Published in final edited form as:
Prostate. 2011 June 15; 71(9): 955–963. doi:10.1002/pros.21311.

Functional annotation of risk loci identified through genome-wide association studies for prostate cancer

Yizhen Lu[1,2,3,4], **Zheng Zhang**[5], **Hongjie Yu**[3,4], **S. Lily Zheng**[5], **William B. Isaacs**[7], **Jianfeng Xu**[3,4,5,6], and **Jielin Sun**[5,†]

[1]James D. Watson Institute of Genome Sciences, Zhejiang University, Hangzhou 310008, China

[2]T-Life Research Center Fudan University

[3]State Key Laboratory of Genetic Engineering, Shanghai, 220 Handan Road, Shanghai 200433,China

[4]Fudan-VARI Center for Genetic Epidemiology Fudan University, Shanghai, 220 Handan Road, Shanghai 200433,China

[5]Center for Cancer Genomics, Winston-Salem, NC

[6]Van Andel Research Institute, Grand Rapids, MI

[7]Department of Urology, Johns Hopkins Medical Institutions, Baltimore, MD

Abstract

Background—The majority of established prostate cancer risk-associated Single Nucleotide Polymorphisms (SNPs) identified from genome-wide association studies do not fall into protein coding regions. Therefore, the mechanisms by which these SNPs affect prostate cancer risk remain unclear. Here, we used a series of bioinformatic tools and databases to provide possible molecular insights into the actions of risk SNPs.

Methodology/Principal Findings—We performed a comprehensive assessment of the potential functional impact of 33 SNPs that were identified and confirmed as associated with PCa risk in previous studies. For these 33 SNPs and additional SNPs in Linkage Disequilibrium (LD) ($r^2 \geq 0.5$), we first mapped them to genomic functional annotation databases, including the Encyclopedia of DNA Elements (ENCODE), eleven genomic regulatory elements databases defined by the University of California Santa Cruz (UCSC) table browser, and Androgen Receptor (AR) binding sites defined by a ChIP-chip technique. Enrichment analysis was then carried out to assess whether the risk SNP blocks were enriched in the various annotation sets. Risk SNP blocks were significantly enriched over that expected by chance in two annotation sets, including AR binding sites (p=0.003), and FoxA1 binding sites (p=0.05). About one third of the 33 risk SNP blocks are located within AR binding regions.

Conclusions/Significance—The significant enrichment of risk SNPs in AR binding sites may suggest a potential molecular mechanism for these SNPs in prostate cancer initiation, and provide guidance for future functional studies.

Keywords

functional annotation; prostate cancer; bioinformatics; genome-wide association study

[†]**Address for correspondence**: Dr. Jielin Sun, Center for Cancer Genomics, Medical Center Blvd, Winston-Salem, NC 27157, Phone: (336) 713-7500, Fax: (336) 713-7566, jisun@wfubmc.edu.

Introduction

Genome-wide association studies have identified ~33 established prostate cancer (PCa) risk-associated Single Nucleotide Polymorphisms (SNPs) (1–9). SNPs are DNA sequence variations occurring when a single nucleotide in the genome differs between individuals. Although these SNP associations have been consistently replicated in multiple studies, the functional roles of these risk SNPs remain uncharacterized, largely due to their location in DNA regions that do not encode proteins. Additionally, almost half of the SNP loci are located in regions which do not harbor any known genes. For example, multiple SNPs at 8q24 reside within a 1.3Mb gene-desert region (10), with the closest gene, *c-Myc,* located ~300kb downstream to rs1447295. Because the knowledge base for non-coding DNA is generally limited, few studies have been performed to evaluate the functional impact of the risk SNPs on PCa etiology (11–12). Indeed, the ability to understand the functional impact of the risk SNPs will most likely require additional emphasis on potential transcriptional regulatory mechanisms on non-coding DNA sequence.

The Encyclopedia of DNA Elements (ENCODE) project aims to provide a comprehensive catalog of biological information for functionally important elements. These elements include non-protein coding DNA sequences which regulate gene transcription, gene expression, and DNA replication. The ENCODE pilot study rigorously analyzed a small proportion (1%) of the human genome using computational and experimental methods. Results of the pilot study highlighted the complexity of transcriptional regulation and demonstrated the knowledge gap in this area (13). Based on the initial success, National Human Genome Research Institute (NHGRI) expanded the human ENCODE project to the whole genome (www.genome/gov/ENCODE). At this time, a primary focus of ENCODE is on the characterization of binding of transcriptional factors (TF) and chromatin structure, which represent two of the major factors involved in transcriptional regulation, using the ChIP-seq technique. ChIP-seq is a state-of-the-art high-throughput approach that involves chromatin immunoprecipitation and high-throughput sequencing of immunoprecipitated DNA. Compared with other methods of characterizing regulatory elements, a key advantage of ChIP-seq is a systematic and nonbiased approach, which does not depend on previous knowledge of canonical promoter regions and allows for evaluation of binding complexes of transcription factors and regulatory elements in a more natural state (14). The availability of this comprehensive catalog may facilitate an improved understanding of the functional role of risk SNPs located in non-coding regions.

