similar results to us have also been previously reported (Fondon et al., 2012; Gymrek, 2016; Huang et al., 2002; Kelkar et al., 2008; Lai & Sun, 2003), although with some minor differences. It should be noted that the set of STRs loci used for analysis is dependent on the reference genome and the methodology of defining STR locations in the reference genome and therefore results may vary in relation to these. All in all, this research adds value by using whole genome analysis to provide in depth knowledge of genotyping accuracy of different types of repeat units, including their size, composition

and repeated sequence length. This showed the accuracy can be even affected by the composition of the motifs which is useful to know when analysing genome-wide STRs data.

Moreover, I investigated how alternations in library preparation can affect the genotyping results and determined that a smaller insert size and larger number of PCR cycles can negatively impact genotyping accuracy. However, genotyping is overall quite accurate. I determined that 93.3% of the time the correct genotype was called for dinucleotides, and over 98% for 3–6 bp repeats. The most common error in the PCR samples have one repeat less or more than the actual length for 2 bp repeats. This is a similar result to previous investigations of stutter noise distributions across different repeat sizes which showed the most frequent stutter allele is one repeat smaller than the truth in all repeat unit sizes (Gymrek, 2016). However, our results for 3–5 bp repeats are in contrast to this research as we observed that the alleles in the PCR dataset are more commonly one repeat larger than a PCR-free sample, not smaller. On the whole, one repeat difference, which this research confirmed, is not a significant concern when analysing non-coding regions. This usually affects 2 bp repeats so the error can be minimised when applying appropriate filtering strategies. In addition, better filtering will lead to more accurate results which might be of high utility in applications including STRs disease association such as those recently performed in Autism studies (Mitra et al., 2021), which also used the repeat-aware filtering tool dumpSTR in their pipeline. Our findings can be used by the community to focus on areas where methods need improvement.

In Chapter 4, I focused on genotyping the known disease-causing STRs loci. Tools such as ExpansionHunter and GangSTR provided a catalogue that allowed the genotyping of 22–23 known pathogenic loci at the time of the analysis. I collected information for all published pathogenic STR loci that have been discovered to date. I found that the number of these loci has grown to 55 and the number of diseases caused by expansions to 58, which is more than reported in the latest review articles of repeat expansion disorders (Chintalaphani et al., 2021; Khristich & Mirkin, 2020). I determined the correct genomic coordinates in the reference genome for all loci and incorporated these findings into a new STRs variant catalogue. The new catalogue enables genotyping of more known pathogenic loci than any of the currently available variant catalogues, at least doubling the total loci from 22–23 to 55.

I used one of the most robust STRs genotyping tool (ExpansionHunter) and validated genotyping of each of the STR loci with series of simulations to provide a reliable variant catalogue. Moreover, for each locus, I empirically determined all similar regions across the genome where reads could misalign and specified these as off-target regions that the tool can use during genotyping. This was done for 43 loci, which is 41 more than the currently available variant catalogue that ExpansionHunter provides and all sets of off-target regions were validated again. As a result, this allowed genotyping of long alleles in these 41 loci with significantly more accuracy than had been possible before.

I genotyped two and half thousand healthy samples to obtain population genotypes for all loci and incorporated this, with the findings from literature, to create a comprehensive online STRs database and made it publicly available. This is a valuable resource for all researchers as it contains more loci with validated genomic coordinates than any other similar resource and provides genotyping data for all pathogenic STRs loci that is not available elsewhere. Additionally, I created an online tool to create custom variant catalogues for ExpansionHunter aiming to facilitate genotyping long alleles and standardise the process of genotyping disease-causing loci. This is important as adding new loci into the catalogue requires knowledge and even minor errors such misspecifying base pair coordinates could result in genotyping errors. For example, not correctly accounting for differences in 0-based and 1-based coordinate system. Variant catalogues created by the online tool also include the set of off-target regions I have determined, enabling the genotyping of long alleles. By using the online tool, researchers can be confident in the provided catalogue that has been validated and will therefore give consistent results between research groups and more confidence in STR analysis into the future.

During the work presented in the current thesis, I also detected some issues with ExpansionHunter regarding genotyping specific loci longer than the read length, replaced and nested type of repeats and long homozygous alleles. As a result, the tool has already received fixes from the maintainers that has resolved the first and most critical issue. Extensive testing of the tool by independent bioinformaticians on different types of repeats and validation of our catalogue will increase user's confidence regarding the performance of the tool.

Finally, I created a straightforward computational tool for genotyping these disease loci named STRipy. It is designed to be used by all researchers and clinicians with or without any skills in bioinformatics to genotype all disease-causing STR loci. This tool has a graphical interface and is easy and intuitive to use, in contrast to conventional bioinformatics tools which generally run on the command line. We find this to be important as setting up bioinformatics tools and interpreting results is complex for anyone not directly in the field. Our tool requires no installation or setup and is able to return a genotype for a suspected pathogenic locus of an individual in only a matter of few minutes after downloading the tool. Hence, we can help improve clinical diagnoses of STR diseases helping to find cases that would have been missed due to the lack of STRs specific analysis because of its analytical complexity.

I have incorporated the population data obtained for each locus into the tool which will assist a user in determining whether an allele is in the pathogenic range. This is especially important for newly discovered and very rare diseases whereby the healthy and pathogenic repeat range is not well defined and no such data is available elsewhere. Furthermore, I have implemented additional features to STRipy that aims to overcome some of the technical limitations of ExpansionHunter, such as using unmapped reads to improve genotyping accuracy of long alleles. STRipy can also report the presence of the pathogenic repeat unit in nested and replaced type of repeats that is necessary to determine whether an expansion could be disease causing.

All things considered, STRipy fills a specific niche in the field and is meant to serve the wider community as the first line screening tool for STR diseases that is available to the wider community. Therefore, it is a complementary software for ExpansionHunter or other similar tools developed, which have their main role in large scale multi-loci analysis and which can also be incorporated into bioinformatics pipelines. The outcome of Chapter 4 creates valuable resources for genotyping STRs and assists in diagnosing patients with rare diseases. The main limitation of this is that all of our validations were based on simulations and using real life data can bring out some unexpected biases or glitches. However, the work on the tool will continue as issues are being reported to provide a reliable and up to date computational tool for the community. From a broader perspective, it may be worth developing more bioinformatics tools in a way that is more accessible to wider community, such as using graphical interfaces and simplifying the whole

procedure. This could improve diagnosing individuals through its broader use as researchers who are not proficient in bioinformatics analysis would be able to use such tools independently in their everyday work to test their hypotheses.

5.2 Discussion

The field of short tandem repeats is still evolving and as new methods and tools for analysing STRs from HTS data emerge, it will also bring us new discoveries of novel disease-causing loci, which is illustrated by over 10 new recent findings over the past few years (Ishiura & Tsuji, 2020). Most of these new discoveries have been in non-coding regions, indicating the importance of introns and untranslated regions for moderating gene function. Among them, there are many findings of the "nested" type of repeats which are not found in the healthy population or in the reference genome, but present in affected individuals. Although the first nested repeat was discovered in 2009 (Sato et al., 2009), no more findings of this type were made before 2017. Eight out of nine replaced or nested type of repeats were discovered between 2017 and 2020. For a long time, our knowledge was limited to the "standard" or "imperfect GCN" type of repeat expansions, but discoveries of nested type repeats could mean that there is a whole plethora of rare disease causing STRs loci which have not been discovered due to the lack of sensitive methods or tools capturing these repeats.

A recently developed tool named ExpansionHunter Denovo (Dolzhenko et al., 2019) is capable of finding novel repeats, which is specifically looking for long expansions genome-wide. On the other hand, as some diseases are only caused by short expansions (for example, all that are caused by imperfect GCN types of repeats) and it is important to scan through the whole genome of a sample to try and find the presence of novel repeats or expansions (including small ones) that are not in the reference. Such analysis that is specifically aiming to find the presence of stretches which are composed of novel repeats genome-wide has not yet been done, but which would be important to do.

New developments for STRipy could also provide the functionality to determine such loci containing novel repeats. For example, a user can specify a custom coordinate and therefore genotype any STRs genome-wide. That would make it possible to genotype STRs loci in genes where and STR expansion is not yet associated with the disease but suspected in the pathogenicity of an individual.

Also, previous analysis of STRs might be focusing too much on overall repeat length while neglecting the interruptions. For instance, tools such as ExpansionHunter used extensively in this thesis, count other repeat units into the total repeat length. When analysing CGG repeats in the FMR1 locus then interruptions, such as AGG will be counted and the combined length of repeats is returned for the locus. However, instead of focusing solely on the repeat length and determining the pathogenicity based on that number, we should be considering biochemical properties of the region. In the case of FMR1, it has been shown that AGG interruptions that are usually present in the locus can stabilise repeat alleles and the length of pure CGG repeats predict the instability better than the total number of repeats which includes AGG as well (Khristich & Mirkin, 2020). Another example, the loss of one CAA repeat in the HTT locus has been shown to increase somatic repeat instability of the CAG stretch and resulting in an earlier age of onset of the Huntington's disease (Wright et al., 2019). These observations may extend to less well studied loci. Therefore, when genotyping an STR locus it would be worth reporting any interruptions in the locus which in turn could help assessing the pathogenicity of the locus. Along with reporting stretches of novel repeats, this is another planned future development for STRipy. The next version of STRipy is aimed to address both of these issues and provide more metrics about any locus under analysis that helps to capture novel repeats as well.

By using PCR-free short-read sequencing data, then the presence, location and number of novel repeats in an STRs locus should be accurately determined for small alleles. Although, as it is common to use PCR during library preparation then this creates a challenge due to stutter noise and low coverage of GC-rich regions. An additional and even more difficult challenge is determining the extent of long stretches of repeats from the short-read sequencing data due to the small length of the reads. This limitation occurring upstream from genotyping analysis tools can be improved by new sequencing technologies. One technology able to overcome many of the current limitations is long read sequencing. Some technologies do not require a PCR step and result in uniform coverage while having reads that are long enough to span lengthy alleles full of repeats (Schadt et al., 2011). It has been shown that long-read sequencing can be used to genotype long and novel STRs (Liu et al., 2020; Logsdon et al., 2020; Sone et al., 2019). Unfortunately, the cost is generally still too high to be routinely used for genotyping clinical samples. Once the technology has matured and becomes more widely accessible,

research on STRs will greatly benefit as we can more accurately than ever determine the sequences in long alleles. That includes information about the presence of interruptions and novel repeats, as well as their exact location that helps to assess the biochemical properties of the resulting proteins (Logsdon et al., 2020). It is worth noting that the current drawback of long-read sequencing technology is the high error rate, being around 15% (van Dijk et al., 2018), but this is likely to be improved in the future and can be mitigated using deeper sequencing (Amarasinghe et al., 2020). There are already tools developed for calling STRs in long-read sequencing data, for example, STRique (Gießelmann et al., 2018), TRiCoLOR (Bolognini et al., 2020) and RepeatHMM (Liu et al., 2020).

Long-read sequencing will be even more important for analysis of variable number tandem repeats (VNTRs), also known as minisatellites. VNTRs have repeat units that are longer than the upper limit of STRs (De Roeck et al., 2019). For example, a recently discovered locus in the ABCA7 gene has 25 bp long motif and has been associated with Alzheimer's disease. Even for healthy individuals the allele is thousands of base pairs long (De Roeck et al., 2018), which would be extremely difficult to genotype by using short-read sequencing data and previously only Southern blotting could be used to determine the approximate length of the allele. With long-read sequencing technology, reads even up to megabases long can be created (Amarasinghe et al., 2020) that spans those long alleles and this will make it possible to perform genotyping for VNTRs too.

There is no consensus on the definition of the upper length of STR motif and the minimum length of VNTR motif. This also raises a question why the analysis of these repeats is often divided based on the length of a repeat unit and not what would be technologically possible. For example, an expansion of dodecamer repeat unit in CSTB gene is pathogenic from 12 repeats or 144 bp (Lalioti et al., 1997), which can be detected and genotyped from short-read sequencing data. Therefore, it would be reasonable to erase the line between STRs and VNTRs and look at tandem repeats as a continuous range while considering the length of repeats we can analyse from our sequencing data. Using short-read sequencing data has its limits and the longer the repeat unit is the less repeats we can accurately genotype. In contrast, long-read sequencing data where the lengths are significantly higher allows STRs and VNTRs to be looked at together. If we increase the

upper limit of the lengths of repeat units then this might also lead to new discoveries of repeats composed of motifs longer than 6 bp.

Moving even further upstream from the sequencing technology, then another challenge we may want to address is analysing genomes obtained from somatic cells. Research has shown that triplet repeats are unstable in somatic cells and somatic repeat expansions have been detected in brains of patients with Huntington's disease, and these changes increased with age (Kennedy & Shelbourne, 2000; Swami et al., 2009). That raises a question whether we can detect all pathogenic STRs expansions considering that the standard DNA sequencing method involves using whole blood. For neurological diseases, it would be the most relevant to genotyping STRs from DNA obtained from brain cells, but this would be an invasive technique that cannot be used for solely diagnosing purposes. Similarly in cancer, repeat instability has been found in 14 types of cancer (Hause et al., 2016) indicating that the mosaicism of highly mutable STRs taking place through mitosis. Despite the improvements of computational tools as well as sequencing technology, the challenge of detecting somatic mosaicism still remains. This should be taken into account when using sequencing data to confirm a diagnose. For example, individuals whose allele length is very close to the pathogenic range might gain a repeat or two during cell divisions and end up with a somatic mosaicism and a disease that results from an expanded allele in brain where it is functional, but which is not detected from DNA sequenced from blood. In the future, single-cell sequencing could be viable tool to detect mutations that are present in a small number of cells and uncover somatic mosaicism.

In conclusion, there are challenges at different stages of analysing STRs, from collecting the sample to analysing sequencing data and interpreting results. Some of these issues will likely be addressed sooner, but others might take longer. This thesis makes a contribution to the downstream part of STRs analysis and assisting researchers and clinicians alike to easily perform STRs genotyping. We give recommendations on how to filter STR call data that will help researchers perform more accurate data processing and receive the maximum amount of high-quality data. Improved methods and the new tools that have been developed in this thesis will facilitate diagnosing STRs diseases from HTS data and standardise the process.

