Published in final edited form as:

Prostate. 2014 May ; 74(6): 579–589.

A Comprehensive Resequencing-Analysis of 250kb Region of 8q24.21 in Men of African Ancestry

Charles C. Chung^{1,2}, Ann W. Hsing¹, Edward Yeboah³, Richard Biritwum⁴, Yao Tettey⁵, Andrew Adjei⁵, Michael B. Cook¹, Angelo De Marzo⁶, George Netto⁶, Evelyn Tay^{7,8}, Joseph F. Boland^{1,2}, Meredith Yeager^{1,2}, and Stephen J. Chanock¹

¹Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland, USA

²Cancer Genomics Research Laboratory, SAIC-Frederick Inc., Frederick National Laboratory for Cancer Research, Frederick, Maryland, USA

³Department of Urology, University of Ghana, Ghana

⁴Department of Epidemiology, University of Ghana, Ghana

⁵Department of Pathology, University of Ghana, Ghana

⁶Department of Pathology, Johns Hopkins University, Baltimore, Maryland, USA

⁷Korle Bu Teaching Hospital, PO BOX 77, Accra, Ghana

⁸University of Ghana Medical School, PO Box 4236, Accra, Ghana

Summary

Background—Genome-wide association studies (GWAS) have identified that a ~1M region centromeric to the *MYC* oncogene on chromosome 8q24.21 harbors at least 5 independent loci associated with prostate cancer risk and additional loci associated with cancers of breast, colon, bladder, and chronic lymphocytic leukemia (CLL). Because GWAS identify genetic markers that may be indirectly associated with disease, fine-mapping based on sequence analysis provides important insights into patterns of linkage disequilibrium (LD) and is critical in defining the optimal variants to nominate for biological follow-up.

Methods—To catalog variation in individuals of African ancestry, we resequenced a region (250kb; chr8:128,050,768-128,300,801, hg19) containing several prostate cancer susceptibility loci as well as a locus associated with CLL. Our samples set included 78 individuals from Ghana and 47 of African-Americans from Johns Hopkins University.

Results—After quality control metrics were applied to next-generation sequence data, 1,838 SNPs were identified. Of these, 285 were novel and not yet reported in any public database. Using genotypes derived from sequencing, we refined the LD and recombination hotspots within the

Correspondence should be addressed to: Stephen Chanock, M.D. Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Advanced Technology Center- NCI, 8717 Grovemont Circle, Bethesda, MD 20892-4605, chanocks@mail.nih.gov, Tel: 301-435-7559, Fax: 301-402-3134.

Disclosure Statement: None of the authors listed have any significant or perceived conflicts of interest relating to the publishing of this manuscript.

region and determined a set of tag SNPs to be used in future fine-mapping studies. Based on LD, we annotated putative risk loci and their surrogates using ENCODE data, which should help guide laboratory studies.

Conclusions—In comparison to the 1000 Genome Project data, we have identified additional variants that could be important in establishing priorities for future functional work designed to explain the biological basis of associations between SNPs and both prostate cancer and chronic lymphocytic leukemia.

Introduction

Prostate cancer (PrCa) is the most commonly diagnosed cancer in men in developed countries, with higher incidence rates in U.S.-based individuals of African ancestry than those of Europeans [1, 2]. Genome-wide association studies (GWAS) have been successful in identifying more than 75 PrCa susceptibility loci [3, 4], most of which were observed in studies of European ancestry, though more recently other ethnic populations have been assessed (e.g., African-Americans and Asians) [5, 6]. Multiple GWAS have identified a region on chromosome 8q24.21 that harbors numerous cancer susceptibility loci including prostate [5, 7-13], chronic lymphocytic leukemia (CLL) [14], breast [15], colon [16, 17], bladder [18], and ovarian [19].

For PrCa, there are at least 5 independent risk loci across approximately 1Mb of 8q24.21 that are defined by recombination hotspots and linkage disequilibrium (LD) within the region (see [20] for review). “Prostate cancer region 2” (chr8:128.073Mb-128.236Mb, hg19, see [21]), includes 8 markers reported to be associated with PrCa risk in individuals of European ancestry, non-European ancestry, or both [5, 7-10, 12, 13]. The region also contains a SNP associated with CLL in Europeans [14]. The associated regions lack known protein-coding genes, however; *MYC* (v-myc myelocytomatosis viral oncogene homolog (avian)) is ~300kb telomeric to the multi-cancer association signals and there is growing evidence that loci within the GWAS-identified regions may affect its function or expression [22-27].

