nature genetics

The contribution of rare variation to prostate cancer heritability

Nicholas Mancuso[1,14], Nadin Rohland[2,3,14], Kristin A Rand[4,5], Arti Tandon[2,3], Alexander Allen[2,3], Dominique Quinque[2,3], Swapan Mallick[2,3], Heng Li[2,3], Alex Stram[4], Xin Sheng[4], Zsofia Kote-Jarai[6], Douglas F Easton[7], Rosalind A Eeles[6,8], the PRACTICAL consortium[9], Loic Le Marchand[10], Alex Lubwama[11], Daniel Stram[4,5], Stephen Watya[11], David V Conti[4,5], Brian Henderson[4,5,13], Christopher A Haiman[4,5,15], Bogdan Pasaniuc[1,12,15] & David Reich[2,3,15]

We report targeted sequencing of 63 known prostate cancer risk regions in a multi-ancestry study of 9,237 men and use the data to explore the contribution of low-frequency variation to disease risk. We show that SNPs with minor allele frequencies (MAFs) of 0.1–1% explain a substantial fraction of prostate cancer risk in men of African ancestry. We estimate that these SNPs account for 0.12 (standard error (s.e.) = 0.05) of variance in risk (~42% of the variance contributed by SNPs with MAF of 0.1–50%). This contribution is much larger than the fraction of neutral variation due to SNPs in this class, implying that natural selection has driven down the frequency of many prostate cancer risk alleles; we estimate the coupling between selection and allelic effects at 0.48 (95% confidence interval [0.19, 0.78]) under the Eyre-Walker model. Our results indicate that rare variants make a disproportionate contribution to genetic risk for prostate cancer and suggest the possibility that rare variants may also have an outsize effect on other common traits.

More than 220,000 men are expected to be diagnosed with prostate cancer and more than 27,000 are expected to die of the disease in the United States alone in 2015 (ref. 1). Approximately 58% of risk for prostate cancer has been estimated to be due to inherited genetic factors[2–6]. Thus far, genome-wide association studies (GWAS) have identified more than 100 common risk variants for prostate cancer that explain ~33% of the familial risk[7–25], leaving the majority of risk unexplained. Because GWAS have primarily investigated common variants (MAF >1%) for association with prostate cancer risk, an unexplored hypothesis is that part of the 'missing heritability' is attributable to rare variants (MAF <1%). To address this hypothesis, we focused on examining rare variation at known susceptibility regions that are only partially tagged by GWAS arrays. The rationale for investigating known risk-associated regions is that, (i) unlike in the rest of the genome, genetic variation in these regions has been established to confer risk and (ii) there are examples of rare and low-frequency variation at known GWAS-identified risk regions being important for a number of common diseases, including prostate cancer (8q24)[26–29].

We carried out targeted sequencing of known prostate cancer GWAS loci to investigate the contribution of low-frequency and rare variation to prostate cancer risk. We targeted all 63 autosomal risk regions for prostate cancer that were known to us at the time of study design (since then, an additional 37 loci have been discovered). For each region, we started with the index SNP previously associated with prostate cancer by GWAS and attempted to tile Agilent SureSelect baits to cover all nucleotides within a block of strong linkage disequilibrium (LD) around the SNP (plus exons and conserved elements within 200 kb of the SNP). We constructed individually barcoded next-generation sequencing libraries for all of the samples, pooled these into sets that typically contained 24 libraries each, and then performed in-solution hybrid enrichment. After removal of duplicated molecules, we achieved an average coverage of 9.3× in 9,237 cases and controls across four ancestry groups (4,006 African, 1,753 European, 1,770 Japanese and 1,708 Latino). We identified 197,786 variants that were present in all ancestry groups, imputed genotypes into all individuals, and then correlated the genotypes to prostate cancer risk.

First, we show that sequencing-based association analysis is able to study a substantially larger fraction of the genetic risk for prostate cancer than studies of common variants alone, as we find that the variance explained in the trait by all the sequenced variants is

[1]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, USA. [2]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [3]Broad Institute, Cambridge, Massachusetts, USA. [4]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA. [5]Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, USA. [6]The Institute of Cancer Research, London, UK. [7]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [8]Royal Marsden National Health Service (NHS) Foundation Trust, London and Sutton, UK. [9]A full list of members and affiliations appears in the **Supplementary Note**. [10]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii, USA. [11]School of Public Health, Makerere University College of Health Sciences, Kampala, Uganda. [12]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, USA. [13]Deceased. [14]These authors contributed equally to this work. [15]These authors jointly supervised this work. Correspondence should be addressed to B.P. (pasaniuc@ucla.edu), C.A.H. (christopher.haiman@med.usc.edu) or D.R. (reich@genetics.med.harvard.edu).

Received 2 July; accepted 20 October; published online 16 November 2015; doi:10.1038/ng.3446

significantly larger than the variance explained by known GWAS variants at the same loci. Second, we find evidence of genetic heterogeneity by ancestry in risk for prostate cancer. Third, we use variance-components methods to partition the SNP heritability across different variant frequency classes and find that a large amount of SNP heritability comes from the rare variant class in men of African ancestry; that is, variants with 0.1% ≤ MAF < 1% explain a point estimate of 0.12 of variance in the trait as compared to an estimate of 0.17 for variants with MAF ≥ 1%. Third, we used the SNP heritability assigned to the rare variant class to make the first relatively precise estimate of the strength of coupling between selection and allelic effect for a common trait. Finally, we replicated association signals at known GWAS loci and used an approach that combines epigenetic annotation (e.g., localization of androgen receptor binding sites in a prostate adenocarcinoma cell line) with the association signal to identify plausible causal variants at some of these loci.

RESULTS

Experimental strategy

To explore the contribution of rare and low-frequency variation to risk of prostate cancer, we targeted 90 index SNPs at 63 autosomal regions that had been associated with prostate cancer risk by GWAS at the time that this study was designed (October 2011). For each index SNP, we used Haploview[30] (HapMap release 24) to visualize the surrounding LD block in European-ancestry individuals. We then manually identified boundaries for target capture on the basis of the region where LD as measured by the absolute value of Lewontin's D′ fell precipitously (**Supplementary Table 1** and **Supplementary Data Set 1**). We also targeted all exons (defined on the basis of RefSeq annotation) within 200 kb of each index SNP together with all conserved noncoding sequences (defined on the basis of a 29-mammal alignment[31]) within 5 kb of each exon and elements >100 bp in length or with conservation scores >75 within the 200-kb window centered on each index SNP. Outside of the targeted GWAS loci, we also included exons and conserved elements of *MYC* and *PVT1* because of their potential importance in prostate cancer. We designed and ordered Agilent SureSelect[32] in-solution enrichment probes to target a total of 12 Mb in two rounds of target design. The total span of the regions we wished to target was 16.7 Mb, but we were not able to design probes for 4.7 Mb owing to the presence of repetitive elements that needed to be masked during probe design (**Supplementary Table 1**).

