Introduction

For most common diseases and traits the genetic basis underlying susceptibility has yet to be completely revealed. While genome-wide association studies (GWAS) have been remarkably successful in identifying common genetic variants associated with risk, the effect sizes of the risk alleles have been modest (relative risk, RR of 1.1–1.4) and in most cases, even in sum, they can explain only a fraction of familial risk or disease heritability. GWAS have relied almost exclusively on Illumina and Affymetrix SNP arrays, with SNP content selected primarily from HapMap to capture a large fraction of common variation in coding and non-coding regions in populations of European ancestry. The vast majority of alleles with frequencies <5%, and in particularly those with frequencies ≤1%, have not been tested. This low allele frequency spectrum of genetic variation represents a very large fraction of all variation in the human genome. Thus, to date, a large fraction of genetic variation has yet to be explored with respect to disease etiology.

It is possible that the majority of less common (1–5%) and rare variants (<1%) will have weak effects, like the GWAS-identified common variants, and if this is the case then very large studies will be required for their discovery. An alternative hypothesis is that less common and rare variants convey larger relative risks than common variants, and indeed this assumption is required in order that rare variants contribute meaningfully to the understanding of inherited susceptibility. Such enhancement of effect sizes for rarer variants has been hypothesized to contribute to disease heritability. Using the Illumina HumanExome SNP array, we successfully genotyped 191,032 common and rare non-synonymous, splice site, or nonsense variants in a multiethnic sample of 2,984 breast cancer cases, 4,376 prostate cancer cases, and 7,545 controls. In breast cancer, the strongest associations included either SNPs in or gene burden scores for genes LDLRAD1, SLC19A1, FGFBP3, CASP5, MMAB, SLC16A6, and INS-IGF2. In prostate cancer, one of the most associated SNPs was in the gene GPRC6A (rs2274911, Pro91Ser, OR = 0.88, P = 1.3 x 10^-5) near to a known risk locus for prostate cancer; other suggestive associations were noted in genes such as F13A1, ANXA4, MANSC1, and GP6. For both breast and prostate cancer, several of the most significant associations involving SNPs or gene burden scores (sum of minor alleles) were noted in genes previously reported to be associated with a cancer-related phenotype. However, only one of the associations (rs145889899 in LDLRAD1, p = 2.5 x 10^-7 only seen in African Americans) for overall breast or prostate cancer risk was statistically significant after correcting for multiple comparisons. In addition to breast and prostate cancer, other cancer-related traits were examined (body mass index, PSA level, and alcohol drinking) with a number of known and potentially novel associations described. In general, these findings do not support there being many protein coding variants of moderate to high risk for breast and prostate cancer with odds ratios over a range that is probably required for protein coding variation to play a truly outstanding role in risk heritability. Very large sample sizes will be required to better define the role of rare and less penetrant coding variation in prostate and breast cancer disease genetics.

Abstract

Rare variation in protein coding sequence is poorly captured by GWAS arrays and has been hypothesized to contribute to disease heritability. Using the Illumina HumanExome SNP array, we successfully genotyped 191,032 common and rare non-synonymous, splice site, or nonsense variants in a multiethnic sample of 2,984 breast cancer cases, 4,376 prostate cancer cases, and 7,545 controls. In breast cancer, the strongest associations included either SNPs in or gene burden scores for genes LDLRAD1, SLC19A1, FGFBP3, CASP5, MMAB, SLC16A6, and INS-IGF2. In prostate cancer, one of the most associated SNPs was in the gene GPRC6A (rs2274911, Pro91Ser, OR = 0.88, P = 1.3 x 10^-5) near to a known risk locus for prostate cancer; other suggestive associations were noted in genes such as F13A1, ANXA4, MANSC1, and GP6. For both breast and prostate cancer, several of the most significant associations involving SNPs or gene burden scores (sum of minor alleles) were noted in genes previously reported to be associated with a cancer-related phenotype. However, only one of the associations (rs145889899 in LDLRAD1, p = 2.5 x 10^-7 only seen in African Americans) for overall breast or prostate cancer risk was statistically significant after correcting for multiple comparisons. In addition to breast and prostate cancer, other cancer-related traits were examined (body mass index, PSA level, and alcohol drinking) with a number of known and potentially novel associations described. In general, these findings do not support there being many protein coding variants of moderate to high risk for breast and prostate cancer with odds ratios over a range that is probably required for protein coding variation to play a truly outstanding role in risk heritability. Very large sample sizes will be required to better define the role of rare and less penetrant coding variation in prostate and breast cancer disease genetics.

Citation: Haiman CA, Han Y, Feng Y, Xia L, Hsu C, et al. (2013) Genome-Wide Testing of Putative Functional Exonic Variants in Relationship with Breast and Prostate Cancer Risk in a Multiethnic Population. PLoS Genet 9(3): e1003419. doi:10.1371/journal.pgen.1003419

Editor: Marshall S. Horwitz, University of Washington, United States of America

Received August 10, 2012; Accepted February 12, 2013; Published March 28, 2013

Copyright: © 2013 Haiman et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The Multiethnic Cohort Study and genotyping and analysis for this project were supported by National Institutes of Health (http://www.nih.gov/) grants R01-CA63464, R37-CA54281, R01-CA132839-01, R01-CA140865, U01-HG00726-01, U01-CA132679, and U19-CA148537 and by the California Breast Cancer Research Program (http://www.cbcrp.org/) grant 15UB-8402. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: haiman@usc.edu (CAH); stram@usc.edu (DOS)
* These authors contributed equally to this work.
Author Summary

For breast and prostate cancer, GWAS have revealed many risk variants (>70 for each cancer as of this report). All together the common variants in these regions explain only a minority of familial risk of these cancers. Using the Illumina HumanExome SNP array, we explored the hypothesis of rare coding variation contributing to breast and prostate cancer risk in a sample of African American, Latino, Japanese, Native Hawaiian, and European American breast and prostate cancer cases and controls from the Multiethnic Cohort study. While only one association exceeded significance thresholds after correcting for multiple comparisons, a number of suggestive associations involving genes previously reported to be associated with a cancer-related phenotype were noted. Our results do not generally support a major role of protein-coding variants with odds ratios over a range that is probably required for protein coding variation to play a truly outstanding role in risk heritability. If very rare and/or less penetrant coding variants underlie disease heritability of these cancers, then very large sample sizes (i.e. consortia) will be required for their discovery.

including rare missense variants in CHEK2, ATM, NBS1, RAD50, BRIP1, and PALB2 in breast cancer [2], rare coding mutations in RAD51D and BRIP1 in ovarian cancer [3,4], as well as rare coding variants in genes implicated in hyperglycieridemia [5] and colorectal cancer adenomas [6]. More recently, whole-genome and candidate gene sequencing studies have revealed rare coding variants in ALDH1A1 for gout [7] and a number of genes (NOD2, IL23R, CARD9, IL18RAP, CUL2, Clorf106, PTNP22 and MEC19) involved in inflammatory bowel disease [8]. Studies in prostate cancer have reported rare gene coding mutations in BRCA2 (found in 2% of cases <55 years) to be associated with greater risk of prostate cancer (RR>4.5) and more aggressive disease [9,10]. For many of these examples, in addition to single SNP association testing, burden of rare variation analyses have been applied to increase the number of observations in the comparison groups (thus the statistical power), and to provide statistical support for the involvement of the gene which is not achieved when examining large number of SNPs in any given gene.

To date, a lack of technology to survey the genome and accurately enumerate and test the variants in large numbers of samples has limited the exploration of less common and rare alleles. In the past year the Illumina Infinium HumanExome array (or “exome chip”) has been developed in collaboration with investigators who combined whole-exome sequencing conducted in >12,000 individuals of primarily European ancestry as well as in small numbers of other racial/ethnic minorities including African Americans, Hispanics, and Asians; the content on the array includes >200,000 putative functional exonic variants and is predicted to alter the protein coding sequence, and which passed quality control procedures (see Methods). Of the 15,837 samples, 14,905 were included in the analysis (3,315 European Americans, 3,854 African Americans, 3,106 Latinos, 3,843 Japanese Americans and 787 Native Hawaiians; see Methods for exclusion criteria). A few mitochondiral SNPs were included on the array (n = 165 SNPs passing quality control) but are not discussed here (no associations with them were seen in the top ranked 1,000 associations for either breast or prostate cancer).

