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

Segun Jung, Hongjian Jin and Ramana V. Davuluri

Division of Health and Biomedical Informatics, Department of Preventive Medicine, Northwestern University Feinberg School of Medicine

Research Objective

Genome-wide association studies (GWAS) have identified numerous single nucleotide polymorphisms (SNPs) associated with disease susceptibility. GWAS have reported more than 70 SNPs associated with Prostate Cancer (PCa) risk. Functional roles of these SNPs, however, are largely unknown. 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. This includes a recently discovered rSNP (rs339331) and 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.

Methods

Dataset

- **TCGA RNA-seq**: 497 tumor and 52 matched normal samples
- **TCGA SNP array**: 500 samples
- **TCGA clinical data**: 369 patients
- **ChIP-seq data** for the HOXB13 transcription factor
- **Microarray profiling** of LNCaP control and HOX13 silencing cells

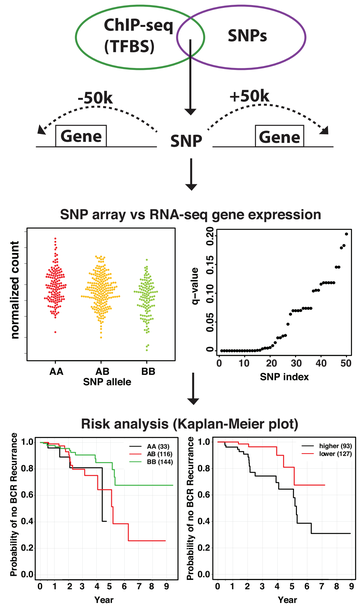
Computational procedure for identifying risk SNP candidates

Step 1: Find SNPs lying in a TF binding site (ChIP-seq peaks) based on genomic position

Step 2: Search neighboring genes of each SNP

Step 3: Identify significant allele-specific SNPs correlated with gene expression using SNP array and RNA-seq data

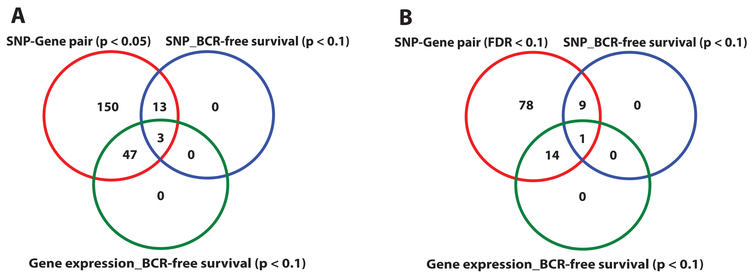
Step 4: Risk analysis based on log-rank test for each SNP and gene in relation to clinical information



- **Step 1.** Performed OverlapSelect using ChIIP-seq (36143 peaks) and SNP array data (905422 SNPs)
- **Step 2.** Ran the following command: bedtools window –a snp.bed –b gene.bed –w 50000 > output.bed
- **Step 3.** Performed ANOVA on the extracted SNP array and gene expression data for each SNP-gene pair
- **Step 4.** Performed risk analysis (Kaplan-Meier) of biochemical relapse using 295 tumor and 52 normal RNA-seq samples RNA-seq and 293 SNP array data

Results

1. 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.

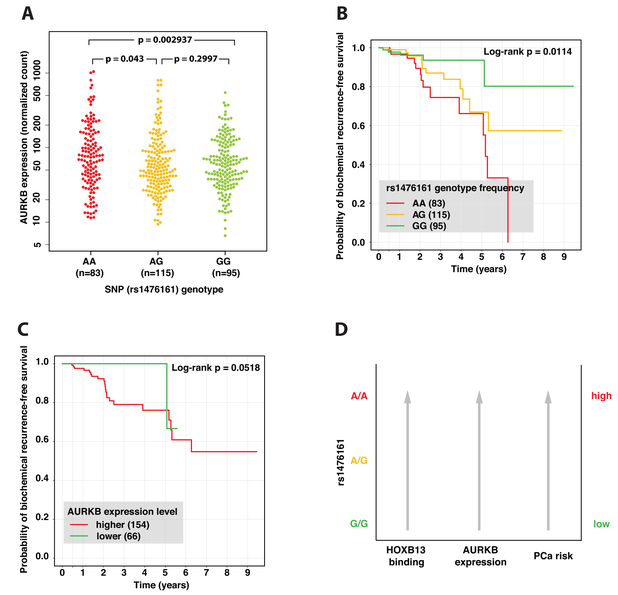
2. 16 SNP-gene list from the PCa data analysis reported as eQTL in other studies.

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
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	2747.7
rs847377	AGR3	1.83E-02	C	T	202	231	61	362.3	429	487.6

3. 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
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

4. Association between rs1476161 genotype, AURKB expression and PCa risk



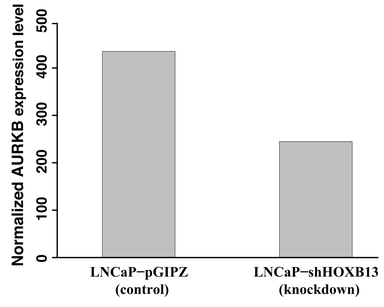
(A) The overall distribution of normalized AURKB read counts shows significant difference between AA and GG (p-value= 2.937E-03, Student's t-test).

(B) Kaplan-Meier plot depicks GG genotype is the risk allele for the PCa progression.

(C) Kaplan-Meier plot illustrates the higher PCa risk with lower AURKB expression level.

(D) AA promotes HOXB13 binding that causes increased AURKB expression resulted in high PCa risk

5. Experimental validation



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

Conclusions

- **We presented an *in silico* methodology in conjunction with an experimental validation for identifying rSNPs located in the TF-bound noncoding regions**
- **We identified a novel rSNP and its target gene pair candidate (rs1476161, AURKB) as a potential biomarker in PCa**