List of references

Abel, H. J. & Duncavage, E. J. (2013). Detection of structural DNA variation from next generation sequencing data: A review of informatic approaches. *Cancer Genetics*, *206*(12), 432–40. https://doi.org/10.1016/j.cancergen.2013.11.002

Albrecht, A. N., Kornak, U., Böddrich, A., Süring, K., Robinson, P. N., Stiege, A. C., Lurz, R., Stricker, S., Wanker, E. E., & Mundlos, S. (2004). A molecular pathogenesis for transcription factor associated poly-alanine tract expansions. *Human Molecular Genetics*, *13*(20), 2351–59. https://doi.org/10.1093/hmg/ddh277

Almeida, P. & Penha-Gonçalves, C. (2004). Long perfect dinucleotide repeats are typical of vertebrates, show motif preferences and size convergence. *Molecular Biology and Evolution*, *21*(7), 1226–33. https://doi.org/10.1093/molbev/msh108

Alonso, I., Jardim, L. B., Artigalas, O., Saraiva-Pereira, M. L., Matsuura, T., Ashizawa, T., Sequeiros, J., & Silveira, I. (2006). Reduced penetrance of intermediate size alleles in spinocerebellar ataxia type 10. *Neurology*, *66*(10), 1602–04. https://doi.org/10.1212/01.wnl.0000216266.30177.bb

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, *21*(1), 1–16. https://doi.org/10.1186/s13059-020-1935-5

Ashizawa, T., Öz, G., & Paulson, H. L. (2018). Spinocerebellar ataxias: prospects and challenges for therapy development. *Nature Reviews Neurology*, *14*(10), 590–605. https://doi.org/10.1038/s41582-018-0051-6

Atsuta, N., Watanabe, H., Ito, M., Banno, H., Suzuki, K., Katsuno, M., Tanaka, F., Tamakoshi, A., & Sobue, G. (2006). Natural history of spinal and bulbar muscular atrophy (SBMA): A study of 223 Japanese patients. *Brain*, *129*(6), 1446–55. https://doi.org/10.1093/brain/awl096

Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, *43*(1), 483–92. https://doi.org/10.1002/0471250953.bi1110s43

Aziz, M., Stathopulu, E., Callias, M., Taylor, C., Turk, J., Oostra, B., Willemsen, R., & Patton, M. (2003). Clinical features of boys with fragile X premutations and intermediate alleles. *American Journal of Medical Genetics - Neuropsychiatric Genetics*, *121 B*(1), 119–27. https://doi.org/10.1002/ajmg.b.20030

Bachinski, L. L., Czernuszewicz, T., Ramagli, L. S., Suominen, T., Shriver, M. D., Udd, B., Siciliano, M. J., & Krahe, R. (2009). Premutation allele pool in myotonic dystrophy type 2. *Neurology*, *72*(6), 490–97. https://doi.org/10.1212/01.wnl.0000333665.01888.33

Bahlo, M., Bennett, M. F., Degorski, P., Tankard, R. M., Delatycki, M. B., & Lockhart, P. J. (2018). Recent advances in the detection of repeat expansions with short-read next-generation sequencing [version 1; referees: 3 approved]. *F1000Research*, *7*, 1–11. https://doi.org/10.12688/f1000research.13980.1

Beck, J., Poulter, M., Hensman, D., Rohrer, J. D., Mahoney, C. J., Adamson, G., Campbell, T., Uphill, J., Borg, A., Fratta, P., Orrell, R. W., Malaspina, A., Rowe, J., Brown, J., Hodges, J., Sidle, K., Polke, J. M., Houlden, H., Schott, J. M., et al. (2013). Large C9orf72 hexanucleotide repeat expansions are seen in multiple neurodegenerative syndromes and are more frequent than expected in the UK population. *American Journal of Human Genetics*, *92*(3), 345–53. https://doi.org/10.1016/j.ajhg.2013.01.011

Bell, G. I. & Jurka, J. (1997). The Length Distribution of Perfect Dimer Repetitive DNA Is Consistent with Its Evolution by an Unbiased Single-Step Mutation Process. *Journal of Molecular Evolution*, *44*(4), 414–21. https://doi.org/10.1007/PL00006161

Benjamini, Y. & Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, *40*(10), 1–14. https://doi.org/10.1093/nar/gks001

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, *27*(2), 573–80.

Berger, S. L., Kouzarides, T., Shiekhattar, R., & Shilatifard, A. (2009). An operational definition of epigenetics. *Genes and Development*, *23*(7), 781–83. https://doi.org/10.1101/gad.1787609

Bienvenu, T., Poirier, K., Friocourt, G., Bahi, N., Beaumont, D., Fauchereau, F., Jeema, L. Ben, Zemni, R., Vinet, M. C., Francis, F., Couvert, P., Gomot, M., Moraine, C., Van Bokhoven, H., Kalscheuer, V., Frints, S., Gecz, J., Ohzaki, K., Chaabouni, H., et al. (2002). ARX, a novel Prd-class-homeobox gene highly expressed in the telencephalon, is mutated in X-linked mental retardation. *Human Molecular Genetics*, *11*(8), 981–91. https://doi.org/10.1093/hmg/11.8.981

Björn, N., Pradhananga, S., Sigurgeirsson, B., & Lundeberg, J. (2018). Comparison of Variant Calls from Whole Genome and Whole Exome Sequencing Data Using Matched Samples. *Next Generat Sequenc & Applic*, *5*(1), . https://doi.org/10.4172/2469-9853.1000154

Bolognini, D., Magi, A., Benes, V., Korbel, J. O., & Rausch, T. (2020). TRiCoLOR: tandem repeat profiling using whole-genome long-read sequencing data. *GigaScience*, *9*(10), 1–7. https://doi.org/10.1093/gigascience/giaa101

Brais, B., Bouchard, J.-P., Xie, Y.-G., Rochefort, D. L., Chrétien, N., Tomé, F. M. S., Lafrentére, R. G., Rommens, J. M., Uyama, E., Nohira, O., Blumen, S., Korcyn, A. D., Heutink, P., Mathieu, J., Duranceau, A., Codère, F., Fardeau, M., & Rouleau, G. A. (1998). Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. *Nature Genetics*, *18*(2), 164–67. https://doi.org/10.1038/ng0298-164

Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J., & Rolf, B. (1998). Mutation Rate in Human Microsatellites: Influence of the Structure and Length of the Tandem Repeat. *The American Journal of Human Genetics*, *62*(6), 1408–15. https://doi.org/10.1086/301869

Brook, J. D., McCurrach, M. E., Harley, H. G., Buckler, A. J., Church, D., Aburatani, H., Hunter, K., Stanton, V. P., Thirion, J. P., Hudson, T., Sohn, R., Zemelman, B., Snell, R. G., Rundle, S. A., Crow, S., Davies, J., Shelbourne, P., Buxton, J., Jones, C., et al. (1992). Molecular basis of myotonic dystrophy: Expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. *Cell*, *68*(4), 799–808. https://doi.org/10.1016/0092-8674(92)90154-5

Brown, L., Paraso, M., Arkell, R., & Brown, S. (2005). In vitro analysis of partial loss-of-function ZIC2 mutations in holoprosencephaly: Alanine tract expansion modulates DNA binding and transactivation. *Human Molecular Genetics*, *14*(3), 411–20. https://doi.org/10.1093/hmg/ddi037

Bruinsma, S., Burgess, J., Schlingman, D., Czyz, A., Morrell, N., Ballenger, C., Meinholz, H., Brady, L., Khanna, A., Freeberg, L., Jackson, R. G., Mathonet, P., Verity, S. C., Slatter, A. F., Golshani, R., Grunenwald, H., Schroth, G. P., & Gormley, N. A. (2018). Bead-linked transposomes enable a normalization-free workflow for NGS library preparation. *BMC Genomics*, *19*(1), 1–16. https://doi.org/10.1186/s12864-018-5096-9

Budiš, J., Kucharík, M., Ďuriš, F., Gazdarica, J., Zrubcová, M., Ficek, A., Szemes, T., Brejová, B., & Radvanszky, J. (2019). Dante: genotyping of known complex and expanded short tandem repeats. *Bioinformatics*, *35*(8), 1310–17. https://doi.org/10.1093/bioinformatics/bty791

Butler, J. M. (2007). Short tandem repeat typing technologies used in human identity testing. *BioTechniques*, *43*(4), . https://doi.org/10.2144/000112582

Cacace, R., Van Cauwenberghe, C., Bettens, K., Gijselinck, I., van der Zee, J., Engelborghs, S., Vandenbulcke, M., Van Dongen, J., Bäumer, V., Dillen, L., Mattheijssens, M., Peeters, K., Cruts, M., Vandenberghe, R., De Deyn, P. P., Van Broeckhoven, C., & Sleegers, K. (2013). C9orf72 G4C2 repeat expansions in Alzheimer's disease and mild cognitive impairment. *Neurobiology of Aging*, *34*(6), 1712.e1-1712.e7. https://doi.org/10.1016/j.neurobiolaging.2012.12.019

Campbell, I. (2010). Cell biology and gene expression. *Anaesthesia and Intensive Care Medicine*, *11*(3), 118–22. https://doi.org/10.1016/j.mpaic.2009.12.004

Cao, M. D., Tasker, E., Willadsen, K., Imelfort, M., Vishwanathan, S., Sureshkumar, S., Balasubramanian, S., & Bodén, M. (2014). Inferring short tandem repeat variation from paired-end short reads. *Nucleic Acids Research*, *42*(3), e16–e16. https://doi.org/10.1093/nar/gkt1313

Caspar, S.M., Dubacher, N., Kopps, A. M., Meienberg, J., Henggeler, C., & Matyas, G. (2018). Clinical sequencing: From raw data to diagnosis with lifetime value. *Clinical Genetics*, *93*(3), 508–19. https://doi.org/10.1111/cge.13190

Caspar, S. M., Dubacher, N., Kopps, A. M., Meienberg, J., Henggeler, C., & Matyas, G. (2018). Clinical sequencing: From raw data to diagnosis with lifetime value. *Clinical Genetics*, *93*(3), 508–19. https://doi.org/10.1111/cge.13190

Cen, Z., Jiang, Z., Chen, Y., Zheng, X., Xie, F., Yang, X., Lu, X., Ouyang, Z., Wu, H., Chen, S., Yin, H., Qiu, X., Wang, S., Ding, M., Tang, Y., Yu, F., Li, C., Wang, T., Ishiura, H., et al. (2018). Intronic pentanucleotide TTTCA repeat insertion in the SAMD12 gene causes familial cortical myoclonic tremor with epilepsy type 1. *Brain*, *141*(8), 2280–88. https://doi.org/10.1093/brain/awy160

Chakraborty, R., Kimmel, M., Stivers, D. N., Davison, L. J., & Deka, R. (1997). Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proceedings of the National Academy of Sciences*, *94*(3), 1041–46. https://doi.org/10.1073/pnas.94.3.1041

Chen, Y.-C., Auer-Grumbach, M., Matsukawa, S., Zitzelsberger, M., Themistocleous, A. C., Strom, T. M., Samara, C., Moore, A. W., Cho, L. T.-Y., Young, G. T., Weiss, C., Schabhüttl, M., Stucka, R., Schmid, A. B., Parman, Y., Graul-Neumann, L., Heinritz, W., Passarge, E., Watson, R. M., et al. (2015). Transcriptional regulator PRDM12 is essential for human pain perception. *Nature Genetics*, *47*(7), 803–08. https://doi.org/10.1038/ng.3308

Chintalaphani, S. R., Pineda, S. S., Deveson, I. W., & Kumar, K. R. (2021). An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. *Acta Neuropathologica Communications*, *9*(1), . https://doi.org/10.1186/s40478-021-01201-x

Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Özen, S., Sanjad, S., Nelson-Williams, C., Farhi, A., Mane, S., & Lifton, R. P. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences*, *106*(45), 19096–101. https://doi.org/10.1073/pnas.0910672106

Corbett, M. A., Kroes, T., Veneziano, L., Bennett, M. F., Florian, R., Schneider, A. L., Coppola, A., Licchetta, L., Franceschetti, S., Suppa, A., Wenger, A., Mei, D., Pendziwiat, M., Kaya, S., Delledonne, M., Straussberg, R., Xumerle, L., Regan, B., Crompton, D., et al. (2019). Intronic ATTTC repeat expansions in STARD7 in familial adult myoclonic epilepsy linked to chromosome 2. *Nature Communications*, *10*(1), 1–10. https://doi.org/10.1038/s41467-019-12671-y

Corral-Juan, M., Serrano-Munuera, C., Rábano, A., Cota-González, D., Segarra-Roca, A., Ispierto, L., Cano-Orgaz, A. T., Adarmes, A. D., Méndez-Del-Barrio, C., Jesús, S., Mir, P., Volpini, V., Alvarez-Ramo, R., Sánchez, I., & Matilla-Dueñas, A. (2018). Clinical, genetic and neuropathological characterization of spinocerebellar ataxia type 37. *Brain*, *141*(7), 1981–97. https://doi.org/10.1093/brain/awy137

Cortese, A., Simone, R., Sullivan, R., Vandrovcova, J., Tariq, H., Yau, W. Y., Humphrey, J., Jaunmuktane, Z., Sivakumar, P., Polke, J., Ilyas, M., Tribollet, E., Tomaselli, P. J., Devigili, G., Callegari, I., Versino, M., Salpietro, V., Efthymiou, S., Kaski, D., et al. (2019). Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. *Nature Genetics*, *51*(4), 649–58. https://doi.org/10.1038/s41588-

019-0372-4

Cortese, A., Tozza, S., Yau, W. Y., Rossi, S., Beecroft, S. J., Jaunmuktane, Z., Dyer, Z., Ravenscroft, G., Lamont, P. J., Mossman, S., Chancellor, A., Maisonobe, T., Pereon, Y., Cauquil, C., Colnaghi, S., Mallucci, G., Curro, R., Tomaselli, P. J., Thomas-Black, G., et al. (2020). Cerebellar ataxia, neuropathy, vestibular areflexia syndrome due to RFC1 repeat expansion. *Brain*, *143*(2), 489–90. https://doi.org/10.1093/brain/awz418

Dashnow, H., Lek, M., Phipson, B., Halman, A., Sadedin, S., Lonsdale, A., Davis, M., Lamont, P., Clayton, J. S., Laing, N. G., MacArthur, D. G., & Oshlack, A. (2018). STRetch: Detecting and discovering pathogenic short tandem repeat expansions. *Genome Biology*, *19*(1), 1–13. https://doi.org/10.1186/s13059-018-1505-2

Daughters, R. S., Tuttle, D. L., Gao, W., Ikeda, Y., Moseley, M. L., Ebner, T. J., Swanson, M. S., & Ranum, L. P. W. (2009). RNA gain-of-function in spinocerebellar ataxia type 8. *PLoS Genetics*, *5*(8), . https://doi.org/10.1371/journal.pgen.1000600

De Baere, E., Beysen, D., Oley, C., Lorenz, B., Cocquet, J., De Sutter, P., Devriendt, K., Dixon, M., Fellous, M., Fryns, J. P., Garza, A., Jonsrud, C., Koivisto, P. A., Krause, A., Leroy, B. P., Meire, F., Plomp, A., Van Maldergem, L., De Paepe, A., et al. (2003). FOXL2 and BPES: Mutational hotspots, phenotypic variability, and revision of the genotype-phenotype correlation. *American Journal of Human Genetics*, *72*(2), 478–87. https://doi.org/10.1086/346118

De Baere, E., Dixon, M. J., Small, K. W., Jabs, E. W., Leroy, B. P., Devriendt, K., Gillerot, Y., Mortier, G., Meire, F., Van Maldergem, L., Courtens, W., Hjalgrim, H., Huang, S., Liebaers, I., Van Regemorter, N., Touraine, P., Praphanphoj, V., Verloes, A., Udar, N., et al. (2001). Spectrum of FOXL2 gene mutations in blepharophimosis-ptosis-epicanthus inversus (BPES) families demonstrates a genotype-phenotype correlation. *Human Molecular Genetics*, *10*(15), 1591–600. https://doi.org/10.1093/hmg/10.15.1591