Because SNP-disease associations identified in GWAS may not necessarily indicate a causal association with the identified SNP, focused work is necessary to catalog a comprehensive set of variants for fine-mapping within the associated regions. Next-generation sequencing (NGS) technology has made it possible to interrogate the genome in depth, discovering rare and novel variants to expand the list of variants to further investigate possible contribution to disease. Recent studies have successfully identified rare variants that confer greater risks for PrCa using NGS [28, 29], including one within 8q24.21 that confers risk (OR = 2.90) in men of European descent [29]. Since the frequencies of variants often vary according to ancestry, it is important to understand linkage disequilibrium patterns in different populations.

In the present study, we targeted a 250kb region flanking 8q24.21 PrCa region 2 [21] to perform deep resequencing using NGS in 125 individuals of African ancestry in order to catalog single nucleotide variants and analyze the genetic architecture of the region. We then performed an *in silico* bioinformatic analysis using the extensive functional annotations that

are available from the ENCODE/GENCODE projects [20, 30], further facilitating the understanding of possible biological basis of risk for each locus by evaluating regions of transcription, transcription factor binding sites, histone modification, and chromatin structure.

Materials and Methods

Samples

We used DNAs from 125 individuals, all of African ancestry, of which 78 individuals were from Ghana (39 prostate cancer cases and 39 controls) and 47 were self-reported African-American cancer-free controls that were collected at Johns Hopkins University (The Baltimore Prostate Health Study, Baltimore, MD).

Region selection

A 250 kb region of chromosome 8q24.21, spanning 128,050,768-128,300,801 (UCSC genome build hg19) was selected for next generation sequence analysis based on observed patterns of LD and recombination hotspots inclusive of prostate cancer region 2. The boundaries of this region within 8q24.21 include 8 reported prostate cancer risk loci [5, 7-10, 12, 13] and a chronic lymphocytic leukemia risk locus [14]. There are no RefSeq genes reported within this region.

Primer design, PCR and sequencing

A Nimblegen capture probe pool was designed to cover the 250 kb targeted region. The capture probes were approximately 60 bp in size and the probe pool was designed for amplicon overlap (100 bp on average). Primers were designed using Nimblegen Proprietary Capture probe design software followed by *in silico* quality assessment for uniqueness, possible sequence paralogy and DNA repeat sequences using BLAT from the UCSC Genome Browser (<http://genome.ucsc.edu/cgi-bin/hgBlat>). To check for primer secondary structure and a PCR efficiency check, NetPrimer (<http://www.premierbiosoft.com/netprimer/index.html>) was used. Primers were ordered from Integrated DNA Technologies (Coralville, IA, USA; <http://www.idtdan.com>). The BED file is available upon request for this region. After performing the Nimblegen solution-based sequence capture method, sequence analysis was performed on the 454 Genome Sequencer FLX system (<http://www.454.com/products-solutions/product-list.asp>).

Detection of Genetic Variation

An in-house automated computational pipeline was developed to process sequence reads generated by 454 FLX Genome Sequencers. Sequence reads from the same sample were pooled based on barcodes provided by Roche/454. A quality check (QC) was performed using vendor-supplied software; sequence reads that passed QC were aligned to the whole genome by Newbler (<http://my454.com/products/analysis-software/index.asp>). The resulting assembly was further realigned to homogenize the positional distribution of potential insertion/deletions (indels) using a utility bamleftalign (<https://github.com/ekg/freebayes>). SNPs were called using two software packages – The Genome Analysis Toolkit (GATK) [31] and FreeBayes, a Bayesian genetic variant detector (<http://bioinformatics.bc.edu/>

[marthlab/FreeBayes](#)) excluding alignments from analysis if they have a mapping quality less than 30, also excluding alleles with base quality less than 20. Indels were not called in this study. We developed a reference-guided filtration strategy to obtain analysis-ready SNPs. Potential SNP loci unanimously identified by FreeBayes, GATK, (Phred-scaled quality value > 30) and Newbler *HCDiffs* were compared with the HapMap (release 28), dbSNP (build 137), the 1000 Genomes pilot data (release V3, 2011), the Illumina Omni2.5 scanned 1000 Genomes data from GATK-bundle_hg19_1.2 (<ftp://gsapubftp-anonymous@ftp.broadinstitute.org>), and the Complete Genomics Public Genome Data [32] to identify novel SNPs and to filter loci with discordant allele contents. Only the unanimous genotype calls by FreeBayes, GATK, and Newbler *HCDiffs*, but not referenced by public datasets were assigned novelty to ensure high quality genotype calls.

