bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Large-scale transcriptome-wide association study identifies new prostate cancer risk regions

Nicholas Mancuso[1], Simon Gayther[2], Alexander Gusev[3], Wei Zheng[4], Kathryn L. Penney[5,6], the PRACTICAL consortium, CRUK, BPC3, CAPS, PEGASUS[*], Zsofia Kote-Jarai[7], Rosalind Eeles[7], Matthew Freedman[8], Christopher Haiman[9], Bogdan Pasaniuc[1,10,11]

1. Department of Pathology & Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA

2. The Center for Bioinformatics and Functional Genomics, Cedars-Sinai Medical Center, Los Angeles, CA, USA

3. Dana Farber Cancer Institute, Boston, MA

4. Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN

5. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts.

6. Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital/Harvard Medical School, Boston, Massachusetts

7. Division of Genetics and Epidemiology, The Institute of Cancer Research & Royal Marsden NHS Foundation Trust, London, UK

8. Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, Massachusetts

9. Department of Preventive Medicine, Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA.

10. Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA

11. Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA

[*] Members from the PRACTICAL Consortium, CRUK, BPC3, CAPS and PEGASUS are provided in the supplementary note.

Abstract

Although genome-wide association studies (GWAS) for prostate cancer (PrCa) have identified more than 100 risk regions, most of the risk genes at these regions remain largely unknown. Here, we integrate the largest PrCa GWAS (N=142,392) with gene expression measured in 45 tissues (N=4,458), including normal and tumor prostate, to perform a multi-tissue transcriptome-wide association study (TWAS) for PrCa. We identify 235 genes at 87 independent 1Mb regions

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

associated with PrCa risk, 9 of which are regions with no genome-wide significant SNP within 2Mb. 24 genes are significant in TWAS only for alternative splicing models in prostate tumor thus supporting the hypothesis of splicing driving risk for continued oncogenesis. Finally, we use a Bayesian probabilistic approach to estimate credible sets of genes containing the causal gene at pre-defined level; this reduced the list of 235 associations to 120 genes in the 90% credible set. Overall, our findings highlight the power of integrating expression with PrCa GWAS to identify novel risk loci and prioritize putative causal genes at known risk loci.

Introduction

Prostate cancer (PrCa) affects ~1 in 7 men during their lifetime and is one of the most common cancers worldwide, with up to 58% of risk due to genetic factors[1; 2]. Genome-wide association studies (GWAS) have identified over 100 genomic regions harboring risk variants for PrCa which explain roughly one third of familial risk[3-7]. With few exceptions[8], the causal variants and target susceptibility genes at most GWAS risk loci have yet to be identified. Multiple studies have shown that PrCa- and other disease-associated variants are enriched near variants that correlate with gene expression levels[9-13]. In fact, recent approaches have integrated expression quantitative trait loci (eQTLs) with GWAS to implicate several plausible genes for PrCa risk (e.g., *IRX4*, *MSMB*, *NCOA4*, *NUDT11* and *SLC22A3*)[5; 14-21]. While overlapping eQTLs and GWAS is powerful, the high prevalence of eQTLs[22] coupled with linkage disequilibrium (LD) renders it difficult to distinguish the true susceptibility gene from spurious co-localization at the same locus[23]. Therefore, disentangling LD is critical for prioritization and causal gene identification at risk loci.

Gene expression imputation followed by a transcriptome-wide association study[24-26] (TWAS) has been recently proposed as a powerful approach to prioritize candidate risk genes underlying complex traits. By taking LD into account across SNPs, the resulting association statistics reflect the underlying effect of steady-state gene or alternative splicing expression levels on disease risk[25; 27], which can be used to identify new regions or to rank genes for functional validation at known risk regions[24-28]. Here we perform a multi-tissue transcriptome-wide association study[24-26] to identify new risk regions and to prioritize genes at known risk regions for PrCa. Specifically, we integrate gene expression data from 48 panels measured in 45 tissues across 4,448 individuals with GWAS of prostate cancer from the OncoArray in 142,392 men[29]. Notably, we include alternatively spliced and total gene expression data measured in tumor prostate to identify genes contributing to prostate cancer risk or to continued oncogenesis. We identify 235 gene-trait associations for PrCa with 24 (11) genes identified uniquely using models of

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

alternative spliced (total) expression in tumor. Significant genes were found in 87 independent 1Mb regions, of which 9 regions are located more than 2Mb away from any OncoArray GWAS significant variants, thus identifying new candidate risk regions. Second, we use TWAS to investigate genes previously reported as susceptibility genes for prostate cancer identified by eQTL-based analyses. We find a significant overlap with 57 out of 104 previously reported genes assayed in our study also significant in TWAS. Third, we use a novel Bayesian prioritization approach to compute credible sets of genes and prioritize 120 genes that explain at least 90% of the posterior density for association signal at TWAS risk regions. One notable example, *IRX4,* had 97% posterior probability to explain the association signal at its region with the remaining 3% explained by 9 neighboring genes. Overall, our findings highlight the power of integrating gene expression data with GWAS and provide testable hypotheses for future functional validation of prostate cancer risk.

Results

Overview of methods

To identify genes associated with PrCa risk, we performed a TWAS using 48 gene expression panels measured in 45 tissues[22; 30-36] integrated with summary data from the OncoArray PrCa GWAS of 142,392 individuals of European ancestry (81,318/61,074 cases/controls; see Methods)[29]. We performed the summary-based TWAS approach as described in ref[25] using the FUSION software (see Methods). Briefly, this approach uses reference linkage-disequilibrium (LD) and reference gene expression panels with GWAS summary statistics to estimate the association between the cis-genetic component of gene expression, or alternative splicing events, and PrCa risk[25]. First, for each panel, FUSION estimated the heritability of steady-state gene and alternative splicing expression levels explained by SNPs local to each gene (i.e. 1Mb flanking window) using the mixed-linear model (see Methods). Genes with nominally significant ($P < 0.05$) estimates of SNP-heritability (cis-h_g^2), are then put forward for training predictive models. Genes with non-significant estimates of heritability are pruned, as they are unlikely to be accurately predicted. Next, FUSION fits predictive linear models (e.g., Elastic Net, LASSO, GBLUP[37], BSLMM[38]) for every gene using local SNPs. The model with the best cross-validation prediction accuracy (out-of-sample R^2) was used for prediction into the GWAS cohort. This was repeated for all expression datasets, resulting in 117,459 tissue-specific models spanning 16,052 unique genes using total expression and 5,140 using alternatively spliced introns for a combined 17,023 unique genes. The average number of models per expression panel was 2397.6 (see Table S1). Gene expression measured in normal prostate tissue from GTEx[22]

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

resulted in only 854 gene models, which can be explained due to smaller sample size ($N = 87$) compared with the average ($N = 234$; see Table S1). Indeed, the number of gene models per panel was highly correlated with sample size, which implies that statistical power to detect genes with cis-regulatory control is limited by sample size (see Figure S1). Focusing only on models capturing total gene expression, genes on average had heritable levels of expression in 6.4 different panels (median 3) with 11,364 / 16,052 genes having heritable expression in at least 2 panels (see Figure 1). Predictive power of linear gene expression models is upper-bounded by heritability; thus, we use a normalized R^2 to measure in-sample prediction accuracy (R^2 / cis-h_g^2). We found the average R^2 / cis-h_g^2 across all tissue-specific models was 61%, which indicates that most of the signal in cis-regulated total expression and alternative splicing levels is captured by the fitted models (see Figure 1). To assess the predictive stability for models of normal prostate gene expression, we compared measured and predicted gene expression for TCGA[36; 39] samples using models fitted in GTEx[22] normal prostate. We found a highly significant replication ($R^2 = 0.07$; $P = 1.5 \times 10^{-29}$), explaining 39% of in-sample cross-validation R^2 (see Figure S2), which is consistent with previous out-of-sample estimates[24; 25]. We performed a cross-tissue analysis within TCGA and found tumor prostate gene expression models replicated in normal prostate (total expression $R^2 = 0.06$; splicing $R^2 = 0.05$; see Table S2). Given the large number of genes having evidence of genetic control across multiple tissues, we next aimed to measure the similarity of different tissue models (see Methods). Across all reference panels for each gene we observed an average $R^2 = 0.64$ (see Figure S3). Similarly, when averaging across genes, reference panels displayed an average cross-tissue $R^2 = 0.52$ (see Figure S4). Together, these results suggest that trained models predict similar levels of cis-regulated expression on average, despite reference panels measuring expression in different tissues, from varying QC, and capture technologies. Next, we performed simulations to measure the statistical power of TWAS under a variety of trait architectures (see Supplementary Note). Consistent with previous work, we found TWAS to be well-powered at various effect-sizes and heritability levels for gene expression. Importantly, we found no inflation under the null when cis-regulated gene expression has no effect on downstream trait (see Figure S5).

