Identification of Candidate Regulatory SNPs by Integrative Analysis for Prostate Cancer Genome Data

Segun Jung
Northwestern University Feinberg School of Medicine
680 N Lake Shore Drive
Chicago, IL 60611
segun.jung@northwestern.edu

Hongjian Jin
Northwestern University Feinberg School of Medicine
680 N Lake Shore Drive
Chicago, IL 60611
hongjian-jin@northwestern.edu

Ramana V Davuluri
Northwestern University Feinberg School of Medicine
680 N Lake Shore Drive
Chicago, IL 60611
ramana.davuluri@northwestern.edu

ABSTRACT

Genome-wide association studies (GWAS) have identified numerous single nucleotide polymorphisms (SNPs), also known as generic variants, associated with disease susceptibility. Prostate cancer (PCa) is a highly heritable disease. GWAS studies have so far reported more than 70 SNPs that are associated with PCa risk. However, most of these SNPs are located in the noncoding genomic regions that little are known about their functional roles. Here we describe an informatics system that performs an integrative analysis of ChIP-seq, RNA-seq, SNP array and clinical data for identifying candidate regulatory SNPs (rSNPs) that could alter transcription factor (TF) binding sites and neighboring gene regulation. By applying the informatics framework on HOXB13 TF in PCa, we identified 213 rSNPs that include a recently discovered rSNP (rs339331) and identified a novel candidate rSNP (rs1476161) associated with the PCa risk. We confirmed rs1476161 by performing the HOXB13 knockout experiment. The expression level the target gene, AURKB, was decreased by about 2-fold in HOXB13-silencing cells compared to the control cells. This indicates the involvement of HOXB13 in altering AURKB gene expression, suggesting a critical role of rs1476161 in allele-specific gene regulation. Taken together, the results demonstrate the feasibility of our system in searching for candidate rSNPs associated with PCa risk.

Categories and Subject Descriptors

J.3 [**LIFE AND MEDICAL SCIENCES**]: Biology and genetics; G.3 [**PROBABILITY AND STATISTICS**]: Survival analysis

General Terms

Algorithms

Keywords

SNP, TF, ChIP-Seq, RNA-Seq, TCGA, Prostate cancer

BCB '15, September 09-12, 2015, Atlanta, GA, USA © 2015 ACM. ISBN 978-1-4503-3853-0/15/09…$15.00 DOI: http://dx.doi.org/10.1145/2808719.2808748

1. INTRODUCTION

Prostate cancer (PCa) is the second most common cause of cancer mortality among men in the Western countries [1]. This is one of the most heritable diseases, and the hereditary genetic factors contribute significantly to its susceptibility [2]. To date, over 70 single-nucleotide polymorphisms (SNPs) have been identified to be associated with PCa predisposition by genome-wide association studies (GWAS), accounting for ~30% of the familial risk in PCa [3]. However, little is known about the molecular basis involving these susceptibility SNPs.

SNPs, associated with disease, in coding region are largely easy to understand their functional roles. However, deciphering the functional implication of SNPs, located outside of coding regions, remains a challenge. Such disease-associated SNPs in noncoding regions are called regulatory SNPs (rSNPs) because they can alter the binding affinity of transcription factors (TFs) to the DNA sequence in the regulatory region that modulate the gene expression level, leading to disease phenotypes.

Most of the rSNPs are located in intronic and intergenic noncoding genomic regions and have long been studied to decipher their functional roles in gene expression and genome/chromatin organization [4, 5]. To aid in the study of rSNPs, several computational approaches have contributed to the annotation of noncoding variants [6-11]. Recent emerging evidence suggests that these rSNPs are the key players in gene regulation programs by modulating key TFs, which interact with critical transcriptional enhancers [12]. Indeed, several important TFs that affect the risk in PCa were reported that include androgen receptor (AR), GATA binding protein 2 (GATA2), Organic Cation Transporter 1 (Oct1), and Homeobox B13 (HOXB13) [13].

The HOXB13 TF is known to play important roles in increased risk of prostate cancer [14-17]. Genetic evidences suggested that HOXB13 promotes PCa risk with unknown mechanisms [15]. Recently, a genetic variant, rs339331, located in the DNA binding site of HOXB13 was reported to be associated with the PCa susceptibility [18, 19]. Here we describe an integrative informatics system for searching regulatory SNPs and their target genes associated with TFs, specifically focusing on HOXB13 in this study (Figure 1). Owing to high-throughput technologies, such as microarrays and next-generation sequencing, we integrated publicly available ChIP-seq, RNA-seq, SNP array and clinical data for identifying rSNP candidates regulating

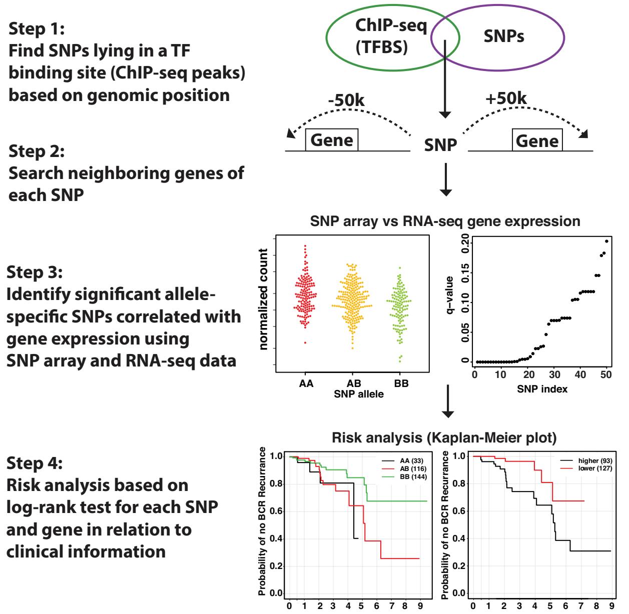
Step 1:
Find SNPs lying in a TF binding site (ChIP-seq peaks) based on genomic position
ChIP-seq (TFBS)
SNPs
-50k
+50k
Gene
SNP
Gene
Step 2:
Search neighboring genes of each SNP
SNP array vs RNA-seq gene expression
Step 3:
Identify significant allele-specific SNPs correlated with gene expression using SNP array and RNA-seq data
normalized count
AA
AB
BB
SNP allele
q-value
0
0.05
0.10
0.15
0.20
0
10
20
30
40
50
SNP index
Risk analysis (Kaplan-Meier plot)
Step 4:
Risk analysis based on log-rank test for each SNP and gene in relation to clinical information
Probability of no BCR Recurrance
1.0
0.8
0.6
0.4
0.2
0.0
AA (33)
AB (116)
BB (144)
higher (93)
lower (127)
0 1 2 3 4 5 6 7 8 9