In addition to the ENCODE project, several recent studies have utilized the high-throughput ChIP-chip method to identify genome-wide binding sites for transcription factors, including Androgen receptor (AR), Forkhead box A1 (FoxA1), and Estrogen receptor (ER) (15,16). Similar to ChIP-seq, ChIP-chip is another high-throughput, global approach for mapping transcription factors. ChIP-chip analysis involves isolating target DNA through chromatin immunoprecipitation, followed by analysis on DNA microarrays that tile the human genome (14). AR is a well-known transcription factor that plays an important role in prostate cancer initiation and progression. Risk SNPs located in putative AR binding sites might change the affinity of androgen-AR complex to binding sequences, thus providing a mechanism leading to modification of PCa risk.

In our study, we performed a comprehensive assessment of the potential functional impact of SNPs that were associated with PCa risk by GWAS studies, utilizing ENCODE genomic annotation databases, as well as ~20 annotation databases from the University of California Santa Cruz table browser (UCSC table browser) (http://genome.ucsc.edu/) and TF binding sites defined by previous studies (15,16). Enrichment analysis was then performed to evaluate whether the risk SNPs were over-represented in any of the annotation sets. To our

knowledge, our study is among the first attempts to comprehensively characterize the potential function of risk SNPs using existing bioinformatic databases. These results assist in the interpretation of the molecular mechanisms of the risk SNPs on PCa etiology and provide guidance for future functional studies.

Methods

Define SNPs that are in Linkage Disequilibrium (LD) with established PCa risk-associated SNPs discovered from GWAS

The causative SNP may be the risk SNP itself or the SNPs in LD with them. Therefore, we identified all SNPs in LD ($r^2 \geq 0.5$) with the 33 risk SNPs discovered by GWAS (SNPs that reached a genome-wide significance level with a p value equal or less than 10^{-7} in previous studies (1–9)) based on the CEU genotype data from the HapMap release #27 (Phase II +PhaseIII) (http://hapmap.ncbi.nlm.nih.gov/). We consider each risk SNP and SNPs that are in LD ($r^2 \geq 0.5$) with it as one risk SNP block.

Overlapping the risk SNP blocks with functionally annotated genomic regions

We mapped SNPs in each risk SNP block to the ENCODE genomic annotation databases (release #2), as well as eleven annotation databases from UCSC (http://genome.ucsc.edu/) and transcription factors defined by previous studies. We defined a risk SNP block as located within a given annotated region if the risk SNP itself, or at least one of the SNPs in LD with the risk SNP, mapped to the annotated region.

Assessment of enrichment of the risk SNP blocks in the annotated genomic regions

We counted the number of risk SNP blocks that mapped to each annotated genomic region. Each risk SNP block was counted only once, even if more than one SNP within the same block mapped to the annotated region. A simulation analysis was used to assess the statistical significance of any potential enrichment for risk SNP blocks within annotated genomic regions, under a null hypothesis that none of these blocks were truly associated with PCa risk. We began the simulation analysis by randomly generating 1,000 sets of 33 SNPs (1,000 replicates) from the ~2.5 million SNPs in the genome with minor allele frequency (MAF)>=0.05 (Hapmap Phase II). We then identified all SNPs in LD with the randomly selected 33 SNPs, and performed the same analysis as for the true risk SNPs, including overlapping the SNP blocks with functionally annotated genomic regions and then counting the number of the SNP blocks that mapped to each annotated genomic region. Next, the mean number of risk SNP blocks that mapped to each annotated region was calculated based on the average counts of the 1,000 replicates. Finally, empirical p-values were calculated based on the number of replicates in which the number of counts was equal or larger than the observed number, divided by the total number of replicates. To reduce the concern of multiple testing, we limited the enrichment analysis to annotation sets with 5 or more mapped risk SNP blocks.

Results

Identification of SNPs in LD with PCa risk SNPs

We identified a total of 972 SNPs in LD with the 33 risk SNPs. A list of these SNPs and pair-wise r^2 for each risk SNP is provided in Supplementary Table 1.