De Roeck, A., Duchateau, L., Van Dongen, J., Cacace, R., Bjerke, M., Van den Bossche, T., Cras, P., Vandenberghe, R., De Deyn, P. P., Engelborghs, S., Van Broeckhoven, C., Sleegers, K., Goeman, J., Crols, R., Nuytten, D., Mercelis, R., Vandenbulcke, M., Sieben, A., De Bleecker, J. L., et al. (2018). An intronic VNTR affects splicing of ABCA7 and increases risk of Alzheimer's disease. *Acta Neuropathologica*, *135*(6), 827–37. https://doi.org/10.1007/s00401-018-1841-z

De Roeck, A., De Coster, W., Bossaerts, L., Cacace, R., De Pooter, T., Van Dongen, J., D'Hert, S., De Rijk, P., Strazisar, M., Van Broeckhoven, C., & Sleegers, K. (2019). NanoSatellite: accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on PromethION. *Genome Biology*, *20*(1), 239. https://doi.org/10.1186/s13059-019-1856-3

Délot, E., King, L. M., Briggs, M. D., Wilcox, W. R., & Cohn, D. H. (1999). Trinucleotide expansion mutations in the cartilage oligomeric matrix protein (COMP) gene. *Human Molecular Genetics*, *8*(1), 123–28. https://doi.org/10.1093/hmg/8.1.123

Demos, M. K., Fullston, T., Partington, M. W., Gécz, J., & Gibson, W. T. (2009). Clinical

study of two brothers with a novel 33 bp duplication in the ARX gene. *American Journal of Medical Genetics, Part A*, *149*(7), 1482–86. https://doi.org/10.1002/ajmg.a.32851

Deng, J., Yu, J., Li, P., Luan, X., Cao, L., Zhao, J., Yu, M., Zhang, W., Lv, H., Xie, Z., Meng, L. C., Zheng, Y., Zhao, Y., Gang, Q., Wang, Q., Liu, J., Zhu, M., Guo, X., Su, Y., et al. (2020). Expansion of GGC Repeat in GIPC1 Is Associated with Oculopharyngodistal Myopathy. *American Journal of Human Genetics*, *106*(6), 793–804. https://doi.org/10.1016/j.ajhg.2020.04.011

Depienne, C. & Mandel, J.-L. (2021). 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? *The American Journal of Human Genetics*, *108*(5), 764–85. https://doi.org/10.1016/j.ajhg.2021.03.011

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–98. https://doi.org/10.1038/ng.806

Doi, K., Monjo, T., Hoang, P. H., Yoshimura, J., Yurino, H., Mitsui, J., Ishiura, H., Takahashi, Y., Ichikawa, Y., Goto, J., Tsuji, S., & Morishita, S. (2014). Rapid detection of expanded short tandem repeats in personal genomics using hybrid sequencing. *Bioinformatics*, *30*(6), 815–22. https://doi.org/10.1093/bioinformatics/btt647

Dolzhenko, E., van Vugt, J. J. F. A., Shaw, R. J., Bekritsky, M. A., van Blitterswijk, M., Narzisi, G., Ajay, S. S., Rajan, V., Lajoie, B. R., Johnson, N. H., Kingsbury, Z., Humphray, S. J., Schellevis, R. D., Brands, W. J., Baker, M., Rademakers, R., Kooyman, M., Tazelaar, G. H. P., van Es, M. A., et al. (2017). Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Research*, *27*(11), 1895–903. https://doi.org/10.1101/gr.225672.117

Dolzhenko, E., Deshpande, V., Schlesinger, F., Krusche, P., Petrovski, R., Chen, S., Emig-Agius, D., Gross, A., Narzisi, G., Bowman, B., Scheffler, K., van Vugt, J. J. F. A., French, C., Sanchis-Juan, A., Ibáñez, K., Tucci, A., Lajoie, B. R., Veldink, J. H., Raymond, F. L., et al. (2019). ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics*, edited by Inanc Birol*35*(22), 4754–56. https://doi.org/10.1093/bioinformatics/btz431

Doyu, M., Sobue, G., Mitsuma, T., Uchida, M., Iwase, T., & Takahashi, A. (1993). Very late onset X-linked recessive bulbospinal neuronopathy: Mild clinical features and a mild increase in the size of tandem CAG repeat in androgen receptor gene. *Journal of Neurology, Neurosurgery and Psychiatry*, *56*(7), 832–33. https://doi.org/10.1136/jnnp.56.7.832

Duffy, K. R., Gurram, N., Peters, K. C., Wellner, G., & Grgicak, C. M. (2017). Exploring STR signal in the single- and multicopy number regimes: Deductions from an in silico model of the entire DNA laboratory process. *Electrophoresis*, *38*(6), 855–68. https://doi.org/10.1002/elps.201600385

Duitama, J., Zablotskaya, A., Gemayel, R., Jansen, A., Belet, S., Vermeesch, J. R., Verstrepen, K. J., & Froyen, G. (2014). Large-scale analysis of tandem repeat variability in the human genome. *Nucleic Acids Research*, *42*(9), 5728–41. https://doi.org/10.1093/nar/gku212

Duyao, M., Ambrose, C., Myers, R., Novelletto, A., Persichetti, F., Frontali, M., Folstein, S., Ross, C., Franz, M., & Abbott, M. (1993). Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nature Genetics*, *4*, 387–92. https://doi.org/10.1038/ng0893-387

Eberle, M. A., Fritzilas, E., Krusche, P., Källberg, M., Moore, B. L., Bekritsky, M. A., Iqbal, Z., Chuang, H. Y., Humphray, S. J., Halpern, A. L., Kruglyak, S., Margulies, E. H., McVean, G., & Bentley, D. R. (2017). A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Research*, *27*(1), 157–64. https://doi.org/10.1101/gr.210500.116

Eichler, E. E. (2019). Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. *New England Journal of Medicine*, *381*(1), 64–74. https://doi.org/10.1056/nejmra1809315

Ellegren, H. (2000a). Heterogeneous mutation processes in human microsatellite DNA sequences. *Nature Genetics*, *24*(4), 400–02. https://doi.org/10.1038/74249

Ellegren, H. (2000b). Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in Genetics*, *16*(12), 551–58. https://doi.org/10.1016/S0168-9525(00)02139-9

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. *Nature Reviews*, *5*(June), 435–45. https://doi.org/10.1038/nrg1348

Fang, W. L., Chang, S. C., Lan, Y. T., Huang, K. H., Chen, J. H., Lo, S. S., Hsieh, M. C., Li, A. F. Y., Wu, C. W., & Chiou, S. H. (2012). Microsatellite instability is associated with a better prognosis for gastric cancer patients after curative surgery. *World Journal of Surgery*, *36*(9), 2131–38. https://doi.org/10.1007/s00268-012-1652-7

Florian, R. T., Kraft, F., Leitão, E., Kaya, S., Klebe, S., Magnin, E., van Rootselaar, A. F., Buratti, J., Kühnel, T., Schröder, C., Giesselmann, S., Tschernoster, N., Altmueller, J., Lamiral, A., Keren, B., Nava, C., Bouteiller, D., Forlani, S., Jornea, L., et al. (2019). Unstable TTTTA/TTTCA expansions in MARCH6 are associated with Familial Adult Myoclonic Epilepsy type 3. *Nature Communications*, *10*(1), 1–14. https://doi.org/10.1038/s41467-019-12763-9

Fondon, J. W., Martin, A., Richards, S., Gibbs, R. A., & Mittelman, D. (2012). Analysis of microsatellite variation in Drosophila melanogaster with population-scale genome sequencing. *PLoS ONE*, *7*(3), 1–9. https://doi.org/10.1371/journal.pone.0033036

Fotsing, S. F., Margoliash, J., Wang, C., Saini, S., Yanicky, R., Shleizer-Burko, S., Goren, A., & Gymrek, M. (2019). The impact of short tandem repeat variation on gene

expression. *Nature Genetics*, *51*(11), 1652–59. https://doi.org/10.1038/s41588-019-0521-9

Fratta, P., Collins, T., Pemble, S., Nethisinghe, S., Devoy, A., Giunti, P., Sweeney, M. G., Hanna, M. G., & Fisher, E. M. C. (2014). Sequencing analysis of the spinal bulbar muscular atrophy CAG expansion reveals absence of repeat interruptions. *Neurobiology of Aging*, *35*(2), 443.e1-443.e3. https://doi.org/10.1016/j.neurobiolaging.2013.07.015

Frésard, L. & Montgomery, S. B. (2018). Diagnosing rare diseases after the exome. *Cold Spring Harbor Molecular Case Studies*, *4*(6), 1–10. https://doi.org/10.1101/mcs.a003392

Frisen, L. (2003). A novel duplication in the HOXA13 gene in a family with atypical hand-foot-genital syndrome. *Journal of Medical Genetics*, *40*(4), 49e – 49. https://doi.org/10.1136/jmg.40.4.e49

Fungtammasan, A., Ananda, G., Hile, S. E., Su, M. S. W., Sun, C., Harris, R., Medvedev, P., Eckert, K., & Makova, K. D. (2015). Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. *Genome Research*, *125*(5), 736–49. https://doi.org/10.1101/gr.185892.114

Galvão, R., Mendes-Soares, L., Câmara, J., Jaco, I., & Carmo-Fonseca, M. (2001). Triplet repeats, RNA secondary structure and toxic gain-of-function models for pathogenesis. *Brain Research Bulletin*, *56*(3–4), 191–201. https://doi.org/10.1016/S0361-9230(01)00651-7

García-Murias, M., Quintáns, B., Arias, M., Seixas, A. I., Cacheiro, P., Tarrío, R., Pardo, J., Millán, M. J., Arias-Rivas, S., Blanco-Arias, P., Dapena, D., Moreira, R., Rodríguez-Trelles, F., Sequeiros, J., Carracedo, Á., Silveira, I., & Sobrido, M. J. (2012). "Costa da Morte" ataxia is spinocerebellar ataxia 36: Clinical and genetic characterization. *Brain*, *135*(5), 1423–35. https://doi.org/10.1093/brain/aws069

Garden, G. A. & La Spada, A. R. (2008). Molecular pathogenesis and cellular pathology of spinocerebellar ataxia type 7 neurodegeneration. *Cerebellum*, *7*(2), 138–49. https://doi.org/10.1007/s12311-008-0027-y

Gemayel, R., Vinces, M. D., Legendre, M., & Verstrepen, K. J. (2010). Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual Review of Genetics*, *44*, 445–77. https://doi.org/10.1146/annurev-genet-072610-155046

Gießelmann, P., Brändl, B., Raimondeau, E., Bowen, R., Rohrandt, C., Tandon, R., Kretzmer, H., Assum, G., Galonska, C., Siebert, R., Ammerpohl, O., Heron, A., Schneider, S. A., Ladewig, J., Koch, P., Schuldt, B. M., Graham, J. E., Meissner, A., & Müller, F. J. (2018). Repeat expansion and methylation state analysis with nanopore sequencing. *BioRxiv*, . https://doi.org/10.1101/480285

Gispert, S., Kurz, A., Waibel, S., Bauer, P., Liepelt, I., Geisen, C., Gitler, A. D., Becker, T., Weber, M., Berg, D., Andersen, P. M., Krüger, R., Riess, O., Ludolph, A. C., & Auburger, G. (2012). The modulation of Amyotrophic Lateral Sclerosis risk by Ataxin-2 intermediate polyglutamine expansions is a specific effect. *Neurobiology*

of Disease, *45*(1), 356–61. https://doi.org/10.1016/j.nbd.2011.08.021

Gonitel, R., Moffitt, H., Sathasivam, K., Woodman, B., Detloff, P. J., Faull, R. L. M., & Bates, G. P. (2008). DNA instability in postmitotic neurons. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(9), 3467–72. https://doi.org/10.1073/pnas.0800048105

Goodman, F. R., Bacchelli, C., Brady, A. F., Brueton, L. A., Fryns, J. P., Mortlock, D. P., Innis, J. W., Holmes, L. B., Donnenfeld, A. E., Feingold, M., Beemer, F. A., Hennekam, R. C. M., & Scambler, P. J. (2000). Novel HOXA13 mutation and the phenotypic spectrum of hand-foot-genital syndrome. *American Journal of Human Genetics*, *67*(1), 197–202. https://doi.org/10.1086/302961

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–51. https://doi.org/10.1038/nrg.2016.49

Gousse, G., Patural, H., Touraine, R., Chabrier, S., Rolland, E., Antoine, J. C., & Perrin, L. (2018). Lethal form of spinocerebellar ataxia type 7 with early onset in childhood. *Archives de Pediatrie*, *25*(1), 42–44. https://doi.org/10.1016/j.arcped.2017.09.029

Grunseich, C., Kats, I. R., Bott, L. C., Rinaldi, C., Kokkinis, A., Fox, D., Chen, K. lian, Schindler, A. B., Mankodi, A. K., Shrader, J. A., Schwartz, D. P., Lehky, T. J., Liu, C. Y., & Fischbeck, K. H. (2014). Early onset and novel features in a spinal and bulbar muscular atrophy patient with a 68 CAG repeat. *Neuromuscular Disorders*, *24*(11), 978–81. https://doi.org/10.1016/j.nmd.2014.06.441

Gupta, A. & Jankovic, J. (2009). Spinocerebellar ataxia 8: Variable phenotype and unique pathogenesis. *Parkinsonism and Related Disorders*, *15*(9), 621–26. https://doi.org/10.1016/j.parkreldis.2009.06.001

Gusella, J. F., Macdonald, M. E., & Lee, J. M. (2014). Genetic modifiers of Huntington's disease. *Movement Disorders*, *29*(11), 1359–65. https://doi.org/10.1002/mds.26001

Gutiérrez Gutiérrez, G., Díaz-Manera, J., Almendrote, M., Azriel, S., Eulalio Bárcena, J., Cabezudo García, P., Camacho Salas, A., Casanova Rodríguez, C., Cobo, A. M., Díaz Guardiola, P., Fernández-Torrón, R., Gallano Petit, M. P., García Pavía, P., Gómez Gallego, M., Gutiérrez Martínez, A. J., Jericó, I., Kapetanovic García, S., López de Munaín Arregui, A., Martorell, L., et al. (2020). Clinical guide for the diagnosis and follow-up of myotonic dystrophy type 1, MD1 or Steinert's disease. *Neurología (English Edition)*, *35*(3), 185–206. https://doi.org/10.1016/j.nrleng.2019.01.008

Gymrek, M. (2017). A genomic view of short tandem repeats. *Current Opinion in Genetics & Development*, *44*, 9–16. https://doi.org/10.1016/j.gde.2017.01.012

Gymrek, M., Willems, T., Guilmatre, A., Zeng, H., Markus, B., Georgiev, S., Daly, M. J., Price, A. L., Pritchard, J. K., Sharp, A. J., & Erlich, Y. (2016). Abundant contribution of short tandem repeats to gene expression variation in humans. *Nature Genetics*, *48*(1), 22–29. https://doi.org/10.1038/ng.3461

Gymrek, M., Golan, D., Rosset, S., & Erlich, Y. (2012). lobSTR: A short tandem repeat profiler for personal genomes. *Genome Research*, *22*(6), 1154–62. https://doi.org/10.1101/gr.135780.111