Descriptive statistics and data quality control

Genotype completion, duplicate check, minor allele frequency (MAF) estimations, deviations from fitness for Hardy-Weinberg proportion (HWP), and genotype concordance check were performed using the GLU software package (<http://code.google.com/p/glu-genetics/>).

LD, recombination hotspots, and Tag SNP Analysis

We assessed the LD patterns among reported PrCa associated loci with SNPs detected within the targeted region and recombination hotspots using SNPs with MAF ≥ 0.05 threshold ($n=785$ Ghanaian SNPs; $n=812$ African-American SNPs), first by estimating phase and calculating background recombination rates using PHASE v2.1 software [33], and then by using SequenceLDhot [34], which uses an approximate marginal likelihood method and calculates likelihood ratio statistics at a set of possible hotspots. Tag SNP analysis was performed for each population ($r^2 \geq 0.8$, MAF ≥ 0.05 with minimum genotype call rate 0.25) using the TagZilla program implemented in the GLU software package. (<http://code.google.com/p/glu-genetics/>)

In silico genomic analysis

Putative functional elements within the resequenced region were assessed using the publically available UCSC genome browser (<http://genome.ucsc.edu/>). Genomic annotation for all sequence-identified SNPs was conducted using ENCODE tools – HaploReg [35] and RegulomeDB [36]. Queries for all SNPs were made using RegulomeDB and further queried for SNPs with $r^2 \geq 0.6$ with the previously reported PrCa SNPs in either or both in Ghanaian or African-American and we cross-examined predicted regulatory DNA elements such as regions of DNase hypersensitivity, binding sites of transcription factors, and promoter regions that have been biochemically characterized by the ENCODE project.

Results

Coverage and Depth

Sequence coverage and median depth over all samples in the targeted genomic region (chr8:128,050,768-128,300,801, hg19) are shown in Figure 1. There were five intervals cumulatively comprising about 6kb with no coverage out of the 250kb targeted region

(2.4%) due to design issues or amplicon failures. The average read depth per locus per sample was approximately 17× (range from 0 to 68.5×, median 16.3×).

Single Nucleotide Polymorphism Detection and Quality Control

Genotypes were called for 5,327 and 9,414 segregating sites in 78 African samples from Ghana by GATK and FreeBayes, respectively, and 10,352 and 6,634 segregating sites in 47 African-American samples from the Johns Hopkins University by GATK and FreeBayes, respectively. The Newbler *HCDiffs* reported 2,402 segregating sites from all 125 samples. Intersecting genotype calls among all three callers (n=1,838) were merged and compared to publicly available datasets (Table 1). Supplementary Table 1 contains detailed information for the variants in the overlapping set. The average per-locus genotype call rate for 1,838 SNPs was 0.975 (Supplementary Table 1). Of the final called genotype set, 280 and 190 SNPs were exclusively detected within the Ghanaian and African-American populations, respectively; 1,368 SNPs were detected in both populations (Supplementary Figure 1).

Within the targeted region, we compared publicly-available variants described by HapMap release 28 and the Illumina Omni2.5 data on the 1000 Genomes Project samples provided by the GATK bundle report. From HapMap release 28, there were 370 SNPs that overlapped our targeted region; we detected 329 (88.9%) of these variants as polymorphic in our set of 125 individuals (Supplementary Table 2). Likewise, from the Illumina Omni2.5 1000 Genomes data, 289 SNPs overlapped the targeted region and we detected 248 (85.8%) of them in our dataset (Supplementary Table 2). Therefore, the majority of the variants detected within our resequencing experiment have been validated elsewhere.

A total of 285 SNPs had not been previously reported and 13 SNPs were only reported in dbSNP b137 (Table 1 and Supplementary Table 1). Of the 285 novel SNPs found in this study, many were rare (MAF = 0.004, median = 0.004, minimum = 0.004, maximum = 0.293) though 30 had MAF > 0.02 in the individuals tested. Within the set of novel SNPs, 85 were observed only in the African-American samples, 135 were observed only in the Ghanaian samples, and 77 were observed in both sets.