Defining the functional annotation databases

We further grouped the genomic annotation databases into six categories (Table 1), majorly based on the potential functionality and techniques used to define the annotation sets: 1)

Yale ENCODE (Yale Transcription Factor Binding Sites (TFBS)) characterizes the binding sites for a series of transcription factors including c-Myc, GATA-2, SIRT6, TCF7L2, STAT1, NK-kB, c-Fox, c-Jun, E2F6, Max and SIRT6; 2) Broad ENCODE (Broad histone) defines genomic regions with chromatic accessibility and histone modifications, including regions that are enriched with histone markers (H3K4m1, H3K4m2, H3K4m3, H3K27ac, and H3K9ac); 3) regulatory elements defined by UCSC table browser (http://genome.ucsc.edu/), which includes 11 genomic regulatory annotation sets; 4) a conserved region annotation set was also retrieved from UCSC phastConsElements28way and phaseConsElements17way table with conservation scores >500 (a conservation score is a measurement of the degree of conservation of a genomic region) ; 5) coding regions and splice sites that include annotation sets for the protein coding regions, and non-protein-coding RNAs (including transfer RNAs, ribosomal RNAs, small nuclear RNAs, and micro (mi) RNAs); 6) annotation sets including AR, ER and FoxA1 binding sites as defined by the ChIP-on-chip technique were obtained from previous studies (14,15).

Mapping of risk SNP blocks to the functional annotation sets

Detailed annotation for each of the 33 risk SNP blocks are shown in Table 2. Only annotation sets with more than one mapped risk SNP block are listed in Table 2. Detailed information about the mapped SNPs that are in LD with the risk SNPs are presented in Supplementary Table 2. Briefly, 10 risk SNP blocks fall into conserved regions. No risk SNP blocks map to coding regions, non-synonymous changes, splice sites, non-protein-coding RNAs, miRNAs, miRNA target regions or methylation sites (data not shown). Based on category #6 genomic annotation databases (defined in preceding paragraph), 11, 4 and 9 risk SNP blocks were found to map to AR, ER and FoxA1 binding sites, respectively.

Enrichment analysis

Risk SNP blocks were significantly enriched in genomic regions containing AR binding sites, with 11 (34.4%) risk SNP blocks mapping to these sites, whereas only 4.0 (12.5%) blocks randomly generated from the genome are located at such sites (p=0.003). Similarly, risk SNP blocks were significantly enriched in regions containing FoxA1 binding sites (7 (21.88%) vs 3.47 (10.84%); p=0.05) (Table 2). Risk SNP blocks were not significantly enriched in any of the other annotation genomic sets that were tested.

Discussion

PCa risk-associated SNPs identified from GWA studies have been consistently replicated and confirmed in a large number of studies (1–9). The clinical utility of predicting an individual's risk for PCa and identification of high risk men for PCa using these SNPs have been extensively discussed and explored (17–19). In contrast, the biological mechanisms by which the risk SNPs affect PCa initiation are poorly understood. In this study, we used bioinformatics tools to provide insight into the potential functional impact of these SNPs. Importantly, risk SNPs were found to be significantly enriched over that expected by chance in two functional annotation sets, consisting of AR binding sites (p=0.003) and FoxA1 binding sites (p=0.05).

AR, a member of the nuclear hormone receptor family, is a well-known transcription factor which plays an important role in prostate cancer initiation, although the precise mechanisms by which androgens promote prostate carcinogenesis remain ill-defined despite years of investigation. It is clear, however, that healthy men receiving preventive drugs that block the conversion of testosterone to dihydrotestosterone, the more potent androgen, experience a significant reduction (~25%) in PCa risk (20,21). Upon androgen binding, the AR-androgen dimer translocates from the cytosol into the cell nucleus. The AR-androgen dimer complex

then binds to specific DNA sequences known as AR binding sites, recruiting coactivators and other factors which direct and regulate target gene expression. In our study, we demonstrated that almost one third of the 33 known risk SNP blocks are located in AR binding regions identified by the ChIP-chip method (15). A total of ~22,000 AR binding regions have been mapped across the prostate cell genome, using the Model-based Analysis of Tiling-arrays (MAT) algorithm (22) and based on a false discovery rate of 15% ($p<10^{-4}$) (15). The average length of the AR binding regions is 911 base pairs (bp), with a range from 299 to 5,554 bp. The significant enrichment of known risk SNP blocks in regions that harbor AR binding regions suggests a molecular mechanism that may explain the associations between the risk SNPs and PCa risk. The statistical significance (P=0.003) remains even after a stringent Bonferroni correction (13 independent tests were performed in enrichment analysis, p=0.05/13=0.0037). Known risk SNPs that are located within the putative AR binding sites may change the binding affinity of the AR-androgen complex for the binding sequence, leading to altered expression of AR target genes that presumably play rate limiting roles in PCa formation.