Gymrek, M. (2016). PCR-free library preparation greatly reduces stutter noise at short tandem repeats. *BioRxiv*, 043448. https://doi.org/10.1101/043448

Haas, J., Barb, I., Katus, H. A., & Meder, B. (2014). Targeted next-generation sequencing: The clinician's stethoscope for genetic disorders. *Personalized Medicine*, *11*(6), 581–92. https://doi.org/10.2217/pme.14.40

Halman, A. (2020a). Source code for the "Accuracy of short tandem repeats genotyping tools in whole exome sequencing data" article. [dataset]. *Zenodo*, V1. https://doi.org/10.5281/zenodo.3695971

Halman, A. (2020b). Supplementary information for the "Accuracy of short tandem repeats genotyping tools in whole exome sequencing data" article [dataset]. *Harvard Dataverse*, V1. https://doi.org/10.7910/DVN/RWTGWK

Hannan, A. J. (2010). Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for "missing heritability." *Trends in Genetics*, *26*(2), 59–65. https://doi.org/10.1016/j.tig.2009.11.008

Hannan, A. J. (2018). Tandem repeats mediating genetic plasticity in health and disease. *Nature Reviews Genetics*, *19*(5), 286–98. https://doi.org/10.1038/nrg.2017.115

Haraksingh, R. R. & Snyder, M. P. (2013). Impacts of variation in the human genome on gene regulation. *Journal of Molecular Biology*, *425*(21), 3970–77. https://doi.org/10.1016/j.jmb.2013.07.015

Hares, D. R. (2015). Selection and implementation of expanded CODIS core loci in the United States. *Forensic Science International: Genetics*, *17*, 33–34. https://doi.org/10.1016/j.fsigen.2015.03.006

Harewood, L. & Fraser, P. (2014). The impact of chromosomal rearrangements on regulation of gene expression. *Human Molecular Genetics*, *23*(R1), 76–82. https://doi.org/10.1093/hmg/ddu278

Harr, B., Todorova, J., & Schlötterer, C. (2002). Mismatch repair-driven mutational bias in D. melanogaster. *Molecular Cell*, *10*(1), 199–205. https://doi.org/10.1016/S1097-2765(02)00575-0

Hause, R. J., Pritchard, C. C., Shendure, J., & Salipante, S. J. (2016). Classification and characterization of microsatellite instability across 18 cancer types. *Nature Medicine*, *22*(11), 1342–50. https://doi.org/10.1038/nm.4191

Hecht, J. T., Nelson, L. D., Crowder, E., Wang, Y., Elder, F. F. B., Harrison, W. R., Francomano, C. A., Prange, C. K., Lennon, G. G., Deere, M., & Lawler, J. (1995). Mutations in exon 17B of cartilage oligomeric matrix protein (COMP) cause pseudoachondroplasia. *Nature Genetics*, *10*(3), 325–29. https://doi.org/10.1038/ng0795-325

Highnam, G., Franck, C., Martin, A., Stephens, C., Puthige, A., & Mittelman, D. (2013). Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Research*, *41*(1), e32–e32. https://doi.org/10.1093/nar/gks981

Hile, S. E. & Eckert, K. A. (2004). Positive correlation between DNA polymerase α-primase pausing and mutagenesis within polypyrimidine/polypurine microsatellite sequences. *Journal of Molecular Biology*, *335*(3), 745–59. https://doi.org/10.1016/j.jmb.2003.10.075

Holmes, S. E., O'hearn, E., Rosenblatt, A., Callahan, C., Hwang, H. S., Ingersoll-Ashworth, R. G., Fleisher, A., Stevanin, G., Brice, A., Potter, N. T., Ross, C. A., & Margolis, R. L. (2001a). A repeat expansion in the gene encoding junctophilin-3 is associated with Huntington disease-like 2. *Nature Genetics*, *29*(4), 377–78. https://doi.org/10.1038/ng760

Holmes, S. E., Hearn, E. O., Ross, C. A., & Margolis, R. L. (2001b). SCA12: An unusual mutation leads to an unusual spinocerebellar ataxia. *Brain Research Bulletin*, *56*(3–4), 397–403. https://doi.org/10.1016/S0361-9230(01)00596-2

Huang, Q. Y., Xu, F. H., Shen, H., Deng, H. Y., Liu, Y. J., Liu, Y. Z., Li, J. L., Recker, R. R., & Deng, H. W. (2002). Mutation patterns at dinucleotide microsatellite loci in humans. *American Journal of Human Genetics*, *70*(3), 625–34. https://doi.org/10.1086/338997

Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2012). ART: A next-generation sequencing read simulator. *Bioinformatics*, *28*(4), 593–94. https://doi.org/10.1093/bioinformatics/btr708

Innis, J. W., Mortlock, D., Chen, Z., Ludwig, M., Williams, M. E., Williams, T. M., Doyle, C. D., Shao, Z., Glynn, M., Mikulic, D., Lehmann, K., Mundlos, S., & Utsch, B. (2004). Polyalanine expansion in HOXA13: Three new affected families and the molecular consequences in a mouse model. *Human Molecular Genetics*, *13*(22), 2841–51. https://doi.org/10.1093/hmg/ddh306

Ishiura, H., Doi, K., Mitsui, J., Yoshimura, J., Matsukawa, M. K., Fujiyama, A., Toyoshima, Y., Kakita, A., Takahashi, H., Suzuki, Y., Sugano, S., Qu, W., Ichikawa, K., Yurino, H., Higasa, K., Shibata, S., Mitsue, A., Tanaka, M., Ichikawa, Y., et al. (2018). Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. *Nature Genetics*, *50*(4), 581–90. https://doi.org/10.1038/s41588-018-0067-2

Ishiura, H., Shibata, S., Yoshimura, J., Suzuki, Y., Qu, W., Doi, K., Almansour, M. A., Kikuchi, J. K., Taira, M., Mitsui, J., Takahashi, Y., Ichikawa, Y., Mano, T., Iwata, A., Harigaya, Y., Matsukawa, M. K., Matsukawa, T., Tanaka, M., Shirota, Y., et al. (2019). Noncoding CGG repeat expansions in neuronal intranuclear inclusion disease, oculopharyngodistal myopathy and an overlapping disease. *Nature Genetics*, *51*(8), 1222–32. https://doi.org/10.1038/s41588-019-0458-z

Ishiura, H. & Tsuji, S. (2020). Advances in repeat expansion diseases and a new concept of repeat motif–phenotype correlation. *Current Opinion in Genetics and*

Development, *65*, 176–85. https://doi.org/10.1016/j.gde.2020.05.029

Jansena, A., Gemayela, R., & Verstrepena, K. J. (2012). Unstable microsatellite repeats facilitate rapid evolution of coding and regulatory sequences. *Repetitive DNA*, *7*, 108–25. https://doi.org/10.1159/000337121

Jaspert, A., Fahsold, R., Grehl, H., & Claus, D. (1995). Myotonic dystrophy: Correlation of clinical symptoms with the size of the CTG trinucleotide repeat. *Journal of Neurology*, *242*(2), 99–104. https://doi.org/10.1007/BF00887824

Jones, L., Houlden, H., & Tabrizi, S. J. (2017). DNA repair in the trinucleotide repeat disorders. *The Lancet Neurology*, *16*(1), 88–96. https://doi.org/10.1016/S1474-4422(16)30350-7

Kaivola, K., Kiviharju, A., Jansson, L., Rantalainen, V., Eriksson, J. G., Strandberg, T. E., Laaksovirta, H., Renton, A. E., Traynor, B. J., Myllykangas, L., & Tienari, P. J. (2019). C9orf72 hexanucleotide repeat length in older population: normal variation and effects on cognition. *Neurobiology of Aging*, *84*, 242.e7-242.e12. https://doi.org/10.1016/j.neurobiolaging.2019.02.026

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–43. https://doi.org/10.1038/s41586-020-2308-7

Kato, M., Saitoh, S., Kamei, A., Shiraishi, H., Ueda, Y., Akasaka, M., Tohyama, J., Akasaka, N., & Hayasaka, K. (2007). A longer polyalanine expansion mutation in the ARX gene causes early infantile epileptic encephalopathy with suppression-burst pattern (Ohtahara syndrome). *American Journal of Human Genetics*, *81*(2), 361–66. https://doi.org/10.1086/518903

Katti, M. V., Ranjekar, P. K., & Gupta, V. S. (2001). Differential Distribution of Simple Sequence Repeats in Eukaryotic Genome Sequences. *Molecular Biology and Evolution*, *18*(7), 1161–67. https://doi.org/10.1093/oxfordjournals.molbev.a003903

Kawaguchi, Y., Okamoto, T., Taniwaki, M., Aizawa, M., Inoue, M., Katayama, S., Kawakami, H., Nakamura, S., Nishimura, M., Akiguchi, I., Kimura, J., Narumiya, S., & Kakizuka, A. (1994). CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. *Nature Genetics*, *8*(3), 221–28. https://doi.org/10.1038/ng1194-221

Kebschull, J. M. & Zador, A. M. (2015). Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Research*, *43*(21), 1–15. https://doi.org/10.1093/nar/gkv717

Kekou, K., Sofocleous, C., Papadimas, G., Petichakis, D., Svingou, M., Pons, R. M., Vorgia, P., Gika, A., Kitsiou-Tzeli, S., & Kanavakis, E. (2016). A dynamic trinucleotide repeat (TNR) expansion in the DMD gene. *Molecular and Cellular Probes*, *30*(4), 254–60. https://doi.org/10.1016/j.mcp.2016.07.001

Kelkar, Y. D., Tyekucheva, S., Chiaromonte, F., & Makova, K. D. (2008). The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Research*, *18*(1), 30–38. https://doi.org/10.1101/gr.7113408

Kennedy, L. & Shelbourne, P. F. (2000). Dramatic mutation instability in HD mouse striatum: does polyglutamine load contribute to cell-specific vulnerability in Huntington's disease? *Human Molecular Genetics*, *9*(17), 2539–44. https://doi.org/10.1093/hmg/9.17.2539

Khristich, A. N. & Mirkin, S. M. (2020). On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. *Journal of Biological Chemistry*, *295*(13), 4134–70. https://doi.org/10.1074/jbc.REV119.007678

Knight, S. J. L., Flannery, A. V., Hirst, M. C., Campbell, L., Christodoulou, Z., Phelps, S. R., Pointon, J., Middleton-Price, H. R., Barnicoat, A., Pembrey, M. E., Holland, J., Oostra, B. A., Bobrow, M., & Davies, K. E. (1993). Trinucleotide repeat amplification and hypermethylation of a CpG island in FRAXE mental retardation. *Cell*, *74*(1), 127–34. https://doi.org/10.1016/0092-8674(93)90300-F

Kobayashi, H., Abe, K., Matsuura, T., Ikeda, Y., Hitomi, T., Akechi, Y., Habu, T., Liu, W., Okuda, H., & Koizumi, A. (2011). Expansion of intronic GGCCTG hexanucleotide repeat in NOP56 causes SCA36, a type of spinocerebellar ataxia accompanied by motor neuron involvement. *American Journal of Human Genetics*, *89*(1), 121–30. https://doi.org/10.1016/j.ajhg.2011.05.015

Koide, R., Kobayashi, S., Shimohata, T., Ikeuchi, T., Maruyama, M., Saito, M., Yamada, M., Takahashi, H., & Tsuji, S. (1999). A neurological disease caused by an expanded CAG trinucleotide repeat in the TATA-binding protein gene: A new polyglutamine disease? *Human Molecular Genetics*, *8*(11), 2047–53. https://doi.org/10.1093/hmg/8.11.2047

Kojima, K., Kawai, Y., Nariai, N., Mimori, T., Hasegawa, T., & Nagasaki, M. (2016). Short tandem repeat number estimation from paired-end reads for multiple individuals by considering coalescent tree. *BMC Genomics*, *17*(Suppl 5), . https://doi.org/10.1186/s12864-016-2821-0

Komure, O., Sano, A., Nishino, N., Yamauchi, N., Ueno, S., Kondoh, K., Sano, N., Takahashi, M., Murayama, N., Kondo, I., Nagafuchi, S., Yamada, M., & Kanazawa, I. (1995). Dna analysis in hereditary dentatorubral-pallidoluysian atrophy: Correlation between cag repeat length and phenotypic variation and the molecular basis of anticipation. *Neurology*, *45*(1), 143–49. https://doi.org/10.1212/WNL.45.1.143

Kozarewa, I., Ning, Z., Quail, M. a, Sanders, M. J., Berriman, M., & Turner, D. J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature Methods*, *6*(4), 291–95. https://doi.org/10.1038/nmeth.1311

Kraus-Perrotta, C. & Lagalwar, S. (2016). Expansion, mosaicism and interruption: mechanisms of the CAG repeat mutation in spinocerebellar ataxia type 1. *Cerebellum & Ataxias*, *3*(1), 20. https://doi.org/10.1186/s40673-016-0058-y

Kristmundsdóttir, S., Sigurpálsdóttir, B. D., Kehr, B., & Halldórsson, B. V. (2017). popSTR: population-scale detection of STR variants. *Bioinformatics (Oxford, England)*, *33*(24), 4041–48. https://doi.org/10.1093/bioinformatics/btw568

Kruglyak, S., Durrett, R. T., Schug, M. D., & Aquadro, C. F. (1998). Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(18), 10774–78. https://doi.org/10.1073/pnas.95.18.10774

La Spada, A. R., Roling, D. B., Harding, A. E., Warner, C. L., Spiegel, R., Hausmanowa-Petrusewicz, I., Yee, W.-C., & Fischbeck, K. H. (1992). Meiotic stability and genotype – phenotype correlation of the trinucleotide repeat in X–linked spinal and bulbar muscular atrophy. *Nature Genetics*, *2*(4), 301–04. https://doi.org/10.1038/ng1292-301

La Spada, A. R. & Taylor, J. P. (2010). Repeat expansion disease: progress and puzzles in disease pathogenesis. *Nature Reviews Genetics*, *11*(4), 247–58. https://doi.org/10.1038/nrg2748

LaCroix, A. J., Stabley, D., Sahraoui, R., Adam, M. P., Mehaffey, M., Kernan, K., Myers, C. T., Fagerstrom, C., Anadiotis, G., Akkari, Y. M., Robbins, K. M., Gripp, K. W., Baratela, W. A. R., Bober, M. B., Duker, A. L., Doherty, D., Dempsey, J. C., Miller, D. G., Kircher, M., et al. (2019). GGC Repeat Expansion and Exon 1 Methylation of XYLT1 Is a Common Pathogenic Variant in Baratela-Scott Syndrome. *American Journal of Human Genetics*, *104*(1), 35–44. https://doi.org/10.1016/j.ajhg.2018.11.005

Lagorio, I., Zara, F., Striano, S., & Striano, P. (2019). Familial adult myoclonic epilepsy: A new expansion repeats disorder. *Seizure*, *67*(March), 73–77. https://doi.org/10.1016/j.seizure.2019.03.009

Lai, Y. & Sun, F. (2003). The Relationship between Microsatellite Slippage Mutation Rate and the Number of Repeat Units. *Molecular Biology and Evolution*, *20*(12), 2123–31. https://doi.org/10.1093/molbev/msg228