Multi-tissue TWAS identifies 235 genes associated with PrCa status

In total, we tested 117,459 tissue-specific gene models of expression for association with PrCa status and observed 932 reaching transcriptome-wide significance ($P_{TWAS} < 4.26 \times 10^{-7}$), resulting in 235 unique genes, of which 118 were significant in more than one panel (see Table S3; Figure 2). On average, we found 16.8 tissue-specific models associated with PrCa per

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reference expression panel (see Table S1). In 1Mb regions with at least 1 transcriptome-wide significant gene, we observed 10.7 tissue-specific associated models on average, and 2.7 associated genes on average, indicating that further refinement of association signal at TWAS risk loci is necessary. To quantify the overlap between non-HLA, autosomal risk loci in the OncoArray PrCa GWAS and our TWAS results, we partitioned GWAS summary data into 1Mb regions and observed 131 harboring at least one genome-wide significant SNP. Of these, 126/131 overlapped at least one gene model in our data and 68/131 overlapped at least one transcriptome-wide significant gene (see Figure S6). Associated genes were the closest gene to the top GWAS SNP 20% of the time when using 26,292 RefSeq genes. This result is consistent with previous reports[9; 25; 26] and suggests that prioritizing genes based on distance to index SNPs is suboptimal. We found gene model associations were largely consistent, further supporting the predictive stability of models using cis-SNPs (see Figure S7; Supplementary Note). We observed little evidence of prediction accuracy introducing biased results (see Figure S8; Supplementary Note). As a partial control, we compared TWAS results with S-PrediXcan, a related method for predicting gene expression into GWAS summary statistics, using independently trained models and observed a strong correlation ($R = 0.87$; see Figure S9; Supplementary Note), further supporting the validity of the TWAS approach.

Most of the gene models captured total expression levels in normal tissues, however as a positive control we included models for total expression in tumor prostate tissue (see Methods). Predicted expression using tumor prostate models accounted only for 42/235 significant genes compared with 6/235 in normal prostate which is likely due to the large difference in sample size between the original reference panels (see Table S1). Given this, we found no significant increase in proportion of tumor prostate associated models compared with normal prostate (Fisher's exact $P = 0.27$). Of the 335 genes with models trained in both reference panels a single shared gene, *MLPH* (OMIM: 606526, a gene whose function is related to melanosome transport[40]), was associated with PrCa risk (see Table S2). 11/42 genes were significant only in tumor prostate models of total expression. 7/11 genes were modeled in other panels but did not reach transcriptome-wide significance while the other 4/11 were not significantly heritable, and thus not testable, in other panels. We also tested models of alternatively spliced introns for association to PrCa risk. We identified predicted expression of alternatively spliced introns in tumor prostate accounted for 69/235 genes, with an average of 2.5 (median 1) alternatively spliced intron associations per gene. We next quantified the amount of overlap between results driven from models of alternative splicing events versus models of total gene expression. 24/69 genes were found only in alternatively spliced introns, and 16/24 genes had models of total

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

gene expression but did not reach transcriptome-wide significance. The remaining 8/24 were tested solely in alternatively spliced introns, due to heritability of total gene expression not reaching significance. Together these results emphasize earlier work demonstrating that sQTLs for a gene commonly capture signal independent of eQTLs[41].

TWAS analysis increases power to find PrCa associations

Most of the power in the TWAS approach can be attributed to large GWAS sample size. However, two other factors can increase power over GWAS. First, TWAS carries a reduced testing burden compared with that of GWAS, due to TWAS having many fewer genes compared with SNPs. 10/235 genes were located at 9 novel independent 1Mb regions (i.e. no overlapping GWAS SNP), all of which remained significant under a summary-based permutation test ($P < 0.05 / 10$; see Table 1; Table S2; Methods). We found this result was stable to increasing region sizes (see Table S4) and unlikely be the result of long-range tagging with known GWAS risk (see Table S5; Supplemental Note). We observed increased association signal for SNPs at these regions compared to the genome-wide background after accounting for similar MAF and LD patterns (see Figure S10), which, together with observed TWAS associations, suggests that GWAS sample size is still a limiting factor in identifying PrCa risk SNPs. As a partially independent check, we performed a multi-tissue TWAS using summary data from an earlier PrCa GWAS ($N = 49{,}346$)[7] and found 2 novel regions. We found both regions to overlap a genome-wide significant SNP within 1Mb in this data further supporting the robustness of TWAS (see Table S6). Second, we expect to observe increased association signal when expression of a risk gene is regulated by multiple local SNPs[25]. We observed 90/932 instances across 31 genes where TWAS association statistics were stronger than the respective top overlapping GWAS SNP statistics (one-sided Fisher's exact $P < 2.2 \times 10^{-16}$; 7% higher χ^2 statistics on average). For example, *GRHL3* (OMIM:608317; a gene associated with suppression of squamous cell carcinoma tumors[42]) exhibited stronger signal in TWAS using expression in prostate tumor ($P_{TWAS} = 9.38 \times 10^{-10}$) compared with the lead SNP signal ($P_{GWAS} = 1.49 \times 10^{-5}$). Similarly, *POLI* (OMIM:605252, a DNA repair gene associated with mutagenesis of cancer cells[43; 44]) resulted in larger TWAS associations ($P_{TWAS} = 2.01 \times 10^{-7}$) compared with the best proximal SNP ($P_{GWAS} = 5.44 \times 10^{-7}$).

TWAS replicates previously reported genes

We next sought to quantify the extent of overlapping results between TWAS and previous studies that integrated eQTL data measured in normal and tumor prostate tissues at PrCa risk

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

regions (see Methods; see Table S7)[5; 14-20]. We considered only autosomal, non-HLA genes which resulted in 130 previously reported genes. We found a significant overlap between reported genes, with 104/130 assayed in our study and 57/104 reaching transcriptome-wide significance in at least one of our panels (Fisher's exact $P < 2.2 \times 10^{-16}$; see Tables S7-S8). For example, *MLPH* was reported in 4/8 studies. We found significant associations suggesting that decreased expression of *MLPH* in normal and tumor prostate tissue increases risk for PrCa (e.g., GTEx prostate *MLPH* $Z_{TWAS} = -5.80;\ P_{TWAS} = 6.69 \times 10^{-9}$; TCGA prostate $Z_{TWAS} = -6.77;\ P_{TWAS} = 1.25 \times 10^{-11}$). Predicted *MLPH* in tumor prostate remained significant under permutation, which suggests that chance co-localization with GWAS risk is unlikely (Table S2). To assess the amount of residual association signal due to genetic variation in the GWAS risk region after accounting for predicted expression of *MLPH* we performed a summary-based conditional analysis (see Methods). We found *MLPH* to explain most of the signal at its region (lead SNP $P_{GWAS} = 4.03 \times 10^{-11}$; conditioned on *MLPH* lead SNP $P_{GWAS} = 1.13 \times 10^{-3}$; see Figure 3). Our findings are consistent with recent work that found decreased expression levels of *MLPH* to be associated with increased PrCa risk[45]. Despite previous eQTL data focusing on normal and tumor prostate tissue, we observed associations in 49 expression panels overlapping the 57 observed genes in total, underscoring earlier works demonstrating the consistency of cross-tissue cis-regulatory effects[46].

Bayesian prioritization pinpoints a single gene for most TWAS risk regions

TWAS genes are indicative of association and do not necessarily reflect causality (e.g., due to co-regulation at the same region). To prioritize genes at regions with multiple TWAS signals (Figure 2), we used a Bayesian formulation to estimate 90%-credible gene sets (see Methods). We found 120 unique genes across 87 non-overlapping 1Mb regions comprising our 90% credible sets (see Tables S9-S10). 71/87 credible sets contained either a single gene or the same gene in multiple tissues. The average number of unique genes per credible set was 1.38 (median 1). 27/120 prioritized genes were previously reported in eQTL analyses[5; 14-20], which supports the hypothesis that TWAS followed by Bayesian prioritization refines associations to relevant disease genes. For example, *MLPH* was the sole gene defining its region's 90% credible set with a posterior probability of 94%. Similarly, *SLC22A3* (OMIM: 604842; a gene involved in polyspecific organic cation transporters[47] and previously implicated in PrCa risk[18]) exhibited > 94% posterior probability to be causal.

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Expression and splicing events predicted in prostate tissue have largest average effect

Given the large number of significant associations observed for non-prostate tissues in our data, we wanted to quantify which tissue is most relevant for PrCa risk. We first grouped TWAS PrCa associations into prostate/non-prostate and tested for enrichment in normal and tumor prostate expression models. Predicted expression and splicing events in normal and tumor prostate made up 223/932 associations with PrCa (see Table S2) which was highly significant compared to the grouping of all other tissues (Fisher's exact $P = 7.8 \times 10^{-9}$). This measure only quantifies the total amount of observed associations and neglects average association strength. Next, we computed the mean TWAS association statistic using all genes predicted from each expression reference panel (see Figure 4). We observed the largest average TWAS associations in genes predicted from normal and tumor prostate tissue, which reaffirms our intuition of expression and slice events in prostate being the most relevant for PrCa risk. We re-ranked mean associations using only genes found to be transcriptome-wide significant and observed a similar ordering with total expression in normal prostate ranked highest (average $\chi^2 = 176.2$; see Figure S11).