We produced a total of 9,237 next-generation Illumina sequencing libraries from four ancestry groups (4,006 African, 1,753 European, 1,770 Japanese and 1,708 Latino) using a high-throughput library construction strategy previously described in ref. 33 (Online Methods). The results of the sequencing are presented in **Table 1**, where information on the mean coverage and the total number of variants discovered is provided for each ancestry group. The total number of megabases targeted, the mean coverage, the number of sites discovered and other metrics for each region are provided in **Supplementary Table 2**. The average coverage across samples was 9.3×, with s.d. of 5.4 across individuals and 5.4 across targeted nucleotides. We identified a total of 197,786 variants, of which 44% were not identified in the 1000 Genomes Project (**Supplementary Table 3**). The coverage we obtained for the great majority of samples was high enough in theory to obtain reliable diploid genotype calls after imputation at most targeted bases[34]. To assess the accuracy of sequencing, we measured the Pearson correlation of these genotype calls with those made using arrays (roughly half of the samples had also been assayed using GWAS arrays). The correlation between the genotype calls from sequencing and arrays was $r^2 = 0.84$ before imputation, increasing to 0.92 after imputation (**Supplementary Fig. 1**).

Sequencing explains additional variance beyond GWAS SNPs

To explore the value of sequencing in explaining additional variance in prostate cancer risk, we fit the genetic data to variance-components models to estimate the contribution of all genetic variants at the sequenced risk loci to the underlying liability of prostate cancer. First, we used simulations starting from the real genotype data to quantify potential biases in variance-components estimation. Consistent with findings of previous studies[35], our simulations show that the approach of using two variance components—one for rare variants (0.1% ≤ MAF < 1%) and one for common variants (MAF ≥ 1%)—estimated from dosage data and fitted jointly using restricted maximum likelihood (REML) as implemented in GCTA[36] produces the least amount of bias when estimating SNP heritability (**Supplementary Figs. 2–9** and **Supplementary Table 4**). We also investigated the performance of fitting the REML equations with AI-REML, a Newton-style approach, versus an EM-based approach, EM-REML, as implemented in GCTA, with AI-REML attaining the least bias in our data (**Supplementary Fig. 10**, **Supplementary Tables 5** and **6**, and **Supplementary Note**). We considered the effect of estimating SNP heritability from best-guess calls rather than imputed dosages and found that these approaches give statistically indistinguishable results. Lastly, we explored the role of adjustment for LD in estimating the genetic relationship matrix (GRM) and observed upward bias for LD-adjusted GRMs when the underlying heritability explained by rare variants ($h^2_{g,\mathrm{rare}}$) was set to 0 in our simulations. This upward bias was also reflected in estimates made using real phenotype data (**Supplementary Table 7**). Similar results were obtained over a variety of simulated disease architectures with various amounts of contribution from rare variation and total numbers of underlying causal variants (**Supplementary Note**).

Motivated by our simulation findings, we estimated the contribution of rare and common variation to risk of prostate cancer by fitting two variance components in GCTA while correcting for the top ten principal components and age; we report heritability estimates on the liability scale (Online Methods; see **Supplementary Fig. 11** for the principal-component analysis plot). We find that the total variance explained by all variants at these loci (SNP heritability $h^2_g = h^2_{g,\mathrm{rare}} + h^2_{g,\mathrm{common}}$) is larger than what is explained by the index variants alone (**Table 2**). For example, we estimate the variance explained by all variants in the African-ancestry sample at 0.30 (s.e. = 0.06), which is significantly larger ($P < 0.05$) than the variance explained by all 84 index variants present in these data (0.06, s.e. = 0.01) (six of the 90 SNPs targeted were not covered by reads passing our analysis filters). This finding is consistent across all ancestry groups, thus emphasizing the usefulness of sequencing in recovering additional signal beyond index GWAS variants[37].

Table 1 Sizes for each ancestry group and the coverage and standard deviation in coverage achieved

	Number of samples		Average coverage per sample (s.d.)	Average coverage per locus (s.d.)	Variants		
					Rare	Common	
Ancestry	Cases	Controls			0.1% ≤ MAF < 1%	MAF ≥ 1%	Total
African	2,054	1,952	8.3 (5.1)	8.4 (5.2)	58,699	63,972	122,671
European	900	853	8.8 (6.0)	8.8 (6.0)	33,606	53,164	86,770
Japanese	914	856	11.8 (5.2)	11.9 (5.2)	29,121	40,742	69,863
Latino	864	844	8.0 (5.7)	8.1 (5.7)	46,374	45,932	92,306
Overall	4,732	4,505	9.3 (5.4)	9.4 (5.4)	–	–	–

Table 2 Estimates of h_g^2 and standard errors using sequencing data

Ancestry	Sample size	h_g^2 index SNPs (s.e.)	$h_{g,\text{rare}}^2$ (s.e.)	*P* value	$h_{g,\text{common}}^2$ (s.e.)	*P* value
African	4,006	0.06 (0.01)	0.12 (0.05)	2.29×10^{-3}	0.17 (0.03)	7.08×10^{-13}
European	1,753	0.10 (0.01)	0.00 (0.06)	5.00×10^{-1}	0.27 (0.06)	5.83×10^{-11}
Japanese	1,770	0.08 (0.01)	0.05 (0.07)	2.68×10^{-1}	0.13 (0.04)	3.09×10^{-5}
Latino	1,708	0.06 (0.01)	0.00 (0.06)	5.00×10^{-1}	0.14 (0.05)	2.38×10^{-5}

The results for index SNPs correspond to h_g^2 contributed solely from the targeted index variants. Estimates for h_g^2 attributable to rare and common components were obtained from joint REML analysis on the underlying liability scale. Rare variants are defined as those with 0.1% ≤ MAF < 1%, whereas common variants are defined as those with MAF ≥ 1%.