Lalioti, M. D., Scott, H. S., Buresi, C., Rossier, C., Bottani, A., Morris, M. A., Malafosse, A., & Antonarakis, S. E. (1997). Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. *Nature*, *386*(6627), 847–51. https://doi.org/10.1038/386847a0

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., McEwan, P., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), . https://doi.org/10.1186/gb-2009-10-3-r25

Lattante, S., Pomponi, M. G., Conte, A., Marangi, G., Bisogni, G., Patanella, A. K., Meleo, E., Lunetta, C., Riva, N., Mosca, L., Carrera, P., Bee, M., Zollino, M., & Sabatelli, M. (2018). ATXN1 intermediate-length polyglutamine expansions are associated with amyotrophic lateral sclerosis. *Neurobiology of Aging*, *64*, 157.e1-157.e5. https://doi.org/10.1016/j.neurobiolaging.2017.11.011

Laumonnier, F., Ronce, N., Hamel, B. C. J., Thomas, P., Lespinasse, J., Raynaud, M., Paringaux, C., Van Bokhoven, H., Kalscheuer, V., Fryns, J. P., Chelly, J., Moraine, C., & Briault, S. (2002). Transcription factor SOX3 is involved in X-linked mental retardation with growth hormone deficiency. *American Journal of Human Genetics*, *71*(6), 1450–55. https://doi.org/10.1086/344661

Lavoie, H., Debeane, F., Trinh, Q. D., Turcotte, J. F., Corbeil-Girard, L. P., Dicaire, M. J., Saint-Denis, A., Pagé, M., Rouleau, G. A., & Brais, B. (2003). Polymorphism, shared functions and convergent evolution of genes with sequences coding for polyalanine domains. *Human Molecular Genetics*, *12*(22), 2967–79. https://doi.org/10.1093/hmg/ddg329

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., & Carey, V. J. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, *9*(8), 1–10. https://doi.org/10.1371/journal.pcbi.1003118

Lebo, M. S., Hao, L., Lin, C. F., & Singh, A. (2020). Bioinformatics in Clinical Genomic Sequencing. *Clinics in Laboratory Medicine*, *40*(2), 163–87. https://doi.org/10.1016/j.cll.2020.02.003

Lèbre, S. & Gascuel, O. (2017). The combinatorics of overlapping genes. *Journal of Theoretical Biology*, *415*(September 2016), 90–101. https://doi.org/10.1016/j.jtbi.2016.09.018

Lee, T. I. & Young, R. A. (2013). Transcriptional Regulation and Its Misregulation in Disease. *Cell*, *152*(6), 1237–51. https://doi.org/10.1016/j.cell.2013.02.014

Leggett, R. M. & Clark, M. D. (2017). A world of opportunities with nanopore sequencing. *Journal of Experimental Botany*, *68*(20), 5419–29. https://doi.org/10.1093/jxb/erx289

Lei, X. X., Liu, Q., Lu, Q., Huang, Y., Zhou, X. Q., Sun, H. Y., Wu, L. W., Cui, L. Y., & Zhang, X. (2019). TTTCA repeat expansion causes familial cortical myoclonic tremor with epilepsy. *European Journal of Neurology*, *26*(3), 513–18. https://doi.org/10.1111/ene.13848

Leopoldino, A. M. & Pena, S. D. J. (2003). The mutational spectrum of human autosomal tetranucleotide microsatellites. *Human Mutation*, *21*(1), 71–79. https://doi.org/10.1002/humu.10153

Lesage, S., Le Ber, I., Condroyer, C., Broussolle, E., Gabelle, A., Thobois, S., Pasquier, F., Mondon, K., Dion, P. A., Rochefort, D., Rouleau, G. A., Dürr, A., & Brice, A. (2013). C9orf72 repeat expansions are a rare genetic cause of parkinsonism. *Brain*, *136*(2), 385–91. https://doi.org/10.1093/brain/aws357

Levinson, G. & Gutman, G. A. (1987). High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in Escherichia coli K-12. *Nucleic Acids Research*, *15*(13), 5323–38. https://doi.org/10.1093/nar/15.13.5323

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Preprint]. 1–3. http://arxiv.org/abs/1303.3997

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–79. https://doi.org/10.1093/bioinformatics/btp352

Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–60. https://doi.org/10.1093/bioinformatics/btp324

Li, Y. C., Korol, A. B., Fahima, T., & Nevo, E. (2004). Microsatellites within genes: Structure, function, and evolution. *Molecular Biology and Evolution*, *21*(6), 991–1007. https://doi.org/10.1093/molbev/msh073

Lin, W. De, Chou, I. C., Wang, C. H., & Tsai, F. J. (2018). Novel mutations in the cartilage oligomeric matrix protein gene identified in two Taiwanese patients with pseudoachondroplasia and multiple epiphyseal dysplasia. *Pediatrics and Neonatology*, *59*(4), 412–14. https://doi.org/10.1016/j.pedneo.2017.10.006

Liquori, C. L. (2001). Myotonic Dystrophy Type 2 Caused by a CCTG Expansion in Intron 1 of ZNF9. *Science*, *293*(5531), 864–67. https://doi.org/10.1126/science.1062125

Liu, Q., Tong, Y., & Wang, K. (2020). Genome-wide detection of short tandem repeat expansions by long-read sequencing. *BMC Bioinformatics*, *21*(Suppl 21), 1–15. https://doi.org/10.1186/s12859-020-03876-w

Liu, S., Hou, W., Sun, T., Xu, Y., Li, P., Yue, B., Fan, Z., & Li, J. (2017). Genome-wide mining and comparative analysis of microsatellites in three macaque species. *Molecular Genetics and Genomics*, *292*(3), 537–50. https://doi.org/10.1007/s00438-017-1289-1

Liu, X., Han, S., Wang, Z., Gelernter, J., & Yang, B. Z. (2013). Variant Callers for Next-Generation Sequencing Data: A Comparison Study. *PLoS ONE*, *8*(9), 1–11. https://doi.org/10.1371/journal.pone.0075619

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, *21*(10), 597–614. https://doi.org/10.1038/s41576-020-0236-x

Maltecca, F., Filla, A., Castaldo, I., Coppola, G., Fragassi, N. A., Carella, M., Bruni, A., Cocozza, S., Casari, G., Servadio, A., & De Michele, G. (2003). Intergenerational instability and marked anticipation in SCA-17. *Neurology*, *61*(10), 1441–43. https://doi.org/10.1212/01.WNL.0000094123.09098.A0

Mantere, T., Kersten, S., & Hoischen, A. (2019). Long-read sequencing emerging in

medical genetics. *Frontiers in Genetics*, *10*(MAY), 1–14. https://doi.org/10.3389/fgene.2019.00426

Mardis, E. R. (2013). Next-Generation Sequencing Platforms. *Annual Review of Analytical Chemistry*, *6*(1), 287–303. https://doi.org/10.1146/annurev-anchem-062012-092628

Margolis, R. L. & Holmes, S. E. (2003). Huntington's disease-like 2: A clinical, pathological, and molecular comparison to Huntington's disease. *Clinical Neuroscience Research*, *3*(3), 187–96. https://doi.org/10.1016/S1566-2772(03)00061-6

Maroilley, T. & Tarailo-Graovac, M. (2019). Uncovering Missing Heritability in Rare Diseases. *Genes*, *10*(4), 1–18. https://doi.org/10.3390/genes10040275

Matsuura, T., Fang, P., Pearson, C. E., Jayakar, P., Ashizawa, T., Roa, B. B., & Nelson, D. L. (2006). Interruptions in the expanded ATTCT repeat of spinocerebellar ataxia type 10: Repeat purity as a disease modifier? *American Journal of Human Genetics*, *78*(1), 125–29. https://doi.org/10.1086/498654

Matsuura, T., Yamagata, T., Burgess, D. L., Rasmussen, A., Grewal, R. P., Watase, K., Khajavi, M., McCall, A. E., Davis, C. F., Zu, L., Achari, M., Pulst, S. M., Alonso, E., Noebels, J. L., Nelson, D. L., Zoghbi, H. Y., & Ashizawa, T. (2000). Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. *Nature Genetics*, *26*(2), 191–94. https://doi.org/10.1038/79911

Mattina, T., Perrotta, C. S., & Grossfeld, P. (2009). Jacobsen syndrome. *Orphanet Journal of Rare Diseases*, *4*(1), 1–10. https://doi.org/10.1186/1750-1172-4-9

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–303. https://doi.org/10.1101/gr.107524.110

McLoughlin, H. S., Moore, L. R., & Paulson, H. L. (2020). Pathogenesis of SCA3 and implications for other polyglutamine diseases. *Neurobiology of Disease*, *134*(July 2019), 104635. https://doi.org/10.1016/j.nbd.2019.104635

McMurray, C. T. (2010). Mechanisms of trinucleotide repeat instability during human development. *Nature Reviews Genetics*, *11*(11), 786–99. https://doi.org/10.1038/nrg2828

McNulty, S. N., Mann, P. R., Robinson, J. A., Duncavage, E. J., & Pfeifer, J. D. (2020). Impact of Reducing DNA Input on Next-Generation Sequencing Library Complexity and Variant Detection. *Journal of Molecular Diagnostics*, *22*(5), 720–27. https://doi.org/10.1016/j.jmoldx.2020.02.003

Menéndez-González, M., Clarimón, J., Rosas-Allende, I., Blázquez, M., San Martín, E. S., García-Fernández, C., Lleó, A., Dols-Icardo, O., Illán-Gala, I., Morís, G., Ribacoba, R., Álvarez, V., & Martínez, C. (2019). HTT gene intermediate alleles in

neurodegeneration: evidence for association with Alzheimer's disease. *Neurobiology of Aging*, *76*, 215.e9-215.e14. https://doi.org/10.1016/j.neurobiolaging.2018.11.014

Metzgar, D., Bytof, J., & Wills, C. (2000). Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Research*, *10*(1), 72–80. https://doi.org/10.1101/gr.10.1.72

Meyer, L. A., Broaddus, R. R., & Lu, K. H. (2009). Endometrial Cancer and Lynch Syndrome : Clinical and Pathologic Considerations. *16*(1), .

Mirkin, S. M. (2007). Expandable DNA repeats and human disease. *Nature*, *447*(7147), 932–40. https://doi.org/10.1038/nature05977

Mishra, S. K., Khosa, S., Trikamji, B., Khanli, H. M., Scheuner, M. T., Jamal, N. I., & Hanssen, A. M. (2019). A case of Fragile X-associated tremor/ataxia syndrome (FXTAS). *Journal of Clinical Neuroscience*, *66*, 269–70. https://doi.org/10.1016/j.jocn.2019.05.044

Mitra, I., Huang, B., Mousavi, N., Ma, N., Lamkin, M., Yanicky, R., Shleizer-burko, S., Lohmueller, K. E., & Gymrek, M. (2021). Patterns of de novo tandem repeat mutations and their role in autism. *Nature*, *589*(February 2020), . https://doi.org/10.1038/s41586-020-03078-7

Molla, M., Delcher, A., Sunyaev, S., Cantor, C., & Kasif, S. (2009). Triplet repeat length bias and variation in the human transcriptome. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(40), 17095–100. https://doi.org/10.1073/pnas.0907112106

Monckton, D. G. (2021). The Contribution of Somatic Expansion of the CAG Repeat to Symptomatic Development in Huntington's Disease: A Historical Perspective. *Journal of Huntington's Disease*, *10*(1), 7–33. https://doi.org/10.3233/JHD-200429

Mortlock, D. P., Sateesh, P., & Innis, J. W. (2000). Evolution of N-terminal sequences of the vertebrate HOXA13 protein. *Mammalian Genome*, *11*(2), 151–58. https://doi.org/10.1007/s003350010029

Mousavi, N., Shleizer-Burko, S., Yanicky, R., & Gymrek, M. (2019). Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Research*, *47*(15), e90–e90. https://doi.org/10.1093/nar/gkz501

Mousavi, N., Margoliash, J., Pusarla, N., Saini, S., Yanicky, R., & Gymrek, M. (2020). TRTools: a toolkit for genome-wide analysis of tandem repeats. *Bioinformatics*, . https://doi.org/10.1093/bioinformatics/btaa736

Mundlos, S., Otto, F., Mundlos, C., Mulliken, J. B., Aylsworth, A. S., Albright, S., Lindhout, D., Cole, W. G., Henn, W., Knoll, J. H. M., Owen, M. J., Mertelsmann, R., Zabel, B. U., & Olsen, B. R. (1997). Mutations involving the transcription factor CBFA1 cause cleidocranial dysplasia. *Cell*, *89*(5), 773–79. https://doi.org/10.1016/S0092-8674(00)80260-3

Murmann, A. E., Yu, J., Opal, P., & Peter, M. E. (2018). Trinucleotide Repeat Expansion Diseases, RNAi, and Cancer. *Trends in Cancer*, *4*(10), 684–700. https://doi.org/10.1016/j.trecan.2018.08.004

Nagafuchi, S., Yanagisawa, H., Sato, K., Shirayama, T., Ohsaki, E., Bundo, M., Takeda, T., Tadokoro, K., Kondo, I., Murayama, N., Tanaka, Y., Kikushima, H., Umino, K., Kurosawa, H., Furukawa, T., Nihei, K., Inoue, T., Sano, A., Komure, O., et al. (1994). Dentatorubral and pallidoluysian atrophy expansion of an unstable CAG trinucleotide on chromosome 12p. *Nature Genetics*, *6*(1), 14–18. https://doi.org/10.1038/ng0194-14

Nageshwaran, S. & Festenstein, R. (2015). Epigenetics and triplet-repeat neurological diseases. *Frontiers in Neurology*, *6*(DEC), 1–9. https://doi.org/10.3389/fneur.2015.00262

Nakamura, K. (2001). SCA17, a novel autosomal dominant cerebellar ataxia caused by an expanded polyglutamine in TATA-binding protein. *Human Molecular Genetics*, *10*(14), 1441–48. https://doi.org/10.1093/hmg/10.14.1441

Nakamura, N., Tsunoda, K., Mitsutake, A., Shibata, S., Mano, T., Nagashima, Y., Ishiura, H., Iwata, A., Toda, T., Tsuji, S., & Sawamura, H. (2020). Clinical Characteristics of Neuronal Intranuclear Inclusion Disease-Related Retinopathy With CGG Repeat Expansions in the NOTCH2NLC Gene. *Investigative Ophthalmology & Visual Science*, *61*(11), 27. https://doi.org/10.1167/iovs.61.11.27

Nallathambi, J., Moumné, L., Baere, E., Beysen, D., Usha, K., Sundaresan, P., & Veitia, R. A. (2007). A novel polyalanine expansion in FOXL2: The first evidence for a recessive form of the blepharophimosis syndrome (BPES) associated with ovarian dysfunction. *Human Genetics*, *121*(1), 107–12. https://doi.org/10.1007/s00439-006-0276-0

Nithianantharajah, J. & Hannan, A. J. (2007). Dynamic mutations as digital genetic modulators of brain development, function and dysfunction. *BioEssays*, *29*(6), 525–35. https://doi.org/10.1002/bies.20589

Ohta, T. & Kimura, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research*, *22*(2), 201–04. https://doi.org/10.1017/S0016672300012994

Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2015). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data.