For breast and prostate cancer cases and controls are shown in Table 1. In this multiethnic sample, 191,032 (88%) putative functional variants were found to be polymorphic in at least one population, with 26,569 (12%) being monomorphic in all five populations (Figure 1). The percentage of monomorphic SNPs ranged from 34.1% in African Americans, 39.6% in European Americans and 43.3% in Latinos to 66.8% in Native Hawaiians and 74.2% in Japanese Americans (Figure S1). Of the polymorphic SNPs, 178,776 (93.4%) were nonsynonymous (NS) variants, 8,308 (4.4%) splice site (SP) variants, and 3,948 (2.1%) nonsense variants which either lead to a gain or loss of a stop codon. Of the polymorphic SNPs, 34,834 (18.2%) were polymorphic in all four of the largest populations (excluding Native Hawaiians), with 81,713 SNPs (42.7%) being polymorphic in African Americans, Latinos and European Americans (Figure 2). African Americans had the largest number of unique polymorphic SNPs (21,908, 11.4%), followed by European Americans (16,653, 8.7%), Japanese Americans (6,776, 3.5%) and Latinos (5,134, 2.7%).

In the pooled sample, 190,662 putative functional (NS, SP, or stop) SNPs had a minor allele frequency (MAF) <1% (56,759<0.01%; 85,897 between 0.01% and 0.1%, and 48,006 between 0.1% and 1%) (Figure 1, Figure S1). The minor allele frequency distributions were similar across three of the five populations with African Americans, European Americans and Latinos having roughly the same number of SNPs with frequencies greater than 0 and less than 1% (100–110 thousand); However there were only 37,979 SNPs with a frequency above zero and less than 1% in Japanese Americans and 52,985 in Native Hawaiians. The number of SNPs with a frequency >1% ranged from approximately 18–35 thousand between sampled populations.

Exonic SNPs and Breast and Prostate Cancer Risk

In the pooled sample, 190,662 putative functional (NS, SP, or stop) SNPs had a minor allele frequency (MAF) <1% (56,759<0.01%; 85,897 between 0.01% and 0.1%, and 48,006 between 0.1% and 1%) (Figure 1, Figure S1). The minor allele frequency distributions were similar across three of the five populations with African Americans, European Americans and Latinos having roughly the same number of SNPs with frequencies greater than 0 and less than 1% (100–110 thousand); However there were only 37,979 SNPs with a frequency above zero and less than 1% in Japanese Americans and 52,985 in Native Hawaiians. The number of SNPs with a frequency >1% ranged from approximately 18–35 thousand between sampled populations.

Inspexion of the distribution of the chi-square (score) tests from models for overall breast or prostate cancer showed evidence of over-dispersion of test statistics (genomic control lambda estimate to be approximately 1.15 for breast and 1.20 for prostate) however when very rare SNPs were removed (MAF<0.1% overall) then the Wald statistics appeared to be sampled from an overall central chi-square distribution (genomic control lambda = 1.00 for breast cancer and lambda = 1.05 for prostate cancer). In the gene burden analyses, the distribution of observed score tests showed mild evidence of over-dispersion (lambda = 1.04 for breast cancer and lambda = 1.06 for prostate cancer). When the single SNP analysis was restricted to estrogen receptor-negative (ER-) breast or advanced prostate cancer, where there were many more controls than cases included in each model, then the behavior of the score test for the single SNP associations was problematic for rare SNPs. For such SNPs we followed up any apparently globally significant associations with exact logistic regression analysis, in order to reduce what appeared to be a proliferation of false positive signals.
The total number of genes having at least one polymorphic functional variant genotyped and passed quality control varied slightly between breast (17,168 genes) and prostate cancer (17,203 genes) due to sampling (i.e. some variants were polymorphic only for breast cancer cases and so were not included in the prostate cancer analyses and vice versa).

![Figure 1. Minor allele frequency for all variants successfully genotyped using the Illumina Human Exome array.](doi:10.1371/journal.pgen.1003419.g001)
Breast Cancer Single SNP Associations

In the ethnic-pooled breast cancer analyses (2,984 cases and 7,545 controls), the most significant predicted protein-altering variant was a rare SP variant rs145889899 at the splice donor site in the second intron of the gene \textit{LDLRAD1} (OR = 3.74, \( p = 2.5 \times 10^{-27} \)), which was almost exclusively seen in African Americans, this variant was statistically significant at our exome-wide level (nominal \( p < 3.9 \times 10^{-7} \), see Methods). Of the top 10 ranked associations, the remaining 9 involved NS variants (\( p \)-values \( \geq 1.3 \times 10^{-6} \), Table 2, Table S1). None of the other associations met the Bonferroni adjustment for multiple comparison testing. All of the 10 most associated variants, were quite rare and present mainly or exclusively in one or two ethnic groups. The genes containing the most significant SNPs for breast cancer ranged widely in apparent function (see Table 2) with GWAS associations reported with SNPs in \textit{CFB} (complement factor B) for age-related macular degeneration [11], \textit{HLA-C2A} for platelet counts [12] and \textit{ACADS} for metabolic traits [13].Table S1 gives information for the 100 most significant associations for breast cancer, both overall and by ethnic group when including all SNPs passing quality control (not just the non-synonymous, splice site and nonsense variants described here).

For ER- breast cancer (n = 441 cases) many associations (358) with very rare SNPs were nominally significant using the score test but the \( p \)-values failed to stand up to further investigation using exact logistic regression (the exact \( p \)-values ranged from 3 \( \times 10^{-5} \) to 0.21). The many small \( p \)-values apparently reflected overly liberal behavior of the score test when alleles are rare and when there are many more cases than controls. In order to reduce discussion of a large number of likely false positive tests we consider in the subtype analyses only SNPs with at least 10 minor alleles seen over all cases and controls. With this restriction we found a total of ten globally significant SNPs (using the score test). However, \( p \)-values from exact logistic regression for these SNPs were again far less striking (ranging from 3 \( \times 10^{-5} \) to 1.5 \( \times 10^{-3} \)).

When restricted to estrogen receptor-positive (ER+) cases (n = 1,688) (and screening out SNPs with less than 10 minor alleles seen) the most significant coding SNP was a rare NS variant in \textit{UMODL1} (exm1573155, Ala542Thr, OR = 7.28, \( p = 9.8 \times 10^{-7} \)) (Table 2, Table S2). This SNP had a frequency of just over 0.2% in African Americans controls and 0 in the other groups. No associations are reported for this gene in the GWAS catalog. Neither this SNP nor any others were significant after correction for multiple testing.
# Table 2. The Most Significant Associations of Single Coding Variants with Breast Cancer Risk.