Figure 1. Computational procedure for identifying risk SNP candidates. This process utilizes ChIP-seq, SNP-array, RNA-seq and clinical data.

neighboring gene expression that may be critical for the increase of PCa risk. Our approach confirmed the recent discovery of a generic variant and identified other rSNP candidates. Finally, we carried out an experimental validation for one of the candidate rSNPs and confirmed its association with altered gene expression.

2. METHODS

2.1 Data sets

We downloaded RNA-seq, SNP array, and clinical data of prostate adenocarcinoma (PRAD) from the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/). These data include 497 tumor and 52 matched normal samples for RNA-seq, 500 samples for SNP array data, and 369 patients' clinical information from which 295 patients are eligible for the risk analysis of biochemical relapse. ChIP-seq data of the HOXB13 transcription factor were downloaded from the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under the accession number PRJEB4865.

The RNA-seq raw counts were prepared by removing genes whose counts were all zero across patient samples. Then, the counts for a gene in each sample were divided by the geometric mean calculated across all samples; the size factors, the median of those ratios from each sample, are used for adjusting differences in the library sizes of sequencing experiments using the DESeq2 R package [20]. Then, the normalized count data were log2-transformed for statistical analysis.

Affymetrix Genome-wide Human SNP array 6.0 data were processed mainly using the corrected robust linear mixture model (crlmm) algorithm for data normalization and genotype calls. Briefly, crlmm estimates the genotype based on a two-stage hierarchical model (M) for the log-ratio of the allele A and B intensities I_A and I_B defined as $M = \log_2 (I_A / I_B)$. The model M is grounded in an empirical Bayes approach, in which the means conditioned on genotype have a multivariate normal distribution while the variances follow an inverse gamma distribution, for the posterior probability estimation of each genotype [21]. The final data are represented as 0 (AA), 1 (AB), and 2 (BB) [22].

ChIP-seq reads were mapped to the human genome (hg19) using Burrows-Wheeler Alignment (BWA) software [23] and peak calling was performed by MACS [24].

Microarray profiling of LNCaP control cell and HOXB13 silencing cells was performed using HumanHT-12 v 4.0 Expression BeadChip (Illumina). Bead-level data were preprocessed using GenomeStudio (Illumina), and the expression values were quantile-normalized using the beadarray R package [25].

Clinical data were processed to extract relevant variables for the risk analysis of biochemical recurrence such as relapsed time and status of biochemical recurrence (corresponding to the column names of "days to psa" and "biochemical recurrence" in the TCGA clinical data file).

2.2 Computational procedure for Identifying regulatory SNP candidates associated with disease

Figure 1 shows the computational procedure for identifying regulatory SNPs involving in disease risk. We describe each of the four steps in details as follows:

Step 1. The bed format files of ChIIP-seq and SNP array data—in total of 36,143 peaks and 905,422 SNPs, respectively—are prepared as an input to the OverlapSelect program from the UCSC Kent source library. This data integration is based on genomic positions that find all the SNPs lying in each of the ChIP-seq peaks.

Step 2. Using the Genome Reference Consortium human genome build 37 (GRCh37) downloaded from the Ensembl website (http://useast.ensembl.org/Homo_sapiens/), we preprocessed to prepare the gene list in a bed format. To search the neighboring genes of the chosen SNPs, we used the bedtools program [26] with the following command: bedtools window –a snp.bed –b gene.bed –w 50000 > output.bed

Note that we only keep the unique pairs of SNP and neighboring genes.

Step 3. For each of the SNP-gene pairs from Step 2, we extract corresponding SNP array and log2-transformed gene expression data from the common patient samples (e.g., 494 samples for PCa). We then applied the Analysis of Variance (ANOVA) test to assess the statistical significance between each SNP genotype and its neighboring gene expression. Although ANOVA can handle small sample sizes, we set an ad hoc minimum allele frequency of 20 for each SNP allele, motivated by the HapMap project targeting SNPs with a minimum minor allele frequency of 5% [27], and paired to log2-transformed expression data.

Step 4. For the association analysis of gene expression and risk of biochemical relapse, we used a collection of 295 RNA-seq tumor samples (202 samples excluded from the original 497 samples due to the absence of clinical data) and 52 RNA-seq normal samples from the TCGA cohort. We transformed the gene expression levels to z-score defined as: z-score= x-μ/σ, where x is the tumor samples, μ is the mean of normal samples, and σ is the standard

Table 1. 16 SNP-gene list from the PCa data analysis reported as eQTL in other studies.

