University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Gusev, A;Shi, H;Kichaev, G;Pomerantz, M;Li, F;Long, HW;Ingles, SA;Kittles, RA;Strom, SS;Rybicki, BA;Nemesure, B;Isaacs, WB;Zheng, W;Pettaway, CA;Yeboah, ED;Tettey, Y;Biritwum, RB;Adjei, AA;Tay, E;Truelove, A;Niwa, S;Chokkalingam, AP;John, EM;Murphy, AB;Signorello, LB;Carpten, J;Leske, MC;Wu, SY;Hennis, AJM;Neslund-Dudas, C;Hsing, AW;Chu, L;Goodman, PJ;Klein, EA;Witte, JS;Casey, G;Kaggwa, S;Cook, MB;Stram, DO;Blot, WJ;Eeles, RA;Easton, D;Kote-Jarai, Z;Al Olama, AA;Benlloch, S;Muir, K;Giles, GG;Southey, MC;Fitzgerald, LM;Gronberg, H;Wiklund, F;Aly, M;Henderson, BE;Schleutker, J;Wahlfors, T;Tammela, TLJ;Nordestgaard, BG;Key, TJ;Travis, RC;Neal, DE;Donovan, JL;Hamdy, FC;Pharoah, P;Pashayan, N;Khaw, KT;Stanford, JL;Thibodeau, SN;McDonnell, SK;Schaid, DJ;Maier, C;Vogel, W;Luedeke, M;Herkommer, K;Kibel, AS;Cybulski, C;Wokolorczyk, D;Kluzniak, W;Cannon-Albright, L;Teerlink, C;Brenner, H;Dieffenbach, AK;Arndt, V;Park, JY;Sellers, TA;Lin, HY;Slavov, C;Kaneva, R;Mitev, V;Batra, J;Spurdle, A;Clements, JA;Teixeira, MR;Pandha, H;Michael, A;Paulo, P;Maia, S;Kierzek, A;Conti, DV;Albanes, D;Berg, C

Title:
Atlas of prostate cancer heritability in European and African-American men pinpoints tissue-specific regulation

Date:
2016-04-07

Citation:
Gusev, A., Shi, H., Kichaev, G., Pomerantz, M., Li, F., Long, H. W., Ingles, S. A., Kittles, R. A., Strom, S. S., Rybicki, B. A., Nemesure, B., Isaacs, W. B., Zheng, W., Pettaway, C. A., Yeboah, E. D., Tettey, Y., Biritwum, R. B., Adjei, A. A., Tay, E. ,... Berg, C. (2016). Atlas of prostate cancer heritability in European and African-American men pinpoints tissue-specific regulation. Nature Communications, 7 (1), https://doi.org/10.1038/ncomms10979.

Persistent Link:
https://hdl.handle.net/11343/262290

License:
CC BY

nature COMMUNICATIONS

ARTICLE

Received 26 Dec 2014 | Accepted 3 Feb 2016 | Published 7 Apr 2016

DOI: 10.1038/ncomms10979 OPEN

# Atlas of prostate cancer heritability in European and African-American men pinpoints tissue-specific regulation

Alexander Gusev *et al.*#

Although genome-wide association studies have identified over 100 risk loci that explain ~33% of familial risk for prostate cancer (PrCa), their functional effects on risk remain largely unknown. Here we use genotype data from 59,089 men of European and African American ancestries combined with cell-type-specific epigenetic data to build a genomic atlas of single-nucleotide polymorphism (SNP) heritability in PrCa. We find significant differences in heritability between variants in prostate-relevant epigenetic marks defined in normal versus tumour tissue as well as between tissue and cell lines. The majority of SNP heritability lies in regions marked by H3k27 acetylation in prostate adenoc7arcinoma cell line (LNCaP) or by DNaseI hypersensitive sites in cancer cell lines. We find a high degree of similarity between European and African American ancestries suggesting a similar genetic architecture from common variation underlying PrCa risk. Our findings showcase the power of integrating functional annotation with genetic data to understand the genetic basis of PrCa.

Correspondence and requests for materials should be addressed to A.G. (email: agusev@hsph.harvard.edu) or to B.P (email: pasaniuc@ucla.edu).
#A full list of authors and their affiliations appears at the end of the paper.

Family history is a well-established risk factor for prostate cancer (PrCa), which has an estimated heritability of 58%—one of the highest across common cancers[1]. Genome-wide association studies (GWAS) have been particularly successful in identifying over 100 risk loci that capture ~33% of the estimated familial risk[2]. Although most of the GWAS PrCa variants overlap prostate-specific regulatory elements (for example, androgen receptor-binding sites (ARBS))[2–8], a quantification of the contribution of genetic variation from various chromatin marks to PrCa risk is currently lacking.

Recent work form the ENCODE/ROADMAP consortia[9] has shown that a large fraction of the genome plays a role in at least one biochemical event, in at least one tissue. Although this functional atlas of the human genome has greatly enhanced our understanding of regulatory elements, such functional elements are often tissue specific[10,11] making their interpretability in the context of PrCa risk challenging. Existing studies that have integrated PrCa GWAS findings with tissue-specific functional annotations have relied only on the GWAS significant variants (~100 in the most recent study) or single-nucleotide polymorphisms (SNPs) tagging them[2,7], thus ignoring loci that do not reach genome-wide significance. Recent methodological advances have shown that the entire polygenic architecture of common traits can be interrogated using variance components across all assayed SNPs (typed and/or imputed) to increase power for detecting trait-specific functional annotations[12]. In addition to offering superior performance relative to methods that evaluate only GWAS SNPs, the variance components methods also allow for comparison of estimates across different studies and sample sizes. This is because variance components yield an unbiased estimate (under standard assumptions) of SNP heritability $(h^2_g)$—the variance in trait explained by SNPs that reside within elements of a given functional category[12–15].

Here, we use targeted and genome-wide SNP array data from 59,089 male PrCa cases and controls of European (BPC3 (ref. 16) and iCOGS (ref. 4), respectively, see Methods) and African American (AAPC (ref. 17), see Methods) ancestry to dissect the genetic risk of PrCa. We estimate the SNP heritability of previously implicated regulatory annotations[7,18] and perform a broad analysis of 544 epigenetic marks from ENCODE/ROADMAP (ref. 9). Our approach interrogates the entire common polygenic architecture of PrCa while accounting for potential correlations between related functional categories. First, we find that SNPs near ARBS assayed in prostate tumour explain significantly more of the heritability of PrCa than ARBS SNPs assayed in prostate normal tissue. Second, we localize most of the heritability of PrCa to regions in the genome marked by three functional categories: (i) H3K27ac histone modifications in prostate adenocarcinoma cell lines (LNCaP; typically marking active enhancers[19]); (ii) androgen receptors in prostate tissue[18]; and (iii) DNase I hypersensitivity sites (DHS) in cancer cell lines. We replicate the LNCaP H3K27ac and DHS results across different ancestries and show that risk prediction from genome-wide SNP data is significantly improved with a predictor that incorporates the functional atlas as prior. Overall, our results suggest a similar genetic architecture from common variation of PrCa risk across men of European and African ancestry and highlight H3k27ac histone mark in LNCaP and ARBS in prostate tissue for follow-up studies of PrCa risk.

## Results

**Partitioning the genetic risk for prostate cancer**. We analysed multiple functional annotations and quantified the fraction of variance in trait explained by SNPs that are localized within each functional class. Our approach models the phenotype (PrCa) of a set of individuals as being drawn from a multivariate normal distribution with variance components estimated based on genetic data (that is, SNPs) plus an environmental term (see Methods)[13,14]. For each functional category $i$, a genetic relationship matrix across all individuals is computed from all the SNPs residing in the given functional category to serve as a variance component. Multiple components are then jointly fitted using the restricted maximum likelihood (REML) as implemented in the GCTA software[14] to estimate variance parameters $\left(\sigma^2_i\right)$ for each component. The SNP heritability for component $i$ is then estimated as $h^2_{g,i} = \sigma^2_i / \sum_j \sigma^2_j$, where the sum in the denominator is across all fitted components including the environmental term. Therefore, we view $h^2_{g,i}$ as an estimate of the variance in trait that can be explained by all the SNPs in the corresponding functional category with a linear model of the trait (that is, SNP heritability)[12]. We expect functional categories that are enriched with casual variants for PrCa to attain a higher estimated SNP heritability as compared with functional categories depleted of causal variants for PrCa. To focus our results on noncoding variation and account for potential confounders because of linkage disequilibrium (LD), we explicitly included coding and coding-proximal regulatory variation as 'background' components whenever we quantified the effect of each functional annotation tested (see Methods).

The variance component model has previously been shown to yield robust estimates under the assumption that causal variants are typed and uniformly sampled from a given component[13,20,21]. Here, we perform additional simulations using the UK10K whole-genome sequence data to confirm the validity of this model for our data, and to assess how representative SNP estimates are of true underlying biology at common sequenced variants. The simulation framework uses real genotype data from the UK10K consortium to generate additive, polygenic phenotypes with a given heritability and then performs heritability estimation with the variance component model (see Methods). Although the UK10K data contains a much smaller set of individuals as the iCOGS data (3,047 versus 42,613 individuals, see Methods), it contains variation from whole-genome sequencing; this allows us to evaluate model performance by simulation when restricting to SNPs genotyped on the iCOGS platform. We focused on the LNCaP: H3k27ac annotation (which was most significant in our data, see below) to evaluate the multiple component models. Over thousands of simulations, we confirmed that the variance components approach correctly recovered the causal contribution to trait from a given functional category when causal variants were typed (Supplementary Table 1, see Methods). Under both null and enriched scenarios the estimates were unbiased and standard errors properly calibrated (Supplementary Table 1). For common sequenced variants not present on the iCOGS platform, relative estimates of noncoding enrichment/depletion were conservative, with the tagged effects distributed across the typed components (Supplementary Table 2). Deviations from the standard variance components model assumptions on the distribution of effect-sizes and ancestry-specific effects in African Americans yielded either well calibrated or conservative estimates of SNP heritability in the focal LNCaP: H3k27ac category (see Methods, Supplementary Tables 1–3).

Our primary functional analyses focus on the densely genotyped iCOGS sample (21,678 cases and 20,935 controls), whose large sample size allowed for highly accurate estimates of component-specific $h^2_g$. Although the iCOGS chip is custom built to oversample risk loci, it provides a broad coverage of the common variation genome wide[4]. To showcase the power of the variance components approach, we estimated the total SNP

heritability of PrCa at 0.28 (s.e. 0.01) in the iCOGS data (not significantly different from the total SNP heritability estimate of 0.26 (s.e. 0.05) in the BPC3 data), a significant increase from the variance explained only by the known GWAS variants $\left(h^2_{GWAS}\right)$ of 0.06 (s.e.m. 0.001) (see Methods; Supplementary Table 4). Interestingly, the total SNP heritability in the African American sample, which was genotyped on a different platform than iCOGS (see Methods), was estimated at 0.32 (s.e. 0.06) indicating a similar aggregate contribution of common variation to PrCa risk across the two ethnicities despite higher overall risk in African Americans[22] (Supplementary Table 4).