Oota, S. (2020). Somatic mutations – Evolution within the individual. *Methods*, *176*(October 2019), 91–98. https://doi.org/10.1016/j.ymeth.2019.11.002

Orr, H. T., Chung, M., Banfi, S., Kwiatkowski, T. J., Servadio, A., Beaudet, A. L., McCall, A. E., Duvick, L. A., Ranum, L. P. W., & Zoghbi, H. Y. (1993). Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nature Genetics*, *4*(3), 221–26. https://doi.org/10.1038/ng0793-221

Orr, H. T. (2012). Polyglutamine neurodegeneration: expanded glutamines enhance

native functions. *Current Opinion in Genetics & Development*, *22*(3), 251–55. https://doi.org/10.1016/j.gde.2012.01.001

Orr, H. T. & Zoghbi, H. Y. (2007). Trinucleotide Repeat Disorders. *Annual Review of Neuroscience*, *30*(1), 575–621. https://doi.org/10.1146/annurev.neuro.29.051605.113042

Pandolfo, M. (2001). Molecular basis of Friedreich ataxia. *Movement Disorders*, *16*(5), 815–21. https://doi.org/10.1002/mds.1162

Patra, K. C. & Shirolkar, M. S. (2015). Childhood-onset (Juvenile) Huntingtons disease: A rare case report. *Journal of Pediatric Neurosciences*, *10*(3), 276–79. https://doi.org/10.4103/1817-1745.165709

Payseur, B. A., Jing, P., & Haasl, R. J. (2011). A genomic portrait of human microsatellite variation. *Molecular Biology and Evolution*, *28*(1), 303–12. https://doi.org/10.1093/molbev/msq198

Pereira, R., Oliveira, J., & Sousa, M. (2020). Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics. *Journal of Clinical Medicine*, *9*(1), 132. https://doi.org/10.3390/jcm9010132

Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T., & Sandhu, M. S. (2018). Long reads: their purpose and place. *Human Molecular Genetics*, *27*(R2), R234–41. https://doi.org/10.1093/hmg/ddy177

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Auwera, G. A. Van der, Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., et al. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, 201178. https://doi.org/10.1101/201178

Press, M. O., Carlson, K. D., & Queitsch, C. (2014). The overdue promise of short tandem repeat variation for heritability. *Trends in Genetics*, *30*(11), 504–12. https://doi.org/10.1016/j.tig.2014.07.008

Quilez, J., Guilmatre, A., Garg, P., Highnam, G., Gymrek, M., Erlich, Y., Joshi, R. S., Mittelman, D., & Sharp, A. J. (2016). Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Research*, *44*(8), 3750–62. https://doi.org/10.1093/nar/gkw219

Radvanszky, J., Surovy, M., Polak, E., & Kadasi, L. (2013). Uninterrupted CCTG tracts in the myotonic dystrophy type 2 associated locus. *Neuromuscular Disorders*, *23*(7), 591–98. https://doi.org/10.1016/j.nmd.2013.02.013

Rančelis, T., Aarasimavičius, J., Ambrozaitytė, L., Kavaliauskienė, I., Domarkienė, I., Karčiauskaitė, D., Kučinskienė, Z. A., & Kučinskas, V. (2017). Analysis of pathogenic variants from the ClinVar database in healthy people using next-generation sequencing. *Genetics Research*, *99*(2017), e6. https://doi.org/10.1017/S0016672317000040

Ranum, L. P. W. & Day, J. W. (2002). Dominantly inherited, non-coding microsatellite expansion disorders. *Current Opinion in Genetics and Development*, *12*(3), 266–71. https://doi.org/10.1016/S0959-437X(02)00297-6

Rauch, R., Hofbeck, M., Zweier, C., Koch, A., Zink, S., Trautmann, U., Hoyer, J., Kaulitz, R., Singer, H., & Rauch, A. (2010). Comprehensive genotype-phenotype analysis in 230 patients with tetralogy of Fallot. *Journal of Medical Genetics*, *47*(5), 321–31. https://doi.org/10.1136/jmg.2009.070391

Renton, A. E., Majounie, E., Waite, A., Simón-Sánchez, J., Rollinson, S., Gibbs, J. R., Schymick, J. C., Laaksovirta, H., van Swieten, J. C., Myllykangas, L., Kalimo, H., Paetau, A., Abramzon, Y., Remes, A. M., Kaganovich, A., Scholz, S. W., Duckworth, J., Ding, J., Harmer, D. W., et al. (2011). A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron*, *72*(2), 257–68. https://doi.org/10.1016/j.neuron.2011.09.010

Rhoads, A. & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*, *13*(5), 278–89. https://doi.org/10.1016/j.gpb.2015.08.002

Rosas, I., Martínez, C., Clarimón, J., Lleó, A., Illán-Gala, I., Dols-Icardo, O., Borroni, B., Almeida, M. R., van der Zee, J., Van Broeckhoven, C., Bruni, A. C., Anfossi, M., Bernardi, L., Maletta, R., Serpente, M., Galimberti, D., Scarpini, E., Rossi, G., Caroppo, P., et al. (2020). Role for ATXN1, ATXN2, and HTT intermediate repeats in frontotemporal dementia and Alzheimer's disease. *Neurobiology of Aging*, *87*, 139.e1-139.e7. https://doi.org/10.1016/j.neurobiolaging.2019.10.017

Rüb, U., Schöls, L., Paulson, H., Auburger, G., Kermer, P., Jen, J. C., Seidel, K., Korf, H. W., & Deller, T. (2013). Clinical features, neurogenetics and neuropathology of the polyglutamine spinocerebellar ataxias type 1, 2, 3, 6 and 7. *Progress in Neurobiology*, *104*, 38–66. https://doi.org/10.1016/j.pneurobio.2013.01.001

Rutherford, N. J., Heckman, M. G., DeJesus-Hernandez, M., Baker, M. C., Soto-Ortolaza, A. I., Rayaprolu, S., Stewart, H., Finger, E., Volkening, K., Seeley, W. W., Hatanpaa, K. J., Lomen-Hoerth, C., Kertesz, A., Bigio, E. H., Lippa, C., Knopman, D. S., Kretzschmar, H. A., Neumann, M., Caselli, R. J., et al. (2012). Length of normal alleles of C9ORF72 GGGGCC repeat do not influence disease phenotype. *Neurobiology of Aging*, *33*(12), 2950.e5-2950.e7. https://doi.org/10.1016/j.neurobiolaging.2012.07.005

Salehi, L. B., Bonifazi, E., Di Stasio, E., Gennarelli, M., Botta, A., Vallo, L., Iraci, R., Massa, R., Antonini, G., Angelini, C., & Novelli, G. (2007). Risk prediction for clinical phenotype in myotonic dystrophy type 1: Data from 2,650 patients. *Genetic Testing*, *11*(1), 84–90. https://doi.org/10.1089/gte.2006.0511

Salipante, S. J., Scroggins, S. M., Hampel, H. L., Turner, E. H., & Pritchard, C. C. (2014). Microsatellite instability detection by next generation sequencing. *Clinical Chemistry*, *60*(9), 1192–99. https://doi.org/10.1373/clinchem.2014.223677

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, *74*(12),

5463–67. https://doi.org/10.1073/pnas.74.12.5463

Sato, N., Amino, T., Kobayashi, K., Asakawa, S., Ishiguro, T., Tsunemi, T., Takahashi, M., Matsuura, T., Flanigan, K. M., Iwasaki, S., Ishino, F., Saito, Y., Murayama, S., Yoshida, M., Hashizume, Y., Takahashi, Y., Tsuji, S., Shimizu, N., Toda, T., et al. (2009). Spinocerebellar Ataxia Type 31 Is Associated with "Inserted" Penta-Nucleotide Repeats Containing (TGGAA)n. *American Journal of Human Genetics*, *85*(5), 544–57. https://doi.org/10.1016/j.ajhg.2009.09.019

Sawaya, S., Bagshaw, A., Buschiazzo, E., Kumar, P., Chowdhury, S., Black, M. A., & Gemmell, N. (2013). Microsatellite Tandem Repeats Are Abundant in Human Promoters and Are Associated with Regulatory Elements. *PLoS ONE*, *8*(2), . https://doi.org/10.1371/journal.pone.0054710

Schadt, E. E., Turner, S., & Kasarskis, A. (2011). Corrigendum: A window into third generation sequencing. *Human Molecular Genetics*, *20*(4), 853. https://doi.org/10.1093/hmg/ddq481

Schloss, J. A. (2008). How to get genomes at one ten-thousandth the cost. *Nature Biotechnology*, *26*(10), 1113–15. https://doi.org/10.1038/nbt1008-1113

Schlötterer, C. & Tautz, D. (1992). Slippage synthesis of simple sequence DNA. *Nucleic Acids Research*, *20*(2), 211–15. https://doi.org/10.1093/nar/20.2.211

Schöls, L., Bauer, I., Zühlke, C., Schulte, T., Kölmel, C., Bürk, K., Topka, H., Bauer, P., Przuntek, H., & Riess, O. (2003). Do CTG expansions at the SCA8 locus cause ataxia? *Annals of Neurology*, *54*(1), 110–15. https://doi.org/10.1002/ana.10608

Seixas, A. I., Loureiro, J. R., Costa, C., Ordóñez-Ugalde, A., Marcelino, H., Oliveira, C. L., Loureiro, J. L., Dhingra, A., Brandão, E., Cruz, V. T., Timóteo, A., Quintáns, B., Rouleau, G. A., Rizzu, P., Carracedo, Á., Bessa, J., Heutink, P., Sequeiros, J., Sobrido, M. J., et al. (2017). A Pentanucleotide ATTTC Repeat Insertion in the Non-coding Region of DAB1, Mapping to SCA37, Causes Spinocerebellar Ataxia. *American Journal of Human Genetics*, *101*(1), 87–103. https://doi.org/10.1016/j.ajhg.2017.06.007

Shoubridge, C. & Gecz, J. (2012). Polyalanine tract disorders and neurocognitive phenotypes. *Advances in Experimental Medicine and Biology*, *769*, 185–203. https://doi.org/10.1007/978-1-4614-5434-2_12

Sibly, R. M., Meade, A., Boxall, N., Wilkinson, M. J., Corne, D. W., & Whittaker, J. C. (2003). The structure of interrupted human AC microsatellites. *Molecular Biology and Evolution*, *20*(3), 453–59. https://doi.org/10.1093/molbev/msg056

Sivan, Y., Zhou, A., Jennings, L. J., Berry-Kravis, E. M., Yu, M., Zhou, L., Rand, C. M., & Weese-Mayer, D. E. (2019). Congenital central hypoventilation syndrome: Severe disease caused by co-occurrence of two PHOX2B variants inherited separately from asymptomatic family members. *American Journal of Medical Genetics, Part A*, *179*(3), 503–06. https://doi.org/10.1002/ajmg.a.61047

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner,

C., Kent, S. B. H., & Hood, L. E. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature*, *321*(6071), 674–79. https://doi.org/10.1038/321674a0

Sone, J., Mori, K., Inagaki, T., Katsumata, R., Takagi, S., Yokoi, S., Araki, K., Kato, T., Nakamura, T., Koike, H., Takashima, H., Hashiguchi, A., Kohno, Y., Kurashige, T., Kuriyama, M., Takiyama, Y., Tsuchiya, M., Kitagawa, N., Kawamoto, M., et al. (2016). Clinicopathological features of adult-onset neuronal intranuclear inclusion disease. *Brain*, *139*(12), 3170–86. https://doi.org/10.1093/brain/aww249

Sone, J., Mitsuhashi, S., Fujita, A., Mizuguchi, T., Hamanaka, K., Mori, K., Koike, H., Hashiguchi, A., Takashima, H., Sugiyama, H., Kohno, Y., Takiyama, Y., Maeda, K., Doi, H., Koyano, S., Takeuchi, H., Kawamoto, M., Kohara, N., Ando, T., et al. (2019). Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. *Nature Genetics*, *51*(8), 1215–21. https://doi.org/10.1038/s41588-019-0459-y

Spada, A. R. La, Wilson, E. M., Lubahn, D. B., Harding, A. E., & Fischbeck, K. H. (1991). Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature*, *352*(6330), 77–79. https://doi.org/10.1038/352077a0

Srivastava, A. K., Choudhry, S., Gopinath, M. S., Roy, S., Tripathi, M., Brahmachari, S. K., & Jain, S. (2001). Molecular and clinical correlation in five Indian families with spinocerebellar ataxia 12. *Annals of Neurology*, *50*(6), 796–800. https://doi.org/10.1002/ana.10048

Stevanin, G., Giunti, P., David, G., Belal, S., Dürr, A., Ruberg, M., Wood, N., & Brice, A. (1998). De novo expansion of intermediate alleles in spinocerebellar ataxia 7. *Human Molecular Genetics*, *7*(11), 1809–13. https://doi.org/10.1093/hmg/7.11.1809

Strand, M., Prolla, T. A., Liskay, R. M., & Petes, T. D. (1993). Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature*, *365*(6443), 274–76. https://doi.org/10.1038/365274a0

Strømme, P., Mangelsdorf, M. E., Shaw, M. A., Lower, K. M., Lewis, S. M. E., Bruyere, H., Lütcherath, V., Gedeon, Á. K., Wallace, R. H., Scheffer, I. E., Turner, G., Partington, M., Frints, S. G. M., Fryns, J. P., Sutherland, G. R., Mulley, J. C., & Gécz, J. (2002). Mutations in the human ortholog of Aristaless cause X-linked mental retardation and epilepsy. *Nature Genetics*, *30*(4), 441–45. https://doi.org/10.1038/ng862

Subramanian, S., Mishra, R. K., & Singh, L. (2003). Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biology*, *4*(2), R13. https://doi.org/10.1186/gb-2003-4-2-r13

Sugiyama, A., Sato, N., Kimura, Y., Fujii, H., Shigemoto, Y., Suzuki, F., Tanei, Z. ichi, Saito, Y., Sasaki, M., Takahashi, Y., Matsuda, H., & Kuwabara, S. (2020). The cerebellar white matter lesions in dentatorubral-pallidoluysian atrophy. *Journal of the Neurological Sciences*, *416*(April), 117040. https://doi.org/10.1016/j.jns.2020.117040

Sun, Q. Y., Xu, Q., Tian, Y., Hu, Z. M., Qin, L. X., Yang, J. X., Huang, W., Xue, J., Li, J. C., Zeng, S., Wang, Y., Min, H. X., Chen, X. Y., Wang, J. P., Xie, B., Liang, F., Zhang, H. N., Wang, C. Y., Lei, L. F., et al. (2020). Expansion of GGC repeat in the human-specific NOTCH2NLC gene is associated with essential tremor. *Brain*, *143*(1), 222–33. https://doi.org/10.1093/brain/awz372

Swami, M., Hendricks, A. E., Gillis, T., Massood, T., Mysore, J., Myers, R. H., & Wheeler, V. C. (2009). Somatic expansion of the Huntington's disease CAG repeat in the brain is associated with an earlier age of disease onset. *Human Molecular Genetics*, *18*(16), 3039–47. https://doi.org/10.1093/hmg/ddp242