SNP ID	Gene symbol	P-value	Allele A	Allele B	Allele frequency			Mean of normalized count		
					AA	AB	BB	AA	AB	BB
rs2742624	UPK3A	2.90E-46	A	G	63	202	229	2894.1	2220.5	786.8
rs2412106	CHURC1	7.95E-17	A	G	193	212	89	2170.1	2530.2	2768.8
rs1045270	WDYHV1	2.07E-13	A	G	210	218	66	722.6	579.8	514.8
rs3825393	KCTD10	2.51E-11	C	T	248	186	60	3321.6	3801.5	4391.2
rs6799720	PLOD2	1.21E-10	G	T	121	247	126	841.7	1257.3	1427.3
rs11689112	RALB	1.68E-10	A	C	244	202	48	4014.2	3536.1	2920.9
rs185397	GOT2	3.08E-10	A	G	65	182	247	7196.4	9322.1	7366.0
rs4325349	KRT86	4.42E-06	C	G	58	218	218	25.2	18.4	11.5
rs7894521	ECHDC3	2.61E-05	G	T	92	106	296	563.5	844.9	944.4
rs3746337	PYGB	3.45E-05	C	T	169	218	107	20172.3	18992.3	16455.2
rs10100297	MMP16	3.38E-04	C	T	97	211	186	50.2	45.2	35.8
rs3897474	GPR180	1.00E-03	A	G	200	204	90	582.1	554.1	508.8
rs11489585	RSBN1L	1.71E-03	A	G	271	187	36	698.7	778.8	836.3
rs2283119	ASAH1	8.46E-03	G	T	151	194	149	11760.6	12907.2	11621.8
rs3821747	RPL22L1	9.57E-03	A	G	315	150	29	2279.2	2896.0	2747.7
rs847377	AGR3	1.83E-02	C	T	202	231	61	362.3	429.0	487.6

deviation of normal samples. Therefore, the scores above and below zero indicate higher and lower expression, respectively.

To stratify the tumors into two groups, high and low gene expression, we first classified them into positive and negative z-scores. We then ranked the tumors based on the z-score in each group. Tumors with high and low gene expression are defined as those in the highest and lowest 75% of tumors with positive and negative z-scores, respectively. Since z-scores close to zero are uninformative and thus not useful in this analysis, we excluded the corresponding tumor samples accordingly. Lastly, we performed Kaplan-Meier analysis using this stratification.

Similarly, we used a collection of 293 SNP array data whose clinical data exist for the association analysis of each SNP allele and risk of biochemical relapse followed by Kaplan-Meier analysis.

All these steps are automated and the external programs in Step 1 and 2 are embedded using the system function in R. In order to accelerate the statistical computation, we made use of 8 cores by using the two R libraries—foreach and doParallel, taking less than 15 minutes running on MacBook Pro (2.6 GHz intel Core i7 processor; 16 GB 1600 MHz DDR3 memory).

3. RESULTS

3.1 Identification of rSNP candidates and their neighboring genes associated with HOXB13

Majority of the transcription factor binding sites is located in intergenic regions, a stretch of DNA sequences outside of protein coding genes, largely 20 - 25kb far from the neighboring genes [28]. Using the HOXB13 ChIP-seq data, we searched for SNPs lying within the potential DNA binding sites of HOXB13 in whole genome. We found 1,946 SNPs using the OverlapSelect program. We then looked for each SNP's neighboring genes within 100kb (±50kb) that returned 8,168 unique SNP-gene pairs; our rationale for setting the range of ±50kb is to inclusively searching for all the potential local SNP-gene associations. For each SNP-gene pair, we extracted SNP array and log2-transformed RNA-seq data from the 494 TCGA patient samples and compared the differences between three different allele groups (two homozygous and one heterozygous) and the gene expression level by using the ANOVA test. Note that we excluded the SNP-gene pairs from the analysis when the minimum number of samples for each allele is less than 20. In this way, the risk alleles we find in this study are present in at least ~5% of population [27]. We identified 213 SNP-gene pairs that are strongly correlated by applying p-value cutoff of 0.05 in which 102 SNP-gene pairs still remained after applying q-value (FDR) cutoff of 0.1 for multiplicity (Figure 2). Notably, our results include the recent discovery of the generic variant (rs339331) and its targeted gene (RFX6) with p-value of 3.24E-02 [18].

Recent studies have reported that expression quantitative trait loci (eQTL) SNPs may be associated with disease through gene regulation [29, 30]. Thus, we searched the 213 SNP-gene pairs for finding eQTL signatures from the Genotype-Tissue Expression (GTEx) portal (www.gtexportal.org/home/). Table 1 shows the 16 SNP-gene list reported as eQTL, and several of these genes are known to play critical roles in human cancers; for instance, UPK3A for bladder cancer [31], MMP16 for melanoma [32] and lung cancer [33], AGR3 for ovarian cancer [34], and RalB in multiple cancers [35]. This implicates that the SNP-genes pairs including eQTL SNPs may serve as potential biomarkers in PCa.