**Enrichment at androgen receptor-binding sites in tumours**. We first focused on SNPs localized in the ARBS: an epigenetic profile causally implicated in prostate tumorigensis. In contrast to typical assays that focus on cell lines, the ARBS were defined by chromatin immunoprecipitation and high-throughput sequencing (ChIP-seq) directly in primary human tissue (seven normal and 13 tumour specimens)[18]. We observed that variants within 5 kb of tumour-specific ARBS explained 17.0% of the genome-wide $h^2_g$ (s.e. 1.7%; $P = 2.6 \times 10^{-16}$ by $Z$-test), whereas the variants near normal-specific ARBS explained 0.0% of the $h^2_g$ (s.e. 0.9%; $P = 0.11$ by $Z$-test) (Fig. 1). The difference between these two groups was highly significant and demonstrates the importance of assaying functional marks in both normal and tumour tissues. We note that the 5 kb extension may also include other regulatory variants near the tumour/normal-specific ARBS (but not heritability from coding/untranslated region (UTR)/promoter variants, which were explicitly modelled, see Methods). Smaller flanking regions were also investigated but did not include enough markers for the variance components model to converge. We also quantified the proportion of SNP heritability explained directly by all ARBS variants (both normal and tumour without 5 kb flanks) at 10.7% of $h^2_g$; significantly different from the SNP heritability of ARBS variants assayed in prostate adenocarcinoma cancer cell line (LNCaP; 3.2% of $h^2_g$) ($P = 4.4 \times 10^{-7}$ for difference by $Z$-test) (Fig. 1). This difference is partially explained by the very low number of SNPs within cell line ARBS making their aggregate contribution small but not empowering us to place a strong bound on the enrichment. Overall, these findings highlight the increased complexity of ARBS in a sample of tissues as compared with the single LNCaP cell line.

**Identification of functional marks relevant to PrCa risk**. Next, we looked for marks that contribute to the heritability of PrCa across a broad spectrum of functional annotations without prior assumptions on relevance to disease. We investigated 544 epigenetic annotations spanning six major classes (DHS; H3k4me1; H3k4me3; H3k9ac; H3k27ac; and computationally predicted functional classes or 'segmentations'[23,24]) averaging 101 cell types per class (see Methods). After accounting for multiple testing, we identified 82 annotations that exhibited statistically significant deviations in SNP heritability from what was expected based on the proportion of the genome covered by that particular annotation (see Fig. 2 and Supplementary Data).

We first focused on 17 functional marks measured in the prostate, of which 14 were statistically significant (Supplementary Table 5). The single most significant enrichment was observed for H3k27ac marks in LNCaP ($P = 1 \times 10^{-32}$ by $Z$-test), which localized 22% of the total $h^2_g$ to the 2.9% of genotyped SNPs within the annotation. This was followed by variants in DHS marks in LNCaP ($P = 2 \times 10^{-18}$ by $Z$-test; 16.7% of $h^2_g$ localized in 3.1% of genome). The DHS annotations allowed us to compare estimates across three major prostate cell lines: LNCaP; normal prostate epithelial (PrEC); and immortalized prostate epithelial (RWPE1) (overlapping by 25–50% with ARBS, Supplementary Fig. 1). We observed heritability explained by LNCaP DHS to be nominally significantly higher than PrEC ($P = 0.01$ by $Z$-test); and both LNCaP and PrEC to be significantly higher than RWPE1 ($P = 1.5 \times 10^{-9}$, $P = 1.2 \times 10^{-5}$, respectively, by $Z$-test) (Fig. 3). More broadly, 10 out of 16 DHS marks measured in cancer cell lines were observed as significant, with colorectal cancer as the next most significant cancer ($P = 6.0 \times 10^{-10}$ by $Z$-test; 9.4% of heritability localized in 2.0% of genome; Supplementary Data). H3k27ac in LNCaP remained the most significantly enriched mark across all 544 annotations (presented in detail in the Supplementary Data). The most depleted categories were repressed regions computationally predicted by Segway-chromHMM in HepG2 cells ($P = 1.3 \times 10^{-19}$ by $Z$-test; 51.9% of $h^2_g$ from 74.3% of SNPs; Supplementary Data), with similar levels of depletion in repressed regions from other cell types. These regions are typically associated with decreased gene expression and repressive histone marks[23–25], further emphasizing the importance of active regulation.

As H3k27ac typically marks active enhancers, we further evaluated variants with respect to their enhancer or 'super'-enhancer status (large clusters of enhancers that are enriched for genes involved in cell identity[26]) (see Methods). We did not observe differences in average heritability explained by SNPs within the two marks across 49 cell lines (see Methods), with an average of 1.51 (1.47)-fold increase over random SNPs for

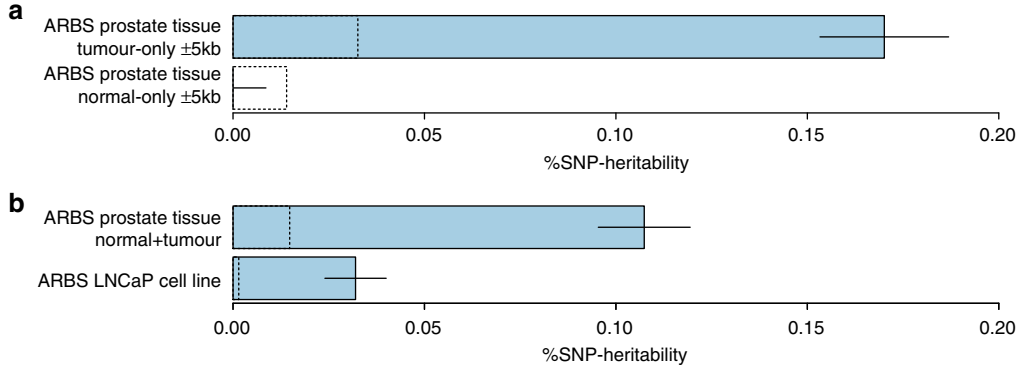



**Figure 1 | Functional partitioning for variants within ARBS for PrCa.** Bars graphs detailing %SNP heritability estimates from two models of PrCa relevant functional annotations. (**a**) Joint comparison of variants within 5 kb of tumour-only and normal-only regions in the ARBS in prostate tissue ($P = 2.1 \times 10^{-19}$ for difference by $Z$-test). (**b**) Estimates from ARBS in prostate tissue (no longer using a 5 kb flank) and ARBS in LNCaP cell lines[7] ($P = 4.4 \times 10^{-7}$ for difference). The null ($\%h^2_g = \%$ SNPs) is labelled by the dashed lines. Error bars show analytical standard error of estimate.

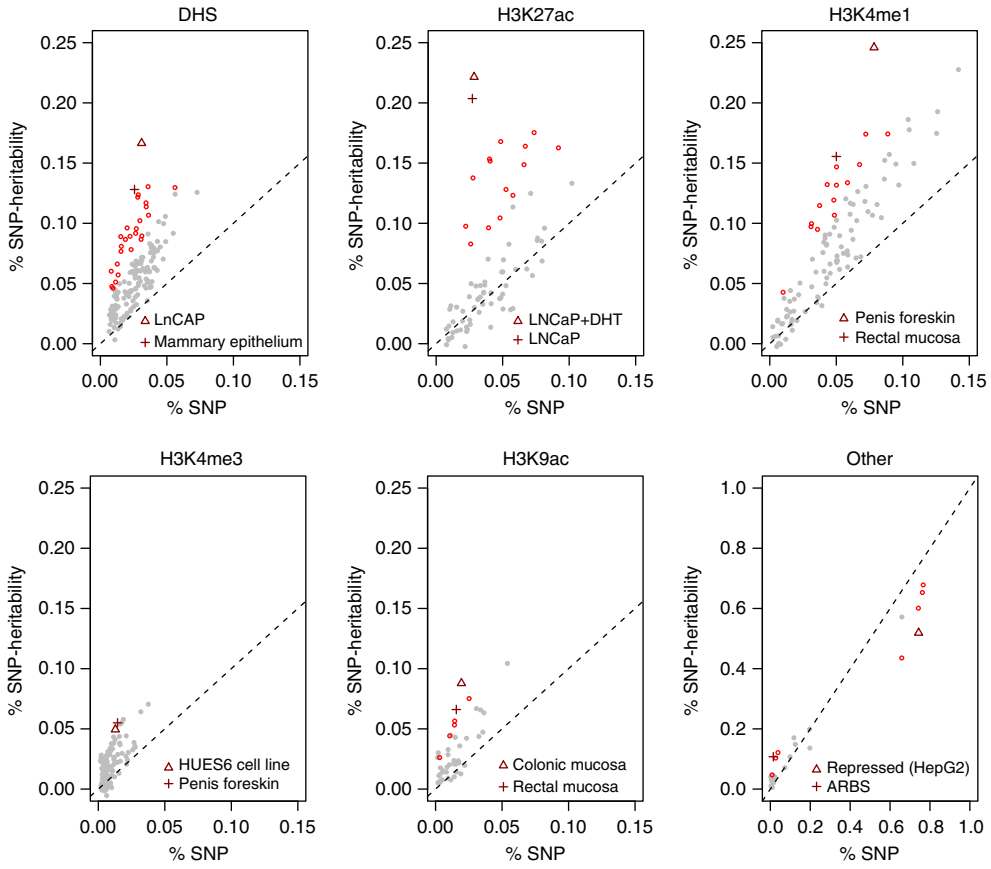



**Figure 2 | Functional partitioning of heritability across six main epigenetic classes.** Each point corresponds to an estimate of % SNP heritability (*y* axis) from SNPs within a cell-type-specific functional annotation versus annotation size (%SNPs, *x* axis). Overall, 544 annotations were tested, and red points indicate significant deviations from the null of $\%h_g^2$ equal to %SNPs after accounting for all tests. The two most significant annotations in each class are shown with triangle/cross, respectively, and labelled in bottom right (see Supplementary Data for all annotations).

enhancers (super enhancers) (Fig. 4). Surprisingly, we observed an individually significant difference only in LNCaP, with 4.9 (1.7)-fold enrichment at enhancers (super enhancers), in contrast to previous hypotheses[26] (Fig. 4).