## All Cases (n = 2,984) vs Controls (n = 7,545)

| SNP ID   | Chr | Position | rs#       | A1/A2 | Type   | Gene       | OR  | P       | AA MAF | NH MAF | JA MAF | LA MAF | EA MAF |
|----------|-----|----------|-----------|--------|--------|------------|-----|---------|--------|--------|--------|--------|--------|
| exm61019 | 1   | 54476084 | rs145889899 | T/C    | 5P     | LDLRAD1    | 3.74| 2.50E-07| 0.0065 | 0      | 0      | 0      |
| exm1579798 | 21  | 46950811 | rs14289279 | A/C    | Val342Phe| SLC19A1    | 12.67| 1.30E-06| 0      | 0.0025 | 0      | 0      |
| exm841657 | 10  | 93668692 | NA        | A/G    | Pro12Leu| FGFBP3     | 2.50E-06| 0      | 0      | 0      | 0      |
| exm952402 | 11  | 104878019 | rs45585319 | C/A    | Leu17Arg| CASP5      | 9.69 | 2.50E-06| 0.00093| 0      | 0      |
| exm53277 | 6   | 31918154 | rs149101394 | G/A    | Lys533Arg| CTBP       | 12.67| 1.30E-06| 0      | 0.0007 | 0      | 0      |
| exm510328 | 5   | 179285781 | NA        | A/G    | Arg6Cys | SLC19A1    | 25.69| 7.10E-06| 0      | 0      | 0      |
| exm510328 | 5   | 179285781 | NA        | A/G    | Arg6Cys | SLC19A1    | 25.69| 7.10E-06| 0      | 0      | 0      |
| exm132287 | 1   | 186313108 | rs58030082 | T/C    | Val1178Ile| TPR        | 1.19 | 1.30E-05| 0.021  | 0.056  | 0.00038| 0      |

## ER+ Cases (n = 1,688) vs Controls (n = 7,545)

| SNP ID   | Chr | Position | rs#       | A1/A2 | Type   | Gene       | OR  | P       | AA MAF | NH MAF | JA MAF | LA MAF | EA MAF |
|----------|-----|----------|-----------|--------|--------|------------|-----|---------|--------|--------|--------|--------|--------|
| exm1573155 | 21  | 43529776 | NA        | A/G    | Ala542Thr| UMODL1     | 7.28 | 9.80E-07| 0      | 0      | 0      | 0      |
| exm85453 | 1   | 114380886 | rs138092829 | C/T    | Asp379Gly| PTPN22     | 12.12| 1.60E-06| 0      | 0      | 0      | 0      |
| exm1358833 | 17  | 74625633 | NA        | G/C    | Gly98Arg | ST6GALNAC1 | 3.26 | 2.40E-06| 0      | 0      | 0      | 0      |
| exm1159729 | 15  | 48734008 | rs113573732 | A/C    | Ala2025Ser| FBN1       | 7.66 | 2.70E-06| 0      | 0      | 0      |
| exm1093791 | 12  | 56998559 | NA        | T/C    | Val927Met| BAZ2A      | 2.50E-06| 0      | 0      | 0      | 0      |
| exm1234521 | 16  | 30795481 | NA        | G/C    | Lys56Asn | ZNF629     | 11.61| 4.20E-06| 0      | 0      | 0      |
| exm551694 | 6   | 44274257 | rs137852769 | G/C    | Glu510Gln| HADHA      | 11.10| 3.80E-06| 0      | 0      | 0      |
| exm178049 | 2   | 26418053 | rs137852769 | G/C    | Glu510Gln| HADHA      | 11.73| 5.20E-06| 0      | 0      | 0      |
| exm412038 | 4   | 88534411 | NA        | A/G    | Arg358Lys| DSPP       | 9.39 | 9.30E-06| 0      | 0      | 0      |

## ER- Cases (n = 441) vs Controls (n = 7,545)

| SNP ID   | Chr | Position | rs#       | A1/A2 | Type   | Gene       | OR  | P       | AA MAF | NH MAF | JA MAF | LA MAF | EA MAF |
|----------|-----|----------|-----------|--------|--------|------------|-----|---------|--------|--------|--------|--------|--------|
| exm1165463 | 15  | 59098004 | rs144893047 | T/C    | Arg35His | ADAM10     | 8.52 | 1.30E-09| 0      | 0      | 0      |
| exm221867 | 2   | 113539232 | rs142134831 | T/G    | Glu90Lys | IL1A       | 9.27 | 1.30E-08| 0      | 0      |
| exm220393 | 2   | 110959008 | rs145427967 | G/T    | Leu42Leu | NPHP1      | 7.40 | 8.40E-08| 0      |
| exm61019 | 1   | 54476084 | rs145889899 | T/C    | SP       | LDLRAD1    | 6.17 | 1.40E-07| 0      |
| exm645918 | 7   | 101183198 | rs190166648 | T/C    | Arg1567Pr | EMID2      | 14.51| 1.50E-07| 0      |
| exm492809 | 5   | 147505341 | NA        | A/G    | Arg356Lys | SPINK5     | 9.57 | 1.60E-07| 0      |
| exm1253047 | 16  | 69724780 | rs145602190 | G/A    | Gln1157Arg| NFAT5      | 5.39 | 1.70E-07| 0      | 0      | 0      |

## Exonic SNPs and Breast and Prostate Cancer Risk
In ethnic-specific analyses of overall breast cancer only one additional SNP (in \textit{FANCI}) met our criteria (p<3.9×10^{-7}) of global significance. This NS variant (rs62020347, \textit{Pro55Leu}) was common in European Americans, African Americans, and Latinos (3–8% frequency) but was only associated with risk among European Americans (MAF 8%, OR = 0.47, p = 1.8×10^{-7}) and was weakly associated with risk overall (p = 0.02) (Table S1).

Breast Cancer Gene Burden Analysis

Table 3 summarizes the most significant findings from the gene burden (sum of coding variants) analysis based on all common and rare (≤1%) functional SNPs in each gene. Further details are given in Table S5. For overall breast cancer no gene burden sum passed
the Bonferroni criteria (3×10^-5) for global significance for testing approximately 17,200 genes (see Methods). The strongest associations were for **MMAB** (p = 5.0×10^-5), **SLC16A6** (p = 5.0×10^-5) and **INS-IGF2** (p = 1.2×10^-4). The **MMAB** gene is close to non-exonic SNPs that have been associated with HDL cholesterol [14] and one of those GWAS SNPs (the intronic variant rs7134594) was among our top 100 single SNP associations with breast cancer (Table S1). **INS-IGF2** contains an intronic SNP that has been associated with type 1 diabetes [15]. Restricting the gene burden analysis to only SNPs with overall frequency ≤1% gave non-significant associations as well (p>8×10^-6) and none of the top five genes in these analyses have globally significant GWAS associations reported. For ER+ breast cancer, the burden of rare SNPs in gene **FGFR2** was nominally globally associated (p = 6×10^-5) although follow-up using exact logistic regression gave a larger p-value (1.0×10^-5). This gene included five rare SNPs and no reports of any GWAS associations for SNPs near this gene are found in the GWAS catalog. When examining ER- breast cancer, the burden of variants in **MMAB** remained one of the strongest associations (p = 2.0×10^-7). The burden of coding SNPs (all of which were rare) in **EGR2** was the leading association in the ER- analysis with a p-value from the score test of 1.2×10^-11. A variant upstream of **EGR2** has been associated in a GWAS of Ewings sarcoma [16]. Rare variant burdens also met our criteria for global significance for **CNR1** (p = 1.7×10^-10), **FKSG83** (p = 1.5×10^-7), **GATM** (p = 4.8×10^-7), and **ACSBG1** (p = 5.3×10^-7). Again as for the single SNP results for ER-disease, these variants were found to be overly liberal compared to an exact test (the smallest exact logistic regression p-value was 2.8×10^-5 for **ACSBG1**).

### Prostate Cancer Single SNP Associations

For overall prostate cancer (4,376 cases and 7,545 controls) none of the single SNP associations with prostate cancer met the Bonferroni adjustment for multiple comparison testing (nominal p<3.9×10^-5). The top two associations found for prostate cancer were for rare NS variants in **F13A1** (rs140712764, Val170Ile, OR = 4.52, p = 6.0×10^-3), **ANXA4** (rs146778617, Val125-Pho, OR = 4.52, p = 6.0×10^-6), Table 4, see also Table S2. Gene **F13A1** is a coagulant factor gene not obviously related to prostate cancer etiology. **ANXA4** encodes a protein that has been discussed as a possible marker for gastric cancer [17]. Of note, the third most significant association was for a common NS variant in **GPRC6A** (rs140712764 in **F13A1**, **ANO1** in **ANXA4**, rs2274911 in **GPRC6A**) and also with rs61746620 in **ZKSCAN2** (Ala574Val, OR = 13.4, p = 1.3×10^-3), although none of these were significant at our Bonferroni criteria.

**Ethnic-specific analyses.** No SNPs were significantly associated with overall prostate cancer in ethnic specific analysis (Table S3).