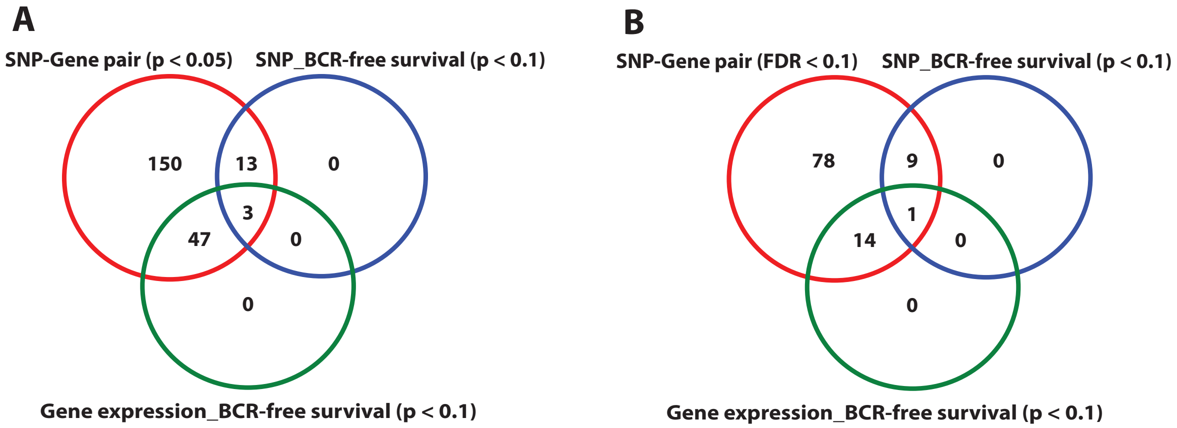
A
SNP-Gene pair (p < 0.05)
SNP_BCR-free survival (p < 0.1)
150
13
0
3
47
0
0
Gene expression_BCR-free survival (p < 0.1)
B
SNP-Gene pair (FDR < 0.1)
SNP_BCR-free survival (p < 0.1)
78
9
0
1
14
0
0
Gene expression_BCR-free survival (p < 0.1)

Figure 2. Association between SNP, gene expression and risk of prostate cancer. (A) 213 SNPs are significantly associated with their nearby genes (p-value <0.05; ANOVA) in which 16 SNPs and 50 genes are correlated with biochemical recurrence (BCR) (p-value < 0.1; log-rank test) and 3 are in common. (B) For multiple comparisons, we used q-value cutoff of 0.1 for the association of SNP and its neighboring gene pair that returned 102 SNPs from which 10 SNPs and 15 genes are correlated to biochemical recurrence, and found 1 in common.

3.2 rSNP candidates within DNA sequence matched to the HOXB13 binding motif

Among the set of SNP-gene pairs, we found that three SNPs are located in the canonical HOXB13 binding motif (Table 2 and Figure 3). However, we noticed that the statistical significance for rs447003_KRT6A and rs4796539_MED31 is mainly due to the expression difference between homozygous and heterozygous, but not between reference (AA) and alternative (BB) allele. Only the rs339331_RFX6 pair is differentially expressed between reference (AA) and alternative (BB) allele; the reference allele T serves as part of the HOXB13 binding motif whereas its alternative allele disrupts HOXB13 DNA binding that decreases the RFX6 expression level by ~5-fold. Indeed, Huang *et al* recently demonstrated that the rs339331 SNP involves in prostate cancer risk by altering the RFX6 gene expression through an interaction with HOXB13 [18].

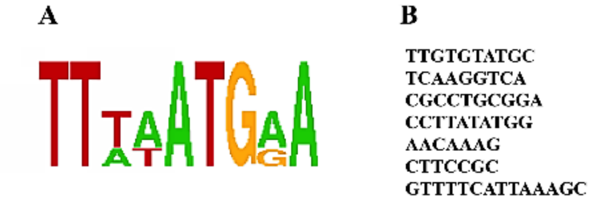

Figure 3. HOXB13 sequence logo (A) and the list of DNA binding motifs (B).

3.3 Association between SNP, gene and prostate cancer risk

To evaluate whether a single generic variant is closely linked to PCa risk by affecting gene expression programs, we performed a statistical analysis to find SNP-gene pair associated with the risk of biochemical relapse based on the log-rank test. We observed that 16 SNPs and 50 genes are highly correlated with the PCa risk from which three SNP-gene pairs are in common (Figure 2 and Table 3). Although not much is known about the three candidates in relation to HOXB13, we found through a literature search that the AURKB gene has been actively investigated in the PCa community to decipher its functional role [36-38]. In fact, AURKB is proven to be associated with other diseases such as glioblastoma [39]. Thus, we investigated further the gene by comparing allele-specific expression level of two homozygous alleles. Indeed, Figure 4A shows that significant expression differences were observed between reference and alternative allele (p= 2.937E-03, Student's t-test) whereas insignificant expression differences between homozygous and heterogynous alleles were observed. Next, we evaluated whether AURKB is associated with the clinical variables that would indicate PCa progression. Using Kaplan-Meier with the log-rank test, we found that rs1476161_AURKB are highly correlated with the risk biochemical recurrence (Figure 4B-C). Figure 4B shows the reference allele A involved with a notable increase in the frequency of biochemical relapse than the alternative allele G

Table 2. **SNP candidates and the neighboring target genes whose sequence contains the canonical HOXB13 DNA-binding motif.**

SNP ID	Gene symbol	Gene name	P-value	Allele A	Allele B	Allele frequency			Mean of normalized count		
						AA	AB	BB	AA	AB	BB
rs447003	KRT6A	Keratin 6A	4.51E-03	C	T	60	235	199	90.4	143.9	102.0
rs4796539	MED31	Mediator Complex Subunit 31	1.04E-02	A	G	89	206	199	289.8	311.1	294.5
rs339331	RFX6	Regulatory Factor X, 6	3.24E-02	T	C	263	186	45	116.6	69.6	22.6