**Genomic functional atlas of prostate cancer SNP heritability.** Although the results above showcase the power of the variance component approach in finding epigenetic marks relevant for PrCa, such marks often overlap making the causal mark difficult to identify (Supplementary Fig. 1). To account for the correlation among marks we grouped the 82 marginally significant annotations into 15 biologically relevant, non-overlapping groups organized by mark and cell line, and partitioned $h_g^2$ across all groups in a joint model (see Methods, Table 1, Fig. 5 and Supplementary Table 6). Five components were nominally significant in the joint model at $P<0.05$; out of the five components three remained significant after accounting for 15 tests: H3k27ac marks in LNCaP ($P=2.5\times10^{-20}$ by *Z*-test); DHS marks in other cancer cell types ($P=3.9\times10^{-5}$ by *Z*-test); and repressed segmentations ($P=2.1\times10^{-20}$ by *Z*-test). To further refine our model, we restricted to the significant annotations (and the background components accounting for LD to coding regions) and re-evaluated them jointly, referred to as the 'selected' model. This selected model localized 51.0% of the $h_g^2$ within 12.1% of SNPs (LNCaP: H3K27ac + ARBS + DHS cancer), whereas coding regions only explained 3.3% (s.e. 1.4%) of $h_g^2$ within 1.8% of SNPs (Supplementary Table 7). The localization was even stronger with imputed data, where 86% of the $h_g^2$ was localized to 8.6% of SNPs (Table 1 and Supplementary Tables 8 and 9). Estimates from imputed markers were more representative of underlying enrichment in our simulations (see Methods, Supplementary Table 2) but may include the effects of nearby markers[12] and so we consider them as an upper bound. None of the estimates changed significantly after adjusting for known GWAS associations[2] (79 of which were typed in this data), underscoring the polygenic nature of this effect.

Having inferred the selected model, we re-analysed each of the 82 marginally significant categories jointly with the selected model (see Methods). Only three marks remained significant: two H3k27ac annotations in the colon crypt and one H3k27ac annotation in pancreas (Supplementary Data). This implies that the marginal enrichment of the 82 annotations was primarily driven by the overlap with functional marks in the selected model. For example, the H3K4me1 mark in penis foreskin keratinocytes that was previously highly significant (24.6% $h_g^2$, $P=3.0\times10^{-16}$ by *Z*-test, Fig. 1) was no longer enriched after conditioning on the selected model (7.1% $h_g^2$, $P=0.29$ by *Z*-test, Supplementary Data). The reduction to a small number of categories in the selected model with limited loss in signal further emphasizes the extent to which the selected model has localized the functional sources of enrichment. Focusing on the two most enriched categories in the selected model, we found that SNPs present in both the prostate tissue ARBS and LNCaP H3k27ac marks yielded significantly higher average heritability per SNP than either mark individually (Supplementary Table 10).

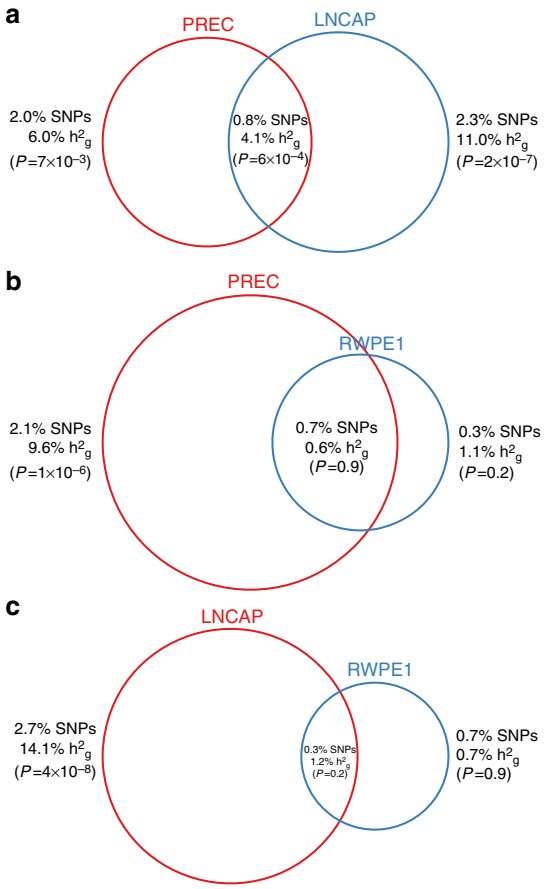



**Figure 3 | Pairwise analysis of DHS marks in three prostate cell types.** Joint model from all pairs of DHS marks shown for: cancer cell line (LNCAP); normal prostate epithelial (PREC); and immortalized prostate epithelial (RWPE1). Circle size corresponds to % SNPs, with % SNP heritability and significance labelled. *P* value was computed for difference between $\%h_g^2$ and %SNP, with bold representing significance after correcting for nine tests. The observed trend is LNCAP > PREC > RWPE1: (**a**) $\%h_g^2$ in LNCaP DHS was nominally significantly higher than PrEC ($P = 0.01$); and $\%h_g^2$ in LNCaP and PrEC was significantly higher than RWPE1 (**b**,**c**; $P = 1.5 \times 10^{-9}$, $P = 1.2 \times 10^{-5}$, respectively). All *P* values computed by *Z*-test using $\%h_g^2$ estimate and analytical standard error.

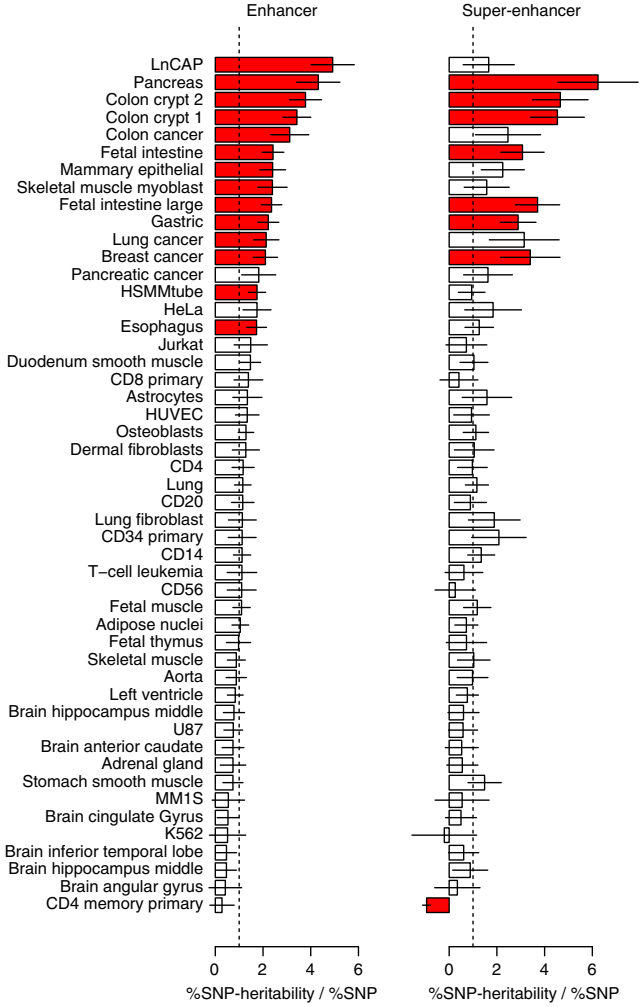



**Figure 4 | Comparison of enhancers and super enhancers across 49 cell types.** Each bar represents the %SNP heritability $\%h_g^2$/ %SNP for enhancers (left) and super enhancers (right) from a given cell type tested marginally. Red indicates significant difference from 1.0 (no enrichment) after accounting for 49 tests. Enhancer LNCAP is most significant, with other cancers also appearing significant and non-cancer tissues least significant. Error bars show analytical s.e. of estimate.

In contrast, the variants specific to ARBS or H3k27ac were comparable in SNP heritability.

**Replication of genomic functional atlas across ancestries.** We evaluated replication of our model using two separate genome-wide SNP data sets of PrCa, one of European ancestry (BPC3; 6,953 samples) and one of African ancestry (AAPC; 9,522 samples) for PrCa (see Methods). To account for the smaller sample size, we focused on the eight-component selected model, only retaining significant components and three coding-proximal classes (coding, UTR, promoter)[12]. Because of platform differences between the populations, we used post-QC imputed variants in each data set, which are most reflective of underlying enrichment in our simulations (see Methods). We replicated the significant deviation in $h_g^2$ at H3k27ac and the repressed loci across both BPC3 and AAPC (Supplementary Tables 11 and 12). However, cancer DHS was only significant in the BPC3 data and ARBS not significant in either (though the estimates were not significantly different from the iCOGS estimate). The enrichment did not change after restricting to very high-quality imputed markers (Supplementary Table 13). Although the relatively small validation sample size did not provide enough power to test differences between the ancestries, the mean SNP heritability for variants within each mark were remarkably similar ($\rho = 0.90$ between AAPC and BPC3 across eight components), suggesting a similar pattern of aggregate contribution to risk coming from common variants marked by epigenetic classes across European and African American ancestries (though individual risk variants themselves may differ).

**H3k27ac mark in LNCaP is specific to PrCa.** As a negative control, we evaluated the selected model with imputed SNPs across 11 common non-cancer diseases from the Wellcome Trust Case Control Consortium (WTCCC) (see Methods, Supplementary Table 14) where we observed two main differences: the LNCaP H3k27ac annotation was no longer significantly enriched (1.1% $h_g^2$ with 2.6% of SNPs); and the repressed regions were much less depleted from the null (28.1%

**Table 1 | Partitioning of heritability across functional classes in prostate cancer.**

| Functional category | %SNPs | Full Model | | Selected model | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | iCOGS genotyped | | iCOGS imputed | | BPC3 imputed | | AAPC imputed | |
| | | % $h_g^2$ | s.e.m. | % $h_g^2$ | s.e.m. | % $h_g^2$ | s.e.m. | % $h_g^2$ | s.e.m. |
| Coding | 1.8 | 3.0 | 1.3 | 0.9 | 2.9 | 0.2 | 10.1 | 3.3 | 11.1 |
| UTR | 1.9 | 1.6 | 1.4 | 3.0 | 3.1 | 21.0 | 11.3 | 5.9 | 11.2 |
| Promoter | 3.4 | *7.8 | 1.8 | 8.9 | 4.1 | 0.0 | 12.7 | 0.0 | 14.7 |
| LNCaP: H3k27ac | 3.2 | **22.3 | 2.1 | **27.0 | 3.8 | *30.3 | 12.1 | *28.9 | 12.7 |
| ARBS | 1.0 | *3.3 | 1.1 | *9.1 | 3.3 | 1.1 | 12.1 | 15.2 | 12.1 |
| LNCaP: FOXA1 | 1.5 | 1.5 | 1.3 | | | | | | |
| LNCaP: H3k4me1 | 2.0 | 1.3 | 1.4 | | | | | | |
| LNCaP: DHS | 2.9 | 5.4 | 1.6 | | | | | | |
| DHS prostate | 1.8 | 2.6 | 1.4 | | | | | | |
| DHS cancer | 4.7 | **14.1 | 2.3 | **49.6 | 6.3 | *47.4 | 21.4 | 46.6 | 22.4 |
| H3k4me1 (other) | 16.3 | 19.6 | 3.5 | | | | | | |
| H3k27ac (other) | 7.3 | 4.1 | 2.4 | | | | | | |
| DHS (other) | 1.8 | 0.2 | 1.3 | | | | | | |
| repressed | 48.7 | **11.0 | 4.1 | **0.3 | 7.0 | **0.0 | 23.8 | **0.0 | 24.5 |
| all other | 1.7 | 0.7 | 1.2 | 0.2 | 2.7 | 0.0 | 9.2 | 0.0 | 7.6 |

ARBS, androgen receptor-binding sites; DHS, DNase I hypersensitivity sites; SNP, single-nucleotide polymorphism; UTR, untranslated region.
Full model denotes a 15-variance components model while 'selected' model denotes a model restricted to the five components attaining significance in the 'full' model (and three components for background). * (**) denotes significant deviation at $P<0.05$ ($P<0.05/15$) of fraction of SNP heritability (% $h_g^2$) from null model of % $h_g^2$ = % SNPs (by $Z$-test; see Supplementary Table 6 for $P$ values).