### Prostate Cancer Gene Burden Analysis

None of the gene burden analyses were significant for overall prostate cancer after correcting for multiple comparisons (p<3×10^-6) either when including common coding variants or when restricting the results to SNPs with frequency ≤1% (Table 5, Table S5). When the analysis was restricted to advanced prostate cancer, four gene burdens (for **SAMD1**, **FOX2**, **NOL1** and **CPF3**) were significant using the score test but not by exact logistic regression (p = 2.5×10^-3, 3.3×10^-3, 5.0×10^-3 and 3.4×10^-6 respectively). No notable findings were observed when only localized prostate cancer was assessed.

### Analyses at Known Risk Loci for Breast and Prostate Cancer

**GWAS loci.** Tables S6 and S7 give results for SNP associations for genes located at known breast and prostate cancer susceptibility regions (e.g. regions harboring globally significant associations) as of the time of this report (73 significant associations for breast cancer and 89 for prostate cancer [19,20]). For each region, we list the genes having one or more genotyped coding variants that lie within 500 kb of the known GWAS SNP and summarize associations (smallest p-value) with coding variants in those genes and with the burden of coding variants (all SNPs and rare SNPs). For breast cancer, we observed limited evidence of associations with rare coding variants in genes proximal to GWAS signals, with 9 genes (PTPN22, **PTPN7**, **MMAB**, **CASP8**, **SLC6A18**, **FOXP2**, **CTSW**, **CCDC88C**, **ZNF404**) having SNPs or gene burdens achieving p-values of p<0.05 (Table S6) after correcting for either the number of nearby (+/−100 kb) SNPs (single SNP analyses) or genes (gene burden) for each GWAS index association. Of SNPs in linkage disequilibrium (LD), r^2>0.3 (in Europeans in 1000 Genomes), with GWAS hits and SNPs in 2 genes (**STXBP4** and **ZNF404**) which were correlated with 2 index GWAS SNPs (rs6504950 and rs3760982) and associated weakly at p<0.05. For prostate cancer, the most significant GWAS-related association, as described above, was with rs2274911 (Pro91Ser) in **GPRC6A**. The next most significant finding was with rs18365251 (Val125Met) in **PMVK** at 1q21 (p = 3.0×10^-4). This SNP was only
### Table 4. The Most Significant Associations of Single Coding Variants with Prostate Cancer Risk.

#### All Cases (n = 4,376) vs Controls (n = 7,545)

| SNP ID   | Chr | Position | rs#    | A1/A2 | Type     | Gene   | OR ± SE | P     | AA MAF ± SE | NH MAF ± SE | JA MAF ± SE | LA MAF ± SE | EA MAF ± SE |
|----------|-----|----------|--------|-------|----------|--------|---------|-------|-------------|-------------|-------------|-------------|-------------|
| exm514211 | 6   | 6266854  | rs140712764 | T/C   | Val170Le | F13A1  | 28.007  | 9.1E-07 | 0.000233    | 0           | 0           | 0           | 0           |
| exm199465 | 2   | 70052624 | rs14678617 | T/G   | Val319He | ANXA4  | 4.523   | 6.0E-06 | 0.002563    | 0           | 0           | 0           | 0           |
| exm574153 | 6   | 117130704| rs2274911  | G/A   | Pro915er | GPRC6A | 0.875   | 1.3E-05 | 0.2379      | 0.2717      | 0.4332      | 0.2657      | 0.2542      |
| exm68152  | 1   | 70896038 | rs14578587 | C/T   | Cys229Arg | CTH   | 9.011   | 3.1E-05 | 0.000693    | 0           | 0           | 0           | 0           |
| exm971959 | 11  | 134128968| NA      |       | SER186ASN | ACAD8  | >999.999| 3.2E-05 | 0           | 0           | 0           | 0           | 0           |

#### Advanced Cases (n = 499) vs Controls (n = 7,545)

| SNP ID   | Chr | Position | rs#    | A1/A2 | Type     | Gene   | OR ± SE | P     | AA MAF ± SE | NH MAF ± SE | JA MAF ± SE | LA MAF ± SE | EA MAF ± SE |
|----------|-----|----------|--------|-------|----------|--------|---------|-------|-------------|-------------|-------------|-------------|-------------|
| exm280349 | 2   | 239049718| NA     | A/G   | Arg108His | KLHL30 | 13.991  | 1.7E-09 | 0           | 0           | 0           | 0           | 0.002081    |
| exm1488544| 19  | 49376683 | rs45533432 | G/A   | Arg65Gly | PPP1R15A | 4.677   | 1.2E-08 | 0.002097    | 0.002481    | 0           | 0           | 0.00384     |
| exm643590 | 7   | 100634145| rs143984295 | A/G   | Ala101Thr | MUC12  | 14.425  | 1.5E-08 | 0           | 0           | 0           | 0           | 0.000384    |
| exm710148 | 8   | 55540419 | rs114797722 | C/G   | Ala329Pro | RP1    | 13.409  | 2.0E-08 | 0.001631    | 0           | 0           | 0           | 0           |
| exm782688 | 9   | 13024593 | rs150292099 | G/A   | VAL157ILE | LRSM1  | 10.488  | 3.5E-07 | 0.002097    | 0           | 0           | 0           | 0           |
| exm2275215| 17  | 58235051 | rs18568468 | T/A   | SP        | CA4    | 7.137   | 5.1E-07 | 0           | 0.002481    | 0.003976    | 0           | 0           |
| exm1321007| 17  | 39520119 | rs150620728 | T/C   | Arg395His | KRT33B | 7.485   | 5.8E-07 | 0.002816    | 0           | 0           | 0           | 0.00152     |
| exm942022 | 11  | 77538984 | rs141331999 | G/A   | Arg139Ser | MOGAT2 | 8.489   | 9.4E-07 | 0.000932    | 0           | 0           | 0           | 0.000384    |
| exm594160 | 6   | 16734185 | rs37516361 | T/G   | Ser221Thr | RNAS1T2 | 8.129   | 1.1E-06 | 0.00303     | 0           | 0           | 0           | 0           |
| exm1607994| 22  | 38483189 | NA     | G/A   | Ser401Pro | BAIAP2L2 | 6.084   | 2.4E-06 | 0.001865    | 0           | 0           | 0           | 0.001244    |

#### Non-Advanced cases (n = 3,666) vs Controls (n = 7,545)

| SNP ID   | Chr | Position | rs#    | A1/A2 | Type     | Gene   | OR ± SE | P     | AA MAF ± SE | NH MAF ± SE | JA MAF ± SE | LA MAF ± SE | EA MAF ± SE |
|----------|-----|----------|--------|-------|----------|--------|---------|-------|-------------|-------------|-------------|-------------|-------------|
| exm514211 | 6   | 6266854  | rs140712764 | T/C   | Val170Le | F13A1  | 28.366  | 8.3E-07 | 0.000233    | 0           | 0           | 0           | 0           |
| exm1228070| 16  | 25255366 | rs1746620 | A/G   | Ala574Val | ZSK-CAN2 | 13.396  | 1.3E-05 | 0.0002331   | 0           | 0           | 0           | 0.0002485  |
| exm199465 | 2   | 70052624 | rs14678617 | T/G   | Val319He | ANXA4  | 4.275   | 3.4E-05 | 0.002563    | 0           | 0           | 0           | 0           |
| exm574153 | 6   | 117130704| rs2274911  | G/A   | Pro915er | GPRC6A | 0.876   | 4.1E-05 | 0.2379      | 0.2717      | 0.4332      | 0.2657      | 0.2542      |
| exm1311040| 17  | 3268826  | rs138527286 | C/T   | Ile65Val | CCL1   | 2.343   | 4.1E-05 | 0.01072     | 0           | 0           | 0           | 0           |
| exm68152  | 1   | 70896038 | rs14578587 | C/T   | Cys229Arg | CTH   | 8.761   | 6.3E-05 | 0.000699    | 0           | 0           | 0           | 0           |
| exm1105738| 14  | 61180657 | rs3742636 | T/G   | His65Pro | SIX4   | 1.127   | 8.6E-05 | 0.4256      | 0.2742      | 0.4688      | 0.2479      | 0.2912      |
common in African Americans (20% frequency; ≤1% in the other populations). An additional eight nearby genes had SNPs with corrected p-values between 0.001 and 0.05: \(\text{ITGA6, VGLL3, TECPR1, TPCN2, FAM83F, PBXIP1, FARP2, and TTLL12 (Table S7).}