Since 48 of the Ghanaian samples were previously scanned for a prostate cancer GWAS using the Illumina 5M chip, we could evaluate genotype concordance by comparing genotypes derived from the GWAS chip data versus genotypes derived from the resequencing experiment. There were 256 overlapping loci available for testing, with an overall genotype concordance rate of 0.989.

LD, recombination hotspots, and Tag SNP Analysis

To date, 9 cancer associated SNPs have been reported within our region of interest in studies of Europeans or non-Europeans [5, 7-10, 12, 13] (Table 2). Our resequencing experiment detected all 8 PrCa associated loci and rs13252298 was monomorphic in the Ghanaian samples. The CLL-associated marker, rs2456449, was detected but was excluded in subsequent analyses due to low-coverage and a low genotype call rate. We assessed the LD patterns among reported PrCa associated loci with SNPs detected within the targeted region and recombination hotspots (Figure 2A and 2B, Supplementary Table 3). Three inferred

recombination hotspots (chr8: 128.069Mb-128.071Mb, 128.103Mb-128.105Mb, and 128.197Mb-128.199Mb) divide the known PrCa associated loci and their surrogates into two regions (~33kb centromeric region, ~94kb telomeric region) in both the Ghanaian (Figure 2A) and African-American (Figure 2B) samples. The centromeric-most region includes rs1016343 and rs13252298, while rs6983561, rs16901966, rs16901979 and rs6987409 are all located within the larger telomeric region. rs1456315 and rs13254738 are located in between within an inferred recombination hotspot (128.103Mb-128.105Mb) with no high-LD surrogates at $r^2 \geq 0.8$.

In order to determine high-LD surrogates for each of the PrCa-associated markers within the region of interest, we estimated pairwise LD among all loci in the dataset. For the 2 SNPs in the centromeric-most region (rs1016343 and rs13252298), very few or no high-LD surrogates at $r^2 \geq 0.6$ were observed (red highlighted, Supplementary Table 3). In contrast, a number of high LD surrogates for the telomeric region associated loci (rs6983561, rs16901966, rs16901979, and rs6987409) were observed (Supplementary Table 3). Additionally, rs6983561 and rs16901979 are highly correlated to each other in both populations (Ghanaian, $r^2=0.846$, African-American, $r^2=0.651$).

We then calculated the number of tag SNPs that would be required to adequately cover all observed variants in a fine-mapping genotyping experiment. In the Ghanaian population, based on loci with MAF ≥ 0.05 (n=785) at $r^2 \geq 0.8$, 340 tag SNPs are required to cover the region at 100% (Supplementary Table 4). In the African-American population, using the same criteria 416 SNPs are required to achieve 100% coverage for loci with MAF ≥ 0.05 (n=812) (Supplementary Table 5).

In silico Genomic Annotation

Within our resequenced target region, there are 3 long non-coding RNAs (lncRNA), RP11-255B23.2, RP11-255B23.3, and RP11-255B23.4 predicted by GENCODE [37, 38], all transcribed on the negative strand (Figure 2). While RP11-255B23.2 (8q24.21:128,084,939-128,094,466, 9.5kb) is located within the ~33kb centromeric region flanked by the predicted recombination hotspots, RP11-255B23.4 (8q24.21:128,197,880-128,215,467, 17.6kb) is located within a cluster of predicted recombination hotspots and RP11-255B23.3 (8q24.21:128,220,111-128,231,333, 11.2kb) is telomeric to the cluster of hotspots, isolating the latter two lncRNA from the ~94kb telomeric region.

SNPs detected from resequencing were queried for their overlap with potential regulatory elements using RegulomeDB [36] and HaploReg [35]. These databases provide integrated information per locus by querying multiple high-throughput ENCODE results from technologies such as ChIP-seq, DNaseI footprint/sensitivity, eQTL, position-weight matrix (PWM) for transcription factor (TF) binding and evolutionary sequence conservation [30]. RegulomeDB provided scores based on the integration of multiple high-throughput datasets for 1,239 SNP loci (no data available for 599 SNP loci, score 7).