Next, we searched for genetic heterogeneity by ancestry in prostate cancer risk using a bivariate REML analysis[38]. Briefly, we computed a single GRM for each unique pair of ancestry groups over the set of SNPs common to both ancestry groups (Online Methods) and estimated the genetic correlation using GCTA[36]. We then tested the hypotheses that there is no shared genetic liability (SNP $r_g = 0$) and that liability is completely shared (SNP $r_g = 1$) (Online Methods). We find significant heterogeneity (after accounting for the six pairs tested) for the African and European ancestry groups (SNP $r_g = 0.56$, s.e. = 0.15; *P* value (SNP $r_g = 1$) = 2.42×10^{-3}; **Table 3**) and nominally significant heterogeneity (*P* value = 0.04) for the Latino and African ancestry groups (**Table 3**).

Having established evidence of heterogeneity, we quantified the contribution to SNP heritability of variants across the MAF spectrum in each ancestry group independently. Rare variants explained a significant amount of SNP heritability (h_g^2) in African-ancestry individuals ($h_{g,\text{rare}}^2 = 0.12$, s.e. = 0.05; $P = 2.29 \times 10^{-3}$); indeed, the heritability explained by these rare variants is comparable to the heritability explained by common variants at these loci ($h_{g,\text{common}}^2 = 0.17$, s.e. = 0.03; $P = 7.08 \times 10^{-13}$; $\frac{h_{g,\text{rare}}^2}{h_g^2} = 0.42$, s.e. = 0.11; Online Methods).

We did not observe significant contribution of rare variation to heritability in the other ancestry groups, although, given the limited sample sizes for the other groups, we cannot exclude the possibility that the fraction of prostate cancer heritability attributable to rare variants is the same in the other groups. In most of the analyses of heritability stratified by variant frequency that follow, we focus on people of African ancestry, as we had the highest power to carry out these studies.

We investigated whether the large contribution from rare variants in men of African ancestry was an artifact of data quality (**Supplementary Note**). We estimated $h_{g,\text{rare}}^2 = 0.13$ (s.e. = 0.06) for the African-American ancestry group after removing any SNPs whose rate of missing data before imputation was associated with the trait ($P \leq 0.01$) (**Supplementary Table 8**). We obtained similar results when estimating SNP heritability directly from the hard genotype calls before imputation, both with and without the differentially missing SNPs for the African-American group ($h_{g,\text{rare}}^2 = 0.11$, s.e. = 0.05; **Supplementary Table 9**). To quantify whether hidden relatedness influenced our results, we estimated heritability at various relatedness thresholds; differences in relatedness did not significantly influence the SNP heritability explained by rare variants ($h_{g,\text{rare}}^2 = 0.13$, s.e. = 0.06; GRM < 0.05; **Supplementary Table 8; see Supplementary Fig. 12** and **Supplementary Table 10** for distribution of pairwise relatedness values; see **Supplementary Tables 11–18** for results for other ancestry groups). We also explored the role of sequencing coverage and estimated SNP heritability from GRMs computed after removing SNPs at various levels of coverage. Overall, we found no significant decrease in $h_{g,\text{rare}}^2$ until a large fraction of the SNPs were discarded (coverage ≥7×; **Supplementary Table 19**). To rule out potential tagging of signal by other loci in the genome, we repeated the SNP heritability estimation including a third variance component that constitutes genotype calls from arrays for the rest of the genome; this approach yielded similar results for $h_{g,\text{rare}}^2$ (**Supplementary Tables 20** and **21**; see **Supplementary Tables 22–27** for results for other ancestry groups). To account for possible confounding from population substructure, we re-estimated the variance attributable to the rare frequency class in the African-ancestry sample, stratifying on the basis of Ugandan and non-Ugandan ancestry as well as the 8q24 locus, which is known to make a large contribution to risk of prostate cancer. Overall, we found no significant difference in $h_{g,\text{rare}}^2$ for the African-American subsets with and without the 8q24 region included in the estimation (**Supplementary Table 28**). We also considered bias in our initial estimates of h_g^2 resulting from potential misspecification of the GRM. Specifically, we estimated variance components using non-standardized genotype data[35] (thereby reducing the impact of rare variants in GRM computation) and found a similar contribution from the rare variant spectrum (Online Methods). We also standardized the GRM on the basis of the expected variance rather than the sample estimate and found no significant

Table 3 Bivariate REML analysis for each pair of ancestry groups

Ancestry pair	Sample size	SNPs in common	h_g^2 (s.e.)	Covariance	SNP r_g (s.e.)	*P* value (SNP $r_g = 0$)	*P* value (SNP $r_g = 1$)
African	4,006	46,332	0.20 (0.03)	0.17	0.56 (0.15)	2.10×10^{-4}	2.42×10^{-3}
European	1,753		0.25 (0.05)				
African	4,006	31,954	0.18 (0.03)	0.13	0.99 (0.16)	9.86×10^{-8}	0.48
Japanese	1,770		0.13 (0.03)				
African	4,006	61,894	0.21 (0.03)	0.10	0.61 (0.20)	7.65×10^{-4}	0.04
Latino	1,708		0.15 (0.05)				
European	1,753	37,871	0.26 (0.05)	0.15	0.88 (0.18)	3.18×10^{-6}	0.27
Japanese	1,770		0.13 (0.04)				
European	1,753	58,708	0.27 (0.06)	0.22	0.94 (0.19)	2.56×10^{-7}	0.38
Latino	1,708		0.18 (0.05)				
Japanese	1,770	39,485	0.13 (0.04)	0.11	1.00 (0.20)	3.12×10^{-6}	0.50
Latino	1,708		0.18 (0.04)				

Estimates of h_g^2 describe the SNP heritability for each ancestry group over a set of SNPs in common for each pair. Estimates are shown of shared genetic variation in tagged SNPs, or SNP correlation (SNP r_g). The last two columns give the *P* value for the model under the null hypotheses that no correlation exists (SNP $r_g = 0$) and that perfect correlation is present (SNP $r_g = 1$).