Seven SNPs were correlated with a GWAS index SNP at \(r^2 = 0.3\) in the 200 kb window and significant at \(p < 0.05\) (\(\text{SLC2A4RG, PDLIM5, RNMTL1, KLK3, MLPH, RTEL1 as well as GPRC6A.}

Given the modest effects noted with the initial GWAS signals as well as observed with these correlated coding SNPs (\(\text{OR per allele of } \sim 1.1\); Table S6 and S7), and the lack of strong signals noted for the index signals across populations [21] conditional analyses will

| Table 5. The Most Significant Associations of Gene Burden of Coding Variants with Prostate Cancer Risk. |
|---|---|---|---|
| Gene | Chr | # of SNPs OR | P |
| Overall prostate cancer, functional SNPs | C6orf165 | 6 | 26 | 0.86 | 0.00000573 |
| | MANSC1 | 12 | 9 | 0.86 | 0.0000611 |
| | GP6 | 19 | 21 | 0.96 | 0.00006642 |
| | SERPING1 | 11 | 7 | 1.51 | 0.0000963 |
| | SIX4 | 14 | 13 | 1.10 | 0.00012 |
| Overall prostate cancer, rare SNPs | C17orf53 | 17 | 24 | 0.44 | 0.0000533 |
| | WDR54 | 2 | 8 | 1.94 | 0.0000112 |
| | TYW1B | 7 | 5 | 1.55 | 0.0000138 |
| | FGFBP3 | 10 | 5 | 18.04 | 0.0000131 |
| | SERPING1 | 11 | 7 | 3.13 | 0.000188 |
| Advanced prostate cancer, functional SNPs | SAMD1 | 19 | 3 | 26.03 | 0.000000122 |
| | FOXF2 | 6 | 2 | 42.26 | 0.000000513 |
| | NOL4 | 18 | 2 | 17.53 | 0.00000213 |
| | ID1 | 10 | 3 | 13.41 | 0.0000124 |
| | CYP11B1 | 8 | 5 | 1999 | 0.0000207 |
| Advanced prostate cancer, rare SNPs | CPA3 | 3 | 19 | 2.26 | 0.000000697 |
| | SAMD1 | 19 | 3 | 26.03 | 0.000000122 |
| | FOXF2 | 6 | 2 | 42.26 | 0.000000513 |
| | NOL4 | 18 | 2 | 17.53 | 0.00000213 |
| | ID1 | 10 | 3 | 13.41 | 0.0000124 |
| Non-advanced prostate cancer, functional SNPs | ATP6V0D2 | 8 | 17 | 0.75 | 0.0000749 |
| | C6orf165 | 6 | 26 | 0.87 | 0.0000913 |
| | MANSC1 | 12 | 9 | 0.86 | 0.000108 |
| | GP6 | 19 | 21 | 0.96 | 0.000113 |
| | GPR125 | 4 | 37 | 1.11 | 0.000143 |
| Non-advanced prostate cancer, rare SNPs | CPA3 | 3 | 19 | 2.26 | 0.000000697 |
| | SAMD1 | 19 | 3 | 26.03 | 0.000000122 |
| | FOXF2 | 6 | 2 | 42.26 | 0.000000513 |
| | NOL4 | 18 | 2 | 17.53 | 0.00000213 |
| | ID1 | 10 | 3 | 13.41 | 0.0000124 |

\[\text{doi:10.1371/journal.pgen.1003419.t005}\]
be needed in much larger samples of the GWAS population (mainly European ancestry) to determine whether these coding SNPs are the biologically functional alleles underlying the GWAS signal. (Our ability to perform informative conditional analysis here is further hampered by the fact that only a minority of the index GWAS hits are included on the Illumina array).

**Extended associations.** Because of the interest in the possibility that rare coding variants with large effect sizes (OR > 1.5 or higher) may underlie GWAS signals and since LD with rare SNPs can extend much further than with common SNPs, we report in Table S6 and S7 the strongest associations for all coding variants in each gene within 500 kb of each GWAS index signal. The strongest single SNP associations with breast cancer (from 100 to 500 kb) were in **RNASEL** (260 kb from rs3191798 on chromosome 11, p = 1.1 x 10^-6), **CCDC91** (450 kb from index signal rs10771399 on chromosome 12, p = 4.7 x 10^-5), **ZFYVE26** (410 kb from rs2588809 on chromosome 14, p = 7.9 x 10^-4), **C16orf46** (444 kb from rs1329835 on chromosome 16, p = 2.8 x 10^-4), **UNC13A** (337 kb from rs1870, p = 5.7 x 10^-4) and **NRIPI** (182 kb away from rs2923093 on chromosome 21, p = 1.3 x 10^-4).

For prostate cancer the strongest such associations were with **SNED1** (412 kb from rs3771570 on chromosome 2, p = 3.5 x 10^-4) and **PASK** (317 kb from the same index SNP on chromosome 2, p = 4.8 x 10^-4). No other associations in this distance range had p < 0.001 for overall breast or prostate cancer.

**High-risk genes.** We also examined genes implicated in family-based studies of breast or prostate cancer (Tables S6 and S7) as they are strong candidates. For breast cancer, we analyzed 11 genes and did not observe an over-representation of associations at p < 0.05 in any gene (observed/tested: **ATM**, 3/62; **BRCA1**, 2/42; **BRCA2**, 3/80; **BRIPI**, 0/16; **CHEK2**, 0/8; **NBX1**, 3/17; **PALB2**, 1/26; **PETN**, 0/1; **RAD50**, 3/24; **STK11**, 1/4; **TP53**, 4/4), or any significant associations (p < 0.05) from gene-based burden testing. The most significant associations in these genes (p < 0.05) were noted with non-synonymous variants: rs56009889 in **ATM** (p = 4.1 x 10^-3, p = 0.0065), rs80357090 in **BRCA1** (p = 0.018), rs1799944 in **BRCA2** (p = 0.033, p = 0.0046), rs115291485 in **NBX1** (p = 0.041, p = 0.0067), rs2230017 in **RAD50** (p = 0.44, p = 0.0069) and rs138796658 in **PALB2** (p = 1.6 x 10^-3).

For prostate cancer, we analyzed 5 genes and did not observe an over-representation of SNP associations at p < 0.05 (observed/tested: **BRCA2**, 2/83; **ELAC2**, 1/9, **HOXB13**, 0/2; **MSR1**, 1/22; **RNASEL**, 2/21). However, we did observe suggestive evidence of associations with burden testing of rare (MAF < 0.01) SNPs in **ELAC2** (p = 1.6 x 10^-3) and in **RNASEL** (p = 1.2 x 10^-3). The most significant associations included a very rare null variant in **ELAC2** that was mainly observed in African Americans (rs149344601, Ile356Val, MAF = 6.6 x 10^-3; OR = 14.0, p = 0.0014), and a nonsense SNP (rs74313564, Glu263Ter) and null variant (rs15129635) in **RNASEL** (both with OR > 2.5, p = 0.012) that were observed in the same individuals. We did not observe significant associations with any of the reported risk variants in these genes. One association for breast cancer, a single SNP in **LDLRAD1**, appeared to pass our established level of global significance (p < 3.9 x 10^-5) when all cases were examined. No associations (either single SNPs or gene burdens) were globally significant for overall prostate cancer. Subset analyses, by ER status for breast cancer.
cancer or advanced/non-advanced for prostate cancer generally failed to show believable associations. While the score test gave many “globally significant” associations these apparently reflected excess type I error of this test when both the number of cases is small compared to the number of controls and when the SNPs were rare. This breakdown in reliability is similar to that seen for the uncorrected Pearson chi-square test (a special case of the score test when no covariates are present), which is well-known to have poor control of type I error when the expected number of cases is very small for a cell. Following-up such associations with exact logistic regression implemented in SAS (Cary, NC) provided larger p-values not globally significant using our criteria.