We focused on RegulomeDB annotations of 8 previously-reported PrCa associated SNP loci and their surrogates $r^2 \geq 0.15$ (provided in Table 3). SNPs that scored 3 or less by this

database (20 novel and 71 known SNPs) show the highest evidence of transcription factor binding and motif change and/or existence of other regulatory elements. Of the $n=112$ SNPs with $r^2 \geq 0.6$ with the previously-reported SNPs (Supplementary Figure 2 and Supplementary Table 6), 4 of 8 PrCa associated SNPs (or their surrogates) showed evidence of transcription factor binding (rs1016343, rs13254738, rs1456315, and rs6987409). rs1016343 is located within RP11-255B23.2 (Figure 2), yet it is predicted to be ‘less likely to affect binding’ with evidence of PRDM1, FOS, STAT3 bindings to the locus in HeLa-S3, HUVEC, and MCF10A-Er-Src. Although it was not found to be specific to prostate cell lines, there was evidence for enhancer by histone marks and DNaseI sensitivity in multiple cell lines, indicating presence of potential regulatory element overlapping this locus. rs13254738, located 9.9kb 5’ of RP11-255B23.2, was predicted to be ‘likely to affect protein binding’ with evidence of androgen receptor (AR) binding to the locus in VCaP prostate cancer cells [39], yet mismatched binding motif for MEF-2 was found on the locus in LNCaP prostate cancer cell line [40, 41], making rs13254738 a plausible potential regulatory SNP in prostate cancer cell lines, but further functional study is necessary. Although minimal, rs1456315 and rs6987409 were also showed evidence of binding for AR and an ETS transcription factor expressed in prostate epithelial cells, SPDEF, in VCaP prostate cancer cell line, respectively. Sequences overlapping rs6983561 and rs16901979 showed predicted binding motifs for GR and GATA, respectively. No PrCa-specific regulatory elements were reported for rs16901966 and rs13252298, yet sequence overlapping rs16901966 showed evidence of evolutionary conservation (Supplementary Table 6). Among highly correlated surrogates ($r^2 \geq 0.6$, $n=112$) for known PrCa associated loci, 2 SNP loci (rs16901996 and rs10453084) showed evidence for transcription factor binding with altered motif and DNase peak (RegulomeDB score 3a). Sequence overlapping rs16901996 ($r^2 = 0.819$ with previously reported PrCa locus rs6987409 in Ghanaians and 0.758 in African-Americans) showed ChIP-seq evidence of AR binding in VCaP prostate cancer cell line [39] while sequence overlapping rs10453084 ($r^2 = 0.948$ with previously reported PrCa locus rs16901966 in Ghanaians and 0.949 in African-Americans) showed RFX3 binding in K562 cell lines.

Discussion

Using a custom Nimblegen capture and the NGS Roche/454 sequencing technology, we resequenced 125 individuals of African ancestry (78 Ghanaians and 47 African Americans) in order to characterize genetic variation across a 250kb region of 8q24.21 (chr8:128.073Mb-128.236Mb, hg19) that has been reported to harbor susceptibility alleles for PrCa and CLL and is sometimes referred to as “PrCa region 2 [21].” After quality filtering and the establishment of a consensus of variants called from 3 methods, a total of 1,838 SNPs were subjected to further analysis. Though the majority of variants had been previously characterized by the HapMap and/or the 1000 Genomes Project, we identified 298 SNPs that were had not been identified ($n=285$) or only in dbSNP ($n=13$), the majority of which are rare (Table 1). It is possible that a proportion of these SNPs are false-positives, though we observed 77 in individuals from both sample sets so many probably represent true variants, and it is possible that the novel population-specific SNPs have not yet been observed due to the relatively small number of individuals of African ancestry across

intergenic regions studied to date. These novel variants thus are worth replication experiments by other platform to further confirm.

Using the genotype data derived from this study, we then determined the genetic architecture of the 250 kb region with respect to patterns of LD and recombination hotspots, with 3 hotspots effectively dividing the region into 2 sections (Figures 2A and 2B, Supplementary Figure 2). Additionally, we estimated the number of tag SNPs that would be required to provide adequate cover in both the Ghanaian population (n=340) and the African-American (n=416) population at an MAF $\geq$ 5% and an $r^2 \geq$ 0.80.

It is possible that there are varying degrees of African ancestry present within the African-American individuals studied, as admixture in individuals who self-identify as African-American is highly variable [42-44]. Since this regional sequencing data is not sufficient to determine ancestry genome-wide, we do not have the power to determine ancestral proportions for these individuals. Since we do not anticipate a bias in our sample selection, our results will be useful in future fine mapping studies of this region of 8q24 in African and African-American men as our African-American results should not greatly deviate from the admixture observed in other African-American studies.