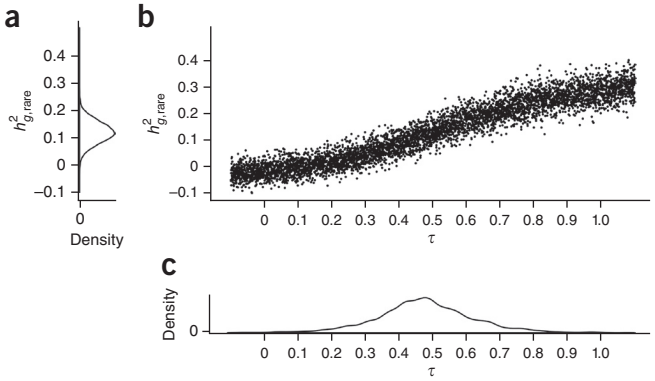

Figure 1 Relationship between strength of selection, the coupling parameter τ and allelic effect sizes in prostate cancer using heritability partitioning for the African-ancestry sample. (**a**) The density estimate for $h^2_{g,\text{rare}}$ obtained from real data. (**b**) The influence of τ on $h^2_{g,\text{rare}}$. Each point represents an estimate of $h^2_{g,\text{rare}}$ given phenotypes simulated from real genotypes under the Eyre-Walker model. (**c**) The estimated empirical density of τ. Estimates were obtained by matching a sampled value of $h^2_{g,\text{rare}}$ from **a** with the closest point estimate from **b**.

change ($h^2_{g,\text{rare}} = 0.12$, s.e. = 0.05). We investigated potential bias in GCTA estimates from linkage across variants of various frequencies by repeating the analysis with three variance components corresponding to rare (0.1% ≤ MAF < 1%), low-frequency (1% ≤ MAF < 5%) and common (MAF ≥ 5%) variants; we observed no significant difference in the amount of variance attributable to the rare variant class (**Supplementary Table 29**). As the standard errors reported by GCTA are asymptotic, we employed a leave-one-out jackknife to estimate $h^2_{g,\text{rare}} = 0.13$ with s.e. = 0.06 in the African-ancestry group (**Supplementary Table 30**). We also randomly sampled 1,753 individuals of African ancestry (corresponding to the size of the European cohort) 100 times and found a mean estimate of $h^2_{g,\text{rare}} = 0.13$ (s.e. = 0.06; **Supplementary Table 31**). To further investigate the significance of $h^2_{g,\text{rare}}$ in African data, we estimated $h^2_{g,\text{rare}}$ in 1,000 simulated phenotypes starting from the real dosage data where the true $h^2_{g,\text{rare}}$ value was set to 0 (all causal variants were set to have MAF ≥ 1%). In none of the 1,000 runs did we observe an estimate of $h^2_{g,\text{rare}} \geq 0.12$, giving an empirical P value <1/1,000. Finally, we performed variance-components analyses using genotypes obtained from best-guess calls, as well as standard unconstrained REML analyses. Overall, we found that most of these potential sources of bias are unlikely to significantly change our results (**Supplementary Figs. 2–9**, **Supplementary Tables 32–35** and **Supplementary Note**).

Evidence of coupling between selection and allelic effects

In the case of neutral genetic variation, alleles that have a MAF <1% account for only a few percent of genetic variation in the population. However, our empirical results from this study show that, at loci known to harbor common variants conferring risk for prostate cancer, variants with MAF < 1% account for an order of magnitude larger heritability for the disease. The only plausible explanation for this observation is that newly arising mutations that confer risk for prostate cancer—especially mutations of strong effect—are often subject to selection that is strong enough to prevent them from becoming common.

To quantify the extent to which selection is driving down the frequency of alleles that confer risk for prostate cancer, we derived a simulation-based pipeline that uses estimates of $h^2_{g,\text{rare}}$ to constrain the value of a parameter τ that Eyre-Walker proposed to measure the coupling between selection coefficients and allelic effect sizes[39] (Online Methods). Briefly, starting from the real genotype data, we simulated phenotypes under Eyre-Walker's model at various values of τ and estimated $h^2_{g,\text{rare}}$ in the simulated trait. We then compared the observed heritability in the real data to the simulations while accounting for sampling noise (Online Methods). We estimated $\tau = 0.48$ with a 95% confidence interval of [0.19, 0.78] for the African-ancestry sample under our mapping procedure (**Fig. 1**). We obtained similar results using a MAF cutoff of 5% in assigning variants to the rare versus common class (**Supplementary Fig. 13**). We found that our procedure was relatively robust to changes in parameters. For example, when adjusting the effective population size for African ancestry to 7,500, we re-estimated $\tau = 0.46$ with a 95% confidence interval of [0.21, 0.78] (**Supplementary Table 36**). Although the small contribution from rare variants together with small sample sizes for the European, Japanese and Latino data sets prohibits us from estimating a tight confidence interval for Eyre-Walker's τ in these populations, the results were roughly consistent across populations (**Table 4** and **Supplementary Fig. 14**). For example, the estimated mean value of τ for the Japanese cohort was 0.38 with a 95% confidence interval of [−0.07, 0.32]. In a meta-analysis over all ancestry groups, we estimated $\tau = 0.42$ [0.22, 0.62], which is similar to the African-ancestry estimates (unsurprisingly, as the African-ancestry data contribute the most to this analysis).

Single-variant association

An advantage of sequencing data—even with a tenfold lower sample size in comparison to the largest current GWAS—is that it interrogates all variants in the analyzed samples and thus has the potential to detect causal variants that are not genotyped or imputed in GWAS. We performed marginal association testing at all sequenced variants (n = 197,786) and replicated most of the GWAS-identified loci (**Supplementary Tables 37–42**). We observed a marginal increase in the association signal when including rare variants with 0.1% ≤MAF < 1% across all populations, as reflected in a decrease in the top $-\log_{10}$ (P value) (**Supplementary Tables 37–42**) and a slight enrichment of low P values in a burden test (**Supplementary Fig. 15**). However, a limitation of the present study is its modest sample size in comparison to the sample size of 87,040 individuals in the most recent GWAS meta-analyses[24]. For example, of the 84 recovered index variants (six of the 90 targeted SNPs were not covered by reads passing our analysis filters), only seven had a P value $<1 \times 10^{-8}$ (most at 8q24) and only 13 had a P value $<1 \times 10^{-4}$. Thus, even though we can directly access alleles not on SNP arrays through our targeted sequencing, the advantage we obtain by directly genotyping SNPs is more than counterbalanced by the tenfold larger GWAS meta-analysis that has conducted imputation for fine mapping of common alleles at these regions. To explore additional signal beyond the known index

Table 4 Estimates of τ for each ancestry group under our simulation-based pipeline with MAF partitioning at 1%

Ancestry	Sample size	Mean τ	95% confidence interval	$h^2_{g,\text{rare}}$
African	4,006	0.48	0.19, 0.78	0.12 (0.05)
European	1,753	0.28	−0.08, 0.90	0.00 (0.06)
Japanese	1,770	0.38	−0.07, 0.92	0.05 (0.07)
Latino	1,708	0.39	−0.08, 1.05	0.00 (0.06)
Meta-analysis	9,237	0.42	0.22, 0.62	0.05 (0.03)

Meta-analysis results were computed using an inverse-weighted variance approach. Similar results were obtained with MAF partitioning at 5% (**Supplementary Fig. 13**).

variants, we performed a conditioning analysis (Online Methods) on the index variants and observed no variants with P value $<1 \times 10^{-8}$ after conditioning; quantile-quantile plots showed residual signal only in the African-ancestry sample, consistent with the hypothesis that there is an additional signal beyond that contributed by the known variants at these loci (either due to better tagging of a single causal variant or the presence of multiple causal alleles[37]) (**Supplementary Figs. 16–18**).