Discussion

Prostate cancer is a common male cancer that is expected to affect more than 180,000 men in the United States in 2017 alone[48]. While GWAS has been successful in localizing risk for PrCa due to genetic variation, the underlying susceptibility genes remain elusive. Here, we have presented results of a transcriptome-wide association study using the OncoArray PrCa GWAS summary statistics for over 142,000 case/control samples. This approach utilizes imputed expression levels and splicing events in the GWAS samples to identify and prioritize putative susceptibility genes. We identified 235 genes whose expression is associated with PrCa risk. These genes localized at 87 genomic regions, of which 9 regions do not overlap with a genome-wide significant SNP in the OncoArray GWAS. We found 24 genes using predictive models for alternatively spliced introns in tumor prostate, which supports the its role in continued risk for tumor oncogenesis. A large fraction of identified genes was confirmed in earlier work, with 57 genes previously reported in eQTL/PrCa GWAS overlap studies. We used a novel Bayesian prioritization approach to refine our associations to credible sets of 120 genes with statistical evidence of causality under standard assumptions. Our results provide a functional map for PrCa risk which can be explored for follow-up and validation.

In this study, we compared our reported TWAS results with genes identified in previous works focusing on expression measured in normal and tumor prostate tissue. Several of these studies

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

considered an eQTL and GWAS risk SNP to overlap if they are in linkage at a specified threshold. While these approaches are sound, they may be limited in statistical power for several reasons. First, if multiple local SNPs independently contribute to risk, overlap studies relying only on the top risk SNP will lose power. Second, earlier overlap studies used thresholds for association signal (i.e., GWAS $P < 5 \times 10^{-8}$) and linkage strength (i.e., LD > 0.5) to consider pairs of SNPs for evidence of expression influencing risk of PrCa. TWAS is largely agnostic to both issues as it jointly considers all SNPs in the region, regardless of reported GWAS association strength. However, when expression of a risk gene is regulated by a single causal SNP, we expect TWAS and earlier overlap approaches to have similar levels in power[25].

Previous works have strongly implicated expression of certain genes in PrCa risk that were not assayed in our study (e.g., MSMB[18; 49]) due to non-significant heritability estimates. TWAS operates by fitting predictive linear models of gene expression based on local genotype data, followed by prediction into large cohorts and subsequent association testing. Expression of genes that are not significantly heritable at current sample sizes are not included in the pipeline. This is the consequence of heritability providing an upper bound on the predictive accuracy under a linear model for genotype; therefore, if a gene has undetectable heritability at a given sample size, it will be difficult to predict using linear combinations of SNPs. To compute TWAS weights for normal prostate tissue, we used samples collected in the GTEx v6 panel ($n = 87$). Thus, our inability to detect heritable levels of gene expression can be explained due to the relatively small number of samples compared with other tissues. Indeed, previous work has shown a strong correlation between sample size in expression panels and the number of identified eGenes[27]; therefore, as sample size increases for relevant tissues, we expect the number of genes included in the TWAS framework to increase. TWAS will lose power in situations where gene expression is a non-linear function of local SNPs, or when trans (or distal) regulation is a major component in modulating expression levels.

We conclude with several caveats and possible future directions. First, while TWAS associations are consistent with models of steady-state gene expression levels altering risk for PrCa, they may be the result of confounding[25; 26]. Imputed gene expression levels are the result of weighted linear combinations of SNPs, many of which may tag non-regulatory mechanisms driving risk and result in inflated association statistics. Second, since genes with eQTLs are common, associations may be the result of chance co-localization between eQTLs and PrCa risk. Lastly, we note recent work has extended TWAS-like methods to expose regulatory mechanisms for susceptibility genes by incorporating chromatin information[50]. An extension to

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

our work would be to pinpoint chromatin variation regulating expression levels at identified risk genes, thus describing a richer landscape of the molecular cascade where SNP → chromatin → expression → PrCa risk.

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

URLs

1000Genomes Phase3: http://www.internationalgenome.org/

Fire Hose v2016_1_28: http://gdac.broadinstitute.org/

FUSION: http://gusevlab.org/projects/fusion/

GCTA v1.26: http://cnsgenomics.com/software/gcta/

GEMMA v0.94: http://www.xzlab.org/software.html

GOseq v1.26: http://bioinf.wehi.edu.au/software/goseq/

MapSplice v2: http://www.netlab.uky.edu/p/bioinfo/MapSplice2

PLINK v1.9: https://www.cog-genomics.org/plink2/

OncoArray: https://epi.grants.cancer.gov/oncoarray/

Funding

CRUK and PRACTICAL consortium

This work was supported by the Canadian Institutes of Health Research, European Commission's Seventh Framework Programme grant agreement n° 223175 (HEALTH-F2-2009-223175), Cancer Research UK Grants C5047/A7357, C1287/A10118, C1287/A16563, 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).

We thank the following for funding support: The Institute of Cancer Research and The Everyman Campaign, The Prostate Cancer Research Foundation, Prostate Research Campaign UK (now Prostate Action), The Orchid Cancer Appeal, The National Cancer Research Network UK, The National Cancer Research Institute (NCRI) UK. We are grateful for support of NIHR funding to the NIHR Biomedical Research Centre at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust and the NIHR Biomedical Research Centre at the University of Cambridge. The Prostate Cancer Program of Cancer Council Victoria also acknowledge grant support from The National Health and Medical Research Council, Australia (126402, 209057, 251533, 396414, 450104, 504700, 504702, 504715, 623204, 940394, 614296), VicHealth, Cancer Council Victoria, The Prostate Cancer Foundation of Australia, The Whitten Foundation, Price Waterhouse Coopers, and Tattersall's.

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genotyping of the OncoArray was funded by the US National Institutes of Health (NIH) [U19 CA 148537 for ELucidating Loci Involved in Prostate cancer SuscEptibility (ELLIPSE) project and X01HG007492 to the Center for Inherited Disease Research (CIDR) under contract number HHSN268201200008I]. Additional analytic support was provided by NIH NCI U01 CA188392 (PI: Schumacher).

Funding for the iCOGS infrastructure came from: the European Community's Seventh Framework Programme under grant agreement n° 223175 (HEALTH-F2-2009-223175) (COGS), Cancer Research UK (C1287/A10118, C1287/A 10710, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007, C5047/A10692, C8197/A16565), 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 Defence (W81XWH-10-1-0341), the Canadian Institutes of Health Research (CIHR) for the CIHR Team in Familial Risks of Breast Cancer, Komen Foundation for the Cure, the Breast Cancer Research Foundation, and the Ovarian Cancer Research Fund.

BPC3

The BPC3 was supported by the U.S. National Institutes of Health, National Cancer Institute (cooperative agreements U01-CA98233 to D.J.H., U01-CA98710 to S.M.G., U01-CA98216 toE.R., and U01-CA98758 to B.E.H., and Intramural Research Program of NIH/National Cancer Institute, Division of Cancer Epidemiology and Genetics).

CAPS

CAPS GWAS study was supported by the Swedish Cancer Foundation (grant no 09-0677, 11-484, 12-823), the Cancer Risk Prediction Center (CRisP; www.crispcenter.org), a Linneus Centre (Contract ID 70867902) financed by the Swedish Research Council, Swedish Research Council (grant no K2010-70X-20430-04-3, 2014-2269)

PEGASUS

PEGASUS was supported by the Intramural Research Program, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health.

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

OncoArray GWAS summary statistics

Genome-wide association summary statistics for the OncoArray PrCa study were obtained from ref[29]. Summary statistics were computed using a fixed-effect meta-analysis for 142,392 total samples of European ancestry from the OncoArray (81,318/61,074 cases/controls), UK stage 1 (1,854/1,894) and UK stage 2 (3,706/3,884), CaPS 1 (474/482) and CaPS 2 (1,458/512), BPC3 (2,068/3,011), NCI PEGASUS (4,600/2,941) and iCOGS (20,219/ 20,440). The initial summary data contained association statistics for 19,726,430 variants. We filtered out summary statistics for SNPs with MAF < 0.01 and any SNPs with ambiguous alternative alleles (e.g., A→T; C→G; or vice-versa). Lastly, we kept only SNPs with rsIDs defined by dbSNP144. Our QC pipeline resulted in association statistics at 10,516,237 SNPs for downstream TWAS analyses.

Previous studies investigating the overlap of eQTL in prostate with risk of PrCa

We collected previous studies that investigated the overlap of eQTLs in normal and tumor prostate tissue at known PrCa risk loci[5; 14-20]. We compared TWAS statistics versus reported eQTL overlap results as aggregated in refs[14; 15]. Across these studies, overlap of eQTLs and PrCa risk loci are computed by one of two possible methods. The first method tests known PrCa risk SNPs for association with expression levels of nearby genes/transcripts. The second method takes a two-step approach. First, genes nearby PrCa risk loci are tested for harboring eQTLs at some significance level. Next, genes with identified eQTL SNPs are tested to be in LD with known PrCa risk variants at some level (e.g., $r^2 > 0.5$).