The results of our study, namely a set of tag SNPs, will be invaluable for further genotyping and fine-mapping studies in other studies of African ancestry. For each of the PrCa loci and their surrogates, we evaluated putative function with respect to the ENCODE data using the RegulomeDB [36] and HaploReg [35] tools, with a number of loci falling within potentially interesting regions that could be explored functionally in laboratory studies. Taken together the results from our study should serve as a guide for further characterization of the complex region of 8q24.21 associated with prostate cancer risk in individuals of non-European ancestry.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Jemal A, Siegel R, Xu J, Ward E. Cancer statistics, 2010. *CA Cancer J Clin.* 2010; 60:277–300. [PubMed: 20610543]
2. Crawford ED. Epidemiology of prostate cancer. *Urology.* 2003; 62:3–12. [PubMed: 14706503]
3. Eeles RA, Olama AA, Benlloch S, Saunders EJ, Leongamornlert DA, Tymrakiewicz M, Ghoussaini M, Luccarini C, Dennis J, Jugurnauth-Little S, et al. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat Genet.* 2013; 45:385–391. [PubMed: 23535732]
4. Goh CL, Schumacher FR, Easton D, Muir K, Henderson B, Kote-Jarai Z, Eeles RA. Genetic variants associated with predisposition to prostate cancer and potential clinical implications. *J Intern Med.* 2012; 271:353–365. [PubMed: 22308973]
5. Takata R, Akamatsu S, Kubo M, Takahashi A, Hosono N, Kawaguchi T, Tsunoda T, Inazawa J, Kamatani N, Ogawa O, et al. Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population. *Nat Genet.* 2010; 42:751–754. [PubMed: 20676098]

6. Xu J, Mo Z, Ye D, Wang M, Liu F, Jin G, Xu C, Wang X, Shao Q, Chen Z, et al. Genome-wide association study in Chinese men identifies two new prostate cancer risk loci at 9q31.2 and 19q13.4. *Nat Genet.* 2012; 44:1231–1235. [PubMed: 23023329]
7. Haiman CA, Patterson N, Freedman ML, Myers SR, Pike MC, Waliszewska A, Neubauer J, Tandon A, Schirmer C, McDonald GJ, et al. Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat Genet.* 2007; 39:638–644. [PubMed: 17401364]
8. Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, Rafnar T, Bergthorsson JT, Agnarsson BA, Baker A, et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet.* 2007; 39:631–637. [PubMed: 17401366]
9. Eeles RA, Kote-Jarai Z, Giles GG, Olama AA, Guy M, Jugurnauth SK, Mulholland S, Leongamornlert DA, Edwards SM, Morrison J, et al. Multiple newly identified loci associated with prostate cancer susceptibility. *Nat Genet.* 2008; 40:316–321. [PubMed: 18264097]
10. Al Olama AA, Kote-Jarai Z, Giles GG, Guy M, Morrison J, Severi G, Leongamornlert DA, Tymrakiewicz M, Jhavar S, Saunders E, et al. Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat Genet.* 2009; 41:1058–1060. [PubMed: 19767752]
11. Gudmundsson J, Sulem P, Gudbjartsson DF, Blondal T, Gylfason A, Agnarsson BA, Benediktsdottir KR, Magnusdottir DN, Orlygsdottir G, Jakobsdottir M, et al. Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. *Nat Genet.* 2009; 41:1122–1126. [PubMed: 19767754]
12. Schumacher FR, Berndt SI, Siddiq A, Jacobs KB, Wang Z, Lindstrom S, Stevens VL, Chen C, Mondul AM, Travis RC, et al. Genome-wide association study identifies new prostate cancer susceptibility loci. *Hum Mol Genet.* 2011; 20:3867–3875. [PubMed: 21743057]
13. Haiman CA, Chen GK, Blot WJ, Strom SS, Berndt SI, Kittles RA, Rybicki BA, Isaacs WB, Ingles SA, Stanford JL, et al. Characterizing genetic risk at known prostate cancer susceptibility loci in African Americans. *PLoS Genet.* 2011; 7:e1001387. [PubMed: 21637779]
14. Crowther-Swanepoel D, Broderick P, Di Bernardo MC, Dobbins SE, Torres M, Mansouri M, Ruiz-Ponte C, Enjuanes A, Rosenquist R, Carracedo A, et al. Common variants at 2q37.3, 8q24.21, 15q21.3 and 16q24.1 influence chronic lymphocytic leukemia risk. *Nat Genet.* 2010; 42:132–136. [PubMed: 20062064]
15. Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, Maranian M, Seal S, Ghoussaini M, Hines S, Healey CS, et al. Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat Genet.* 2010; 42:504–507. [PubMed: 20453838]
16. Tomlinson IP, Webb E, Carvajal-Carmona L, Broderick P, Howarth K, Pittman AM, Spain S, Lubbe S, Walther A, Sullivan K, et al. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat Genet.* 2008; 40:623–630. [PubMed: 18372905]
17. Tenesa A, Farrington SM, Prendergast JG, Porteous ME, Walker M, Haq N, Barnetson RA, Theodoratou E, Cetnarskyj R, Cartwright N, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet.* 2008; 40:631–637. [PubMed: 18372901]
18. Kiemeny LA, Sulem P, Besenbacher S, Vermeulen SH, Sigurdsson A, Thorleifsson G, Gudbjartsson DF, Stacey SN, Gudmundsson J, Zanon C, et al. A sequence variant at 4p16.3 confers susceptibility to urinary bladder cancer. *Nat Genet.* 2010; 42:415–419. [PubMed: 20348956]
19. Goode EL, Chenevix-Trench G, Song H, Ramus SJ, Notaridou M, Lawrenson K, Widschwendter M, Vierkant RA, Larson MC, Kjaer SK, et al. A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. *Nat Genet.* 2010; 42:874–879. [PubMed: 20852632]
20. Chung CC, Chanock SJ. Current status of genome-wide association studies in cancer. *Hum Genet.* 2011; 130:59–78. [PubMed: 21678065]
21. Witte JS. Multiple prostate cancer risk variants on 8q24. *Nat Genet.* 2007; 39:579–580. [PubMed: 17460686]