To investigate sequenced SNPs as plausible causal alleles, we integrated epigenetic and genetic data using PAINTOR[40] to estimate posterior probabilities for causality at each SNP. We used the meta-analysis results for SNPs with MAF ≥ 1.0% (as the Wald statistic is unreliable at MAF < 1% and therefore not well suited to estimation within the PAINTOR framework). First, we ran PAINTOR independently for each of the 20 functional categories that have previously been implicated in prostate cancer[41] and found a significant enrichment for causal variants in FOXA1-binding sites assayed in the LNCaP cell lines as well as at binding sites for androgen receptor[41] (**Supplementary Fig. 19**). Second, we selected the functional categories with significant enrichment (at a nominal level of $P \leq 0.05$) for a joint PAINTOR model to estimate posterior probabilities that each SNP is causal. Of the 24,840 common variants found in all ancestry groups, we identified nine variants with PAINTOR posterior probability >0.90 as causal. In particular, two variants (rs78416326 and rs10486567) exhibited posterior probabilities >0.99 owing to a combination of strong association signal and overlap with functional elements (**Supplementary Fig. 20** and **Supplementary Table 43**). Although biological causality cannot be proven on the basis of statistical association alone, we highlight the variants with high posterior probability for follow-up validation.

DISCUSSION

We have used large-scale targeted sequencing to study the contribution of rare variants to the heritability of prostate cancer for individuals of diverse ancestry. We find that the total variance in the trait contributed by these regions is significantly greater than the variance localized to known GWAS variants, thus showing that large-scale sequencing can uncover missing heritability. We also provide evidence of heterogeneity by ancestry as well as the first direct evidence of which we are aware of rare variants contributing a disproportionate fraction of the genetic heritability for a common disease. On first principles, there are reasons to think that our results actually underestimate the fraction of heritability due to rare variants. First, our study does not have a sample size sufficient to interrogate extremely rare variants (frequency << 0.1%). Second, we focused on known GWAS-identified regions that were ascertained on the basis of harboring an association with a common variant, thus guaranteeing that common variants would be responsible for a substantial fraction of prostate cancer risk at these locations.

Our finding that 42% (95% confidence interval = 21–63%) of the genetic risk for prostate cancer is due to variants in the MAF range of 0.1–1% is striking, given that only a couple percent of neutral variation is due to SNPs in this frequency range. These results suggest that selection has placed downward pressure on the frequencies of many alleles contributing risk for prostate cancer, and we have quantified this coupling of selection and prostate cancer risk. Prostate cancer is a late-onset disease that primarily affects people after reproductive age. For diseases with younger onset, it is plausible that the coupling of selection to disease risk could be even higher, and we predict that this will be observed for other diseases when sequencing studies of large sample size are performed and analyzed using methods like the ones we report here that are capable of partitioning heritability by frequency[42]. Already, associations with rare variants have been found at both the gene and individual-SNP levels[43–45] as well as through sequencing of known GWAS risk loci[46]. Because we have shown that rare variation is capable of explaining a substantial portion of SNP heritability for prostate cancer, we expect that it will be useful to incorporate rare variants into statistical models for prediction of disease risk. Taken together, these results motivate further large sequencing efforts in diverse populations to fully explore the abundance of rare variants that might contribute a substantial fraction of the heritability for at least some important human phenotypes.

We conclude with several caveats. Although we genotyped the majority of variants at the risk-associated regions in the regions we targeted in sequencing, we were not able to sequence a subset of the regions owing to the fact that the technology we used could not enrich for sequences at repetitive regions. Second, the part of the genome we analyzed in this study is non-random: we analyzed loci discovered by common variant association methods, where the fraction of genetic heritability due to common variants is likely to be overestimated owing to the fact that the regions were discovered on account of containing common variants. Thus, it is plausible that the true fraction of heritability for prostate cancer that is due to rare variants is a conservative underestimate of the true proportion across the genome. Third, assaying SNP heritability using variance components makes a number of simplifying assumptions; although we could not identify any source of bias that could explain our results artifactually, it is important to recognize that the analyses we have performed are statistically complex and there might be biases we have not appreciated. An important direction for future work will be to carry out whole-genome sequencing studies in much larger sample sizes, which will provide sufficient statistical power to allow a direct SNP-by-SNP understanding of the contribution of variants in the MAF range of 0.1–1% that this study suggests make a major contribution to human genetic risk for prostate cancer.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The data reported in this study are available at the database of Genotypes and Phenotypes (dbGaP) under accession phs000306.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work is supported in part by the US National Institutes of Health (R01 CA165862, U19 CA148537, UM1 CA164973, RC2 CA148085, U01 CA1326792, R21 CA182821 and U01 CA188392). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. Many of the risk regions examined were discovered through contributions from: P. Hall (COGS), D.F.E., P. Pharoah, K. Michailidou, M.K. Bolla and Q. Wang (BCAC), A. Berchuck (OCAC), R.A.E., D.F.E., A.A. Al Olama, Z.K.-J. and S. Benlloch (PRACTICAL), G. Chenevix-Trench, A. Antoniou, L. McGuffog, F. Couch and K. Offit (CIMBA), J. Dennis, A.M. Dunning, A. Lee, E. Dicks, C. Luccarini and the staff of the Centre for Genetic Epidemiology Laboratory, J. Benitez, A. Gonzalez-Neira and the staff of the CNIO genotyping unit, J. Simard, D.V.C. Tessier, F. Bacot, D. Vincent, S. LaBoissière, F. Robidoux and the staff of the McGill University and Génome Québec Innovation Centre, S.E. Bojesen, S.F. Nielsen, B.G. Nordestgaard and the staff of the Copenhagen DNA laboratory, and J.M. Cunningham, S.A. Windebank, C.A. Hilker, J. Meyer and the staff of the Mayo Clinic Genotyping Core Facility. Funding for the iCOGS infrastructure came from the European Community's Seventh Framework Programme under grant agreement 223175 (HEALTH-F2-2009-223175) (COGS), Cancer Research UK (C1287/A10118,