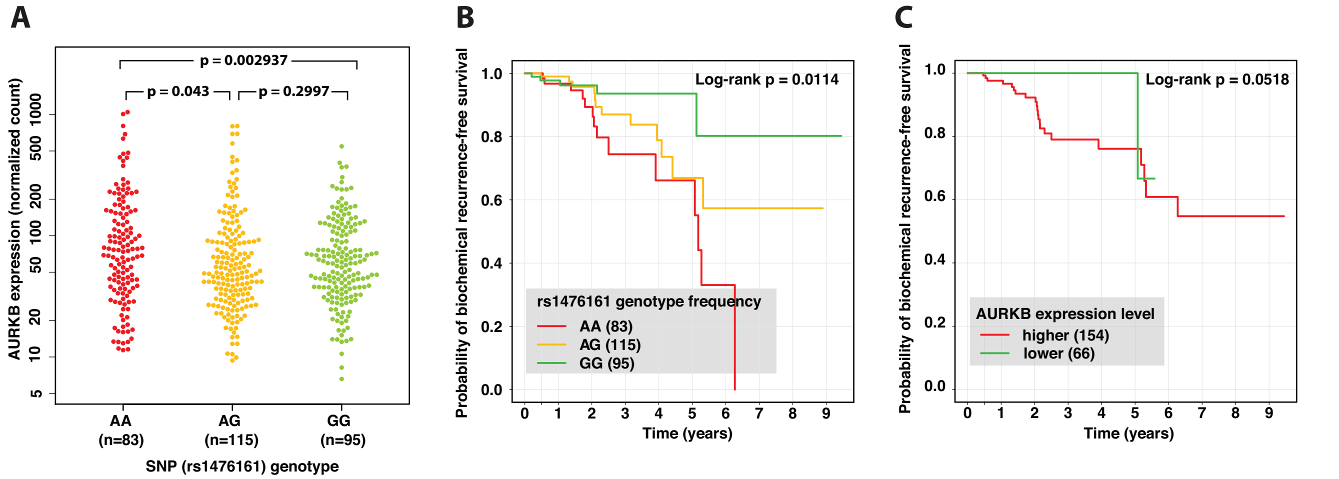

Figure 4. Association between rs1476161 genotype, AURKB expression and PCa risk. (A) Allele-specific gene expression. The overall distribution of normalized AURKB read counts for the AA (83 samples) genotype is higher than the AG (115 samples) and GG (95 samples) genotypes. In particular, significant difference between AA and GG was observed with p-value of 2.937E-03 (Student's t-test). (B) Kaplan-Meier plot for evaluating the risk of biochemical recurrence with respect to each genotype. GG genotype is the risk allele for the PCa progression whereas AA genotype depicts the PCa progression-free. (C) Kaplan-Meier plot for analyzing the risk of biochemical recurrence based on AURKB expression. Higher (top 75% of AURKB-upregulated samples) and lower expression (bottom 75% of AURKB–-downregulated expression) illustrates that the high PCa risk is with lower AURKB expression level whereas gradual decrease of PCa risk is with higher AURKB expression.

(P=1.14E-02). Note that the allele frequencies between the two alleles are not much different (A: 56.290% (2819 / 5008); G: 43.710% (2189 / 5008)) according to the samples submitted to dbSNP [40]. In addition, Figure 4C represents that higher AURKB expression is correlated with the elevated rate of biochemical relapse occurrence.

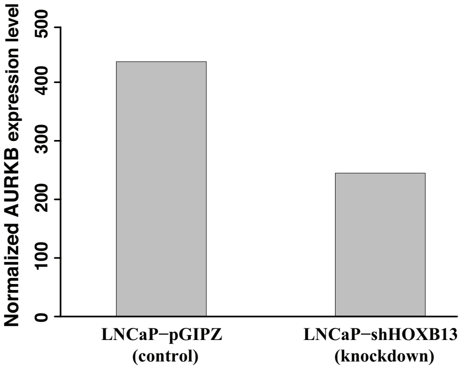

Figure 5. Knockout of HOXB13 diminishes AURKB gene expression level. LNCaP-pGIPZ and LNCaP-shHOXB13 are the control and HOXB13 repressed cell, respectively.

3.4 The rs1476161_AURKB pair potentially regulated by HOXB13

Despite the absence of the HOXB13 binding motif in AURKB, we hypothesized that the gene may still be regulated by HOXB13 through, for example, multiple transcription factor binding. Thus, we investigated whether rs1476161 can affect the AURKB expression in relation to HOXB13. We first knocked down HOXB13 by short hairpin RNA (shRNA) and then examined the change of AURKB expression level. Interestingly, we observed that the expression level of AURKB was decreased by about 2-fold in HOXB13-silencing cells compared to the control cells (Figure 5), indicating the involvement of HOXB13 in the AURKB gene expression. We envision a potential molecular mechanism that HOXB13 regulates other transcription factors such as androgen receptor that bind to this locus [41]. Further experimental investigation can delineate the detailed mechanisms of how transcription factors contribute for regulating AURKB

Table 3. List of SNP candidates and their neighboring target genes significantly correlated among SNP, gene expression and risk of prostate cancer (BCR).

SNP ID	Gene symbol	Gene name	p-value		
			SNP-Gene	Gene-BCR	SNP-BCR
rs7992643	CLYBL	Citrate Lyase Beta Like	2.20E-08	6.54E-02	5.68E-02
rs12938215	DUSP14	Dual Specificity Phosphatase 14	1.47E-02	6.18E-02	5.71E-04
rs1476161	AURKB	Aurora Kinase B	2.33E-02	5.18E-02	1.13E-02

expression in PCa risk.

4. CONCLUSION AND DISCUSSION

We presents an *in silico* methodology in conjunction with an experimental validation for identifying rSNPs located in the TF-bound noncoding regions. These rSNPs affect the TF binding affinities to DNA that in turn alter gene expression associated with increasing the risk of disease. Specifically, we focused on HOXB13, one of the important TFs for PCa progression and development. By integrating various high-throughput sequencing data such as ChIP-seq, RNA-seq and SNP array along with clinical data, we identified 213 rSNP candidates that are highly correlated with neighboring gene's expression level. Some of these were reported as eQTL in previously published studies and as biomarkers in human cancers (Table 1). Notably, one of the top rSNP candidates located in the HOXB13 binding motif along its target gene was recently confirmed as PCa risk allele [18]. In addition, we identified a novel rSNP and its target gene pair candidate (rs1476161, AURKB) that was further confirmed by *in vitro* validation; this would be a potential biomarker in PCa. Put together, our approach is feasible in identifying rSNPs and their target genes in diseases.