$h_g^2$ with 87.8% of SNPs) compared with the 0.3% of $h_g^2$ observed in iCOGS imputed data ($P = 2.2 \times 10^{-4}$ for difference by $Z$-test). Interestingly, although ARBS were significantly enriched in all 11 traits, the enrichment was no longer significant after excluding autoimmune traits. Overall, these differences indicate that the LNCaP H3k27ac mark is uniquely informative for PrCa, whereas variants near the ARBS and DHS cancer elements (which overlap other DHS annotations by 56%; Supplementary Fig. 2) may play a generally important role across other common diseases[12].

**Genomic functional atlas improves polygenic risk prediction.** To validate our SNP heritability genomic atlas, we compared the accuracy of predicting case/control status from genetic data with or without the functional atlas. We evaluated three distinct prediction models in the iCOGS sample: (i) a genetic risk score (GRS) from the genome-wide significant SNPs; (ii) the single best linear unbiased predictor (BLUP) using a single variance component from all SNPs; and (iii) the weighted sum of individual BLUPs from each epigenetic category in the selected model (multi-BLUP; see Methods). Evaluated by cross-validation, the GRS yielded an $R^2 = 0.029$ with true phenotype, whereas the single BLUP yielded an $R^2 = 0.065$ and the multi-BLUP had an $R^2 = 0.071$ (Supplementary Table 15). In a joint model with all three predictors, the multi-BLUP was highly significant ($P = 5.3 \times 10^{-31}$ from multiple regression). When we constructed the GRS from SNPs recently discovered in a much larger PrCa GWAS (ref. 2), the resulting prediction $R^2$ increased to 0.084. However, including the single BLUP or the multi-BLUP as an additional predictor still increased the prediction $R^2$ to 0.096 (joint $P = 6.7 \times 10^{-4}$ from multiple regression) and 0.098 (joint $P = 1.3 \times 10^{-23}$ from multiple regression), respectively (Supplementary Table 15). The consistent statistical significance and increased prediction accuracy confirms the validity of the selected model in this data and in larger GWAS.

## Discussion

Using large-scale genotype data from over 59,089 men of European and African American ancestries jointly with epigenetic annotations, we identified highly significant differences in SNP heritability ($h_g^2$) of PrCa across variants from different epigenetic classes, tissue types and cell lines. Focusing on marks measured in prostate, we observed significantly higher $h_g^2$ around tumour-specific ARBS; ARBS measured in primary tissue relative to cell line; and DHS measured in PrCa cell line relative to prostate epithelial cell line. The enrichment at tumour-specific ARBS was consistent with recent findings showing that these sites were enriched for nearby genes highly expressed in tumours[18]. These analyses are comprehensive and cover most commonly studied prostate cell lines except for vertebral cancer of the prostate, which were not well represented in the ENCODE/ROADMAP. A search across 544 diverse functional annotations restricted most of the $h_g^2$ to a small fraction of the genome marked by prostate regulatory elements. Consistent with previous findings in common disease, functionally repressed regions were significantly depleted in heritability, highlighting the role of active regulation in PrCa susceptibility. Subsequent model selection localized the enrichment from 82 individually significant annotations to six that remained significant in a joint model. In particular, the abundance of enrichment in H3k27ac marks (active enhancers) relative to H3k4me1/H3k4me3 (poised enhancers/promoters) underscores their role in PrCa, though further enrichment in super enhancers was not observed. The enrichment within LNCaP: H3K27ac and depletion at repressed regions was replicated across different ancestries and yielded significant improvements in polygenic risk prediction.

With most GWAS associations falling outside coding regions, our analyses offer an important resource for prioritizing potential loci and focusing future studies on the most heritable genomic regions[27]. The marginal analyses provide a ranking of 544 common functional assays, while the selected model localizes heritability to only those functional classes that are independently enriched. Emerging functional categories may further refine this signal or reveal other relevant epigenetic marks, though little enrichment beyond the selected model was observed in the comprehensive sampling of functional data analysed here. In general, the variance component model offers an opportunity to evaluate biological hypotheses *in silico* and without strictly relying on individually significant SNPs. However, as with any analysis of array-based data, the $h_g^2$ estimates will not include the contribution of SNPs that are untyped or poorly tagged, such

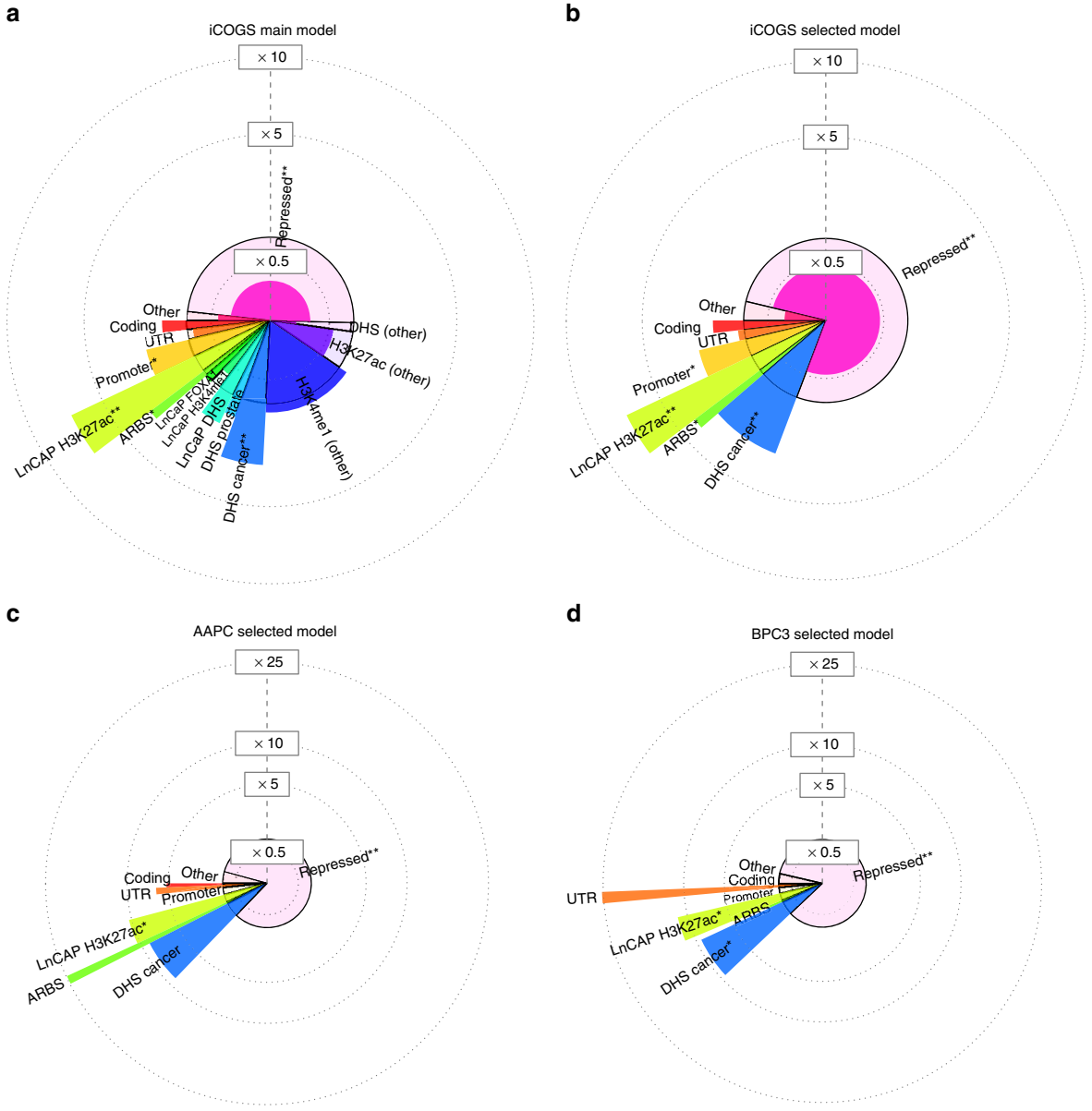



**Figure 5 | Partitioning of heritability across functional classes in prostate cancer.** Visual representation of heritability enrichment in three studies **a**,**b**: iCOGS; **c**: AAPC; **d**: BPC3 (shown numerically in Table 1). Each subplot corresponds to an analysis of the listed joint model, with coloured slices representing the functional annotations evaluated. Volume of each interior (light coloured) pie-chart slice represents the %SNP for the functional annotation, which is equal to the expected $\%h_g^2$ under the null of no enrichment. Volume of each shaded pie-chart slice represents the actual $\%h_g^2$ inferred by the model. Slices extending outside/inside the middle pie correspond to enrichment/depletion in SNP heritability, as indicated by the dotted lines. Colour coding is consistent across all subpanels. * (**) denotes significant deviation at $P<0.05$ ($P<0.05/15$) of fraction of SNP heritability ($\%h_g^2$ from null model of $\%h_g^2 = \%$SNPs by $Z$-test; see Supplementary Table 6 for $P$ values).

as rare variants or other contributors to the missing heritability. Future analyses of whole-genome sequencing, additional functional annotations, and larger sample sizes can yield important insights into functional mechanisms that are still not localized. Overall, our results suggest similar patterns of functional enrichment across men of European and African American ancestry and highlight ARBS, H3k27ac marks in LNCaP cell lines and DHS in cancer cell lines for follow-up studies of PrCa risk.