Tahmasebi, S., Khoutorsky, A., Mathews, M. B., & Sonenberg, N. (2018). Translation deregulation in human disease. *Nature Reviews Molecular Cell Biology*, *19*(12), 791–807. https://doi.org/10.1038/s41580-018-0034-x

Tang, H., Kirkness, E. F., Lippert, C., Biggs, W. H., Fabani, M., Guzman, E., Ramakrishnan, S., Lavrenko, V., Kakaradov, B., Hou, C., Hicks, B., Heckerman, D., Och, F. J., Caskey, C. T., Venter, J. C., & Telenti, A. (2017). Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. *The American Journal of Human Genetics*, *101*(5), 700–15. https://doi.org/10.1016/j.ajhg.2017.09.013

Tang, H. & Nzabarushimana, E. (2017). STRScan: Targeted profiling of short tandem repeats in whole-genome sequencing data. *BMC Bioinformatics*, *18*(S11), 398. https://doi.org/10.1186/s12859-017-1800-z

Tankard, R. M., Bennett, M. F., Degorski, P., Delatycki, M. B., Lockhart, P. J., & Bahlo, M. (2018). Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data. *The American Journal of Human Genetics*, *103*(6), 858–73. https://doi.org/10.1016/j.ajhg.2018.10.015

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., & Prins, P. (2015). Sambamba: Fast processing of NGS alignment formats. *Bioinformatics*, *31*(12), 2032–34. https://doi.org/10.1093/bioinformatics/btv098

Tautz, D. (1989). Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Research*, *17*(16), 6463–71. https://doi.org/10.1093/nar/17.16.6463

Teive, H. A. G., Munhoz, R. P., Arruda, W. O., Raskin, S., Werneck, L. C., & Ashizawa, T. (2011). Spinocerebellar ataxia type 10 - A review. *Parkinsonism and Related Disorders*, *17*(9), 655–61. https://doi.org/10.1016/j.parkreldis.2011.04.001

Testa, C. M. & Jankovic, J. (2019). Huntington disease: A quarter century of progress since the gene discovery. *Journal of the Neurological Sciences*, *396*(April 2018), 52–68. https://doi.org/10.1016/j.jns.2018.09.022

Treangen, T. J. & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*, *46*(1), 36–46. https://doi.org/10.1038/nrg3164

Twerdi, C. D., Boyer, J. C., & Farber, R. A. (1999). Relative rates of insertion and deletion mutations in a microsatellite sequence in cultured cells. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(6), 2875–79. https://doi.org/10.1073/pnas.96.6.2875

Unseld, M., Marienfeld, J., Brandt, P., & Brennicke, A. (1997). Instability of highly expanded CAG repeats in mice transgenic for the Huntington's disease mutation. *Nature Genetics*, *15*, 57–61. https://doi.org/10.1038/ng0797-270

Utsch, B., Becker, K., Brock, D., Lentze, M. J., Bidlingmaier, F., & Ludwig, M. (2002). A novel stable polyalanine [poly(A)] expansion in the HOXA13 gene associated with hand-foot-genital syndrome: proper function of poly(A)-harbouring transcription factors depends on a critical repeat length? *Human Genetics*, *110*(5), 488–94. https://doi.org/10.1007/s00439-002-0712-8

van Blitterswijk, M., Mullen, B., Heckman, M. G., Baker, M. C., DeJesus-Hernandez, M., Brown, P. H., Murray, M. E., Hsiung, G. Y. R., Stewart, H., Karydas, A. M., Finger, E., Kertesz, A., Bigio, E. H., Weintraub, S., Mesulam, M., Hatanpaa, K. J., White, C. L., Neumann, M., Strong, M. J., et al. (2014). Ataxin-2 as potential disease modifier in C9ORF72 expansion carriers. *Neurobiology of Aging*, *35*(10), . https://doi.org/10.1016/j.neurobiolaging.2014.04.016

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends in Genetics*, *34*(9), 666–81. https://doi.org/10.1016/j.tig.2018.05.008

van Kuilenburg, A. B. P., Tarailo-Graovac, M., Richmond, P. A., Drögemöller, B. I., Pouladi, M. A., Leen, R., Brand-Arzamendi, K., Dobritzsch, D., Dolzhenko, E., Eberle, M. A., Hayward, B., Jones, M. J., Karbassi, F., Kobor, M. S., Koster, J., Kumari, D., Li, M., MacIsaac, J., McDonald, C., et al. (2019). Glutaminase Deficiency Caused by Short Tandem Repeat Expansion in GLS. *New England Journal of Medicine*, *380*(15), 1433–41. https://doi.org/10.1056/NEJMoa1806627

Van Mossevelde, S., Engelborghs, S., Van Der Zee, J., & Van Broeckhoven, C. (2018). Genotype-phenotype links in frontotemporal lobar degeneration. *Nature Reviews Neurology*, *14*(6), 363–78. https://doi.org/10.1038/s41582-018-0009-8

Van Mossevelde, S., van der Zee, J., Cruts, M., & Van Broeckhoven, C. (2017). Relationship between C9orf72 repeat size and clinical phenotype. *Current Opinion in Genetics & Development*, *44*, 117–24. https://doi.org/10.1016/j.gde.2017.02.008

Wang, Q., Pierce-Hoffman, E., Cummings, B. B., Alföldi, J., Francioli, L. C., Gauthier, L. D., Hill, A. J., O'Donnell-Luria, A. H., Armean, I. M., Banks, E., Bergelson, L., Cibulskis, K., Collins, R. L., Connolly, K. M., Covarrubias, M., Daly, M. J., Donnelly, S., Farjoun, Y., Ferriera, S., et al. (2020). Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. *Nature Communications*, *11*(1), 1–13. https://doi.org/10.1038/s41467-019-12438-5

Watson, J. D. & Crick, F. H. C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, *171*(4356), 737–38. https://doi.org/10.1038/171737a0

Wei, X. & Zhang, J. (2014). A simple method for estimating the strength of natural selection on overlapping genes. *Genome Biology and Evolution*, *7*(1), 381–90. https://doi.org/10.1093/gbe/evu294

Wessels, M. W., Kuchinka, B., Heydanus, R., Smit, B. J., Dooijes, D., De Krijger, R. R., Lequin, M. H., De Jong, E. M., Husen, M., Willems, P. J., & Casey, B. (2010). Polyalanine expansion in the ZIC3 gene leading to X-linked heterotaxy with VACTERL association: A new polyalanine disorder? *Journal of Medical Genetics*, *47*(5), 351–55. https://doi.org/10.1136/jmg.2008.060913

Wieben, E. D., Aleff, R. A., Tosakulwong, N., Butz, M. L., Highsmith, W. E., Edwards, A. O., & Baratz, K. H. (2012). A Common Trinucleotide Repeat Expansion within the Transcription Factor 4 (TCF4, E2-2) Gene Predicts Fuchs Corneal Dystrophy. *PLoS ONE*, *7*(11), 5–12. https://doi.org/10.1371/journal.pone.0049083

Wierdl, M., Dominska, M., & Petes, T. D. (1997). Microsatellite instability in yeast: Dependence on the length of the microsatellite. *Genetics*, *146*(3), 769–79.

Willems, T., Zielinski, D., Yuan, J., Gordon, A., Gymrek, M., & Erlich, Y. (2017). Genome-wide profiling of heritable and de novo STR variations. *Nature Methods*, *14*(6), 590–92. https://doi.org/10.1038/nmeth.4267

Willems, T., Gymrek, M., Highnam, G., Mittelman, D., & Erlich, Y. (2014). The landscape of human STR variation. *Genome Research*, *24*(11), 1894–904. https://doi.org/10.1101/gr.177774.114

Winnepenninckx, B., Debacker, K., Ramsay, J., Smeets, D., Smits, A., FitzPatrick, D. R., & Kooy, R. F. (2007). CGG-repeat expansion in the DIP2B gene is associated with the fragile site FRA12A on chromosome 12q13.1. *American Journal of Human Genetics*, *80*(2), 221–31. https://doi.org/10.1086/510800

Woerner, S. M., Yuan, Y. P., Benner, A., Korff, S., Von Knebel Doeberitz, M., & Bork, P. (2009). SelTarbase, a database of human mononucleotide-microsatellite mutations and their potential impact to tumorigenesis and immunology. *Nucleic Acids Research*, *38*(SUPPL.1), 682–89. https://doi.org/10.1093/nar/gkp839

Woods, K. S., Cundall, M., Turton, J., Rizotti, K., Mehta, A., Palmer, R., Wong, J., Chong, W. K., Al-Zyoud, M., El-Ali, M., Otonkoski, T., Martinez-Barbera, J. P., Thomas, P. Q., Robinson, I. C., Lovell-Badge, R., Woodward, K. J., & Dattani, M. T. (2005). Over- and underdosage of SOX3 is associated with infundibular hypoplasia and hypopituitarism. *American Journal of Human Genetics*, *76*(5), 833–49. https://doi.org/10.1086/430134

Woollacott, I. O. C. & Mead, S. (2014). The C9ORF72 expansion mutation: Gene structure, phenotypic and diagnostic issues. *Acta Neuropathologica*, *127*(3), 319–32. https://doi.org/10.1007/s00401-014-1253-7

Wright, G. E. B., Collins, J. A., Kay, C., McDonald, C., Dolzhenko, E., Xia, Q., Bečanović, K., Drögemöller, B. I., Semaka, A., Nguyen, C. M., Trost, B., Richards, F., Bijlsma, E. K., Squitieri, F., Ross, C. J. D., Scherer, S. W., Eberle, M. A., Yuen, R. K. C., & Hayden, M. R. (2019). Length of Uninterrupted CAG, Independent of

Polyglutamine Size, Results in Increased Somatic Instability, Hastening Onset of Huntington Disease. *American Journal of Human Genetics*, *104*(6), 1116–26. https://doi.org/10.1016/j.ajhg.2019.04.007

Xin, Q., Li, L., Li, J., Qiu, R., Guo, C., Gong, Y., & Liu, Q. (2012). Eight-alanine duplication in homeobox D13 in a Chinese family with synpolydactyly. *Gene*, *499*(1), 48–51. https://doi.org/10.1016/j.gene.2012.02.046

Xu, L., Haasl, R. J., Sun, J., Zhou, Y., Bickhart, D. M., Li, J., Song, J., Sonstegard, T. S., Van Tassell, C. P., Lewin, H. A., & Liu, G. E. (2016). Systematic profiling of short tandem repeats in the cattle genome Genome Biology and Evolution Advance Access. *Genome Biol. Evol.*, *20*, 20–31. https://doi.org/10.1093/gbe/evw256

Xu, Xin, Peng, M., Fang, Z., & Xu, Xiping. (2000). The direction of microsatellite mutations is dependent upon allele length. *Nature Genetics*, *24*(4), 396–99. https://doi.org/10.1038/74238

Yeetong, P., Pongpanich, M., Srichomthong, C., Assawapitaksakul, A., Shotelersuk, Varote, Tantirukdham, N., Chunharas, C., Suphapeetiporn, K., & Shotelersuk, Vorasuk. (2019). TTTCA repeat insertions in an intron of YEATS2 in benign adult familial myoclonic epilepsy type 4. *Brain*, *142*(11), 3360–66. https://doi.org/10.1093/brain/awz267

Zarouchlioti, C., Sanchez-Pintado, B., Hafford Tear, N. J., Klein, P., Liskova, P., Dulla, K., Semo, M., Vugler, A. A., Muthusamy, K., Dudakova, L., Levis, H. J., Skalicka, P., Hysi, P., Cheetham, M. E., Tuft, S. J., Adamson, P., Hardcastle, A. J., & Davidson, A. E. (2018). Antisense Therapy for a Common Corneal Dystrophy Ameliorates TCF4 Repeat Expansion-Mediated Toxicity. *The American Journal of Human Genetics*, *102*(4), 528–39. https://doi.org/10.1016/j.ajhg.2018.02.010

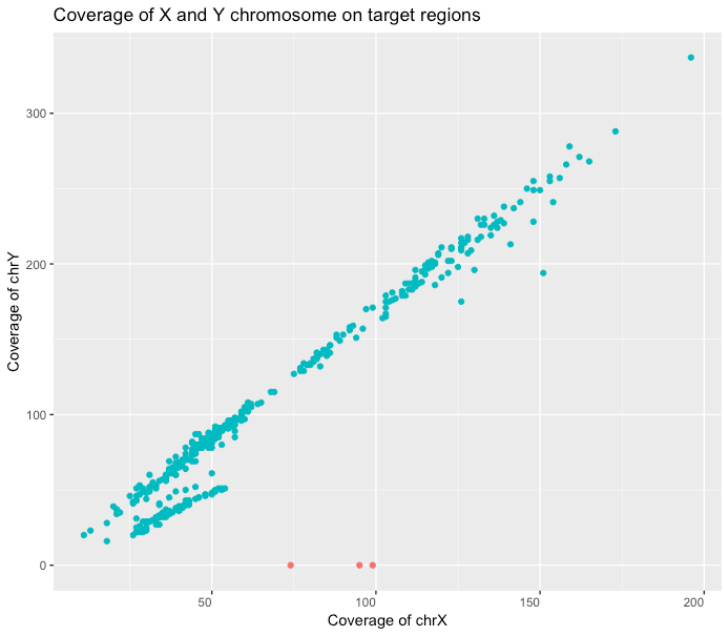
Zeman, A., Stone, J., Porteous, M., Burns, E., Barron, L., & Warner, J. (2004). Spinocerebellar ataxia type 8 in Scotland: Genetic and clinical features in seven unrelated cases and a review of published reports. *Journal of Neurology, Neurosurgery and Psychiatry*, *75*(3), 459–65. https://doi.org/10.1136/jnnp.2003.018895

Zhu, L., Li, Z., Wang, Y., Zhang, C., Liu, Y., & Qu, X. (2015). Microsatellite instability and survival in gastric cancer: A systematic review and meta-analysis. *Molecular and Clinical Oncology*, *3*(3), 699–705. https://doi.org/10.3892/mco.2015.506

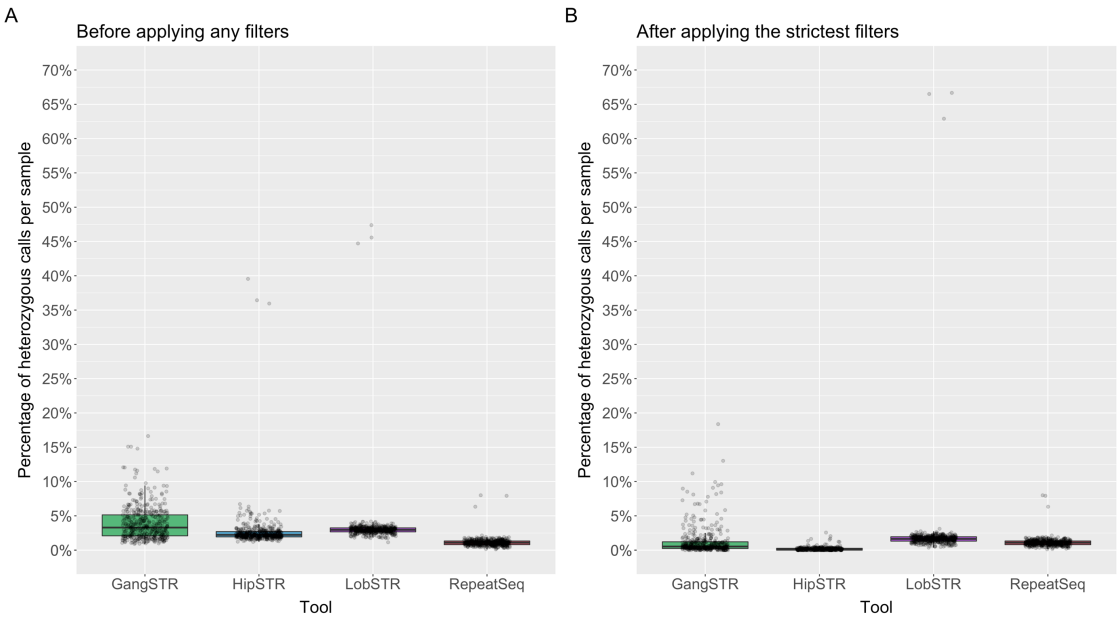
Zoghbi, H. Y. & Orr, H. T. (2000). Glutamine Repeats and Neurodegeneration. *Annual Review of Neuroscience*, *23*(1), 217–47. https://doi.org/10.1146/annurev.neuro.23.1.217