Reference gene expression data sets and predictive models of expression

We downloaded the FUSION software (see URLs) along with its prepackaged weights for gene expression data. FUSION is an R package that implements the TWAS scheme described in ref[25]. Weights for gene expression measured using RNA sequencing data were obtained from the CommonMind Consortium[30] (dorsolateral prefrontal cortex, $n = 452$), the Genotype-Tissue Expression Project[22] (GTEx; 44 tissues; $n = 449$), the Metabolic Syndrome in Men study[32; 33] (adipose, $n = 563$), and The Cancer Genome Atlas (TCGA; prostate adenocarcinoma, $n = 483$)[39]. Expression microarray data were obtained from the Netherlands Twins Registry[35] (NTR; blood, $n = 1{,}247$), and the Young Finns Study[31; 34] (YFS; blood, $n = 1{,}264$). All non-TCGA expression panel individuals were PrCa controls. Detailed description of quality control procedures on measured gene expression and genotype information for all non-TCGA reference panels are described in refs[25; 27]. TCGA genotype, gene expression, and exon-

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

junction data for 525 samples were downloaded using the Broad GDAC FireHose version 2016_1_28 (see URLs). Genotypes were imputed to the Haplotype Reference Consortium[51] and restricted to well-imputed (INFO > 0.9) HapMap3[52] sites. Genes (exon junctions) missing in more than half of samples were removed. RPKM and log-adjusted gene expression levels were estimated in a generalized linear model controlling for 3 gene-expression PCs and rank-normalized. We estimated alternatively spliced introns using the software MapSplice version 2 (see URLs). A total of 482 samples passed quality control procedures in both genotype and gene expression data.

We filtered genes that did not exhibit cis-genetic regulation at current samples sizes by keeping only genes with nominally significant ($P < 0.05$) estimates of cis-SNP heritability (cis-h_g^2), which resulted in 117,459 total tissue-gene pairs from 17,023 unique genes. We refrain from reporting genes from the HLA region due to complicated LD patterns.

To train predictive models, FUSION defines gene expression for n samples ($\boldsymbol{y}_{GE}$) as a linear function of p SNPs ($\boldsymbol{X}$) in a 1Mb region flaking the gene as

$$\boldsymbol{y}_{GE} = \boldsymbol{C\beta} + \boldsymbol{X w}_{GE} + \boldsymbol{\epsilon},$$

where $\boldsymbol{w}_{GE}$ are the p SNP weights, $\boldsymbol{C\beta}$ are covariates (e.g., sex, age, genotype principal components, genotyping platform, PEER factors) and their effects, and $\boldsymbol{\epsilon}$ is random environmental noise. FUSION estimated weights for expression of a gene in a tissue using multiple penalized linear models. Generally, FUSION optimizes for

$$\begin{bmatrix} \widehat{\boldsymbol{w}}_{GE} \\ \widehat{\boldsymbol{\beta}} \end{bmatrix} = \arg\min_{\boldsymbol{w}_{GE},\boldsymbol{\beta}} \|\boldsymbol{y}_{GE} - \boldsymbol{X w}_{GE} - \boldsymbol{C\beta}\|_2^2 + f(\boldsymbol{w}_{GE}),$$

where $f(\boldsymbol{w}_{GE})$ is a parameterized penalty function specific to each model (e.g., GBLUP[37], LASSO, the Elastic Net). The exception to this optimization criterion is the Bayesian sparse linear mixed model (i.e. BSLMM)[38] which fits the posterior mean for $\boldsymbol{w}_{GE}$ using MCMC in the GEMMA v 0.94 software (see URLs) to obtain weights. To determine which model has the best prediction accuracy for a given gene-tissue pair, FUSION computes out-of-sample R^2 by performing 5-fold cross-validation for each model. Weights from the model with the largest R^2 were used to compute TWAS association statistics. We compute the normalized prediction accuracy for a gene as $\min\left(\frac{R^2}{h_g^2}, 1\right)$.

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Cis-heritability of gene expression

FUSION reports the estimated SNP-heritability (i.e. h_g^2) for measured gene expression levels explained by SNPs in the cis-region (1 Mb region surrounding the TSS). This is modeled under a mixed-linear model as

$$\mathrm{var}(\boldsymbol{y}'_{GE}) = \boldsymbol{A}\sigma_g^2 + \boldsymbol{I}\sigma_e^2,$$

where $\boldsymbol{y}'_{GE}$ is the residual gene expression after regressing out fixed-effect covariates $\boldsymbol{C}$, $\boldsymbol{A}$ is the estimated kinship matrix from SNPs in the cis-region and σ_g^2 (σ_e^2) is the variance explained by the cis-SNPs (environment). SNP-heritability is then defined to be ratio of genotypic variance and total trait variance as, $h_g^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$. Variance parameters are estimated using the AI-REML algorithm implemented in GCTA v1.26 (see URLs) with the top 3 genotypic principal components, sex, age, genotyping platform, and PEER factors as covariates.

Measuring cross-tissue similarity in predicted expression

We took an unbiased approach to identify susceptibility genes for PrCa by using gene expression panels measured in various tissues. To quantify how similar predicted expression levels are for the same gene across different tissues we measured the squared Pearson correlation (R^2). This value represents how well predicted expression from one tissue may be used to predict expression in another tissue. To dissect similarities and differences of tissue-specific models, the ideal scenario would be to inspect effects at individual SNPs defining the models. In practice this is not possible due to predictive models not including the same set of SNPs due to QC and technological differences in the original studies. Therefore, as a proxy we predict gene expression into the 489 samples of European ancestry from 1000 Genomes[53] and compute R^2 across shared genes for pairs of tissues (see Supplementary Note).

Transcriptome-wide association study using GWAS summary statistics

FUSION estimates the strength of association between predicted expression of a gene and PrCa (z_{TWAS}) as function of the vector of GWAS summary Z-scores at a given cis locus $\boldsymbol{z}_{GWAS}$ (i.e. vector of SNP association Wald statistics) and the LD-adjusted weights vector learned from the gene expression data $\boldsymbol{w}_{GE}$ as

$$z_{TWAS} = \frac{\boldsymbol{w}'_{GE}\boldsymbol{z}_{GWAS}}{\sqrt{var(\boldsymbol{w}'_{GE}\boldsymbol{z}_{GWAS})}} = \frac{\boldsymbol{w}'_{GE}\boldsymbol{z}_{GWAS}}{\sqrt{\boldsymbol{w}'_{GE}\boldsymbol{V}\boldsymbol{w}_{GE}}}$$

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where $\boldsymbol{V}$ is a correlation matrix across SNPs at the locus (i.e. LD) and "'" indicates transpose. A P-value for z_{TWAS} is obtained using a two-tailed test under $N(0,1)$. In this work, we estimated $\boldsymbol{V}$ using 489 samples of European ancestry in 1000 Genomes[53]**.** To account for the large number of hypotheses tested, we perform a conservative Bonferroni correction at $\alpha = 0.05 / M$, where $M = 117{,}459$ is the number of predictive models. As reported by ref[25], there may be inflation at GWAS risk loci, due to chance co-varying of SNP effects between expression and PrCa. The same work described a permutation procedure that assesses likelihood of observing association by chance conditioned on GWAS signal. The algorithm works by permuting the eQTL weights $\boldsymbol{w}_{GE}$ while keeping $\boldsymbol{z}_{GWAS}$ fixed and computing $z_{TWAS,perm}$. FUSION implements an adaptive procedure that stops once enough scores (i.e. $|z_{TWAS,perm}| \geq |z_{TWAS}|$) have been observed such that the empirical null cannot be rejected at a specified level. We define novel risk regions as a flanking region around a transcriptome-wide significant gene (splicing event; $P_{TWAS} < 4.26 \times 10^{-7}$) that does not harbor a genome-wide significant SNP ($P_{GWAS} < 5 \times 10^{-8}$). We consider 2Mb windows by default (i.e. TSS ± 1Mb) and show that the results are robust to the choice of window size (see Table S4).