22. Wasserman NF, Aneas I, Nobrega MA. An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer. *Genome Res.* 2010; 20:1191–1197. [PubMed: 20627891]
23. Sur IK, Hallikas O, Vaharautio A, Yan J, Turunen M, Enge M, Taipale M, Karhu A, Aaltonen LA, Taipale J. Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science.* 2012; 338:1360–1363. [PubMed: 23118011]
24. Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T, Bjorklund M, Wei G, Yan J, Niittymäki I, et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet.* 2009; 41:885–890. [PubMed: 19561604]
25. Ahmadiyeh N, Pomerantz MM, Grisanzio C, Herman P, Jia L, Almendro V, He HH, Brown M, Liu XS, Davis M, et al. 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proc Natl Acad Sci U S A.* 2010; 107:9742–9746. [PubMed: 20453196]
26. Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H, Beckwith CA, Chan JA, Hills A, Davis M, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet.* 2009; 41:882–884. [PubMed: 19561607]
27. Wright JB, Brown SJ, Cole MD. Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. *Mol Cell Biol.* 2010; 30:1411–1420. [PubMed: 20065031]
28. Ewing CM, Ray AM, Lange EM, Zuhlke KA, Robbins CM, Tembe WD, Wiley KE, Isaacs SD, Johng D, Wang Y, et al. Germline mutations in HOXB13 and prostate-cancer risk. *N Engl J Med.* 2012; 366:141–149. [PubMed: 22236224]
29. Gudmundsson J, Sulem P, Gudbjartsson DF, Masson G, Agnarsson BA, Benediksdottir KR, Sigurdsson A, Magnusson OT, Gudjonsson SA, Magnúsdóttir DN, et al. A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nat Genet.* 2012; 44:1326–1329. [PubMed: 23104005]
30. Consortium EP, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012; 489:57–74. [PubMed: 22955616]
31. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010; 20:1297–1303. [PubMed: 20644199]
32. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science.* 2010; 327:78–81. [PubMed: 19892942]
33. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet.* 2001; 68:978–989. [PubMed: 11254454]
34. Fearnhead P. SequenceLDhot: detecting recombination hotspots. *Bioinformatics.* 2006; 22:3061–3066. [PubMed: 17060358]
35. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* 2012; 40:D930–934. [PubMed: 22064851]
36. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 2012; 22:1790–1797. [PubMed: 22955989]
37. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zaidis A, Searle S, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 2012; 22:1760–1774. [PubMed: 22955987]
38. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, et al. GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* 2006; 7 Suppl 1:S4 1–9. [PubMed: 16925838]