Nevertheless a number of suggestive findings were observed that are worthy of further attempts at replication: The splice site variant rs45889899 in **JDLRAD1** (our top finding for overall breast cancer) is found in low frequency (<1%) in African American controls (higher of course in cases since this is nominally a risk variant), and only seen among cases in the other groups. No associations with any disease or phenotype have to date been reported for this gene. Among the other genes highlighted in Table 2 or Table 3, associations have been reported for SNPs in **SLC19A1** and **CASP5** for renal cancer [31,32]; **BAZ2A** has been reported to be up-regulated in CLL patients [33]. Also notable is a strong link between SNPs in **EGR2** (ER- association) and risk of Ewing’s sarcoma [16].

For prostate cancer (all cases) the third strongest association result was for a common NS coding variant (rs2274911) in **GRPC6A** that is in very high LD with the known intronic GWAS variant rs339331. In our data the NS variant was slightly more associated (Table S3) with prostate cancer risk (p = 1.3 × 10^{-5}) than was rs339331 (p = 2.1 × 10^{-5}). The coding SNP is arguably a more likely causal variant than the intronic SNP since expression of **GRPC6A** is substantially increased in prostate cancer cell lines, and mice deficient in **GRPC6A** show retarded prostate cancer progression [34]. In addition, **GRPC6A** deficiency in mice also attenuates the rapid signaling responses to testosterone, an androgen that is critical for initiation and progression of prostate cancer [35].

Other suggestive findings for prostate cancer include SNPs in a variety of genes such as **F13A1** expression of which has been associated with bone metastasis in prostate cancer [36], **AXH4** which is up-regulated in gastric and other cancers [17], **ND1** where cryptic translocations may be involved in AML occurrence [37] and **MUC12**, expression of which has been reported to be a prognostic marker in colon cancer [38]. The burden of rare SNPs in **FGFBP6** (one of the stronger association seen for breast cancer) was also among the top associations for overall prostate cancer (Table 5, p = 1.5 × 10^{-5}).

We evaluated also associations in regions surrounding known (GWAS) risk alleles as a partial fine-mapping exercise; we specifically focused upon (1) coding alleles reported to be in high LD (in Europeans using 1000 Genomes data) with the index marker, and (2) other (generally less common) coding alleles within 500 kb of the GWAS alleles, that might show associations that could underlie (by synthetic association [30]) GWAS associations. A number of GWAS risk alleles are in reasonable LD (r^2>0.3) with coding SNPs on the array and several of the latter show nominal associations (p<0.05) with breast cancer risk including SNPs in **STXBPA**, **ZNF415**, and **ZNF404** which are all worth evaluating as candidate loci potentially explaining the index GWAS associations. For prostate cancer, a similar observation is made most notably for **GRPC6A** but also for **MLPH** (GWAS index = rs75814330, chromosome 2, p = 0.003), **PDLMS** (rs12500426, chromosome 4, p = 0.019), **RNMTLI** (rs684232, chromosome 17, p = 0.024), **KLK3** (rs2735839, chromosome 19, p = 0.0046), and **RTEL1** (rs6082509, chromosome 20, p = 0.001). Previous reports [24,39] have highlighted the NS SNP rs17632542 in **KLK3** as highly associated with PSA level and a highly significant risk variant in fine-mapping of the locus near rs2735839 [39]; while no report for prostate cancer exists for coding SNPs in **RTEL1**, another NS SNP, rs3200908, in **RTEL1** has been found to be associated with glioma risk [40].

Other coding SNPs that could include causal variants producing synthetic associations (associations of rare with common SNPs of high penetrance) include SNPs in genes **INS-IGF2**, **ZFYTE26**, **C16orf46**, **UNCI134**, **NRP1** and **CCDC91** for breast cancer and SNPs in **SND1** and **PASK** for prostate cancer. These do not have high r^2 with the GWAS variants as they are mostly rare (and are >100 kb away from the index signal) but their nominally strong associations (p-values<1×10^{-5}) might possibly be indicative of signals extending for many thousands of base pairs, although it will take much larger studies to verify or refute this.

We found little evidence that the NS, SP, or nonsense variants captured by the HumanExome SNP array that fall within known or suspected high risk genes for breast or prostate cancer are meaningfully associated with either cancer. The Illumina array does not directly interrogate the rare, high-risk mutations, such as frameshiftn mutations in **BRCA1** or **BRCA2** (e.g. c.68_69delAG) [41], as very few indels are included on this array (just 136 were examined here). The inability to address frameshift mutations either within known risk genes or more widely is a limitation of this report. Other limitations include the focus on Europeans in the development of the array (as seems to be particularly reflected in the relatively small fraction of SNPs found to be polymorphic in Japanese Americans), and the loss of some targeted SNPs in the manufacturing process and in our QC procedures. In addition, this technology (unlike exome sequencing) cannot address the role of either private variation or of variants too rare to have been reliably identified during the discovery phase of the development of the array.

Genotyping cases and controls from our prospective cohort allowed us an opportunity to examine other cancer-related phenotypes and traits for which data and specimens had been collected prior to breast or prostate cancer diagnosis. While two of these endpoints (BMI, alcohol) were based on self-report, we were able to strongly replicate a number of known associations such as rs671 in **ALDH2** with alcohol intake which is proof of principle that the exome array has the potential to reveal biologically relevant coding variants. Apparently novel findings for PSA, BMI, and alcohol consumption will need to be replicated in large-scale exome association analyses; hopefully making the results from these preliminary analyses in a multiethnic population broadly available will contribute to novel discoveries and further understanding the genetic basis of these traits.

In order for rare variants to play an important role in explaining missing heritability [42] even in composite they must have effects that are larger in magnitude than those observed for common SNPs. Roughly speaking, for a given allele the contribution to additive heritability (under a liability model for example [43]) is proportional to 2b^2p(1-p) where b is the log odds ratio (OR) and p is the frequency for that allele. Under simplifying assumptions (such as limited selection and constant population sizes) population genetics theory [44] indicates that there should be approximately as many variants “moderately rare” with frequency in the range 0.1 to 1% as there are the common variants in the range 5 to 50% that have been the targets of GWAS studies to date. However, in order that variants in the frequency range from 0.1 to 1% have the same composite effects on risk as do those in the frequency range
from 5 to 50% then the magnitude of effect sizes must be considerably larger than for the common variants; if ORs for common variants lie in the range from 1.1 to 1.3 then ORs in the range from 2 to 6 are needed for the rare and common alleles to have similarly sized roles in disease susceptibility (assuming that the same fraction of all rare alleles are risk variants as for common alleles). Moreover, under the hypothesis that the coding regions of the genome (~1% of the total genome) by themselves play a profound role in disease susceptibility these ORs would likely need to be skewed even higher — i.e. if rarer variation in 1% of the genome was to play as much a role as does common variation over the entire genome then the existence of ORs above 10 or even greater for such variation may arguably be a necessary consequence.

Realistically our study only begins the assessment of whether a range of effects for “moderately rare” coding variants is possible: the detectable ORs in this study range from approximately 3 to 13 for alleles with frequency 1 to 0.1%, respectively. While these are large ORs the above argument indicates that such effect sizes are not unreasonable if rarer protein coding variation plays a similar role in the heritability of risk as does common variation genome-wide. Our failure to find such ORs for the rarer alleles may be providing evidence against coding variation having a predominant role in breast and prostate cancer heritability and risk (outside of high risk families).