Risk SNP blocks were also enriched in FoxA1 binding sites with a nominal p value of 0.05. FoxA1 acts as an AR collaborating cofactor and assists nuclear receptor binding in certain genomic regions (23,24). The coupling of FoxA1 to binding sites is required for AR binding to enhancers in multiple AR-targeted genes (15). SNPs that reside in FoxA1 binding sites may affect the binding affinity of FoxA1 protein and lead to increased risk for PCa. However, the importance of risk SNPs and FoxA1 binding need to be interpreted with caution since the enrichment in FoxA1 did not reach statistical significance after Bonferroni correction (P = 0.65 after Bonferroni correction).

One advantage of our study is the use of a comprehensive and unbiased approach to evaluate TF binding regions defined by ChIP-seq and ChIP-chip techniques, which allowed us to evaluate the regulatory elements on a genome-wide level. Regulation of gene expression is a complicated process and current knowledge in this area is very limited. The pilot study of the ENCODE project revealed that only about 25% of the regulatory elements are located near previously identified transcription starting sites (TSS). This suggests the ChIP-on-chip method is able to identify a large number of promoter regions and regulatory elements that were previously unknown and distant from the classical TSS (13). Although we did not observe any significant enrichment of risk SNP blocksin the ENCODE annotation sets, this might be due to tissue-specific patterns of TF and TF regulation. Currently, a variety of cell lines, mainly Hela and GM06990 have been used for the identification of regulatory elements in the ENCODE project. The LNCaP cell line is currently proposed for study by ENCODE consortium in Tier 3 (http://genome.ucsc.edu/ENCODE/cellTypes.html). Prostate cancer tissue-specific TF binding may provide valuable information for evaluating the molecular mechanisms of risk SNPs on PCa risk.

The fact that no risk SNP blocks are located in regions that code proteins may be due to two reasons. First, we only evaluated SNPs that are characterized by the Hapmap project, in which unknown SNPs that are located in protein coding regions are not evaluated. Secondly, the current resources that are commonly used for gene annotation, RefSeq and ENSEMBLE, likely represents an incomplete catalogue of human genes. SNPs may be located within exons that are not yet identified. In fact, about 60% of GENCODE exons (GENCODE is a sub-project of ENCODE, which aimed to provide a reference annotation of all protein-coding genes within the pilot study of the ENCODE project) are not annotated in RefSeq and ENSEMBL (25). This fact indicates that a high number of alternative slice forms with unique exons exist across the genome (25). With the completion of ENCODE in the near future, a richer and more complete annotation of human genes should provide more insight.

We also did not observe significant enrichment of risk SNP blocks in the regulatory annotation sets defined by UCSC table browsers. However, the null results for these annotation sets need to be interpreted with caution. The majority of the regulatory elements annotation sets are defined by computational algorithms, rather than by biological experiments. In addition, the regulatory elements predicted in those annotation sets are not tissue specific.

In summary, our study is among the first to comprehensively evaluate the potential functional impact of risk loci identified for PCa through GWAS studies. The fact that about one third of 33 SNP blocks fall within AR binding regions, and that the risk SNPs were statistically enriched in AR regions, may suggest a potential molecular mechanism by which risk SNPs contribute to PCa initiation. These results also provide a guidance for future functional studies. In addition, the databases used for bioinformatics annotation could also be used to annotate and prioritize variants identified through GWAS and whole-genome sequencing.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Grants: National Cancer Institute (CA129684, CA140262, CA148463 to J.X.) and Department of Defense (W81XWH-09-1-0488 to J.S.)

References

1. Yeager M, Orr N, Hayes RB, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet. 2007; 39:645–649. [PubMed: 17401363]
2. Gudmundsson J, Sulem P, Steinthorsdottir V, et al. Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. Nat Genet. 2007; 39:977–983. [PubMed: 17603485]
3. Duggan D, Zheng SL, Knowlton M, et al. Two genome-wide association studies of aggressive prostate cancer implicate putative prostate tumor suppressor gene DAB2IP. J Natl Cancer Inst. 2007; 99:1836–1844. [PubMed: 18073375]
4. Thomas G, Jacobs KB, Yeager M, et al. Multiple loci identified in a genome-wide association study of prostate cancer. Nat Genet. 2008; 40:310–315. [PubMed: 18264096]
5. Gudmundsson J, Sulem P, Rafnar T, et al. Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. Nat Genet. 2008; 40:281–283. [PubMed: 18264098]
6. Eeles RA, Kote-Jarai Z, Giles GG, et al. Multiple newly identified loci associated with prostate cancer susceptibility. Nat Genet. 2008; 40:316–321. [PubMed: 18264097]
7. Yeager M, Chatterjee N, Ciampa J, et al. Identification of a new prostate cancer susceptibility locus on chromosome 8q24. Nat Genet. 2009; 41:1055–1057. [PubMed: 19767755]
8. Gudmundsson J, Sulem P, Gudbjartsson DF, et al. Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. Nat Genet. 2009; 41:1122–1126. [PubMed: 19767754]
9. Eeles RA, Kote-Jarai Z, Al Olama AA, et al. Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. Nat Genet. 2009; 41:1116–1121. [PubMed: 19767753]
10. Ghoussaini M, Song H, Koessler T, Al Olama AA, Kote-Jarai Z, Driver KE, Pooley KA, Ramus SJ, Kjaer SK, Hogdall E, DiCioccio RA, Whittemore AS, Gayther SA, Giles GG, Guy M, Edwards SM, Morrison J, Donovan JL, Hamdy FC, Dearnaley DP, Ardern-Jones AT, Hall AL, O'Brien LT, Gehr-Swain BN, Wilkinson RA, Brown PM, Hopper JL, Neal DE, Pharoah PD, Ponder BA, Eeles RA, Easton DF, Dunning AM. UK Genetic Prostate Cancer Study