39. Wei GH, Badis G, Berger MF, Kivioja T, Palin K, Enge M, Bonke M, Jolma A, Varjosalo M, Gehrke AR, et al. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J.* 2010; 29:2147–2160. [PubMed: 20517297]
40. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* 2011; 21:447–455. [PubMed: 21106904]
41. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* 2006; 34:D108–110. [PubMed: 16381825]
42. Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, Froment A, Bodo JM, Wambebe C, Tishkoff SA, Bustamante CD. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci U S A.* 2010; 107:786–791. [PubMed: 20080753]
43. Jin W, Wang S, Wang H, Jin L, Xu S. Exploring population admixture dynamics via empirical and simulated genome-wide distribution of ancestral chromosomal segments. *Am J Hum Genet.* 2012; 91:849–862. [PubMed: 23103229]
44. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo JM, Doumbo O, et al. The genetic structure and history of Africans and African Americans. *Science.* 2009; 324:1035–1044. [PubMed: 19407144]
45. Luna A, Nicodemus KK. snp.plotter: an R-based SNP/haplotype association and linkage disequilibrium plotting package. *Bioinformatics.* 2007; 23:774–776. [PubMed: 17234637]

Figure 1. Sequence coverage and median depth over all samples in the targeted genomic region (chr8:128,050,768-128,300,801, hg19)

Figure 2. Correlations (r^2) among all SNPs with published prostate cancer susceptibility, linkage disequilibrium (LD) heatmaps and estimated recombination hotspots for **A.** Ghanaian and **B.** African-American individuals. Locations of lincRNAs are shown as well as ENCODE predicted transcriptionally active sites (from bottom to top, transcription factor ChIP-seq from ENCODE, digital DNaseI hypersensitivity clusters in 125 cell types from ENCODE, and layered H3K27Ac mark on 7 cell lines from ENCODE). The left x-axis depict correlation by r^2 scale for colored diamonds and the right x-axis depict likelihood ratio statistics for recombination hotspot estimation (blue line graph). LD heatmap was drawn using *snplotter* R program [45].

Table 1
MAF distribution of resequencing-detected SNPs with regard to public data inclusion

Population	HapMap/Illumina Omni2.5 ^a			1000 Genomes/Complete Genomics ^b			dbSNP b137/novel ^c		
	MAF			MAF			MAF		
	Count	Average	(Min-Max), Median	Count	Average	(Min-Max), Median	Count	Average	(Min-Max), Median
African-American	447	0.191	(0.011-0.500), 0.170	948	0.087	(0.011-0.500), 0.044	163	0.041	(0.011-0.422), 0.011
Ghana	436	0.189	(0.006-0.500), 0.139	999	0.082	(0.006-0.481), 0.039	213	0.029	(0.006-0.442), 0.007
Total ^d	454	0.186	(0.004-0.500), 0.137	1086	0.076	(0.004-0.500), 0.036	298	0.022	(0.004-0.385), 0.004

^a HapMap release 28 reported SNPs and/or Illumina Omni2.5 genotyped SNPs on the 1000 Genomes samples available from GATK bundle

^b 1000 Genomes v3 release reported SNPs and/or Complete Genomics Public Dataset reported SNPs

^c SNPs not reported elsewhere but dbSNP b137 or novel

^d Total counts and MAF calculation considered Ghanaian and African American individuals combined

Table 2
Resequencing detection of previously reported cancer associated loci within the targeted region (chr8:128,050,768-128,300,801, hg19)

Cancer risk SNP	Location (hg19)	Associated Cancer ^a	Associated Population ^d	Reference PMIDs ^b	Ghana ^c	African-American ^c
rs1016343	128093297	PrCa	EUR/ASN	18264097, 19767752, 21743057, 20676098	0.253-T	0.250-T
rs13252298	128095156	PrCa	EUR/ASN	21743057, 20676098	0	0.098-G
rs1456315	128103937	PrCa	ASN	20676098	0.429-C	0.468-C
rs13254738	128104343	PrCa	AA/EUR	17401364, 21743057	0.295-A	0.340-A
rs6983561	128106880	PrCa	AA/ASN	17401364, 20676098	0.436-A	0.533-A
rs16901966	128110252	PrCa	ASN	20676098	0.429-G	0.304-G
rs16901979	128124916	PrCa	AA/EUR	17401364, 17401366, 19767754	0.449-C	0.553-C
rs6987409	128150161	PrCa	AA	21637779	0.186-T	0.174-T
rs2456449	128192981	CLL	EUR	20062064	N/A	N/A

^a PrCa, Prostate cancer; CLL, Chronic Lymphocytic Leukemia

^b PMID, PubMed ID

^c Allele frequency and allele

^d EUR, European; ASN, East Asian; AA, African-American