## Methods

**Epigenetic annotations.** Sample collection and processing for functional annotations was made publically available by the ENCODE/ROADMAP consortia[28]. DHS, H3k4me1, H3k4me3, H3k9ac annotations and genome segmentations[20,29], enhancers and super enhancers[26] and PrCa-specific annotations[7,18] were assay and processed as detailed in the original studies. Tumour-only and normal-only ARBS were defined in seven normal and 13 tumour specimens in the original study[18]. All annotations curated for this paper (ENCODE/ROADMAP; Pomerantz *et al.*; and Hazelett *et al.*) are available at https://data.broadinstitute.org/alkesgroup/ANNOTATIONS/PRCA/. The full list of individual annotations with web-links to the corresponding boundary definitions is provided in the Supplementary Data. Some functional marks are listed multiple times due to multiple independent assays or laboratory protocols.

**ARBS ChIP-seq in human tissue specimens.** The ARBS assay was performed as described in REF (ref. 18) and summarized here. Fourteen subjects of European American ancestry were selected for ChIP analysis. Their chromatin was incubated overnight with 6 µg antibody AR (N-20, Santa Cruz Biotechnology, Dallas, TX)

bound to protein A and protein G beads (Life Technologies, Carlsbad, CA). A fraction of the sample was not exposed to antibody to be used as control (input). The samples were de-crosslinked, treated with RNase and proteinase K, and DNA was extracted. The samples were then re-sheared to 100–300 base pairs using the Covaris ultra-sonicator, and concentrations of the ChIP DNA were quantified by Qubit Fluorometer (Life Technologies). DNA sequencing libraries were prepared using the ThruPLEX-FD Prep Kit (Rubicon Genomics, Ann Arbor, MI). Libraries were sequenced using 50-base pair reads on the Illumina platform (Illumina, San Diego, CA) at Dana-Farber Cancer Institute. AR binding sites were generated using Model-Based Analysis of ChIP-seq 2 (MACS2), with a qvalue (false discovery rate, FDR) threshold of 0.01.

The 13 tumours used in this study were androgen dependent and not exposed to androgen deprivation therapies. All of the tumours were specimens obtained from radical prostatectomies, derived from men with early stage disease. These samples were not selected based on any specific features; therefore, we would expect that the distribution of risk variants would be similar to a random sampling of PrCa cases. Large-scale genetic surveys have shown that somatically acquired alterations in primary localized prostate tumours (the type of tumour evaluated in this study) are infrequent. Based on these previous results, we believe that somatically acquired genetic events in regions related to androgen biology are not common and, therefore, do not influence our results.

**Patient material.** Informed consent was obtained from all subjects and all studies were approved by local Research Ethics Committees and/or Institutional Review Boards.

**Data quality control.** Quality control is crucial for accurate heritability estimation, where many small artifacts can add up to large biases. All data sets went through a stringent QC process with the following exclusion criteria: minor allele frequency (MAF) $<1\%$; fraction of missing/uncalled SNPs $>5\%$; Hardy–Weinberg equilibrium $P$ value $<0.01$; case–control missingness $P$ value $<0.05$; imputation INFO score $>0.30$. In addition, close relatives were pruned such that no pair of individuals had genetic relatedness (GRM) coefficients $>0.05$. The top 10 principal components and a coded study label were always included as fixed-effects. All analysed samples, cases and controls, were males.

**iCOGS data.** The iCOGS consortium genotyped balanced cases and controls on a custom targeted array[4]. After quality control, 42,613 samples and 153,621 genotyped SNPs remained. Imputation was performed to the 1000 Genomes reference panel using HAPI-UR (ref. 30) for phasing and IMPUTE2 (ref. 31) for imputation. Overall, 1,910,827 imputed and genotyped SNPs passed QC. Because of computational restrictions, the heritability estimation was carried out in two equally sized halves of the ICOGS, with total effects computed by inverse-variance meta-analysis. We partitioned the genotyped SNP heritability by MAF but observed no trend and only slight enrichment of % $h_g^2$ at high-frequency variants (Supplementary Table 16).

**BPC3 data.** The National Cancer Institute Breast & Prostate Cancer Cohort Consortium (BPC3) consortium genotyped individuals on the Illumina HumanHap610 quad array[32]. After quality control, 6,953 samples and 4,004,229 genotyped and imputed SNPs remained. Age was available for all samples and additionally included as a covariate.

**AAPC data.** The AAPC consortium genotyped individuals of African ancestry on the Illumina Human1M array[2,33,34]. After quality control, 9,522 samples and 10,468,389 genotyped and imputed SNPs remained.

**WTCCC data.** The Wellcome Trust Case Control Consortium Genotyping genotyped cases for 11 traits as well as shared controls on multiple Illumina and Affymetrix arrays[35–37]. The phenotypes analysed here were ankylosing spondylitis (AS); bipolar disorder (BD); coronary artery disease (CAD); Crohn's disease (CD); hypertension (HT); multiple sclerosis (MS); rheumatoid arthritis (RA); schizophrenia (SP); type 1 diabetes (T1D); type 2 diabetes (T2D); and ulcerative colitis (UC). After quality control, a total of 47,053 samples and 4–5 million genotyped and imputed SNPs remained. Reported $h_g^2$ values were estimated for each phenotype separately and meta-analysed using inverse-variance weighting.

**UK10K data.** The UK10K whole-genome sequence data from ALSPAC and TWINSUK (http://www.uk10k.org) was used only for simulation, and so stringent quality control was not applied. After relatedness filtering, 3,047 samples and 15,691,225 non-singleton variants were retained.

**Heritability estimation of individual annotations.** We estimated the SNP heritability ($h_g^2$) captured by functional categories in a joint variance component model using GCTA as described in REF (ref. 20). Briefly, this model assumes the phenotype is drawn from a multivariate normal distribution with variance-covariance modelled by components computed from the SNPs and a normal residual. For each functional category (for example, DHS) $i=1..M$ where $M$ is the total number of categories in the model, a GRM across all pairs of individuals is computed restricting to SNPs within the functional category. Variance components for all GRMs in the model are then fitted using REML as implemented in GCTA to estimate a variance parameter $\left(\sigma_i^2\right)$ used to compute $\%\, h_i^2 = \sigma_i^2 / \sum_{j=1}^{M} \sigma_j^2$. The $h_i^2$ corresponds to the fraction of trait variance that can be explained by the BLUP restricted to SNPs in the corresponding functional category (or annotation). For a given functional annotation, SNPs were categorized into a hierarchy of seven non-overlapping components: (1) coding; (2) UTR; (3) promoter (functional annotation of interest); (4) DHS; (5) intron; and (6) intergenic. SNPs belonging to multiple categories were partitioned explicitly into the first category in this list. The coding and coding-proximal components were included to ensure that the annotation heritability was not inflated by SNPs that were in high LD with coding variation. A genetic relatedness matrix was computed for each component by first standardizing the corresponding SNPs and then computing a SNP covariance over all pairs of samples. Component-specific $\sigma^2$ and errors were fitted iteratively using the Average Information algorithm[38]. The analytical standard error for $\%\, h_i^2$ was estimated by transforming the GCTA-inferred $\sigma_i^2$ and error covariance matrix using the delta method. As in REF (ref. 20) statistical significance was evaluated by comparing the $\%\, h_g^2$ explained by the category and it's standard error to the %SNPs in the category using a $Z$-test (comparing nested models using a likelihood ratio test yielded similar results). Total $h_g^2$ estimates were computed as $h_g^2 = \sum_{j=1}^{M} \sigma_j^2 / \sum_{j=1}^{M+1} \sigma_j^2$ after transforming to the liability scale assuming a prevalence of 0.14 and using the study-specific case/control ratio.

**Hierarchical joint models.** For specific models of interest, we extended the individual annotation model described above to test intersecting and non-intersecting components. This allowed us to evaluate precisely which sub-annotations of overlapping components were likely to be causal. For the tumour/normal model, we expanded each tumour/normal mark by 5 kb in both directions from the center to capture nearby genes and other regulatory regions so that tumour (normal) covered 3.3% (1.4%) of the SNPs, respectively. We estimated $h_g^2$ from the joint hierarchical model: (1) coding; (2) UTR; (3) promoter; (4) normal-only; (5) tumour-only; (6) DHS; (7) intron; and (8) Other. When comparing ARBS from tissue and ARBS LNCaP from cell line, only 59 SNPs (0.03%) overlapped between the two categories, and so we tested two separate models: (1) coding; (2) UTR; (3) promoter; (4) (ARBS tissue/ARBS LNCaP); (5) DHS; (6) intron; and (7) other. For comparisons between LNCAP, PREC and RWPE1 using DHS we tested each pair of cell lines using the joint model: (1) coding; (2) UTR; (3) promoter; (4) DHS particular to one cell line; (5) DHS common to both cell lines; (6) DHS particular to other cell line; (7) DHS other cell lines; (8) Intron; and (9) Other. For comparisons between enhancers and super enhancers, we used the 86 cell-type-specific annotations from REF (ref. 26), testing each enhancer or super enhancer separately in the following joint model: (1) coding; (2) UTR; (3) promoter, (4) (enhancer/super enhancer for cell-type of interest); (5) DHS; (6) intron; (7) other. Of these, 49 cell types yielded model convergence for both the enhancer and corresponding super enhancer and were used to estimate means and correlation. The order and grouping of marginally significant annotations into epigenetic mark and cell type (for example, in Table 1) are listed in the Supplementary Data. For each of the 82 individually significant annotations, we re-evaluated them jointly with the selected model in the following hierarchical joint model: (1) coding; (2) UTR; (3) promoter; (4) LNCaP:H3k27ac; (5) ARBS; (6) DHS cancer; (7) (functional annotation of interest); (8) DHS; (9) intron; and (10) other. Only functional annotations that converged were reported in the Supplementary Data.

**Accuracy of $h_g^2$ estimates from typed variants in simulations.** The variance component model has previously been shown to yield robust estimates under the assumption that causal variants are typed and uniformly sampled from a given functional category[13,20,21]. Here, we perform simulations using the UK10K whole-genome sequence data to confirm the validity of this model for our annotations, and to assess how representative SNP estimates are of true underlying biology at common sequenced variants. Overall, the simulations involve using real markers to generate additive, polygenic phenotypes with a given heritability and then estimating the heritability with the variance component model. We evaluated the UK10K data for three types of SNPs: (i) common sequenced variants (7,534,538 SNPs); (ii) UK10K SNPs typed by the iCOGS platform (178,509; 95% of iCOGS SNPs); and (iii) UK10K SNPs typed and imputed by the iCOGS platform (1,655,723; 87% of the iCOGS imputed SNPs). We focused on the LNCaP:H3k27ac annotation (which was most significant in our data) to evaluate the main joint model. All phenotypes were simulated by drawing 5,000 causal variants randomly from the specified categories and sampling causal effect-sizes from a normal distribution such that SNPs either explain equal variance (the model assumption) or variance in proportion to their MAF. The phenotype was then generated as the dot product of genotype and effect-size with random noise added to fix heritability at 50%. Phenotypes were simulated thousands of times until the standard error over simulations was low enough to evaluate unbiasedness.