Appendices

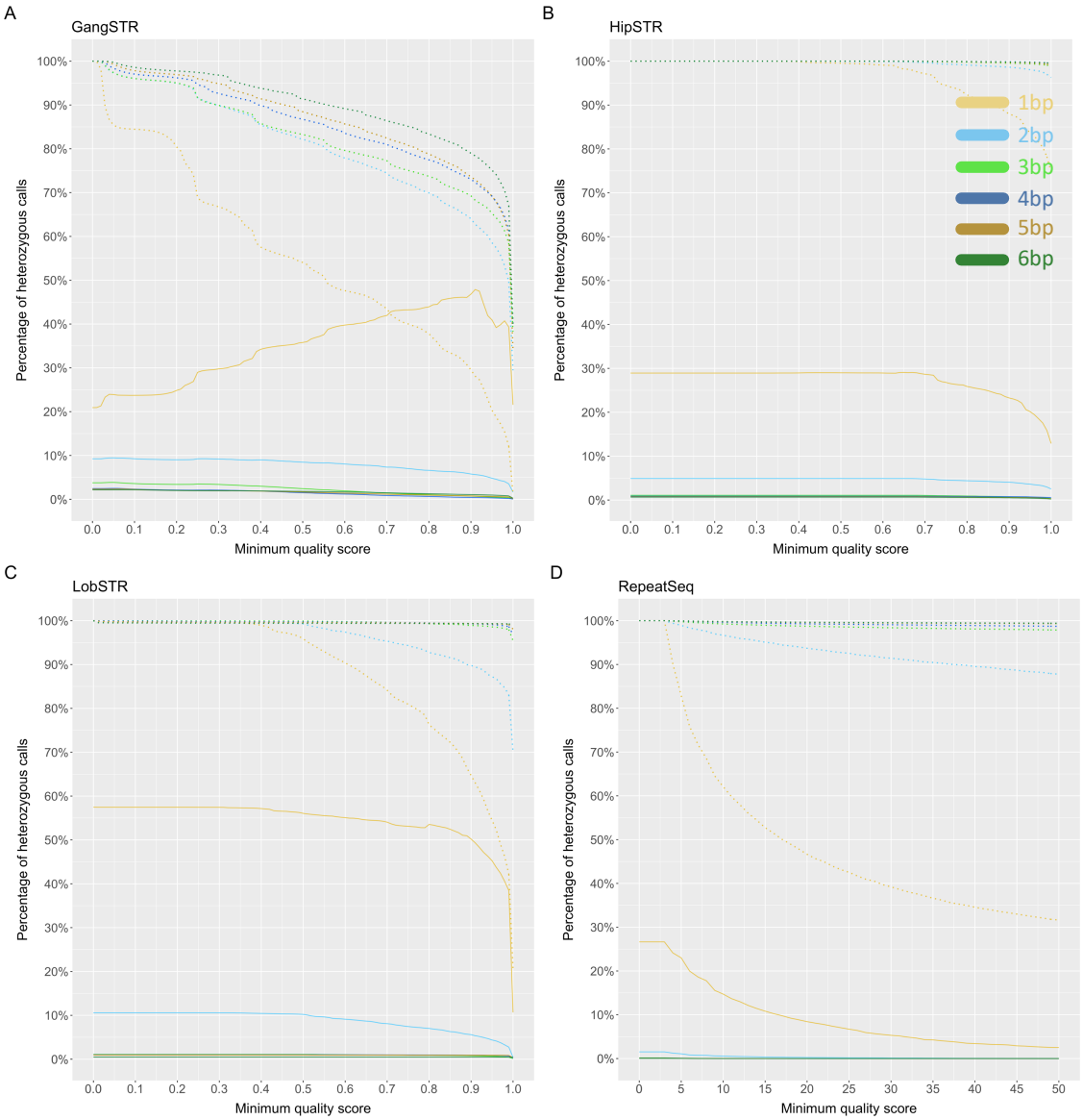
Appendix A: Supplementary information for Chapter 2



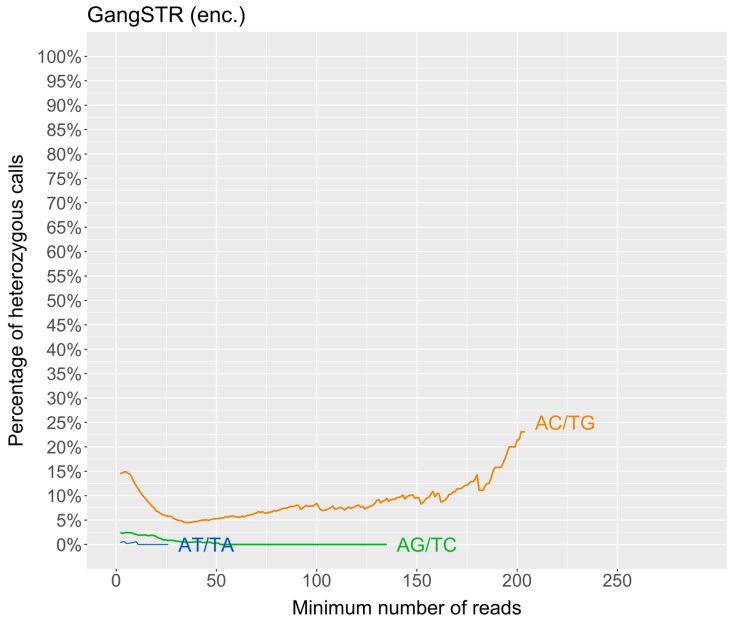
Supplemental Figure 1. Coverage of X and Y chromosome of samples marked as male in the metadata.



Supplemental Figure 2. (A) Percentage of heterozygous calls over all samples (each dot is a sample) - no filters applied. (B) Percentage of heterozygous calls over all samples after applied the strictest recommended filters for the tools. Since no filters were recommended for RepeatSeq then it has the same values on both plots.



Supplemental Figure 3. Percentage of heterozygous calls as a function of minimum quality score. (A) GangSTR, (B) HipSTR, (C) LobSTR, (D) RepeatSeq. Solid line represents the percentage of heterozygous calls as a function of minimum quality score and dotted line the percentage of remaining calls as a function of minimum quality score. Dotted line represents the percentage of remaining calls as a function of minimum number of reads. Heterozygous calls are represented in percentages, but the total number of calls is different for each tool, where 100% is 796775 calls for GangSTR, 757432 calls for HipSTR, 848252 calls for LobSTR and 775030 calls for RepeatSeq.



Supplemental Figure 4. Percentage of heterozygous calls per minimum number of reads (dinucleotides).

Supplemental Table 1. List of samples included in the analysis.

SRR1272193	SRR1272230	SRR1272231	SRR1272233	SRR1272235	SRR1272237
SRR1272240	SRR1272241	SRR1272242	SRR1272244	SRR1272245	SRR1272246
SRR1272248	SRR1272250	SRR1272251	SRR1272253	SRR1272254	SRR1272255
SRR1272257	SRR1272258	SRR1272259	SRR1272263	SRR1272267	SRR1272271
SRR1272274	SRR1272275	SRR1272278	SRR1272280	SRR1272282	SRR1272284
SRR1272285	SRR1272287	SRR1272288	SRR1272289	SRR1272291	SRR1272293
SRR1272295	SRR1272296	SRR1301226	SRR1301227	SRR1301228	SRR1301229
SRR1301230	SRR1301232	SRR1301233	SRR1301234	SRR1301236	SRR1301237
SRR1301238	SRR1301240	SRR1301241	SRR1301242	SRR1301245	SRR1301246
SRR1301248	SRR1301250	SRR1301252	SRR1301253	SRR1301254	SRR1301256
SRR1301257	SRR1301258	SRR1301262	SRR1301264	SRR1301266	SRR1301270
SRR1301272	SRR1301273	SRR1301274	SRR1301276	SRR1301277	SRR1301280
SRR1301281	SRR1301283	SRR1301284	SRR1301286	SRR1301287	SRR1301289
SRR1301290	SRR1301291	SRR1301293	SRR1301295	SRR1301297	SRR1301299
SRR1301303	SRR1301307	SRR1301310	SRR1301313	SRR1301314	SRR1301316
SRR1301318	SRR1301320	SRR1301322	SRR1301324	SRR1301326	SRR1301328
SRR1301329	SRR1301330	SRR1301332	SRR1301334	SRR1301336	SRR1301337
SRR1301338	SRR1301340	SRR1301341	SRR1301342	SRR1301344	SRR1301345
SRR1301346	SRR1301348	SRR1301350	SRR1301352	SRR1301353	SRR1301354
SRR1301356	SRR1301357	SRR1301359	SRR1301361	SRR1301362	SRR1301363
SRR1301364	SRR1301365	SRR1301366	SRR1301368	SRR1301370	SRR1301372
SRR1301373	SRR1301374	SRR1301376	SRR1301378	SRR1301380	SRR1301382
SRR1301384	SRR1301385	SRR1301387	SRR1301389	SRR1301392	SRR1301394
SRR1301395	SRR1301396	SRR1301398	SRR1301399	SRR1301400	SRR1301403
SRR1301406	SRR1301407	SRR1301409	SRR1301411	SRR1301413	SRR1301414
SRR1301416	SRR1301417	SRR1301418	SRR1301422	SRR1301424	SRR1301425
SRR1301426	SRR1301430	SRR1301432	SRR1301434	SRR1301436	SRR1301438
SRR1301440	SRR1301441	SRR1301442	SRR1301444	SRR1301446	SRR1301448
SRR1301450	SRR1301452	SRR1301454	SRR1301457	SRR1301459	SRR1301460
SRR1301461	SRR1301463	SRR1301465	SRR1301468	SRR1301470	SRR1301471
SRR1301475	SRR1301477	SRR1301479	SRR1301481	SRR1301483	SRR1301485
SRR1301486	SRR1301487	SRR1301489	SRR1301491	SRR1301493	SRR1301495
SRR1301498	SRR1301500	SRR1301502	SRR1301505	SRR1301507	SRR1301508
SRR1301509	SRR1301511	SRR1301512	SRR1301513	SRR1301515	SRR1301516

SRR1301517 SRR1301520 SRR1301521 SRR1301523 SRR1301524 SRR1301525
SRR1301527 SRR1301529 SRR1301531 SRR1301533 SRR1301535 SRR1301537
SRR1301539 SRR1301541 SRR1301543 SRR1301544 SRR1301545 SRR1301547
SRR1301549 SRR1301551 SRR1301552 SRR1301555 SRR1301557 SRR1301558
SRR1301559 SRR1301561 SRR1301563 SRR1301565 SRR1301566 SRR1301567
SRR1301570 SRR1301571 SRR1301573 SRR1301574 SRR1301575 SRR1301577
SRR1301578 SRR1301579 SRR1301581 SRR1301583 SRR1301585 SRR1301587
SRR1301590 SRR1301591 SRR1301593 SRR1301595 SRR1301597 SRR1301599
SRR1301601 SRR1301603 SRR1301607 SRR1301611 SRR1301613 SRR1301615
SRR1301617 SRR1301618 SRR1301619 SRR1301621 SRR1301623 SRR1301625
SRR1301627 SRR1301629 SRR1301631 SRR1301633 SRR1301634 SRR1301635
SRR1301639 SRR1301641 SRR1301644 SRR1301645 SRR1301647 SRR1301649
SRR1301651 SRR1301652 SRR1301653 SRR1301655 SRR1301657 SRR1301659
SRR1301661 SRR1301663 SRR1301664 SRR1301665 SRR1301669 SRR1301673
SRR1301675 SRR1301677 SRR1301679 SRR1301680 SRR1301681 SRR1301683
SRR1301685 SRR1301688 SRR1301691 SRR1301692 SRR1301694 SRR1301695
SRR1301697 SRR1301698 SRR1301700 SRR1301701 SRR1301702 SRR1301704
SRR1301705 SRR1301706 SRR1301709 SRR1301711 SRR1301712 SRR1301714
SRR1301717 SRR1301718 SRR1301721 SRR1301722 SRR1301724 SRR1301725
SRR1301727 SRR1301728 SRR1301729 SRR1301732 SRR1301733 SRR1301736
SRR1301738 SRR1301740 SRR1301742 SRR1301743 SRR1301744 SRR1301748
SRR1301750 SRR1301752 SRR1301758 SRR1301759 SRR1301761 SRR1301762
SRR1301763 SRR1301765 SRR1301767 SRR1301769 SRR1301770 SRR1301771
SRR1301773 SRR1301775 SRR1301777 SRR1301778 SRR1301782 SRR1301783
SRR1301786 SRR1301788 SRR1301789 SRR1301790 SRR1301794 SRR1301796
SRR1301797 SRR1301801 SRR1301803 SRR1301805 SRR1301807 SRR1301809
SRR1301811 SRR1301812 SRR1301813 SRR1301815 SRR1301817 SRR1301819
SRR1301821 SRR1301825 SRR1301827 SRR1301828 SRR1301829 SRR1301833
SRR1301835 SRR1301837 SRR1301840 SRR1301841 SRR1301845 SRR1301848
SRR1301849 SRR1301851 SRR1301853 SRR1301856 SRR1301858 SRR1301859
SRR1301860 SRR1301863 SRR1301864 SRR1301867 SRR1301868 SRR1301872
SRR1301876 SRR1301880 SRR1301883 SRR1301885 SRR1301886 SRR1301887
SRR1301889 SRR1301891 SRR1301894 SRR1301898 SRR1301902 SRR1301904
SRR1301906 SRR1301909 SRR1301910 SRR1301914 SRR1301916 SRR1301918
SRR1301921 SRR1301922 SRR1301926 SRR1301929 SRR1301930 SRR1301934
SRR1301938 SRR1301940 SRR1301942 SRR1301945 SRR1301947 SRR1301948
SRR1301949 SRR1301951 SRR1301952 SRR1313245 SRR1313246 SRR1313247
SRR1313250 SRR1313252 SRR1313254 SRR1313256 SRR1515929 SRR1515931
SRR1515933