Collaborators/British Association of Urological Surgeons' Section of Oncology; UK ProtecT Study Collaborators. Multiple loci with different cancer specificities within the 8q24 gene desert. J Natl Cancer Inst. 2008 Jul 2; 100(13):962–966. Epub 2008 Jun 24. [PubMed: 18577746]

11. Lou H, Yeager M, Li H, Bosquet JG, Hayes RB, Orr N, Yu K, Hutchinson A, Jacobs KB, Kraft P, Wacholder S, Chatterjee N, Feigelson HS, Thun MJ, Diver WR, Albanes D, Virtamo J, Weinstein S, Ma J, Gaziano JM, Stampfer M, Schumacher FR, Giovannucci E, Cancel-Tassin G, Cussenot O, Valeri A, Andriole GL, Crawford ED, Anderson SK, Tucker M, Hoover RN, Fraumeni JF Jr, Thomas G, Hunter DJ, Dean M, Chanock SJ. Fine mapping and functional analysis of a common variant in MSMB on chromosome 10q11.2 associated with prostate cancer susceptibility. Proc Natl Acad Sci U S A. 2009 May 12; 106(19):7933–7938. Epub 2009 Apr 21. [PubMed: 19383797]
12. Chang BL, Cramer SD, Wiklund F, Isaacs SD, Stevens VL, Sun J, Smith S, Pruett K, Romero LM, Wiley KE, Kim ST, Zhu Y, Zhang Z, Hsu FC, Turner AR, Adolfsson J, Liu W, Kim JW, Duggan D, Carpten J, Zheng SL, Rodriguez C, Isaacs WB, Grönberg H, Xu J. Fine mapping association study and functional analysis implicate a SNP in MSMB at 10q11 as a causal variant for prostate cancer risk. Hum Mol Genet. 2009 Apr 1; 18(7):1368–1375. Epub 2009 Jan 19. [PubMed: 19153072]
13. The ENCODE Project Consortium Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007 Jun 14; 447(7146):799–816. [PubMed: 17571346]
14. Euskirchen GM, Rozowsky JS, Wei CL, Lee WH, Zhang ZD, Hartman S, Emanuelsson O, Stolc V, Weissman S, Gerstein MB, Ruan Y, Snyder M. Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies. Genome Res. 2007 Jun 17.6:898–909. [PubMed: 17568005]
15. Wang Q, Li W, Zhang Y, Yuan X, Xu K, Yu J, et al. Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer. Cell. 2009 Jul 23; 138(2):245–256. [PubMed: 19632176]
16. Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, Eeckhoute J, Brodsky AS, Keeton EK, Fertuck KC, Hall GF, Wang Q, Bekiranov S, Sementchenko V, Fox EA, Silver PA, Gingeras TR, Liu XS*, Brown M*. Genome-wide analysis of estrogen receptor binding sites. Nat. Genet. 2006; 38:1289–1297. [PubMed: 17013392]
17. Xu J, Sun J, Kader AK, Lindström S, Wiklund F, Hsu FC, Johansson JE, Zheng SL, Thomas G, Hayes RB, Kraft P, Hunter DJ, Chanock SJ, Isaacs WB, Grönberg H. Estimation of absolute risk for prostate cancer using genetic markers and family history. Prostate. 2009 Oct 1; 69(14):1565–1572. [PubMed: 19562736]
18. Hsu FC, Sun J, Zhu Y, Kim ST, Jin T, Zhang Z, Wiklund F, Kader AK, Zheng SL, Isaacs W, Grönberg H, Xu J. Comparison of two methods for estimating absolute risk of prostate cancer based on single nucleotide polymorphisms and family history. Cancer Epidemiol Biomarkers Prev. 2010 Apr; 19(4):1083–1088. [PubMed: 20332264]
19. Sun J, Kader AK, et al. Inherited genetic markers discovered to date are able to identify a significant number of men at considerably elevated risk for prostate cancer. Prostate. (In press).
20. Thompson IM, Goodman PJ, Tangen CM, et al. The influence of finasteride on the development of prostate cancer. N Engl J Med. 2003; 349:215–224. [PubMed: 12824459]
21. Andriole GA, Bostwick D, Brawley OW. The influence of dutasteride on the risk of biopsy-detectable prostate cancer: Outcomes of the REduction by DUtasteride of Prostate Cancer Events (REDUCE) study. N Engl J Med. 2010 Apr 1; 362(13):1192–1202. [PubMed: 20357281]
22. Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, Brown M, Liu XS. Model-based analysis of tiling-arrays for ChIP-chip. Proc Natl Acad Sci U S A. 2006 Aug 15; 103(33):12457–12462. [PubMed: 16895995]
23. Carroll JS, Liu XS, Brodsky AS, Li W, Meyer CA, Szary AJ, Eeckhoute J, Shao W, Hestermann EV, Geistlinger TR, Fox EA, Silver PA, Brown M. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. Cell. 2005 Jul 15; 122(1):33–43. [PubMed: 16009131]
24. Wang Q, Li W, Liu XS, Carroll JS, Jänne OA, Keeton EK, Chinnaiyan AM, Pienta KJ, Brown M. A hierarchical network of transcription factors governs androgen receptor-dependent prostate cancer growth. Mol Cell. 2007 Aug 3; 27(3):380–392. [PubMed: 17679089]

25. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R. GENCODE: producing a reference annotation for ENCODE. Genome Biol. 2006; 7 Suppl 1:S4.1–S4.9. Epub 2006 Aug 7. [PubMed: 16925838]
26. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. J.Mol.Biol. 1987 Jul 20; 196(2):261–282. [PubMed: 3656447]
27. Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermüller J, Hertel J, Lindemeyer M, Reiche K, Tanzer A, Ucla C, Wyss C, Antonarakis SE, Denoeud F, Lagarde J, Drenkow J, Kapranov P, Gingeras TR, Guigó R, Snyder M, Gerstein MB, Reymond A, Hofacker IL, Stadler PF. Structured RNAs in the ENCODE selected regions of the human genome. Genome Res. 2007 Jun; 17(6):852–864. [PubMed: 17568003]
28. Down TA, Hubbard TJP. Computational detection and location of transcription start sites in mammalian genomic DNA. Genome Res. 2002 Mar; 12(3):458–461. [PubMed: 11875034]
29. Davuluri RV, Grosse I, Zhang MQ. Computational identification of promoters and first exons in the human genome. Nat Genet. 2001 Dec; 29(4):412–417. Erratum in: Nat Genet 2002 Nov;32(3): 459. [PubMed: 11726928]
30. Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W, van Steensel B. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature. 2008 June 12.453:948–951. [PubMed: 18463634]
31. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, et al. In vivo enhancer analysis of human conserved non-coding sequences. Nature. 2006 Nov 23; 444(7118):499–502. [PubMed: 17086198]
32. Cheng Y, Miura RM, Tian B. Prediction of mRNA polyadenylation sites by support vector machine. Bioinformatics. 2006; 22:2320–2325. [PubMed: 16870936]
33. Benjamin, P Lewis; Christopher, B Burge; David, P Bartel. Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. Cell. 2005; 120:15–20. [PubMed: 15652477]
34. Andrew, Grimson; Kyle, Kai-How Farh; Wendy, K Johnston; Philip, Garrett-Engele; Lee, P Lim; David, P Bartel. MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing Molecular. Cell. 2007; 27:91–105.
35. Robin, C Friedman; Kyle, Kai-How Farh; Christopher, B Burge; David, P Bartel. Most Mammalian mRNAs Are Conserved Targets of MicroRNAs. Genome Research. 2009; 19:92–105. [PubMed: 18955434]
36. Down TA, Hubbard TJ. Computational detection and location of transcription start sites in mammalian genomic DNA. Genome research. 2002; 12(3):458–461. [PubMed: 11875034]