We confirmed that estimates of $h_g^2$ from a polygenic trait were accurate under the model where causal variants are typed (Supplementary Table S1). Under the null, the LNCaP H3k27ac component is expected to explain 3.22% of the SNP heritability, and the model estimated 3.50% (0.22%) and 3.68% (0.21%) under a low-frequency and high-frequency disease architecture, respectively (Supplementary Table S1). None of the estimates were significantly different from the truth given the number of components tested. Under a scenario where LNCaP H3k27ac explains 50% of the $h_g^2$, the model estimated 51.13% (0.40%) and 46.98% (0.35%) under a low-frequency and high-frequency disease architecture, respectively (Supplementary Table S1). Although the high-frequency architecture (where common variants explain more variance in trait than rare variants) represents a substantial model misspecification, our simulations show that this does not introduce substantial bias and is likely to slightly underestimate the SNP heritability at the focal chromatin mark. In all cases, the empirical standard deviation over 500 simulations was similar to the average analytical s.e.m. computed by GCTA (REML algorithm), thus showing that that analytical standard error is well calibrated (Supplementary Table S1). We note that the standard error is inversely related to the sample size[39,40], and is therefore much higher in these simulations than in the iCOGS data which is 14-fold larger.

Lastly, we performed the real data partitioned analysis in subsets of individuals to evaluate biasedness and power to detect significant enrichment. We confirmed that no significant differences were observed between estimates from the entire study compared with those averaged across subsets of the study (Supplementary Fig. 3). As such, we can confidently report estimates and bounds on the enrichment observed in the entire study that will hold for larger studies. Furthermore, all but one of the significant components from the main model remained significant in smaller samples (ARBS), making it unlikely that they were affected by winner's curse. Recent work has quantified the theoretical relationship between estimation error and effective sample size for individual components[39,40].

**Causal variants not tagged on the iCOGS genotyping platform.** We used the sequenced UK10K common variants to evaluate how well the iCOGS genotyped and imputed SNPs captured underlying heritability by simulating phenotypes using causal variants from sequencing and estimating heritability from the iCOGS SNPs (that is, hiding variants that were not genotyped or imputed, Supplementary Table S2). 83% of common UK10K SNPs lie within 100 kb of an iCOGS SNP, so some common variation is likely to be partially tagged by the chip. If the imputed and/or genotyped SNPs served as a good proxy for the common sequence variation, then we would expect their estimates of $\%\, h_g^2$ to match the simulated fractions. When no functional category was enriched with causal variants, small but significant differences were observed for genotyped coding variants (4.75% $h_g^2$ estimated as compared with simulated 0.67%) and imputed intergenic variants (56.09% $h_g^2$ as compared with 50.52% simulated) but not the focal LNCaP:H3k27ac category. Similar deviations were observed for the disease architecture where common variants explain more variance in trait than rare variants (Supplementary Table S2). When causal variants where enriched within LNCaP:H3k27ac category, deviations between simulated and estimated SNP heritability were larger (Supplementary Table S2). Most of this deviation was due to a significant underestimate at LNCaP:H3k27ac, which was simulated to explain 50% of $h_g^2$ but explained only 12.55% (s.e.m.0.92%) and 30.92% (s.e.m. 1.09%) from genotyped and imputed SNPs, respectively. This heritability was distributed across all the remaining components, particularly in intergenic SNPs for the genotyped estimate and DHS SNPs for the imputed estimate, which tend to be nearby.

Overall, our simulations showed that the model is highly accurate when all causal variants are typed. When considering enrichment from untyped causal variants, the imputed estimate was consistently closer to the truth than the genotyped estimate. Most importantly, the estimate from the focal category (LNCaP H3k27ac in our simulations) was shown to be highly conservative both in the null and in the enriched scenario and unlikely to be biased due to tagging of untyped markers. We note that previous work has shown estimates from imputed SNPs (but not genotyped SNPs) may be contaminated by markers very close to an enriched annotation[12]; as such we focused our results on the densely genotyped iCOGS variants which are expected to be conservative, and primarily used imputed data for validation across data sets.

**Estimates of $h_g^2$ from African American samples.** To assess potential biases in estimating $h_g^2$ from an admixed population, we performed separate simulations in the AAPC data where causal variants were specifically sampled from varying $F_{ST}$ bins. This framework evaluated the potential bias resulting from markers that had drifted to different frequencies in the two populations. The $F_{ST}$ was estimated out-of-sample in the HapMap CEU European and YRI Yoruba populations. We tested the null six-component model (Coding, UTR, Promoter, DHS, Intron, Other) and observed no significant deviations from the null under any class of differentiated SNPs (Supplementary Table S3). However, we note that total $h_g^2$ was simulated at 0.50 but was inferred at 0.38–0.66 across increasing quintiles of causal $F_{ST}$ (Supplementary Table 3), indicating that even with well-calibrated estimates of enrichment the total estimate may be biased upwards if the causal SNPs are highly differentiated (observed in this simulation when mean causal $F_{ST} >= 0.35$).

**Genetic prediction.** We sought to validate the utility of our functional atlas by applying it to genetic prediction. The aim of genetic prediction is to use training individuals with genetics (for example, SNPs) and diagnosed phenotype to accurately predict the phenotype into individuals with only genetic data available[41,42]. Here, we focus on correlation of predicted phenotype with true phenotype ($R^2$), as it has a natural relationship to SNP heritability[12,42]. Intuitively, better localization of the true effect-sizes will reduce noise in training the predictor and increase accuracy. If the functional atlas identified regions with increased heritability, this information should significantly improve the prediction. We evaluated three standard models of risk prediction: GRS; BLUP (ref. 43); and multi-component BLUP (ref. 14). The GRS was computed as a sum over SNPs of the log odds-ratios from the training sample[41]. The set of SNPs used was either the genome-wide significant markers in the training set (restricted to one per 1 MB locus) or the genome-wide significant markers identified in a recent large GWAS of PrCa[2]. In contrast to the GRS, the BLUP used all markers in the data to form the prediction. The standard BLUP was estimated using GCTA over all SNPs. The multi-component BLUP was estimated using the components in the selected model (jointly) to compute a single score equal to the sum of the predictions from each component weighted by their component-specific $h_g^2$. This is analogous to specifying a different prior on the effect-size variance in each component. All predictions were carried out by cross-validation in the full iCOGS data, removing 1,000 individuals in each fold. Prediction $R^2$ was then computed from a regression of phenotype on the predictor score with 10 PCs included as covariates to account for ancestry, subsequently subtracting the $R^2 = 0.021$ from a model with PCs only. $P$ values were estimated for each of the coefficients in the multiple regression of phenotype $\sim$ GRS + single-BLUP + multi-BLUP + PCs. To ensure that prediction across data sets was independent, we carefully removed all iCOGS individuals with a GRM value of $>0.05$ to any individual in the BPC3 when computing BLUP coefficients. We separately analysed the predictor in 26,000 iCOGS samples that had age at diagnosis, but did not observe significant differences before/after including age as a covariate.

## References


1. Hjelmborg, J. B. *et al.* The heritability of prostate cancer in the Nordic twin study of cancer. *Cancer Epidemiol. Biomarkers Prev.* **23,** 2303–2310 (2014).
2. Al Olama, A. A. *et al.* A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nat. Genet.* **46,** 1103–1109 (2014).
3. Castro, E. *et al.* Germline BRCA mutations are associated with higher risk of nodal involvement, distant metastasis, and poor survival outcomes in prostate cancer. *J. Clin. Oncol.* **31,** 1748–1757 (2013).
4. Eeles, R. A. *et al.* Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat. Genet.* **45,** 385–391 (2013).
5. Saunders, E. J. *et al.* Fine-mapping the HOXB region detects common variants tagging a rare coding allele: evidence for synthetic association in prostate cancer. *PLoS Genet.* **10,** e1004129 (2014).
6. Ewing, C. M. *et al.* Germline mutations in HOXB13 and prostate-cancer risk. *N. Engl. J. Med.* **366,** 141–149 (2012).
7. Hazelett, D. J. *et al.* Comprehensive functional annotation of 77 prostate cancer risk loci. *PLoS Genet.* **10,** e1004102 (2014).
8. Hazelett, D. J., Coetzee, S. G. & Coetzee, G. A. A rare variant, which destroys a FoxA1 site at 8q24, is associated with prostate cancer risk. *Cell Cycle* **12,** 379–380 (2013).
9. ENCODE Project Consortium *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
10. Stamatoyannopoulos, J. A. What does our genome encode? *Genome Res.* **22,** 1602–1611 (2012).
11. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337,** 1190–1195 (2012).
12. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95,** 535–552 (2014).
13. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42,** 565–569 (2010).
14. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88,** 76–82 (2011).
15. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43,** 519–525 (2011).
16. Schumacher, F. R. *et al.* Genome-wide association study identifies new prostate cancer susceptibility loci. *Hum. Mol. Genet.* **20,** 3867–3875 (2011).
17. Haiman, C. A. *et al.* Characterizing genetic risk at known prostate cancer susceptibility loci in African Americans. *PLoS Genet.* **7,** e1001387 (2011).
18. Pomerantz, M. *et al.*The androgen receptor cistrome is extensively reprogrammed in human prostate tumorigenesis. *Nat. Genet.* **47,** 1346–1351 (2015).
19. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15,** 272–286 (2014).
20. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95,** 535–552.
21. Lee, S. H. *et al.* Estimation of SNP heritability from dense genotype data. *Am. J. Hum. Genet.* **93,** 1151–1155 (2013).
22. Cancer Facts & Figures for African Americans 2009–2010. Accessed on: December 2015.

23. Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* **41,** 827–841 (2013).
24. Haiman, C. A. *et al.* Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.* **39,** 638–644 (2007).
25. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9,** 473–476 (2012).
26. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155,** 934–947 (2013).
27. Farh, K. K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–343 (2014).
28. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
29. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94,** 559–573 (2014).
30. Williams, A. L., Patterson, N., Glessner, J., Hakonarson, H. & Reich, D. Phasing of many thousands of genotyped samples. *Am. J. Hum. Genet.* **91,** 238–251 (2012).
31. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1,** 457–470 (2011).
32. Schumacher, F. R. *et al.* Genome-wide association study identifies new prostate cancer susceptibility loci. *Hum. Mol. Genet.* **20,** 3867–3875 (2011).
33. Haiman, C. A. *et al.* Characterizing genetic risk at known prostate cancer susceptibility loci in African Americans. *PLoS Genet.* **7,** e1001387 (2011).
34. Kolonel, L. N. *et al.* A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am. J. Epidemiol.* **151,** 346–357 (2000).
35. Barrett, J. C. *et al.* Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nat. Genet.* **41,** 1330–1334 (2009).
36. International Multiple Sclerosis Genetics Consortium *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476,** 214–219 (2011).
37. Burton, P. R. *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678 (2007).
38. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46,** 100–106 (2014).
39. Visscher, P. M. & Goddard, M. E. A general unified framework to assess the sampling variance of heritability estimates using pedigree or marker-based relationships. *Genetics* **199,** 223–232 (2015).
40. Visscher, P. M. *et al.* Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLoS Genet.* **10,** e1004269 (2014).
41. International Schizophrenia, C. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460,** 748–752 (2009).
42. Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14,** 507–515 (2013).
43. Robinson, G. K. That BLUP is a good thing: the estimation of random effects. *Stat. Sci.* 15–32 (1991).