GWAS analyses conditional on predicted expression

To assess the extent of residual association of SNP with PrCa risk after accounting for predicted gene expression levels, FUSION estimates conditional SNP association scores using GWAS summary statistics. Namely, define $\boldsymbol{V}$ as LD for SNPs in the region, $\boldsymbol{V_{GE}}$ as the correlation between predicted expression levels, and $\boldsymbol{C}$ as the correlation between SNPs and predicted expression. The least-squares estimates of $\boldsymbol{z}_{GWAS}|\boldsymbol{z}_{TWAS}$ are determined by,

$$\boldsymbol{z}_{GWAS}|\boldsymbol{z}_{TWAS} = \boldsymbol{z}_{GWAS} - \boldsymbol{C}\boldsymbol{V}_{\boldsymbol{GE}}^{-\boldsymbol{1}}\,\boldsymbol{z}_{TWAS}.$$

The variance of the residual association strength is given by,

$$var[\boldsymbol{z}_{GWAS}|\boldsymbol{z}_{TWAS}] = var[\boldsymbol{z}_{GWAS}] - var[\boldsymbol{C}\boldsymbol{V}_{\boldsymbol{GE}}^{-\boldsymbol{1}}\,\boldsymbol{z}_{TWAS}] = \boldsymbol{V} - \boldsymbol{C}\boldsymbol{V}_{\boldsymbol{GE}}^{-\boldsymbol{1}}\boldsymbol{C}'.$$

This results in the final conditional association score for the ith SNP as,

$$z_i = \left[\boldsymbol{z}_{GWAS} - \boldsymbol{C}\boldsymbol{V}_{\boldsymbol{GE}}^{-\boldsymbol{1}}\,\boldsymbol{z}_{TWAS}\right]_i / \sqrt{diag\left[\boldsymbol{V} - \boldsymbol{C}\boldsymbol{V}_{\boldsymbol{GE}}^{-\boldsymbol{1}}\boldsymbol{C}'\right]_{ii}}.$$

Bayes factors and posterior inference of causal genes

Complex correlations between predicted expression levels at a given region can yield multiple associated genes in TWAS (see Figure 2). Thus, for the vast majority of risk regions it remains unclear which gene is causally influencing PrCa risk. Here we model under the assumption of a

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

single causal gene per risk region and relying on the central limit theorem for normality, we can compute the Bayes Factor that the ith gene in a region is causal as,

$$BF_i = \frac{N(z_{TWAS,i}|0, 1 + n\sigma_\alpha^2)}{N(z_{TWAS,i}|0,1)} = (1 + n\sigma_\alpha^2)^{-1/2} \exp\left(\frac{z_{TWAS,i}^2}{2}\frac{n\sigma_\alpha^2}{1 + n\sigma_\alpha^2}\right),$$

where $z_{TWAS,i}^2$ is the squared TWAS association statistic for the ith gene, n is the GWAS sample size, and σ_α^2 is prior effect-size variance for gene expression on PrCa risk (see Supplementary Note). This model is structurally similar in form to earlier works[54-56] describing Bayes Factors for fine mapping SNPs at GWAS risk regions. The important distinction is that here, we formulate a Bayes Factor for genes at TWAS risk regions. The Bayes Factor for each gene quantifies the amount of evidence in favor of the causal model (ith gene drives risk) versus the null (ith gene has no causal effect). We extend individual Bayes Factors for k genes at a PrCa risk region to compute the posterior probability that a gene is causal as,

$$\Pr(\text{gene } i \text{ is causal} \mid \boldsymbol{z}_{TWAS}, n\sigma_\alpha^2) = \frac{BF_i}{\sum_k BF_k}.$$

Equipped with our definition of posterior probability for each gene being causal, we define ρ-credible gene sets for a PrCa risk region. Formally, a set of indices $i \in I$ defines a ρ-credible gene set if

$$\rho = \sum_{i \in I} \Pr(\text{gene } i \text{ is causal} \mid \boldsymbol{z}_{TWAS}, n\sigma_\alpha^2).$$

For a fixed ρ we optimize over k genes at a region by greedily adding genes until the total density is at least ρ.

To ensure that our ρ-credible sets are well-calibrated we performed simulations by predicting expression levels into 489 samples of European ancestry from 1000 Genomes[53] and estimating the local correlation structure to sample TWAS Z-scores directly (see Supplementary Note). Under the assumption of a single causal gene at a risk region, we sampled TWAS Z-scores for 1000 independent regions. We then performed Bayesian prioritization at each region and computed ρ-credible sets for various levels of ρ while counting the proportion of causal genes identified across all simulations.

Pathway analyses

To determine which pathways may be enriched with genes identified from our Bayesian prioritization approach, we used the R package GOseq[57] which internally links gene identifiers

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to GO terms (GO db: 2017-09-02). We categorized all 17,023 genes into prioritized/not-prioritized and ran the analysis using custom R scripts linking GOseq. GOseq obtains P-values for overrepresented genes using the Wallenius approximation to the non-central hypergeometric distribution. We limited analysis to Gene Ontology Biological Pathways (GO:BP). GOSeq drops genes without GO categories from analysis. We observed 5,005 genes dropped from analyses resulting in 12,018 genes put forward for enrichment tests (see Table S10; Supplementary Note).

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Tables

Table 1. Novel risk loci. TWAS associations that did not overlap a genome-wide significant SNP (i.e. ± 1Mb transcription start site). Study denotes the original expression panel used to fit weights. P-value for TWAS computed under the null of no association between gene expression levels and PrCa risk under a Normal(0, 1) distribution. An asterisk (*) indicates associations that are nominally significant ($P < 0.05/10$) under a permutation test.

Gene	Chr	Tx Start	Tx End	Exon/Exon Junction	Expression Reference	BEST GWAS SNP	BEST GWAS
GRHL3	1	24645811	24690970	-	TCGA.PRAD.TUMOR	rs11589294	1.49E-
				chr1:24668763:24669184	TCGA.PRAD_SP.TUMOR		
RHOA	3	49396578	49449526	-	GTEx.Adipose_Visceral_Omentum	rs34890793	1.17E-
FAM83H	8	144806102	144815914	-	CMC.BRAIN.RNASEQ	rs7831467	3.32E-
TLE4	9	82186687	82341796	-	TCGA.PRAD.TUMOR	rs10117770	2.47E-
				chr9:82189851:82191048	TCGA.PRAD_SP.TUMOR		
				chr9:82268990:82319698	TCGA.PRAD_SP.TUMOR		
				chr9:82319817:82320804	TCGA.PRAD_SP.TUMOR		
				chr9:82320857:82321662	TCGA.PRAD_SP.TUMOR		
				chr9:82321814:82323033	TCGA.PRAD_SP.TUMOR		
				chr9:82323165:82323508	TCGA.PRAD_SP.TUMOR		
				chr9:82323701:82324538	TCGA.PRAD_SP.TUMOR		
				chr9:82324614:82333637	TCGA.PRAD_SP.TUMOR		
				chr9:82333886:82334961	TCGA.PRAD_SP.TUMOR		
				chr9:82335208:82336656	TCGA.PRAD_SP.TUMOR		
				chr9:82336803:82337366	TCGA.PRAD_SP.TUMOR		
				chr9:82337516:82337874	TCGA.PRAD_SP.TUMOR		
				chr9:82337950:82339952	TCGA.PRAD_SP.TUMOR		
STXBP1	9	130374485	130454995	-	TCGA.PRAD.TUMOR	rs1318074	1.79E-
				chr9:130374719:130413882	TCGA.PRAD_SP.TUMOR		
				chr9:130413931:130415994	TCGA.PRAD_SP.TUMOR		
				chr9:130416075:130420654	TCGA.PRAD_SP.TUMOR		
				chr9:130420730:130422309	TCGA.PRAD_SP.TUMOR		
				chr9:130422387:130423381	TCGA.PRAD_SP.TUMOR		
				chr9:130423484:130425484	TCGA.PRAD_SP.TUMOR		
				chr9:130425632:130427526	TCGA.PRAD_SP.TUMOR		
				chr9:130428575:130430359	TCGA.PRAD_SP.TUMOR		
				chr9:130430466:130432177	TCGA.PRAD_SP.TUMOR		
				chr9:130432237:130434330	TCGA.PRAD_SP.TUMOR		
				chr9:130434395:130435460	TCGA.PRAD_SP.TUMOR		
				chr9:130435540:130438083	TCGA.PRAD_SP.TUMOR		
				chr9:130438221:130438923	TCGA.PRAD_SP.TUMOR		
				chr9:130439032:130440710	TCGA.PRAD_SP.TUMOR		
MIR3911	9	130452965	130453074	chr9:130444839:130453054	TCGA.PRAD_SP.TUMOR		
RP11-57H14.2	10	114710405	114711634	-	GTEx.Esophagus_Muscularis	rs11196152	1.61E-
				-	GTEx.Lung		
				-	GTEx.Nerve_Tibial		
				-	GTEx.Pituitary		
				-	GTEx.Thyroid		
				-	GTEx.Whole_Blood		
TM7SF3	12	27124505	27167339	chr12:27129290:27132717	TCGA.PRAD_SP.TUMOR	rs16931510	3.06E-