AURKB is known to play critical roles in human cancers and has been of great interest in the PCa community for elucidating its functional role in the disease. Figure 4 shows that AURKB mRNA level is higher with the reference allele AA of rs1476161 in the patient samples than the alternative allele GG. The risk prediction based on BCR illustrates that AA is the risk allele associated with an increased PCa development. This is coherent with the gene expression-based risk analysis demonstrating higher risk of PCa with the overexpression of AURKB. Indeed, similar observation in other human cancers has been reported, for example, in primary non-small cell lung carcinoma [42], acute myeloid leukemia [43], and thyroid carcinoma [44]. Thus, we speculate rs1476161 as the risk allele for causing increased AURKB expression.

Although our experimental data supports the involvement of HOXB13 in altering AURKB mRNA level, it is not possible to rule out other potential cofactors (e.g., AR, FOXA1, GATA2). Previous studies uncovered that FOXA1 directly regulates HOXB13 [45] and AR-binding to target genes [46, 47]. Thus, we performed additional simulations with FOXA1, AR and GATA2 ChIP-seq data, respectively. Interestingly, we recovered the rs1476161 and AURKB pair (P=2.3E-02) present only in the outcome of FOXA1. Although high expression of AURKB can cause genomic instability due to tetraplodidy [48, 49], TF-dependent overexpression of the gene leading to tumorigenesis remains unknown. We envision rs1476161 affecting at least HOXB13 and FOXA1 binding that eventually alters the AURKB expression.

While TCGA offers unprecedented opportunity for integrative analyses of multiple –omics datasets, some limitations still exist, particularly for the lack of independent validation dataset. This may create a potential bias in the computational predictions. In the future studies, we plan to focus on validating the findings in the independent PCa datasets [50] that will strengthen the *in silico* predictions and minimize the number of false positive predictions.

5. ACKNOWLEDGMENTS

This work was funded by the United-4 A Cure foundation and in part by the National Library of Medicine of the National Institutes of Health (NIH) under the Award Number R01LM011297. The consent is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

6. REFERENCES

[1] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, and M. J. Thun, "Cancer statistics, 2009," *CA: a cancer journal for clinicians,* vol. 59, no. 4, pp. 225-49, Jul-Aug, 2009.

[2] P. Lichtenstein, N. V. Holm, P. K. Verkasalo, A. Iliadou, J. Kaprio, M. Koskenvuo, E. Pukkala, A. Skytthe, and K. Hemminki, "Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland," *The New England journal of medicine,* vol. 343, no. 2, pp. 78-85, Jul 13, 2000.

[3] R. A. Eeles, A. A. Olama, S. Benlloch, E. J. Saunders, D. A. Leongamornlert, M. Tymrakiewicz, M. Ghoussaini, C. Luccarini, J. Dennis, S. Jugurnauth-Little *et al.*, "Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array," *Nature genetics,* vol. 45, no. 4, pp. 385-91, 391e1-2, Apr, 2013.

[4] J. L. Rinn, and H. Y. Chang, "Genome regulation by long noncoding RNAs," *Annual review of biochemistry,* vol. 81, pp. 145-66, 2012.

[5] V. G. Cheung, and R. S. Spielman, "Genetics of human gene expression: mapping DNA variants that influence gene expression," *Nature reviews. Genetics,* vol. 10, no. 9, pp. 595-604, Sep, 2009.

[6] A. P. Boyle, E. L. Hong, M. Hariharan, Y. Cheng, M. A. Schaub, M. Kasowski, K. J. Karczewski, J. Park, B. C. Hitz, S. Weng *et al.*, "Annotation of functional variation in personal genomes using RegulomeDB," *Genome research,* vol. 22, no. 9, pp. 1790-7, Sep, 2012.

[7] L. D. Ward, and M. Kellis, "HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants," *Nucleic acids research,* vol. 40, no. Database issue, pp. D930-4, Jan, 2012.

[8] Y. Fu, Z. Liu, S. Lou, J. Bedford, X. J. Mu, K. Y. Yip, E. Khurana, and M. Gerstein, "FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer," *Genome biology,* vol. 15, no. 10, pp. 480, 2014.

[9] G. R. Ritchie, I. Dunham, E. Zeggini, and P. Flicek, "Functional annotation of noncoding sequence variants," *Nature methods,* vol. 11, no. 3, pp. 294-6, Mar, 2014.

[10] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure, "A general framework for estimating the relative pathogenicity of human genetic variants," *Nature genetics,* vol. 46, no. 3, pp. 310-5, Mar, 2014.

[11] K. K. Farh, A. Marson, J. Zhu, M. Kleinewietfeld, W. J. Housley, S. Beik, N. Shoresh, H. Whitton, R. J. Ryan, A. A. Shishkin *et al.*, "Genetic and epigenetic fine mapping of causal autoimmune disease variants," *Nature,* vol. 518, no. 7539, pp. 337-43, Feb 19, 2015.

[12] L. D. Ward, and M. Kellis, "Interpreting noncoding genetic variation in complex traits and human disease," *Nature biotechnology,* vol. 30, no. 11, pp. 1095-106, Nov, 2012.