## Acknowledgements

This work was supported by NIH fellowship F32 GM106584 (AG), NIH grants R01 MH101244(A.G.), R01 CA188392 (B.P.), U01 CA194393(B.P.), R01 GM107427 (M.L.F.), R01 CA193910 (M.L.F./M.P.) and Prostate Cancer Foundation Challenge Award (M.L.F./M.P.). This study makes use of data generated by the Wellcome Trust Case Control Consortium and the Wellcome Trust Sanger Institute. A full list of the investigators who contributed to the generation of the Wellcome Trust Case Control Consortium data is available on www.wtccc.org.uk. Funding for the Wellcome Trust Case Control Consortium project was provided by the Wellcome Trust under award 076113. This study makes use of data generated by the UK10K Consortium. A full list of the investigators who contributed to the generation of the data is available online (http://www.UK10K.org). The PRACTICAL consortium was supported by the following grants: European Commission's Seventh Framework Programme grant agreement n° 223175 (HEALTH-F2-2009-223175), Cancer Research UK Grants C5047/A7357, C1287/A10118, C5047/A3354, C5047/A10692, C16913/A6135 and The National Institute of Health (NIH) Cancer Post-Cancer GWAS initiative Grant: no. 1 U19 CA 148537-01 (the GAME-ON initiative); Cancer Research UK (C1287/A10118, C1287/A 10710, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007 and C5047/ A10692), the National Institutes of Health (CA128978) and Post-Cancer GWAS initiative (1U19 CA148537, 1U19 CA148065 and 1U19 CA148112—the GAME-ON initiative), the Department of Defense (W81XWH-10-1-0341), A Linneus Centre (Contract ID 70867902), Swedish Research Council (grant no K2010-70X-20430-04-3), the Swedish Cancer Foundation (grant no 09-0677), grants RO1CA056678, RO1CA082664 and RO1CA092579 from the US National Cancer Institute, National Institutes of Health; US National Cancer Institute (R01CA72818); support from The National Health and Medical Research Council, Australia (126402, 209057, 251533, 396414, 450104, 504700, 504702, 504715, 623204, 940394 and 614296); NIH grants CA63464, CA54281 and CA098758; US National Cancer Institute (R01CA128813, PI: J.Y. Park); Bulgarian National Science Fund, Ministry of Education and Science (contract DOO-119/2009; DUNK01/2–2009; DFNI-B01/28/2012); Cancer Research UK grants [C8197/A10123] and [C8197/A10865]; grant code G0500966/75466; NIHR Health Technology Assessment Programme (projects 96/20/06 and 96/20/99); Cancer Research UK grant number C522/A8649, Medical Research Council of England grant number G0500966, ID 75466 and The NCRI, UK; The US Dept of Defense award W81XWH-04-1-0280; Australia Project Grant [390130, 1009458] and Enabling Grant [614296 to APCB]; the Prostate Cancer Foundation of Australia (Project Grant [PG7] and Research infrastructure grant [to APCB]); NIH grant R01 CA092447; Vanderbilt-Ingram Cancer Center (P30 CA68485); Cancer Research UK [C490/A10124] and supported by the UK National Institute for Health Research Biomedical Research Centre at the University of Cambridge; Competitive Research Funding of the Tampere University Hospital (9N069 and X51003); Award Number P30CA042014 from the National Cancer Institute.

## Author contributions

A.G., A.L.P., M.L.F., C.A.H. and B.P. planned the study and wrote the paper. A.G., H.S., G.K. and B.P. performed primary analysis. All authors contributed to study design, data collection and analysis of individual data and annotations.

## Additional information

**Supplementary Information** accompanies this paper at http://www.nature.com/naturecommunications

**Competing financial interests:** The authors declare no competing financial interests.

**Reprints and permission** information is available online at http://npg.nature.com/reprintsandpermissions/

**How to cite this article:** Gusev, A. *et al.* Atlas of prostate cancer heritability in European and African-American men pinpoints tissue-specific regulation. *Nat. Commun.* 7:10979 doi: 10.1038/ncomms10979 (2016).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

Alexander Gusev[1,2], Huwenbo Shi[3], Gleb Kichaev[3], Mark Pomerantz[4], Fugen Li[5,6], Henry W. Long[4,5], Sue A. Ingles[7], Rick A. Kittles[8], Sara S. Strom[9], Benjamin A. Rybicki[10], Barbara Nemesure[11], William B. Isaacs[12], Wei Zheng[13], Curtis A. Pettaway[14], Edward D. Yeboah[15,16], Yao Tettey[15,16], Richard B. Biritwum[15,16], Andrew A. Adjei[15,16], Evelyn Tay[15,16], Ann Truelove[17], Shelley Niwa[17], Anand P. Chokkalingam[18], Esther M. John[19,20], Adam B. Murphy[21], Lisa B. Signorello[1,22], John Carpten[23], M Cristina Leske[11], Suh-Yuh Wu[11], Anslem J.M Hennis[11,24], Christine Neslund-Dudas[10], Ann W. Hsing[19,20], Lisa Chu[19,20], Phyllis J. Goodman[25], Eric A. Klein[26], John S. Witte[27,28], Graham Casey[7], Sam Kaggwa[29], Michael B. Cook[30], Daniel O. Stram[7],

William J. Blot[13,22], Rosalind A. Eeles[31,32], Douglas Easton[33], ZSofia Kote-Jarai[31], Ali Amin Al Olama[33], Sara Benlloch[33], Kenneth Muir[34,35], Graham G. Giles[36,37], Melissa C. Southey[38], Liesel M. Fitzgerald[36], Henrik Gronberg[39], Fredrik Wiklund[39], Markus Aly[39,40], Brian E. Henderson[41], Johanna Schleutker[42,43], Tiina Wahlfors[43], Teuvo L.J Tammela[44], Børge G. Nordestgaard[45,46], Tim J. Key[47], Ruth C. Travis[47], David E. Neal[48,49], Jenny L. Donovan[50], Freddie C. Hamdy[51,52], Paul Pharoah[53], Nora Pashayan[53,54], Kay-Tee Khaw[55], Janet L. Stanford[56,57], Stephen N. Thibodeau[58], Shannon K. McDonnell[58], Daniel J. Schaid[58], Christiane Maier[59], Walther Vogel[59], Manuel Luedeke[60], Kathleen Herkommer[61], Adam S. Kibel[62], Cezary Cybulski[63], Dominika Wokolorczyk[63], Wojciech Kluzniak[63], Lisa Cannon-Albright[64,65], Craig Teerlink[64,65], Hermann Brenner[66,67], Aida K. Dieffenbach[66,67], Volker Arndt[66], Jong Y. Park[68], Thomas A. Sellers[68], Hui-Yi Lin[69], Chavdar Slavov[70], Radka Kaneva[71], Vanio Mitev[71], Jyotsna Batra[72], Amanda Spurdle[73], Judith A. Clements[72], Manuel R. Teixeira[74,75], Hardev Pandha[76], Agnieszka Michael[76], Paula Paulo[74], Sofia Maia[74], Andrzej Kierzek[76], the PRACTICAL Consortium[†], David V. Conti[77], Demetrius Albanes[78], Christine Berg[79], Sonja I. Berndt[30], Daniele Campa[80], E David Crawford[81], W Ryan Diver[82], Susan M. Gapstur[82], J Michael Gaziano[1,83,84], Edward Giovannucci[1,85], Robert Hoover[30], David J. Hunter[1], Mattias Johansson[86,87], Peter Kraft[1,88], Loic Le Marchand[89], Sara Lindström[1,88], Carmen Navarro[90,91], Kim Overvad[79], Elio Riboli[92], Afshan Siddiq[93], Victoria L. Stevens[82], Dimitrios Trichopoulos[1,94,95], Paolo Vineis[96,97], Meredith Yeager[30], Gosia Trynka[98,99], Soumya Raychaudhuri[2,98,100], Frederick R. Schumacher[77], Alkes L. Price[1,2], Matthew L. Freedman[2,4,5], Christopher A. Haiman[77]& Bogdan Pasaniuc[3,101,102]