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

				-	NTR.BLOOD.RNAARR		
				-	GTEx.Adipose_Subcutaneous		
				-	GTEx.Artery_Aorta		
				-	GTEx.Artery_Tibial		
				-	GTEx.Brain_Cerebellar_Hemisphere		
				-	GTEx.Brain_Cerebellum		
				-	GTEx.Brain_Putamen_basal_ganglia		
				-	GTEx.Breast_Mammary_Tissue		
				-	GTEx.Cells_EBV-transformed_lymphocytes		
POLI	18	51795773	51824604	-	GTEx.Colon_Sigmoid	rs11083046	5.44E-
				-	GTEx.Esophagus_Gastroesophageal_Junction		
				-	GTEx.Esophagus_Mucosa		
				-	GTEx.Esophagus_Muscularis		
				-	GTEx.Heart_Atrial_Appendage		
				-	GTEx.Nerve_Tibial		
				-	GTEx.Spleen		
				-	GTEx.Testis		
				-	GTEx.Thyroid		
				-	GTEx.Whole_Blood		
				-	METSIM.ADIPOSE.RNASEQ		
				-	YFS.BLOOD.RNAARR		
KDSR	18	60994959	61034743	-	GTEx.Adipose_Subcutaneous	rs1541296	3.98E-

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figures

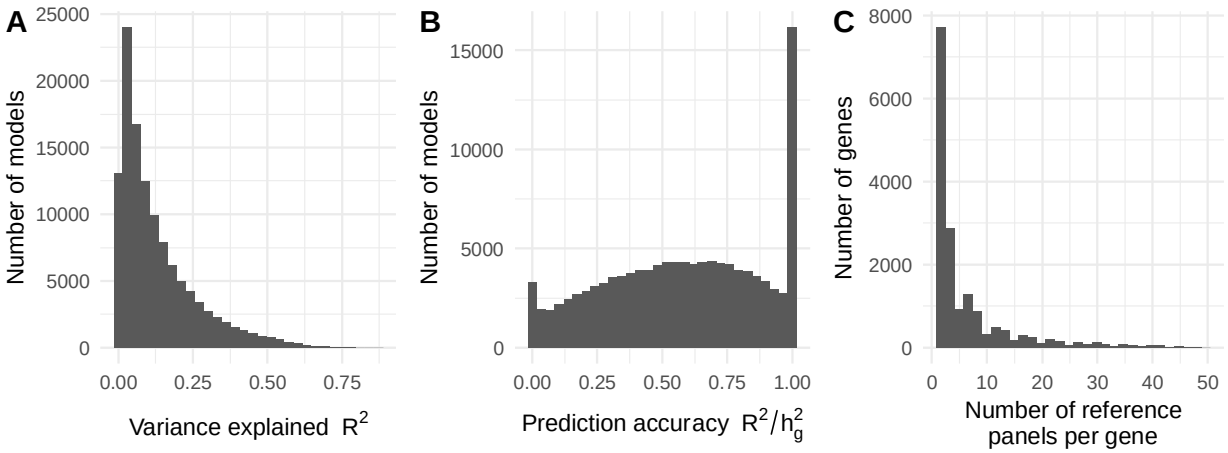

Figure 1. Tissue-specific predictive models for gene expression. A) Cross-validation prediction accuracy of cis-regulated expression and splicing events (R^2) for all 117,459 tissue-specific models. B) Normalized prediction accuracy ($R^2/cis\text{-}h_g^2$) for all 117,459 tissue-specific models. C) Histogram of the number of reference panels per gene. The majority of genes were heritable in a small number of tissues, but many genes exhibited heritable levels across many tissues.

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

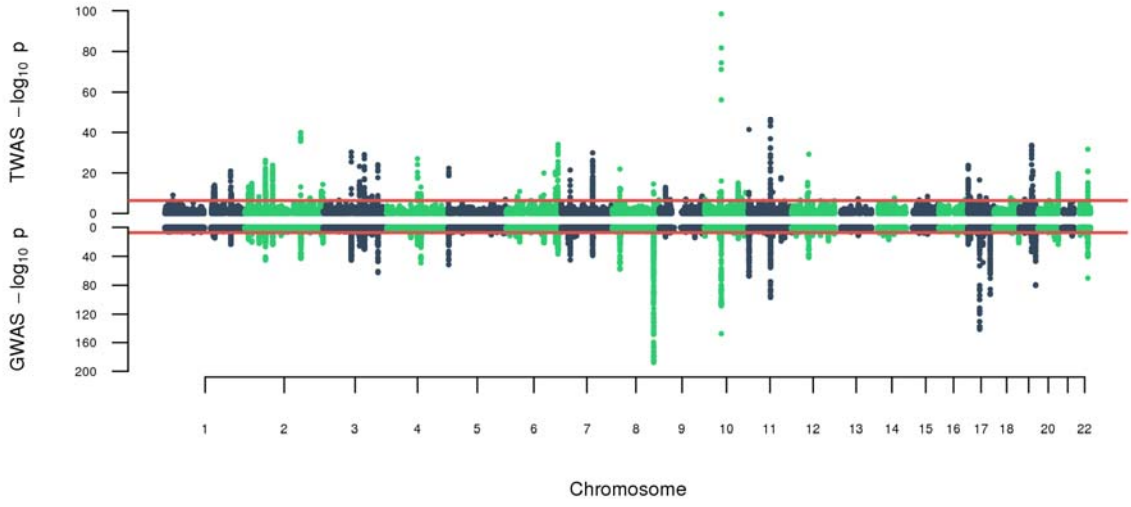

Figure 2. OncoArray PrCa TWAS and GWAS. The top figure is the TWAS Manhattan plot. Each point corresponds to an association test between predicted gene expression with PrCa risk. The red line represents the boundary for transcriptome-wide significance (4.26×10^{-7}). The bottom figure is the GWAS Manhattan plot where each point is the result of a SNP association test with PrCa risk. The red line corresponds to the traditional genome-wide significant boundary (5×10^{-8}).

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

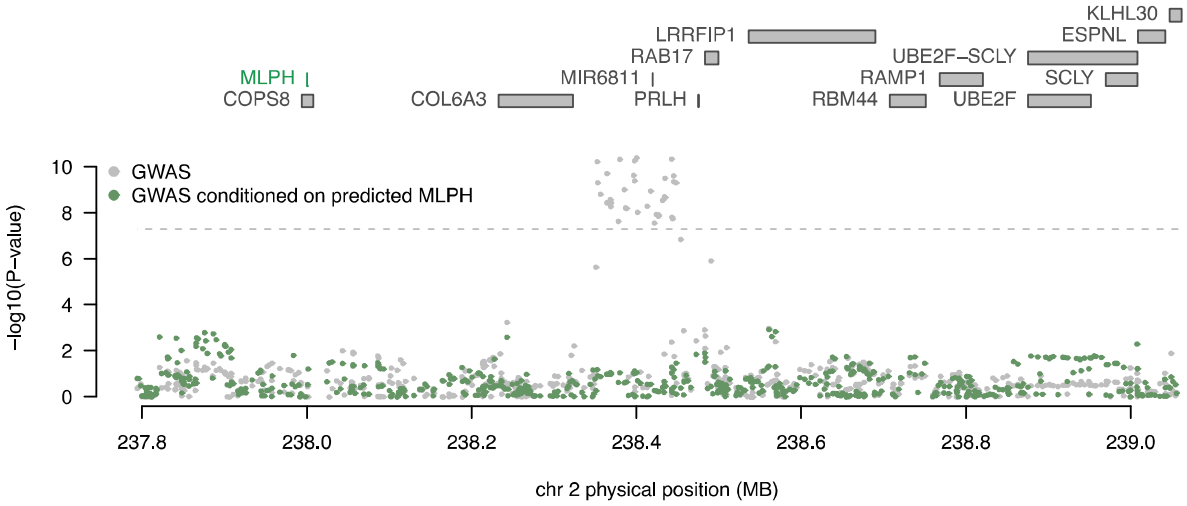

Figure 3. Predicted expression of *MLPH* explains majority of GWAS signal at its genomic region. Each point corresponds to the association between SNP and PrCa status. Gray points indicate the marginal association of a SNP with PrCa status (i.e. GWAS association). Green points indicate the association of the same SNPs with PrCa after conditioning on predicted expression of *MLPH* using models trained from normal prostate (GTEx) and tumor prostate (TCGA). The dashed gray line corresponds to the genome-wide significant threshold (i.e. $P = 5 \times 10^{-8}$). *MLPH* was discussed in previous works as a possible susceptibility gene for PrCa. Association between total expression of *MLPH* and PrCa risk was transcriptome-wide significant in normal and tumor prostate tissue.