Table 1

Annotation databases will be used for the bioinformatics analysis

Annotation type	Brief explanation	Sources and references
Yale TFBS		
Transcription factors	Transcription factor binding sites as defined by Encode project, including c-Myc, GATA-2, SIRT6, TCF7L2, STAT1, NK-kB, c-Fox, c-Jun, E2F6, Max, SIRT6	Encode project performed by Yale University, UCSC Table browser
Broad Histone		
Histone modifications	Genomic regions with chromatic accessibility and histone modifications, including regions that are enriched with histone marks (H3K4m1, H3K4m2, H3K4m3, H3K27ac, and H3K9ac)	ENCODE project performed by Broad Institute, UCSC Table browser
Regulatory elements defined by UCSC table browser		
CpgIslandExt	Targeted methylation regions (CpG island) predicted by computation algorithms	UCSC Table browser, ref 26
encodeUViennaRnaz	Noncoding RNA region predicted by three computational algorithms (EvoFold, RNAz and AlifoldZ) in ENCODE regions	UCSC Table browser, ref 27
eponine	Transcription start sites (TSS) predicted by a probabilistic method (Eponine), with good specificity and excellent positional accuracy.	UCSC Table browser, ref 28
firstEF	Computationally predicted promoter regions and first exons in the human genome	UCSC Table browser, ref 29
lamB1	Genomics regions interacting with nuclear lamina may contribute to the spatial organization of chromosomes inside the nucleus	UCSC Table browser, ref 30
oregano	The Open REGulatory ANNOtation database (ORegAnno) is an open database for the known regulatory elements from scientific literature (with various biological experimentally supported regulatory regions)	UCSC Table browser
polyaDb	mRNA ploly (A) sites that are mapped by cDNA/EST sequences	UCSC Table browser, ref 31
switchDbTss	Computationally predicted transcription start sites (TSS) based on cDNA alignment	UCSC Table browser, ref 32
targetScanS	TargetScan predicts biological targets of micro (mi) RNAs by searching for the presence of conserved 8mer and 7mer sites that match the seed region of each miRNA	UCSC Table browser, ref 33–35
tfbsConsSites	Location and score of transcription factor binding sites conserved in the human/mouse/rat alignment; data are computed with the Transfac Matrix Database (v7.0) and are purely computational	UCSC Table browser
VistaEnhancers	Defines distant-acting transcriptional enhancers by combining computational approach and a moderate throughput mpise transgenesis enhancer assay	UCSC Table browser, ref 36
Conserved region		
Conserved region	Genomic region that are conserved across different species	UCSC phastConsElements28way and phaseConsElements17way table with conservation score >500
Coding region and splicing sites		
Coding	Genomic regions coding for genes	UCSC Table browser
Non-Synonymous change	Genomic regions where a nucleotide substitution leads to an amino acid change	UCSC Table browser
Splice sites	Functional element that affect the splicing process	UCSC snp129
Non-protein-coding RNAs	Transfer RNAs, ribosomal RNAs, small nuclear RNAs, and miRNAs	UCSC Table browser (rnaGene)
Transcription factor binding sites defined by ChIP-on-chip technique		
AR binding	Androgen Receptor binding regions defined by ChIP-on-chip technology	ref 15

Annotation type	Brief explanation	Sources and references
ER binding	Estrogen Receptor binding regions defined by ChIP-on-chip technology	ref 16
FoxA1 binding	FoxA1 binding regions defined by ChIP-on-chip technology	ref 15

Table 2

n on 33 risk SNP blocks based on various genomic databases

Note	BP*	genes	Yale ENCODE	Broad ENCODE	regulatory elements defined by UCSC annotation						Conserved	trascription factor binding sites defined by previous papers (13,14)		
					cpgIslandExt	encodeUViennaRnaz	firstEF	laminB1	oreganno	tfbsConsSites	Region	AR	ER	FoxA
w 2p21	43,407,453	THADA		H3K4me1				Yes						
2p15	62,985,235	EHBP1		H3K4me2,H3K27ac,H3K9ac,				Yes		Yes	Yes			
2q31.1	173,019,799	ITGA6	c-Fos,Max,TCF7L2,STAT1	H3K4me1,H3K4me2,H3K27ac,H3K4me3				Yes			Yes	Yes		Yes
3p12	87,193,364		c-Fos,STAT1	H3K4me1				Yes			Yes			
3q21.3	129,521,063		STAT1,GATA-2,ZNF263	H3K4me1,H3K4me2				Yes		Yes	Yes		Yes	Yes
4q22.3	95,781,900	PDLIM5		H3K4me1,H3K4me2						Yes	Yes	Yes	Yes	Yes
w 4q24	106,280,983	TET2	JunD,NFKB,STAT1	H3K4me2,H3K4me3				Yes			Yes			
6q25	160,753,654		c-Fos	H3K4me1,H3K4me2,H3K4me3,				Yes						
7p15	27,943,088	JAZF1		H3K4me1,H3K4me2				Yes			Yes	Yes		
7q21	97,654,263	LMTK2	NFKB,Pol2,c-Myc,SIRT6,ZNF263	H3K4me1,H3K4me2,H3K4me3				Yes	Yes		Yes		Yes	Yes
8p21.2	23,494,920	cTBP2		H3K4me1,H3K4me2				Yes			Yes	Yes		
8p21.2	23,582,408	NKX3.1	Pol2,c-Myc,HA-E2F1,TCF7L2,GATA-2,NF-YB,SIRT6,ZNF263	H3K4me1,H3K27ac,H3K4me3,H3K9ac,	Yes		Yes	Yes			Yes	Yes		Yes
w 8q24 (5)	128,081,119			H3K4me1,H3K4me3				Yes			Yes			
q24(2)	128,194,098			H3K4me1,H3K4me2				Yes			Yes	Yes		
New q24.21	128,389,528													
w 8q24 (4)	128,404,855		TCF7L2	H3K4me2				Yes		Yes	Yes	Yes		Yes
q24(3)	128,482,487		TCF7L2,STAT1,Rad21								Yes			Yes
q24(1)	128,554,220											Yes		
9q33	123,467,194							Yes						
10q11	51,219,502	MSMB	ZNF263	H3K4me1,H3K4me2,H3K27ac,H3K9ac				Yes	Yes		Yes	Yes	Yes	Yes