[13] Q. Wang, W. Li, X. S. Liu, J. S. Carroll, O. A. Janne, E. K. Keeton, A. M. Chinnaiyan, K. J. Pienta, and M. Brown, "A hierarchical network of transcription factors governs androgen receptor-dependent prostate cancer growth," *Molecular cell,* vol. 27, no. 3, pp. 380-92, Aug 3, 2007.

[14] Y. R. Kim, I. J. Kim, T. W. Kang, C. Choi, K. K. Kim, M. S. Kim, K. I. Nam, and C. Jung, "HOXB13 downregulates intracellular zinc and increases NF-kappaB signaling to promote prostate cancer metastasis," *Oncogene,* vol. 33, no. 37, pp. 4558-67, Sep 11, 2014.

[15] C. M. Ewing, A. M. Ray, E. M. Lange, K. A. Zuhlke, C. M. Robbins, W. D. Tembe, K. E. Wiley, S. D. Isaacs, D. Johng, Y. Wang *et al.*, "Germline mutations in HOXB13 and prostate-cancer risk," *The New England journal of medicine,* vol. 366, no. 2, pp. 141-9, Jan 12, 2012.

[16] C. Jung, R. S. Kim, S. J. Lee, C. Wang, and M. H. Jeng, "HOXB13 homeodomain protein suppresses the growth of prostate cancer cells by the negative regulation of T-cell factor 4," *Cancer research,* vol. 64, no. 9, pp. 3046-51, May 1, 2004.

[17] C. Jung, R. S. Kim, H. J. Zhang, S. J. Lee, and M. H. Jeng, "HOXB13 induces growth suppression of prostate cancer cells as a repressor of hormone-activated androgen receptor signaling," *Cancer research,* vol. 64, no. 24, pp. 9185-92, Dec 15, 2004.

[18] Q. Huang, T. Whitington, P. Gao, J. F. Lindberg, Y. Yang, J. Sun, M. R. Vaisanen, R. Szulkin, M. Annala, J. Yan *et al.*, "A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding," *Nature genetics,* vol. 46, no. 2, pp. 126-35, Feb, 2014.

[19] R. Takata, S. Akamatsu, M. Kubo, A. Takahashi, N. Hosono, T. Kawaguchi, T. Tsunoda, J. Inazawa, N. Kamatani, O. Ogawa *et al.*, "Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population," *Nature genetics,* vol. 42, no. 9, pp. 751-4, Sep, 2010.

[20] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome biology,* vol. 15, no. 12, pp. 550, 2014.

[21] B. Carvalho, H. Bengtsson, T. P. Speed, and R. A. Irizarry, "Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data," *Biostatistics,* vol. 8, no. 2, pp. 485-99, Apr, 2007.

[22] R. B. Scharpf, R. A. Irizarry, M. E. Ritchie, B. Carvalho, and I. Ruczinski, "Using the R Package crlmm for Genotyping and Copy Number Estimation," *Journal of statistical software,* vol. 40, no. 12, pp. 1-32, May 1, 2011.

[23] H. Li, and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics,* vol. 25, no. 14, pp. 1754-60, Jul 15, 2009.

[24] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li *et al.*, "Model-based analysis of ChIP-Seq (MACS)," *Genome biology,* vol. 9, no. 9, pp. R137, 2008.

[25] M. J. Dunning, M. L. Smith, M. E. Ritchie, and S. Tavare, "beadarray: R classes and methods for Illumina bead-based data," *Bioinformatics,* vol. 23, no. 16, pp. 2183-4, Aug 15, 2007.

[26] A. R. Quinlan, and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics,* vol. 26, no. 6, pp. 841-2, Mar 15, 2010.

[27] C. International HapMap, "A haplotype map of the human genome," *Nature,* vol. 437, no. 7063, pp. 1299-320, Oct 27, 2005.

[28] X. Zhang, A. J. Cal, and J. O. Borevitz, "Genetic architecture of regulatory variation in Arabidopsis thaliana," *Genome research,* vol. 21, no. 5, pp. 725-33, May, 2011.

[29] D. L. Nicolae, E. Gamazon, W. Zhang, S. Duan, M. E. Dolan, and N. J. Cox, "Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS," *PLoS genetics,* vol. 6, no. 4, pp. e1000888, Apr, 2010.

[30] M. A. Schaub, A. P. Boyle, A. Kundaje, S. Batzoglou, and M. Snyder, "Linking disease associations with regulatory information in the human genome," *Genome research,* vol. 22, no. 9, pp. 1748-59, Sep, 2012.

[31] K. Matsumoto, T. Satoh, A. Irie, J. Ishii, S. Kuwao, M. Iwamura, and S. Baba, "Loss expression of uroplakin III is associated with clinicopathologic features of aggressive bladder cancer," *Urology,* vol. 72, no. 2, pp. 444-9, Aug, 2008.

[32] O. Tatti, E. Gucciardo, P. Pekkonen, T. Holopainen, R. Louhimo, P. Repo, P. Maliniemi, J. Lohi, V. Rantanen, S. Hautaniemi *et al.*, "MMP16 Mediates a Proteolytic Switch to Promote Cell-Cell Adhesion, Collagen Alignment, and Lymphatic Invasion in Melanoma," *Cancer research,* vol. 75, no. 10, pp. 2083-94, May 15, 2015.

[33] A. Hadchouel, F. Decobert, M. L. Franco-Montoya, I. Halphen, P. H. Jarreau, O. Boucherat, E. Martin, A. Benachi, S. Amselem, J. Bourbon *et al.*, "Matrix metalloproteinase gene polymorphisms and bronchopulmonary dysplasia: identification of MMP16 as a new player in lung development," *PloS one,* vol. 3, no. 9, pp. e3188, 2008.