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

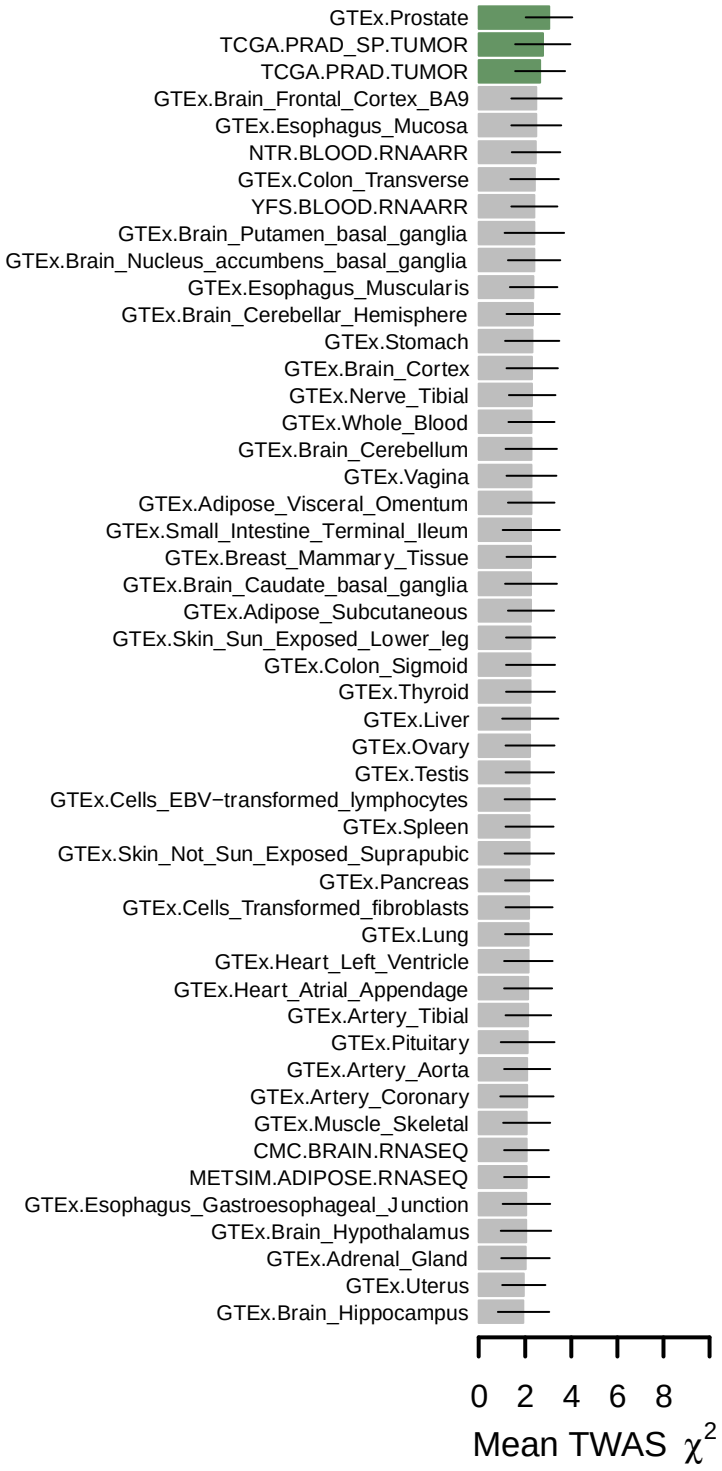

Figure 4. Average TWAS association statistics for genes predicted in each expression panel. Each bar plot corresponds to the average TWAS association statistic using all gene models from a given expression reference panel. Lines represent 1 standard-deviation

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

estimated using the median absolute deviation under normality assumptions. Normal and tumor prostate tissues are marked in green.

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Hjelmborg, J.B., Scheike, T., Holst, K., Skytthe, A., Penney, K.L., Graff, R.E., Pukkala, E., Christensen, K., Adami, H.-O., Holm, N.V., et al. (2014). The Heritability of Prostate Cancer in the Nordic Twin Study of Cancer. Cancer Epidemiology Biomarkers & Prevention 23, 2303.
2. Mucci, L.A., Hjelmborg, J.B., Harris, J.R., and et al. (2016). Familial risk and heritability of cancer among twins in nordic countries. JAMA 315, 68-76.
3. Eeles, R.A., Olama, A.A.A., Benlloch, S., Saunders, E.J., Leongamornlert, D.A., Tymrakiewicz, M., Ghoussaini, M., Luccarini, C., Dennis, J., Jugurnauth-Little, S., et al. (2013). Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. Nat Genet 45, 385-391.
4. Amin Al Olama, A., Dadaev, T., Hazelett, D.J., Li, Q., Leongamornlert, D., Saunders, E.J., Stephens, S., Cieza-Borrella, C., Whitmore, I., Benlloch Garcia, S., et al. (2015). Multiple novel prostate cancer susceptibility signals identified by fine-mapping of known risk loci among Europeans. Human Molecular Genetics 24, 5589-5602.
5. Al Olama, A.A., Kote-Jarai, Z., Berndt, S.I., Conti, D.V., Schumacher, F., Han, Y., Benlloch, S., Hazelett, D.J., Wang, Z., Saunders, E., et al. (2014). A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. Nat Genet 46, 1103-1109.
6. Al Olama, A.A., Kote-Jarai, Z., Giles, G.G., Guy, M., Morrison, J., Severi, G., Leongamornlert, D.A., Tymrakiewicz, M., Jhavar, S., Saunders, E., et al. (2009). Multiple loci on 8q24 associated with prostate cancer susceptibility. Nat Genet 41, 1058-1060.
7. Eeles, R.A., Kote-Jarai, Z., Giles, G.G., Olama, A.A.A., Guy, M., Jugurnauth, S.K., Mulholland, S., Leongamornlert, D.A., Edwards, S.M., Morrison, J., et al. (2008). Multiple newly identified loci associated with prostate cancer susceptibility. Nat Genet 40, 316-321.
8. Spisak, S., Lawrenson, K., Fu, Y., Csabai, I., Cottman, R.T., Seo, J.-H., Haiman, C., Han, Y., Lenci, R., Li, Q., et al. (2015). CAUSEL: an epigenome- and genome-editing pipeline for establishing function of noncoding GWAS variants. Nat Med 21, 1357-1363.
9. Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. PLoS Genet 6, e1000888.
10. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science 337, 1190-1195.
11. Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature 518, 317-330.
12. Hazelett, D.J., Rhie, S.K., Gaddis, M., Yan, C., Lakeland, D.L., Coetzee, S.G., Henderson, B.E., Noushmehr, H., Cozen, W., Kote-Jarai, Z., et al. (2014). Comprehensive Functional Annotation of 77 Prostate Cancer Risk Loci. PLoS Genet 10, e1004102.
13. 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., et al. (2016). Atlas of prostate cancer heritability in European and African-American men pinpoints tissue-specific regulation. 7, 10979.

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

14. Thibodeau, S.N., French, A.J., McDonnell, S.K., Cheville, J., Middha, S., Tillmans, L., Riska, S., Baheti, S., Larson, M.C., Fogarty, Z., et al. (2015). Identification of candidate genes for prostate cancer-risk SNPs utilizing a normal prostate tissue eQTL data set. Nature Communications 6, 8653.
15. Whitington, T., Gao, P., Song, W., Ross-Adams, H., Lamb, A.D., Yang, Y., Svezia, I., Klevebring, D., Mills, I.G., Karlsson, R., et al. (2016). Gene regulatory mechanisms underpinning prostate cancer susceptibility. Nat Genet 48, 387-397.
16. Penney, K.L., Sinnott, J.A., Tyekucheva, S., Gerke, T., Shui, I.M., Kraft, P., Sesso, H.D., Freedman, M.L., Loda, M., Mucci, L.A., et al. (2015). Association of Prostate Cancer Risk Variants with Gene Expression in Normal and Tumor Tissue. Cancer Epidemiology Biomarkers & Prevention 24, 255.
17. Li, Q., Stram, A., Chen, C., Kar, S., Gayther, S., Pharoah, P., Haiman, C., Stranger, B., Kraft, P., and Freedman, M.L. (2014). Expression QTL-based analyses reveal candidate causal genes and loci across five tumor types. Human Molecular Genetics 23, 5294-5302.
18. Grisanzio, C., Werner, L., Takeda, D., Awoyemi, B.C., Pomerantz, M.M., Yamada, H., Sooriakumaran, P., Robinson, B.D., Leung, R., Schinzel, A.C., et al. (2012). Genetic and functional analyses implicate the NUDT11, HNF1B, and SLC22A3 genes in prostate cancer pathogenesis. Proceedings of the National Academy of Sciences 109, 11252-11257.
19. Xu, X., Hussain, W.M., Vijai, J., Offit, K., Rubin, M.A., Demichelis, F., and Klein, R.J. (2014). Variants at IRX4 as prostate cancer expression quantitative trait loci. Eur J Hum Genet 22, 558-563.
20. Huang, Q., Whitington, T., Gao, P., Lindberg, J.F., Yang, Y., Sun, J., Vaisanen, M.-R., Szulkin, R., Annala, M., Yan, J., et al. (2014). A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding. Nat Genet 46, 126-135.
21. Pomerantz, M.M., Shrestha, Y., Flavin, R.J., Regan, M.M., Penney, K.L., Mucci, L.A., Stampfer, M.J., Hunter, D.J., Chanock, S.J., Schafer, E.J., et al. (2010). Analysis of the 10q11 Cancer Risk Locus Implicates MSMB and NCOA4 in Human Prostate Tumorigenesis. PLOS Genetics 6, e1001204.
22. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet 45, 580-585.
23. Chun, S., Casparino, A., Patsopoulos, N.A., Croteau-Chonka, D.C., Raby, B.A., De Jager, P.L., Sunyaev, S.R., and Cotsapas, C. (2017). Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. Nat Genet advance online publication.
24. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Consortium, G.T., Nicolae, D.L., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat Genet 47, 1091-1098.
25. Gusev A, K.A., Shi H, Bhatia G, Chung W, Penninx B, Jansen R, de Geus E, Boomsma DI, Wright FA, Sullivan PF, Nikkola E, Alvarez M, Civelek M, Lusis AJ, Lehtimäki T, Raitoharju E, Kähönen M, Seppälä I, Raitakari OT, Kuusisto J, Laakso M, Price AL, Pajukanta P, Pasaniuc B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nature Genetics.