In summary, the analyses and methods described here do not support NS variants on the current exome chip as conveying moderate to high risk for breast and prostate cancer. While some suggestive findings are noted it is likely that very large sample sizes of the order that can be only developed through collaborative efforts such as those now engaged in the NCI GAME-ON post-GWAS meta-analysis of common variants, will be required in order to further the understanding of the role of rare NS and other coding variation in disease genetics. Exome sequencing of high-risk families will continue to be important to reveal biologically relevant coding variants for these cancers, both for insertion/deletion variants that were not covered by the current array, and to capture rarer variation (including private variants) that cannot be captured except by sequencing.

Materials and Methods

Ethics Statement

This work has been performed according to relevant national and international guidelines. Written consent was obtained at the time of DNA sample collection. The Institutional Review Boards at the University of Southern California and University of Hawaii approved of the study protocol.

Study Population

The MEC consists of more than 215,000 men and women in California and Hawaii aged 45–75 at recruitment, and comprises mainly five self-reported racial/ethnic populations: African Americans, Japanese, Latinx, Native Hawaiians, and European Americans [45]. Between 1993 and 1996, adults enrolled in the study by completing a 26-page mailed questionnaire asking detailed information about demographic factors, personal behaviors, and prior medical conditions. Potential participants were identified through driver’s license files from Departments of Motor Vehicles, voter registration lists, and Health Care Financing Administration data files. Incident breast and prostate cancer, as well as stage and hormone receptor status was identified by linkage of the cohort to the Surveillance, Epidemiology, and End Results cancer registries covering Hawaii and California. Between 1995 and 2006, blood specimens were collected prospectively from ~67,000 participants for genetic and biomarker analyses. Currently, the breast cancer case-control study nested in the MEC includes 5,141 women diagnosed with invasive breast cancer and 5,721 frequency-matched controls without breast cancer, matched by race/ethnicity and age (in 5-year age categories). The case-control study of prostate cancer includes 4,675 men diagnosed with incident prostate cancer and 4,300 male controls without prostate cancer. The Institutional Review Boards at the University of Southern California and University of Hawaii approved of the study protocol.

Genotyping and Quality Control

Genotyping of the Illumina Human Exome BeadChip (n = 247,895 SNPs) was conducted at the USC Genomics Core Laboratory.

DNA extraction of buffy coat fractions was conducted using the Qiagen protocol. Cases and controls were randomly placed across ethnic-specific plates for each cancer type. All samples had DNA concentrations >10 ng/ul. Initial genotype definitions were based on auto-clustering 6,404 samples across all populations which had call rate >0.99 (African American 1883, Japanese American 1823, Latino 1008, European American 1690) using the GenomeStudio software (V2011.1). Following genotype calling on all samples (>16,000), manual inspection was conducted of the following SNPs: 1) SNPs with call rate <0.98 (n = 3,317), 2) monomorphic SNPs with call rate <1 (n = ~15,000), 3) SNPs with minor allele frequency between 0 and 0.001 and call rate <1 (n = ~31,500), 4) SNPs with >1 replicate error based on sample duplicates (~1,000, discussed below), 5) SNPs with apparent differences in minor alleles frequencies >15% across ethnic-specific 96 sample plates (n = 798), or other evidence of batch/plate effects on allele frequency (n = 18,188), 6) all mitochondridial SNPs and all SNPs on the X and Y chromosomes (n = 5,574), and 7) autosomal SNPs out of Hardy-Weinberg Equilibrium in more than one ethnic group with p value<0.001 and at least one ethnic group with p value<0.00001 (n = 827). During the inspections we in total inspected cluster plots for approximately 70,000 SNPs (counting overlapping SNPs in the categories above) and genotypes were manually edited for 27,506 SNPs.

Of the 15,837 samples described above genotyping was successful with call rates ≥98% for 15,573 samples; of these we removed 17 samples for which reported sex conflicted with assessment of X chromosome heterozygosity, and 651 samples based on relatedness. Relatedness was determined using the IBD calculation in plink [46], and we removed one of each estimated MZ twin, sibling, parent-offspring, half sibling, or first cousin pairs. In the analysis, we also removed SNPs with <98% call rates (n = 2,331). To assess genotyping reproducibility we included 338 replicate samples which passed genotyping QC; among these samples the concordance rate of heterozygote calls, number concordant/(number concordant+number discordant), was 99.6% or greater for all replicate samples (average 99.99%). The final analysis dataset included 245,339 SNPs genotyped on 2,984 breast cancer cases and 3,568 controls, and 4,376 prostate cancer cases and 3,977 controls.

Statistical Analysis

We relied on documentation files obtained from the University of Michigan posted on ftp://share.sph.umich.edu/exomeChip/ IlluminaDesigns/ for the assessment of SNP type (i.e. NS, SP), and the amino acid affected. The array also includes SNPs that do not code for protein changes including synonymous SNPs, and other intergenic SNPs including ancestry informative markers, and
GWAS identified risk SNPs for a number of diseases and outcomes. All SNPs were analyzed and their results shown in Tables S1, S2, S3, S4, S5, S6, S7, S8, S9. However our primary analysis focused on the 191,032 putative functional variants in the following categories (NS, SP and stop gain or loss) that passed quality control procedures discussed above.

We estimated principal components in the entire sample using EIGENSTRAT [47] based on 2,957 autosomal ancestry informative markers on the array. We adjusted for the top 10 principal components in all analyses.

**Association testing of single markers.** For all analyses except those of the X and Y chromosomes all controls (men and women combined) were utilized in the analysis of each cancer in order to increase statistical power. Only controls of the same sex were used to analyze X or Y chromosome variants. Analyses were performed overall and within each racial/ethnic group. For each genotyped SNP, odds ratios (OR) and 95% confidence intervals (95% CI) were estimated using unconditional logistic regression of case/control status adjusting for age at diagnosis (cases) or blood draw (controls), and reported race/ethnicity in the overall analyses, and the first 10 eigenvectors in both overall and ethnic-specific analyses. For each SNP, we tested for allele dosage effects through a 1 d.f. score chi-square trend test. When exposures are rare but with very strong effects the score test can be more powerful than the usual Wald test for reasons described in Hauck and Donner [48]. However, we found the score test to be overly liberal when both the exposures are rare and when (as in analysis of advanced prostate and ER-positive breast cancer) the number of cases in a given analysis is small compared to the number of controls. Therefore we followed up any apparently globally significant findings found with the score test by rerunning that analysis using the exact logistic regression procedure implemented in SAS (Cary NC); when using the exact test we dropped the eigenvectors above. We performed the gene burden analyses for binary or continuous phenotypes. The 100 most statistically significant results for each phenotype are presented in Tables S1, S2, S3, S4, S5.

**Evaluation of the known risk loci for breast and prostate cancer.** We also examined whether known risk alleles (generally intergenic or intronic) from GWAS studies of breast or prostate cancer may be reflecting an underlying signal from a nearby protein-altering variant. In these analyses for each GWAS SNP (73 for breast, 83 for prostate cancer) we initially interrogated nearby SNPs known to be or likely to be in LD with the index signal. Because LD data is not yet available for the majority of the SNPs on the HumanExome array, we expanded the associations considered to be all those within a 100 kb region on either side of the index signal, since LD between common SNPs can sometimes extend this far. In this region we highlighted in the results section and discussion, SNPs with modest signals of association (p<0.05) as well as more strongly significant SNPs. Here the common SNPs are likely to be in high LD with the (generally common) GWAS variants, and the rare SNPs could be producing synthetic associations. We then relaxed this 200 kb region to 1 mb (500 kb on either side of the index signal) in order to expand our examination of possible synthetic associations between rare SNPs and the index GWAS findings, since LD with rare SNPs can extend considerably further than with common SNPs.