C1287/A 10710, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007, C5047/A10692 and C8197/A16565), the US National Institutes of Health (CA128978) and Post-Cancer GWAS initiative (1U19 CA148537, 1U19 CA148065 and 1U19 CA148112–GAME-ON initiative), the US Department of Defense (W81XWH-10-1-0341), the Canadian Institutes of Health Research (CIHR) for the CIHR Team in Familial Risks of Breast Cancer, the Komen Foundation for the Cure, the Breast Cancer Research Foundation and the Ovarian Cancer Research Fund. D.R. is an Investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

N.R., C.A.H. and D.R. defined the regions of interest. N.R., S.M. and D.R. designed the in-solution capture reagent. N.R., A.A. and D.Q. prepared libraries. N.R. performed capture and quality control sequencing. N.R., A.T. and S.M. performed sequence analyses. N.M. performed statistical analyses and simulations. K.A.R., A.T., H.L., A.S., X.S., Z.K.-J., D.F.E., R.A.E., the PRACTICAL consortium, L.L.M., A.L., D.S., S.W., D.V.C. and B.H. generated data and analysis tools. C.A.H., B.P. and D.R. supervised the work. All authors reviewed, revised and wrote feedback for the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Siegel, R., Ma, J., Zou, Z. & Jemal, A. Cancer statistics, 2014. *CA Cancer J. Clin.* **64**, 9–29 (2014).
2. Lin, K., Croswell, J.M., Koenig, H., Lam, C. & Maltz, A. in *Prostate-Specific Antigen–Based Screening for Prostate Cancer: An Evidence Update for the U.S. Preventive Services Task Force* (Agency for Healthcare Research and Quality, 2011).
3. Melnikow, J., LeFevre, M., Wilt, T.J. & Moyer, V.A. Counterpoint: randomized trials provide the strongest evidence for clinical guidelines: The US Preventive Services Task Force and Prostate Cancer Screening. *Med. Care* **51**, 301–303 (2013).
4. Gomella, L.G. *et al.* Screening for prostate cancer: the current evidence and guidelines controversy. *Can. J. Urol.* **18**, 5875–5883 (2011).
5. Lichtenstein, P. *et al.* Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **343**, 78–85 (2000).
6. 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).
7. Al Olama, A.A. *et al.* Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat. Genet.* **41**, 1058–1060 (2009).
8. Eeles, R.A. *et al.* Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat. Genet.* **41**, 1116–1121 (2009).
9. Eeles, R.A. *et al.* Multiple newly identified loci associated with prostate cancer susceptibility. *Nat. Genet.* **40**, 316–321 (2008).
10. 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).
11. Kote-Jarai, Z. *et al.* Seven prostate cancer susceptibility loci identified by a multi-stage genome-wide association study. *Nat. Genet.* **43**, 785–791 (2011).
12. Schumacher, F.R. *et al.* Genome-wide association study identifies new prostate cancer susceptibility loci. *Hum. Mol. Genet.* **20**, 3867–3875 (2011).
13. Amundadottir, L.T. *et al.* A common variant associated with prostate cancer in European and African populations. *Nat. Genet.* **38**, 652–658 (2006).
14. Gudmundsson, J. *et al.* Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat. Genet.* **39**, 631–637 (2007).
15. Gudmundsson, J. *et al.* Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. *Nat. Genet.* **41**, 1122–1126 (2009).
16. Gudmundsson, J. *et al.* Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nat. Genet.* **40**, 281–283 (2008).
17. Gudmundsson, J. *et al.* Two variants on chromosome 17 confer prostate cancer risk, and the one in *TCF2* protects against type 2 diabetes. *Nat. Genet.* **39**, 977–983 (2007).
18. Sun, J. *et al.* Evidence for two independent prostate cancer risk-associated loci in the *HNF1B* gene at 17q12. *Nat. Genet.* **40**, 1153–1155 (2008).
19. Thomas, G. *et al.* Multiple loci identified in a genome-wide association study of prostate cancer. *Nat. Genet.* **40**, 310–315 (2008).
20. Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39**, 645–649 (2007).
21. Duggan, D. *et al.* Two genome-wide association studies of aggressive prostate cancer implicate putative prostate tumor suppressor gene *DAB2IP*. *J. Natl. Cancer Inst.* **99**, 1836–1844 (2007).
22. Haiman, C.A. *et al.* Genome-wide association study of prostate cancer in men of African ancestry identifies a susceptibility locus at 17q21. *Nat. Genet.* **43**, 570–573 (2011).
23. Takata, R. *et al.* Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population. *Nat. Genet.* **42**, 751–754 (2010).
24. 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).
25. Witte, J.S., Visscher, P.M. & Wray, N.R. The contribution of genetic variants to disease depends on the ruler. *Nat. Rev. Genet.* **15**, 765–776 (2014).
26. Gudmundsson, J. *et al.* A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nat. Genet.* **44**, 1326–1329 (2012).
27. Cropp, C.D. *et al.* 8q24 risk alleles and prostate cancer in African-Barbadian men. *Prostate* **74**, 1579–1588 (2014).
28. 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).
29. Haiman, C.A. *et al.* Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.* **39**, 638–644 (2007).
30. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
31. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
32. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).
33. Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22**, 939–946 (2012).
34. Li, Y., Sidore, C., Kang, H.M., Boehnke, M. & Abecasis, G.R. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.* **21**, 940–951 (2011).
35. Lee, S.H. *et al.* Estimation of SNP heritability from dense genotype data. *Am. J. Hum. Genet.* **93**, 1151–1155 (2013).
36. 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).
37. Gusev, A. *et al.* Quantifying missing heritability at known GWAS loci. *PLoS Genet.* **9**, e1003993 (2013).
38. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M. & Wray, N.R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism–derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540–2542 (2012).
39. Eyre-Walker, A. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proc. Natl. Acad. Sci. USA* **107**, 1752–1756 (2010).
40. Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722 (2014).
41. Hazelett, D.J. *et al.* Comprehensive functional annotation of 77 prostate cancer risk loci. *PLoS Genet.* **10**, e1004102 (2014).
42. Bhatia, G. *et al.* Haplotypes of common SNPs can explain missing heritability of complex diseases. *bioRxiv* doi:10.1101/022418 (12 July 2015).
43. Huffman, J.E. *et al.* Rare and low-frequency variants and their association with plasma levels of fibrinogen, FVII, FVIII, and vWF. *Blood* **126**, e19–e29 (2015).
44. Peloso, G.M. *et al.* Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am. J. Hum. Genet.* **94**, 223–232 (2014).
45. Lange, L.A. *et al.* Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *Am. J. Hum. Genet.* **94**, 233–245 (2014).
46. Service, S.K. *et al.* Re-sequencing expands our understanding of the phenotypic impact of variants at GWAS loci. *PLoS Genet.* **10**, e1004147 (2014).