Note	BP*	genes	Yale ENCODE	Broad ENCODE	regulatory elements defined by UCSC annotation						Conserved	trascription factor binding sites defined by previous papers (13,14)		
					cpgIslandExt	encodeUViennaRnaz	firstEF	laminB1	oreganno	tfbsConsSites	Region	AR	ER	FoxA
10q26	126,686,862	CTBP2	Pol2,E2F6,ZNF263	H3K4me1,H3K4me3				Yes			Yes			
New 1P15.5	2,190,150	IGF2, IGF2AS, INS, TH	Pol2	H3K4me1		Yes								
q13(2)	68,691,995	AL137479, BC043531	c-Fos,Max,NF-YA,NF-YB	H3K4me1										
q13(1)	68,751,243		Max,c-Myc	H3K4me1,H3K4me2					Yes					
q12 (2)	33,149,092		ZNF263	H3K4me1,H3K4me3				Yes		Yes	Yes			Yes
q12(1)	33,172,153	TCF2	Pol2,STAT1	H3K4me1				Yes						
7q24.3	66,620,348			H3K4me1,H3K4me2				Yes			Yes	Yes		
New 9q13.2	43,427,453		c-Myc	H3K4me1,H3K9ac				Yes			Yes			
19q13	46,677,464	10 Mb to KLK3	NF-YB	H3K27ac,				Yes						
19q13 KLK3)	56,056,435	KLK3												
y22q13	38,782,065		GATA-2	H3K4me1,H3K4me2				Yes				Yes		
New 2q13.2	41,830,156	TTLL1, BIK, MCAT, PACSIN2	c-Fos,Max,c-Jun,c-Myc,TCF7L2,E2F6,GATA-2,SIRT6	H3K4me1,H3K4me2										
Xp11	51,258,412	NUDT10, NUDT11, LOC340602												

d on NCBI build 36

one annotation sets, transcription factor binding sites and regions that are enriched with specific histone markers were provided for each risk SNP block. For the remaining yes" means the risk SNP block overlapped with the annotation category. Empty cells mean the risk SNP block does not overlap with the annotation category.

Table 3

Enrichment analysis of the 33 risk SNP blocks

Annotation type*	# of counts and frequency (%) in the PRAS blocks	# of counts and frequency in the randomly generated SNP blocks	P-value#
Yale TFBS			
c-Myc	5 (15.63)	3.43 (0.11)	0.26
TCF7L2	5 (15.63)	3.08 (0.1)	0.18
STAT1	6 (18.75)	3.61 (0.11)	0.13
c-Fos	5 (15.63)	3.58 (0.11)	0.28
Broad Histone			
H3K4me1	23 (71.88)	19.88 (0.62)	0.18
H3K4me2	16 (50.00)	14.04 (0.44)	0.32
H3K4me3	8 (25.00)	5.72 (0.18)	0.19
H3K27ac	5 (15.63)	4.29 (0.13)	0.43
Regulatory elements defined by UCSC table browser			
laminB1	24 (75.00)	24.61 (76.91)	0.53
tfbsConsSites	6 (18.75)	8.93 (0.28)	0.82
Conserved region			
Conserved region	10 (31.25)	8.69 (27.16)	0.38
Transcription factor binding sites defined by ChIP-chip technology			
AR binding	11 (34.38)	3.99 (12.47)	**0.003**
Fox A1 binding	7 (21.88)	3.47 (10.84)	**0.05**

*Enrichment analysis was only performed for annotation sets with 5 or more mapped risk SNP blocks

#p-value is based on 1,000 simulation replicates