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

26. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet advance online publication.
27. Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A., and Pasaniuc, B. (2017). Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. The American Journal of Human Genetics 100, 473-487.
28. Pavlides, J.M.W., Zhu, Z., Gratten, J., McRae, A.F., Wray, N.R., and Yang, J. (2016). Predicting gene targets from integrative analyses of summary data from GWAS and eQTL studies for 28 human complex traits. Genome Medicine 8, 1-6.
29. Schumacher, F.R., et. al. (2017). Prostate cancer meta-analysis of more than 140,000 men identifies 63 novel prostate cancer susceptibility loci. Nature Genetics.
30. Fromer, M., Roussos, P., Sieberts, S.K., Johnson, J.S., Kavanagh, D.H., Perumal, T.M., Ruderfer, D.M., Oh, E.C., Topol, A., Shah, H.R., et al. (2016). Gene Expression Elucidates Functional Impact of Polygenic Risk for Schizophrenia. bioRxiv.
31. Raitakari, O.T., Juonala, M., Rönnemaa, T., Keltikangas-Järvinen, L., Räsänen, L., Pietikäinen, M., Hutri-Kähönen, N., Taittonen, L., Jokinen, E., Marniemi, J., et al. (2008). Cohort Profile: The Cardiovascular Risk in Young Finns Study. International Journal of Epidemiology 37, 1220-1226.
32. Stančáková, A., Civelek, M., Saleem, N.K., Soininen, P., Kangas, A.J., Cederberg, H., Paananen, J., Pihlajamäki, J., Bonnycastle, L.L., Morken, M.A., et al. (2012). Hyperglycemia and a Common Variant of GCKR Are Associated With the Levels of Eight Amino Acids in 9,369 Finnish Men. Diabetes 61, 1895-1902.
33. Stančáková, A., Javorský, M., Kuulasmaa, T., Haffner, S.M., Kuusisto, J., and Laakso, M. (2009). Changes in Insulin Sensitivity and Insulin Release in Relation to Glycemia and Glucose Tolerance in 6,414 Finnish Men. Diabetes 58, 1212-1221.
34. Nuotio, J., Oikonen, M., Magnussen, C.G., Jokinen, E., Laitinen, T., Hutri-Kähönen, N., Kähönen, M., Lehtimäki, T., Taittonen, L., Tossavainen, P., et al. (2014). Cardiovascular risk factors in 2011 and secular trends since 2007: The Cardiovascular Risk in Young Finns Study. Scandinavian Journal of Public Health 42, 563-571.
35. Wright, F.A., Sullivan, P.F., Brooks, A.I., Zou, F., Sun, W., Xia, K., Madar, V., Jansen, R., Chung, W., Zhou, Y.-H., et al. (2014). Heritability and genomics of gene expression in peripheral blood. Nat Genet 46, 430-437.
36. The Cancer Genome Atlas Research, N., Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45, 1113-1120.
37. de los Campos, G., Vazquez, A.I., Fernando, R., Klimentidis, Y.C., and Sorensen, D. (2013). Prediction of Complex Human Traits Using the Genomic Best Linear Unbiased Predictor. PLoS Genet 9, e1003608.
38. Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic Modeling with Bayesian Sparse Linear Mixed Models. PLoS Genet 9, e1003264.
39. Abeshouse, A., Ahn, J., Akbani, R., Ally, A., Amin, S., Andry, Christopher D., Annala, M., Aprikian, A., Armenia, J., Arora, A., et al. The Molecular Taxonomy of Primary Prostate Cancer. Cell 163, 1011-1025.

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

40. Matesic, L.E., Yip, R., Reuss, A.E., Swing, D.A., O'Sullivan, T.N., Fletcher, C.F., Copeland, N.G., and Jenkins, N.A. (2001). Mutations in Mlph, encoding a member of the Rab effector family, cause the melanosome transport defects observed in leaden mice. Proceedings of the National Academy of Sciences of the United States of America 98, 10238-10243.
41. Li, Y.I., van de Geijn, B., Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y., and Pritchard, J.K. (2016). RNA splicing is a primary link between genetic variation and disease. Science 352, 600.
42. Darido, C., Georgy, Smitha R., Wilanowski, T., Dworkin, S., Auden, A., Zhao, Q., Rank, G., Srivastava, S., Finlay, Moira J., Papenfuss, Anthony T., et al. Targeting of the Tumor Suppressor GRHL3 by a miR-21-Dependent Proto-Oncogenic Network Results in PTEN Loss and Tumorigenesis. Cancer Cell 20, 635-648.
43. Yang, J., Chen Z Fau - Liu, Y., Liu Y Fau - Hickey, R.J., Hickey Rj Fau - Malkas, L.H., and Malkas, L.H. (2004). Altered DNA polymerase iota expression in breast cancer cells leads to a reduction in DNA replication fidelity and a higher rate of mutagenesis. Cancer Research.
44. Yuan, F., Xu, Z., Yang, M., Wei, Q., Zhang, Y., Yu, J., Zhi, Y., Liu, Y., Chen, Z., and Yang, J. (2013). Overexpressed DNA Polymerase Iota Regulated by JNK/c-Jun Contributes to Hypermutagenesis in Bladder Cancer. PLOS ONE 8, e69317.
45. Bu, H., Narisu, N., Schlick, B., Rainer, J., Manke, T., Schäfer, G., Pasqualini, L., Chines, P., Schweiger, M.R., Fuchsberger, C., et al. (2016). Putative Prostate Cancer Risk SNP in an Androgen Receptor-Binding Site of the Melanophilin Gene Illustrates Enrichment of Risk SNPs in Androgen Receptor Target Sites. Human Mutation 37, 52-64.
46. Gutierrez-Arcelus, M., Ongen, H., Lappalainen, T., Montgomery, S.B., Buil, A., Yurovsky, A., Bryois, J., Padioleau, I., Romano, L., Planchon, A., et al. (2015). Tissue-Specific Effects of Genetic and Epigenetic Variation on Gene Regulation and Splicing. PLoS Genet 11, e1004958.
47. Verhaagh, S., Schweifer, N., Barlow, D.P., and Zwart, R. (1999). Cloning of the Mouse and Human Solute Carrier 22a3 (Slc22a3/SLC22A3) Identifies a Conserved Cluster of Three Organic Cation Transporters on Mouse Chromosome 17 and Human 6q26–q27. Genomics 55, 209-218.
48. Siegel, R.L., Miller, K.D., and Jemal, A. (2016). Cancer statistics, 2016. CA: A Cancer Journal for Clinicians 66, 7-30.
49. Sutcliffe, S., De Marzo, A.M., Sfanos, K.S., and Laurence, M. (2014). MSMB variation and prostate cancer risk: Clues towards a possible fungal etiology. The Prostate 74, 569-578.
50. Gusev, A., Mancuso, N., Finucane, H.K., Reshef, Y., Song, L., Safi, A., Oh, E., McCaroll, S., Neale, B., Ophoff, R., et al. (2016). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. bioRxiv.
51. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 48, 1279-1283.
52. Consortium, T.I.H. (2010). Integrating common and rare genetic variation in diverse human populations. Nature 467, 52-58.
53. The Genomes Project, C. (2015). A global reference for human genetic variation. Nature 526, 68-74.

bioRxiv preprint doi: https://doi.org/10.1101/345736; this version posted June 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

54. Hormozdiari, F., Kichaev, G., Yang, W.-Y., Pasaniuc, B., and Eskin, E. (2015). Identification of causal genes for complex traits. Bioinformatics 31, i206-i213.
55. Maller, J.B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J.M.M., Auton, A., Myers, S., Morris, A., et al. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat Genet 44, 1294-1301.
56. Chen, W., Larrabee, B.R., Ovsyannikova, I.G., Kennedy, R.B., Haralambieva, I.H., Poland, G.A., and Schaid, D.J. (2015). Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. Genetics 200, 719.
57. Young, M.D., Wakefield, M.J., Smyth, G.K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biology 11, R14.