ONLINE METHODS

Data sets. *The Multiethnic Cohort.* The Multiethnic Cohort (MEC) consists of over 215,000 men and women enrolled from Hawaii and the Los Angeles region between 1993 and 1996 (ref. 47). Participants are primarily of Native Hawaiian, Japanese, European-American, African-American or Latino ancestry and were between the ages of 45 and 75 years at baseline when they completed a detailed questionnaire to collect information on demographics and lifestyle factors, including diet and medical conditions. Over 65,000 blood samples were collected from study participants for genetic analysis. To obtain information on cancer status, stage and severity of disease, MEC participants were referenced against population-based Surveillance, Epidemiology and End Results (SEER) registries in California and Hawaii. Unaffected cohort participants with blood samples were selected as controls (for case-control sample sizes, see **Table 1**; for stage and grade of cases, see **Supplementary Table 44**).

Uganda Prostate Cancer Study. The Uganda Prostate Cancer Study (UGPCS) is a case-control study of prostate cancer in Kampala, Uganda, that was initiated in 2011. Patients diagnosed with prostate cancer were enrolled from the Urology unit at Mulago Hospital, whereas undiagnosed men (controls) were enrolled from other clinics (for example, surgery) within the hospital. All consenting patients who satisfied strict inclusion criteria (cases, >39 years of age; controls, >39 years of age, PSA level <4 ng/ml to dismiss possible undiagnosed prostate cancer) were recruited into the study. Written consent was obtained, and two identical informed consent forms translated into Luganda were provided to each participant for them to read or to be read to them, sign or thumb print. Descriptive and prostate cancer risk factor information was collected from interviews conducted with patients using a standardized questionnaire. Biospecimens were collected using Oragene saliva collection kits. The Institutional Review Boards at the University of Southern California and at Makerere University approved the study protocol.

Library preparation and target enrichment. We prepared next-generation sequencing libraries from all DNA samples following a cost-effective library preparation protocol developed for this study, which makes it possible to perform multiplexed hybridization enrichment[33]. DNA samples from cases and controls were randomly distributed over 96-well plates to avoid plate effects confounding the results. Each sample was molecularly barcoded during the library preparation stage in 96-well plates to allow us to pool many samples for hybrid capture enrichment and subsequent sequencing. We typically pooled 24 samples in equimolar ratio per capture reaction using the custom SureSelect capture reagent described above. In short, we defined the target region to consist of LD blocks surrounding all prostate cancer risk variants known at the time of design (October 2011), all coding sequences surrounding the variants within a 200-kb window on either side and evolutionarily conserved elements defined by a 29-mammal alignment[31]. This resulted in a total target size of 16.7 Mb, of which probes could be designed for 12 Mb. The missing 4.7 Mb constituted non-unique regions of the genome that were filtered out according to Agilent design recommendations. An overview table of targeted genes, the variants, the size of the targeted region for each variant and the size of the baited region is given in **Supplementary Table 1**. Sequencing was performed at Illumina using HiSeq 2000 instruments for 100 cycles of paired-end sequencing. Using this approach, we covered 78% of the targeted regions (**Supplementary Table 2**), of which 26 regions (41%) had mean coverage ≥10×.

Alignment and genotype calling. Sequences were aligned to the human genome reference sequence (hg19) using Burrows-Wheeler Aligner (BWA) version 0.6.1 (ref. 48). Variants were called using the Genome Analysis Toolkit (GATK) best-practices workflow[49], including mapping the raw reads to the reference genome, base recalibration and compression, and joint calling and variant recalibration. After quality control, 11.3 Mb of autosomal sequence was considered; because of complexities in the analysis, we disregarded data on the X chromosome. Starting from the GATK likelihoods, we applied LD-aware genotype calling using Beagle[50,51] version 3.3.2 with 1000 Genomes Project v3 data as the reference. Variants that displayed low-quality calling ($r^2 < 0.6$) or MAF <0.1% were dropped from the analysis (n = 588,410), resulting in 197,786 SNPs across all ancestry groups. To take advantage of the lower error rate of the GWAS arrays, before LD-aware calling, overlapping sequenced SNPs were replaced with their array counterparts. This resulted in 6,028 replaced calls for 2,042 individuals in the African group, 5,395 replaced calls for the European group, 2,642 replaced calls for the Japanese group and 2,805 replaced calls for Latinos. To compute accuracy of the LD-aware calling, we used 1,172 African samples for which we had GWAS array data that was not used to replace calls before the LD-aware calling. The first ten principal components for each ancestry group were computed using GCTA[36] from the sequenced common variants (MAF ≥ 1.0%) after LD pruning ($r^2 < 0.2$) (ref. 52).

Genotype array design. To capture SNP heritability tagged outside of the targeted regions, we assayed individuals using the Illumina 1M-Duo BeadChip for the African-ancestry group, the Illumina 660W-Quad BeadChip for the Latino and Japanese groups, and the Illumina Human610 BeadChip for Europeans. The number of samples genotyped by array was n = 3,078 for the African group, 1,627 for the European group, 1,674 for the Japanese group and 1,642 for the Latinos. For quality control, we removed any SNP with missingness >0.10. To remove any confounding from tagged variants within the targeted sequenced regions, we removed any SNP within 0.5 Mb of any region and any SNP with LD >0.2 with respect to index variants. We further pruned the set to remove any variants with pairwise LD >0.3. This resulted in n = 251,919, 182,983, 96,711 and 109,118 array-based SNPs for the Africans, Europeans, Japanese and Latinos, respectively (**Supplementary Fig. 21**).

Association analyses. Each variant was subjected to an unconditional marginal case-control association test adjusting for age, Ugandan ancestry for the African group and the top ten principal components under a log-additive model performed by PLINK 1.9 (ref. 53). All reported P values are asymptotic estimates from the Wald statistic. We extended the unconditional association test by incorporating the known associated variants (index SNPs) as covariates for each SNP at a given locus. Conditional association tests were implemented in Python 2.7 with the package statsmodels version 0.5. A meta-analysis combining individual population results was performed using METAL[54] version 2011-03-25. Of the 197,786 SNPs analyzed, 183 were removed from the meta-analysis because they had multiallelic values when compared across all populations. To perform SKAT-O tests for the African-ancestry group, we used a non-overlapping sliding window approach to group rare SNPs into bins containing at most 100 variants across each targeted region, resulting in a total of 601 bins. Tests were performed using the software PLINKSEQ version 0.10. To predict the total risk from sequenced variants, we performed BLUP prediction in GCTA version 1.24 over a single variance component. Predicted effects were partitioned into rare and common variants and risk scores computed using the predicted allelic effects with PLINK. Training and prediction was performed using tenfold cross-validation over samples for each ancestry group (**Supplementary Table 45** and **Supplementary Note**).