[1] Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts 02115, USA. [2] Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. [3] Bioinformatics Interdepartmental Program, University of California Los Angeles, Los Angeles, California 90095, USA. [4] Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, Massachusetts 02115, USA. [5] Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. [6] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. [7] Department of Preventative Medicine, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, California 90033, USA. [8] University of Arizona College of Medicine and University of Arizona Cancer Center, Tucson, Arizona 85721, USA. [9] Department of Epidemiology, University of Texas M.D. Anderson Cancer Center, Houston, Texas 77030, USA. [10] Department of Public Health Sciences, Henry Ford Hospital, Detroit, Michigan 48202, USA. [11] Department of Preventive Medicine, Stony Brook University, Stony Brook, New York 11794, USA. [12] James Buchanan Brady Urological Institute, Johns Hopkins Hospital and Medical Institution, Baltimore, Maryland 21205, USA. [13] Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA. [14] Department of Urology, University of Texas M.D. Anderson Cancer Center, Houston, Texas 77030, USA. [15] Korle Bu Teaching Hospital, Accra, Ghana. [16] University of Ghana Medical School, Accra, Ghana. [17] Westat, Rockville, Maryland 20850, USA. [18] School of Public Health, University of California, Berkeley, California 94720, USA. [19] Cancer Prevention Institute of California, Fremont, California 94538, USA. [20] Stanford University School of Medicine and Stanford Cancer Institute, Palo Alto, California 94305, USA. [21] Department of Urology, Northwestern University, Chicago, Illinois 60611, USA. [22] International Epidemiology Institute, Rockville, Maryland 20850, USA. [23] The Translational Genomics Research Institute, Phoenix, Arizona 85004, USA. [24] Chronic Disease Research Centre and Faculty of Medical Sciences, University of the West Indies, Bridgetown, Barbados. [25] SWOG Statistical Center, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA. [26] Glickman Urological & Kidney Institute, Cleveland Clinic, Cleveland, Ohio 44195, USA. [27] Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California 94118, USA. [28] Institute for Human Genetics, University of California, San Francisco, San Francisco, California 94118, USA. [29] Department of Surgery, Makerere University College of Health Sciences, Kampala 94118, Uganda. [30] Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland 20892, USA. [31] The Institute of Cancer Research, Sutton SM2 5NG, UK. [32] Royal Marsden National Health Service (NHS) Foundation Trust, London and Sutton, UK. [33] Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Strangeways Laboratory, Worts Causeway, Cambridge CB1 8RN, UK. [34] Institute of Population Health, University of Manchester, Manchester M13 9PL, UK. [35] Warwick Medical School, University of Warwick, Coventry CV4 7AL, UK. [36] Cancer Epidemiology Centre, The Cancer Council Victoria, 615 St Kilda Road, Melbourne, Victoria 3004, Australia. [37] Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Victoria 3004, Australia. [38] Genetic Epidemiology Laboratory, Department of Pathology, The University of Melbourne, Grattan Street, Parkville, Victoria 3010, Australia. [39] Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm 171 77, Sweden. [40] Department of Clinical Sciences at Danderyds Hospital, Stockholm 171 77, Sweden. [41] Department of Preventive Medicine, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, California 90007, USA. [42] Department of Medical Biochemistry and Genetics Institute of Biomedicine Kiinamyllynkatu 10, University of Turku, Turku FI-20014, Finland. [43] BioMediTech, University of Tampere and FimLab Laboratories, Tampere 33200, Finland. [44] Department of Urology, Tampere University Hospital and Medical School, University of Tampere, Tampere 33200, Finland. [45] Department of Clinical Biochemistry, Herlev Hospital, Copenhagen University Hospital, Herlev Ringvej 75, Herlev DK-2730, Denmark. [46] Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 1165, Densmark. [47] Cancer Epidemiology, Nuffield Department of Population Health; University of Oxford, Oxford OX3 7LF, UK. [48] University of Cambridge, Department of Oncology, Addenbrooke's Hospital, Box 279, Hills Road, Cambridge CB2 0QQ. [49] Cancer Research UK Cambridge Research Institute, Li Ka Shing Centre, Cambridge, UK. [50] School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol BS8 2PS, UK. [51] Department of Public Health, Section for Epidemiology, Aarhus University, Aarhus OX1 3PN, Denmark. [52] Faculty of Medical Science, University of Oxford, John Radcliffe Hospital, Oxford OX1 3PN, UK. [53] Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Strangeways Laboratory, Worts Causeway, Cambridge CB1

8RN, UK. [54] University College London, Department of Applied Health Research, 1-19 Torrington Place, London WC1E 7HB, UK. [55] Clinical Gerontology Unit, University of Cambridge, Cambridge CB1 8RN, UK. [56] Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109-1024, USA. [57] Department of Epidemiology, School of Public Health, University of Washington, Seattle, Washington 98109, USA. [58] Mayo Clinic, Rochester, Minnesota 55905, USA. [59] Institute of Human Genetics, University Hospital Ulm, 89081 Ulm, Germany. [60] Department of Urology, University Hospital Ulm, 89081 Ulm, Germany. [61] Department of Urology, Klinikum rechts der Isar der Technischen Universitaet Muenchen, 81675 Munich, Germany. [62] Division of Urologic Surgery, Brigham and Womens Hospital, Dana-Farber Cancer Institute, 75 Francis Street, Boston, Massachusetts 02115, USA. [63] International Hereditary Cancer Center, Department of Genetics and Pathology, Pomeranian Medical University, Szczecin, Poland. [64] Division of Genetic Epidemiology, Department of Medicine, University of Utah School of Medicine, Salt Lake City, Utah 84132, USA. [65] George E. Wahlen Department of Veterans Affairs Medical Center, Salt Lake City, Utah 84132, USA. [66] Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany. [67] German Cancer Consortium (DKTK), Heidelberg 69120, Germany. [68] Department of Cancer Epidemiology, Moffitt Cancer Center, 12902 Magnolia Drive, Tampa, Florida 33612, USA. [69] Biostatistics Program, Moffitt Cancer Center, 12902 Magnolia Drive, Tampa, Florida 33612, USA. [70] Department of Urology and Alexandrovska University Hospital, Medical University, Sofia 1431, Bulgaria. [71] Department of Medical Chemistry and Biochemistry, Molecular Medicine Center, Medical University, Sofia, 2 Zdrave Str., Sofia 1431, Bulgaria. [72] Australian Prostate Cancer Research Centre-Qld, Institute of Health and Biomedical Innovation and School of Biomedical Science, Queensland University of Technology, Brisbane, Queensland 4000, Australia. [73] Molecular Cancer Epidemiology Laboratory, Queensland Institute of Medical Research, Brisbane, Queensland 4000, Australia. [74] Department of Genetics, Portuguese Oncology Institute, Porto 4200, Portugal. [75] Biomedical Sciences Institute (ICBAS), University of Porto, Porto 4200, Portugal. [76] The University of Surrey, Guildford, Surrey GU2 7XH, UK. [77] Department of Preventive Medicine, Norris Cancer Center, University of Southern California Keck School of Medicine, Los Angeles, California 90033, USA. [78] Nutritional Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, US National Institute of Health, Bethesda, Maryland 20892, USA. [79] Department of Radiation Oncology and Molecular Radiation Sciences, Johns Hopkins Medicine, Baltimore, Maryland 21287, USA. [80] Genomic Epidemiology Group, German Cancer Research Center (DKFZ), 69121 Heidelberg, Germany. [81] Urologic Oncology, University of Colorado at Denver Health Sciences Center, Denver, Colorado 80230, USA. [82] Epidemiology Research Program, American Cancer Society, Atlanta, Georgia 30303, USA. [83] Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA. [84] Division of Aging, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. [85] Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts 02115, USA. [86] International Agency for Research on Cancer, Lyon 69008, France. [87] Department of Surgical and Perioperative Sciences, Urology and Andrology, Umeå University, Umeå 907 36, Sweden. [88] Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA. [89] Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii 96813, USA. [90] Department of Epidemiology, Regional Health Authority, Murcia 30009, Spain. [91] CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona 28029, Spain. [92] Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London SW7 2AZ, UK. [93] Department of Genomics of Common Disease, School of Public Health, Imperial College London, London SW7 2AZ, UK. [94] Bureau of Epidemiologic Research, Academy of Athens, Athens 106 79, Greece. [95] Hellenic Health Foundation, Athens 106 79, Greece. [96] HuGeF Foundation, Torino 10126, Italy. [97] School of Public Health, Imperial College London, London SW7 2AZ, UK. [98] Divisions of Genetics and Rheumatology, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. [99] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SA, UK. [100] Institute of Inflammation and Repair, University of Manchester, Manchester M13 9PT, UK. [101] Departments of Pathology and Laboratory Medicine, University of California Los Angeles, Los Angeles, California, USA. [102] Department of Human Genetics, University of California Los Angeles, Los Angeles, California 90095, USA. † A full list of consortium members appears at the end of the paper.

## The PRACTICAL consortium

Margaret Cook[33], Michelle Guy[31], Koveela Govindasami[31], Daniel Leongamornlert[31], Emma J. Sawyer[31], Rosemary Wilkinson[31], Edward J. Saunders[31], Malgorzata Tymrakiewicz[31], Tokhir Dadaev[31], Angela Morgan[31], Cyril Fisher[31], Steve Hazel[31], Naomi Livni[31], Artitaya Lophatananon[34,35], John Pedersen[103], John L. Hopper[37], Jan Adolfson[39], Paer Stattin[39], Jan-Erik Johansson[39], Carin Cavalli-Bjoerkman[39], Ami Karlsson[39], Michael Broms[39], Anssi Auvinen[104], Paula Kujala[105], Liisa Maeaettaenen[106], Teemu Murtola[107,108], Kimmo Taari[109], Maren Weischer[45], Sune F. Nielsen[45,46], Peter Klarskov[110], Andreas Roder[111], Peter Iversen[111], Hans Wallinder[112], Sven Gustafsson[112], Angela Cox[113], Paul Brown[50], Anne George[50], Gemma Marsden[50], Athene Lane[50], Michael Davis[50], Wei Zheng[114], Lisa B. Signorello[115], William J. Blot[116,117], Lori Tillmans[58], Shaun Riska[58], Liang Wang[58], Antje Rinckleb[60], Jan Lubiski[63], Christa Stegmaier[118], Julio Pow-Sang[68], Hyun Park[68], Selina Radlein[68], Maria Rincon[68], James Haley[68], Babu Zachariah[68], Darina Kachakova[71], Elenko Popov[70], Atanaska Mitkova[71], Aleksandrina Vlahova[76], Tihomir Dikov[76], Svetlana Christova[76], Peter Heathcote[72], Glenn Wood[72], Greg Malone[72], Pamela Saunders[72], Allison Eckert[72], Trina Yeadon[72], Kris Kerr[72], Angus Collins[72], Megan Turner[72], Srilakshmi Srinivasan[72], Mary-Anne Kedda[72], Kimberly Alexander[72], Tracy Omara[72], Huihai Wu[119], Rui Henrique[74], Pedro Pinto[74], Joana Santos[74]& Joao Barros-Silva[74]

[103] Tissupath Pty Ltd., Melbourne,Victoria 3122, Australia. [104] Department of Epidemiology, School of Health Sciences, University of Tampere, Tampere 33014, Finland. [105] Fimlab Laboratories, Tampere University Hospital, Tampere, Finland. [106] Finnish Cancer Registry, Helsinki, Finland. [107] School of Medicine, University of Tampere, Tampere, Finland. [108] Department of Urology, Tampere University Hospital, Tampere, Finland. [109] Department of Urology, Helsinki University Central Hospital and University of Helsinki, Helsinki 00100, Finland. [110] Department of Urology, Herlev Hospital, Copenhagen University Hospital, Herlev Ringvej 75, Herlev DK-230, Denmark. [111] Copenhagen Prostate Cancer Center, Department of Urology, Rigshospitalet, Copenhagen University Hospital, Tagensvej 20, 7521, Copenhagen DK-2200, Denmark. [112] Department of Epidemiology and Biostatistics, School of Public Health, Imperial College, London SW7 2AZ, UK. [113] CR-UK/YCR Sheffield Cancer Research Centre, University of Sheffield, Sheffield S10 2TN, UK. [114] Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, 2525 West End Avenue, Suite 800, Nashville, Tennessee 37232, USA. [115] National Cancer Institute, NIH, 9609 Medical Center Drive, Suite 2W-172, MSC 9712, Bethesda, MD 20892-9712 (mail), Rockville, Maryland 20850 (FedEx/Courier), USA. [116] International

Epidemiology Institute, 1555 Research Blvd., Suite 550, Rockville, Maryland 20850, USA. [117] Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA. [118] Saarland Cancer Registry, Saarbrücken 66119, Germany. [119] The University of Surrey, Guildford, Surrey GU2 7XH