**Power Analyses**

Recognizing that many variants are only polymorphic in a few racial/ethnic groups, we give power analysis for a study with 1,000 cases and 2,000 controls (roughly the number of cases and controls in each of the four largest populations) by odds ratio (1–200) and allele frequencies ranging from 0.0001 to 1 (Figure S2). The Bonferroni criteria for significance in this study is calculated to be 0.05 divided by the largest number of polymorphic SNPs in any population (African Americans, ~125,000) or roughly 3.9×10⁻⁷. For the gene burden analysis the Bonferroni criteria is 0.05 divided by the number of genes considered or roughly 3×10⁻⁶. We had 80% power to detect odds ratios of 3.3 or above for SNPs with a frequency of 0.01 and odds ratios in the range 13 or above for SNPs of frequency 0.001 in single SNP analyses. Power for the gene burden analysis depends upon the number of polymorphic SNPs in a given gene. Using a Poisson approximation (i.e. with variance assumed to be equal to the mean) a gene with 10 variants each of frequency 0.001 gives power of 80% to detect a per minor allele OR of 3.1. For genes with many more variants (100) of the same frequency detectable ORs per minor allele are 1.6 or greater. For common variants present in all ethnic groups we had much greater power to detect associations, for example we had 80% power to detect a 20% allele with an OR of 1.24 in the global analyses; for the region-specific analyses we have 80% power to detect a 20% allele with an OR of 1.17 in a region with 100 variants and 1.14 in a region with 10 variants.

**Supporting Information**

**Figure S1** Allele frequency of putative functional SNPs for a. All ethnicities combined; b. European American; c. African American; d. Latino; e. Japanese American; f. Native Hawaiian (DOCX).

**Figure S2** Statistical power for single SNP analyses. (DOCX)

**Table S1** One hundred most significant single SNP associations with breast cancer; over all ethnic groups (S1.1) and by ethnic group (S1.2–6). (XLSX)
Table S2 | One hundred most significant associations between single SNPs and ($S2.1$) ER-positive Breast cancer and ($S2.2$) ER-negative breast cancer. (XLSX)

Table S3 | One hundred most significant single SNP associations with prostate cancer; over all ethnic groups ($S3.1$) and by ethnic group ($S3.2$–$S3.6$). (XLSX)

Table S4 | One hundred most significant associations between single SNPs and ($S4.1$) advanced prostate cancer and ($S4.2$) localized prostate cancer. (XLSX)

Table S5 | Gene burden analyses. One hundred strongest associations with ($S5.1$) Overall breast cancer, ($S5.2$) ER-positive breast cancer, ($S5.3$) ER-negative breast cancer, ($S5.4$) Overall prostate cancer, ($S5.5$) Advanced prostate cancer and ($S5.6$) Non-advanced prostate cancer. (XLSX)

Table S6 | Relationship between SNPs or genes known to be associated with breast cancer and coding SNPs on the exome array. Summary of nearest coding snps and gene burden analyses for ($S6.1$) GWAS associations and ($S6.2$) High risk genes. (XLSX)

Table S7 | Relationship between SNPs or genes known to be associated with prostate cancer and coding snps on the exome array. Summary of nearest coding snps and gene burden analyses for ($S7.1$) GWAS associations and ($S7.2$) High risk genes. (XLSX)

Table S8 | Summary statistics for other phenotypes examined: BMI, alcohol intake, and PSA. (XLSX)

Table S9 | Most significant single SNP association results for other phenotypes examined. (XLSX)

Acknowledgments
We thank the men and women who volunteered to participate in the MEC. We also thank Andrea Holbrook for technical support.

Author Contributions
Conceived and designed the experiments: CAH BEH LLM DOS. Performed the experiments: LX XS LCP DVDB EC KP. Analyzed the data: YH YF YP CH CAH DOS. Contributed reagents/materials/analysis tools: LNK LLM BEH. Wrote the paper: CAH DOS YHF.

References
1. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acid Res 33: D154–D157.
2. Stratton MR, Rahman N (2008) The emerging landscape of breast cancer susceptibility. Nat Genet 40: 17–22.
3. Lovelady C, Turnbull C, Ramsay E, Hughes D, Raerk S, et al. Germline mutations in RAD51D confer susceptibility to ovarian cancer. Nat Genet 45: 879–882.
4. Rajpert T, Guddbjartsson DF, Sulem P, Jonsson A, Sigurdsson A, et al. Mutations in BRIP1 confer high risk of ovarian cancer. Nat Genet 45: 1104–1107.
5. Johannsen CT, Wang J, Lanktree MB, Cao H, McIntyre AD, et al. Excess of rare variants in genes identified by genome-wide association study of hyperlipoproteinemia. Nat Genet 42: 684–687.
6. Fearnhead NS, Wilding JL, Winney B, Tonks S, Bartlett S, et al. Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. Proc Natl Acad Sci U S A 103: 15992–15997.
7. Sulem P, Guddbjartsson DF, Walters GB, Helgadottir HT, Helgason A, et al. Identification of low-frequency variants associated with gout and serum uric acid levels. Nat Genet 45: 1127–1130.
8. Ricoux MA, Beaudou M, Gardet A, Stevens C, Sharma Y, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet 43: 1066–1073.
9. Edwards SM, Kote-Jarai Z, Meitz J, Hamoudi R, Hope Q, et al. Two hundred strongest single SNP association results for advanced prostate cancer and ($S2.1$) ER-positive Breast cancer and ($S2.2$) ER-negative Breast cancer. (XLSX)
10. Laramie JM, Wilk JB, Williamson SL, Nagle MW, Latourelle JC, et al. Fine mapping SNPs and Breast and Prostate Cancer Risk.
32. Dong LM, Brennan P, Karami S, Hung RJ, Menashe I, et al. (2009) An analysis of growth, differentiation and apoptosis genes with risk of renal cancer. PLoS ONE 4: e4895. doi:10.1371/journal.pone.0004895

33. Hanlon K, Rudin CE, Harries LW (2009) Investigating the targets of MIR-15a and MIR-16-1 in patients with chronic lymphocytic leukemia (CLL). PLoS ONE 4: e7169. doi:10.1371/journal.pone.0007169

34. Pi M, Quarles LD (2012) GPRC6A regulates prostate cancer progression. Prostate 72: 399–409.

35. Pi M, Parrill AL, Quarles LD (2010) GPRC6A mediates the non-genomic effects of steroids. J Biol Chem 285: 39893–39904.

36. Morrissey C, True LD, Roodier MP, Coleman IM, Hasley S, et al. (2008) Differential expression of angiogenesis associated genes in prostate cancer bone, liver and lymph node metastases. Clin Exp Metastasis 25: 377–388.

37. Hollink IH, van den Heuvel-Eibrink MM, Arentsen-Peters ST, Pratcorona M, Abbas S, et al. (2011) NUP98/NSD1 characterizes a novel poor prognostic group in acute myeloid leukemia with a distinct HOX gene expression pattern. Blood 118: 3645–3656.

38. Matsuyama T, Ishikawa T, Mogushi K, Yoshida T, Iida S, et al. (2010) MUC12 mRNA expression is an independent marker of prognosis in stage II and stage III colorectal cancer. Int J Cancer 127: 2292–2299.

39. Kote-Jarai Z, Amin Al Olama A, Leongamornlert D, Tymrakiewicz M, Saunders E, et al. (2011) Identification of a novel prostate cancer susceptibility variant in the KLK3 gene transcript. Hum Genet 129: 607–694.

40. Egan KM, Thompson RC, Nabors LB, Olson JJ, Brat DJ, et al. (2011) Cancer susceptibility variants and the risk of adult glioma in a US case-control study. J Neurooncol 104: 535–542.

41. King MC, Marks JH, Mandell JB (2003) Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. Science 302: 643–646.

42. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753.

43. Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: A Tool for Genome-wide Complex Trait Analysis. The American Journal of Human Genetics 88: 76–82.

44. Kimura M (1983) Rare variant alleles in the light of the neutral theory. Mol Biol Ecol 1: 84–93.

45. Kolonel LN, Henderson BE, Hankin JH, Nomura AMY, Wilkens LR, et al. (2000) A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. American Journal of Epidemiology 151: 346–357.

46. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575.

47. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis correctly for stratification in genome-wide association studies. Nat Genet 38: 904–909.

48. Hauck W, Donner A (1977) Wald's Test as Applied to Hypotheses in Logit Analysis. JASA 72: 831–853.