Heritability analyses. We estimated the GRM as $A = \frac{1}{m} ZZ^t$, where Z is the standardized genotype matrix and m is the number of SNPs. For each sample, two GRMs corresponding to rare (0.1% ≤ MAF < 1%) and common (MAF ≥ 1%) SNPs were created using GCTA version 1.24. GCTA assumes a linear mixed model where the contribution from each SNP is the result of a random effect given by $y = X\beta + \sum_i g_i + \varepsilon$ where y is a vector of phenotypes, X is a covariate matrix (for example, age), β is a vector of fixed effects and g_i is a vector of random genetic effects for the ith component (we partition into g_{rare} and g_{common}). The variance of y is given by $\text{var}(y) = A_{\text{rare}}\sigma^2_{\text{rare}} + A_{\text{common}}\sigma^2_{\text{common}} + I\sigma^2_\varepsilon$, where A_{rare} and A_{common} correspond to the GRMs for rare and common SNPs, respectively. Creation of the GRMs was done directly from the dosage data (similar results were obtained using best-guess calls; **Supplementary Tables 33** and **34**). We estimate the SNP heritability contributed from rare variants as

$$h^2_{g,\text{rare}} = \frac{\sigma^2_{\text{rare}}}{\sigma^2_{\text{rare}} + \sigma^2_{\text{common}} + \sigma^2_\varepsilon}.$$

Given estimates for $\sigma^2_{g,\text{rare}}$, $\sigma^2_{g,\text{common}}$, σ^2_ε and their covariance matrix S, we use the delta method[55] to approximate the standard error for

$$\frac{\sigma^2_{\text{rare}}}{\sigma^2_{\text{rare}} + \sigma^2_{\text{common}}} = \frac{h^2_{g,\text{rare}}}{h^2_{g,\text{rare}} + h^2_{g,\text{common}}}$$

that is, the proportion of total SNP heritability explained by rare SNPs. The SNP heritability analysis was performed on the dichotomous case-control phenotype using constrained REML in GCTA with a prevalence of 0.19 for the African-ancestry group, 0.14 for the European- and Latino groups, and 0.10 for Japanese (SEER; see URLs). Hence, all reported values of h_g^2 are on the underlying liability scale. To estimate the contribution of the known index variants to SNP heritability, we computed a GRM restricted to the 84 known variants. The covariate matrix for each ancestry group consisted of age and the first ten principal components (with an additional binary variable indicating Ugandan ancestry for the African-ancestry group). LD-adjusted GRMs were computed using LDAK[56] version 4.2. *P* values were estimated from a likelihood-ratio test by dropping one component and comparing against the reduced model (as implemented in GCTA). To estimate GRMs from array data, we removed any SNP within 0.5 Mb of the targeted regions and further pruned for pairwise LD >0.2 in addition to any remaining variants in LD with index SNPs (**Supplementary Fig. 22**). For bivariate REML analysis, we define the GRM for samples over two ancestry groups as

$$A = \frac{1}{m}\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \cdot \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}^t$$

where Z_i is the standardized genotype matrix for ancestry group i and m is the number of SNPs shared by both groups[57].

Coupling selection with allelic effect size. We investigated the relationship between selection and marginal effect sizes on prostate cancer risk using the Eyre-Walker model[39], which sets allelic effect sizes $\beta = \left(4N_e|s|\right)^{\tau}(1+\varepsilon)$. Here N_e is the effective population size (set to 10,000 for our analyses[58]), s is the selection coefficient of the allele and ε is normally distributed noise ($\sigma_\varepsilon = 0.5$; varying this parameter does not significantly affect underlying rare/common variation[39]). As τ increases, we expect the allelic effects and, thus, the contribution to h_g^2 from rare variants, to increase as a result of rare SNPs experiencing stronger selective pressure than common SNPs (**Supplementary Figs. 22–24**). To determine how τ has a possible role in the underlying architecture for prostate cancer, we followed a five-step simulation procedure: (i) we randomly select a set of 10,000 SNPs to be causal; (ii) we assign selection coefficients to each causal variant by mapping their allele frequency to selection coefficients[59]; (iii) we simulate allelic effects under the Eyre-Walker model given selection coefficients, τ and $\sigma_\varepsilon = 0.5$; (iv) we simulate a continuous trait starting from the real genotype data with total SNP heritability matching the SNP heritability estimated from real data; and (v) we perform joint REML analysis in GCTA to estimate rare and common SNP heritability for the simulated trait. We repeated this procedure for 5,000 values of τ uniformly distributed over the interval [−0.1, 1.1]. To match the observed results in real data to the results from the simulations, we sampled 10,000 values from $N\left(h^2_{g,\text{rare}}, \widehat{\text{s.e.}}\right)$ and identified the closest estimate observed from the simulation pipeline and recorded its simulated value for τ. This enabled us to convert the statistical noise around the estimate of the proportion as obtained by GCTA into a variance around τ for each of the ancestry samples.

47. Kolonel, L.N. *et al.* A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am. J. Epidemiol.* **151**, 346–357 (2000).
48. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
49. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
50. Browning, B.L. & Browning, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009).
51. Browning, B.L. & Browning, S.R. A fast, powerful method for detecting identity by descent. *Am. J. Hum. Genet.* **88**, 173–182 (2011).
52. Baran, Y., Quintela, I., Carracedo, Á., Pasaniuc, B. & Halperin, E. Enhanced localization of genetic samples through linkage-disequilibrium correction. *Am. J. Hum. Genet.* **92**, 882–894 (2013).
53. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
54. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
55. Oehlert, G.W. A note on the delta method. *Am. Stat.* **46**, 27–29 (1992).
56. Speed, D., Hemani, G., Johnson, M.R. & Balding, D.J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012).
57. Yang, L. *et al.* Polygenic transmission and complex neuro developmental network for attention deficit hyperactivity disorder: genome-wide association study of both common and rare variants. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **162B**, 419–430 (2013).
58. Takahata, N. Allelic genealogy and human evolution. *Mol. Biol. Evol.* **10**, 2–22 (1993).
59. Lohmueller, K.E. The impact of population demography and selection on the genetic architecture of complex traits. *PLoS Genet.* **10**, e1004379 (2014).

doi:10.1038/ng.3446