[34] E. R. King, C. S. Tung, Y. T. Tsang, Z. Zu, G. T. Lok, M. T. Deavers, A. Malpica, J. K. Wolf, K. H. Lu, M. J. Birrer *et al.*, "The anterior gradient homolog 3 (AGR3) gene is associated with differentiation and survival in ovarian cancer," *The American journal of surgical pathology,* vol. 35, no. 6, pp. 904-12, Jun, 2011.

[35] G. Oxford, C. R. Owens, B. J. Titus, T. L. Foreman, M. C. Herlevsen, S. C. Smith, and D. Theodorescu, "RalA and RalB: antagonistic relatives in cancer cell migration," *Cancer research,* vol. 65, no. 16, pp. 7111-20, Aug 15, 2005.

[36] M. Fadri-Moskwik, K. N. Weiderhold, A. Deeraksa, C. Chuang, J. Pan, S. H. Lin, and L. Y. Yu-Lee, "Aurora B is regulated by acetylation/deacetylation during mitosis in prostate cancer cells," *FASEB journal : official publication of the Federation of American Societies for Experimental Biology,* vol. 26, no. 10, pp. 4057-67, Oct, 2012.

[37] K. J. Niermann, L. Moretti, N. J. Giacalone, Y. Sun, S. M. Schleicher, P. Kopsombut, L. R. Mitchell, K. W. Kim, and B. Lu, "Enhanced radiosensitivity of androgen-resistant prostate cancer: AZD1152-mediated Aurora kinase B inhibition," *Radiation research,* vol. 175, no. 4, pp. 444-51, Apr, 2011.

[38] E. Nna, J. Madukwe, E. Egbujo, C. Obiorah, C. Okolie, G. Echejoh, A. Yahaya, J. Adisa, and I. Uzoma, "Gene

expression of Aurora kinases in prostate cancer and nodular hyperplasia tissues," *Medical principles and practice : international journal of the Kuwait University, Health Science Centre,* vol. 22, no. 2, pp. 138-43, 2013.

[39] W. F. Zeng, K. Navaratne, R. A. Prayson, and R. J. Weil, "Aurora B expression correlates with aggressive behaviour in glioblastoma multiforme," *Journal of clinical pathology,* vol. 60, no. 2, pp. 218-21, Feb, 2007.

[40] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "dbSNP: the NCBI database of genetic variation," *Nucleic acids research,* vol. 29, no. 1, pp. 308-11, Jan 1, 2001.

[41] J. D. Norris, C. Y. Chang, B. M. Wittmann, R. S. Kunder, H. Cui, D. Fan, J. D. Joseph, and D. P. McDonnell, "The homeodomain protein HOXB13 regulates the cellular response to androgens," *Molecular cell,* vol. 36, no. 3, pp. 405-16, Nov 13, 2009.

[42] S. L. Smith, N. L. Bowers, D. C. Betticher, O. Gautschi, D. Ratschiller, P. R. Hoban, R. Booton, M. F. Santibanez-Koref, and J. Heighway, "Overexpression of aurora B kinase (AURKB) in primary non-small cell lung carcinoma is frequent, generally driven from one allele, and correlates with the level of genetic instability," *British journal of cancer,* vol. 93, no. 6, pp. 719-29, Sep 19, 2005.

[43] S. A. Hartsink-Segers, C. M. Zwaan, C. Exalto, M. W. Luijendijk, V. S. Calvert, E. F. Petricoin, W. E. Evans, D. Reinhardt, V. de Haas, M. Hedtjarn *et al.*, "Aurora kinases in childhood acute leukemia: the promise of aurora B as therapeutic target," *Leukemia,* vol. 27, no. 3, pp. 560-8, Mar, 2013.

[44] R. Sorrentino, S. Libertini, P. L. Pallante, G. Troncone, L. Palombini, V. Bavetsias, D. Spalletti-Cernia, P. Laccetti, S. Linardopoulos, P. Chieffi *et al.*, "Aurora B overexpression associates with the thyroid carcinoma undifferentiated phenotype and is required for thyroid carcinoma cell proliferation," *The Journal of clinical endocrinology and metabolism,* vol. 90, no. 2, pp. 928-35, Feb, 2005.

[45] R. P. McMullin, A. Dobi, L. N. Mutton, A. Orosz, S. Maheshwari, C. S. Shashikant, and C. J. Bieberich, "A FOXA1-binding enhancer regulates Hoxb13 expression in the prostate gland," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 107, no. 1, pp. 98-103, Jan 5, 2010.

[46] H. J. Jin, J. Kim, and J. Yu, "Androgen receptor genomic regulation," *Translational andrology and urology,* vol. 2, no. 3, pp. 157-177, Sep, 2013.

[47] H. J. Jin, J. C. Zhao, L. Wu, J. Kim, and J. Yu, "Cooperativity and equilibrium with FOXA1 define the androgen receptor transcriptional program," *Nature communications,* vol. 5, pp. 3972, 2014.

[48] B. Goldenson, and J. D. Crispino, "The aurora kinases in cell cycle and leukemia," *Oncogene,* vol. 34, no. 5, pp. 537-45, Jan 29, 2015.

[49] T. Fujiwara, M. Bandi, M. Nitta, E. V. Ivanova, R. T. Bronson, and D. Pellman, "Cytokinesis failure generating tetraploids promotes tumorigenesis in p53-null cells," *Nature,* vol. 437, no. 7061, pp. 1043-7, Oct 13, 2005.

[50] B. S. Taylor, N. Schultz, H. Hieronymus, A. Gopalan, Y. Xiao, B. S. Carver, V. K. Arora, P. Kaushik, E. Cerami, B. Reva *et al.*, "Integrative genomic profiling of human prostate cancer," *Cancer cell,* vol. 18, no. 1, pp. 11-22, Jul 13, 2010.