The landscape of RNA polymerase II associated chromatin interactions in prostate cancer

Susmita G. Ramanand, … , Michael Q. Zhang, Ram S. Mani


Graphical abstract

Find the latest version:
https://jci.me/134260/pdf
The landscape of RNA polymerase II associated chromatin interactions in prostate cancer

Susmita G. Ramanand¹, Yong Chen²,³, Jiabei Yuan¹, Kelly Daescu², Maryou Lambros⁴, Kathleen E Houlanah⁵,⁶,⁷, Suzanne Carreira⁴, Wei Yuan⁴, GuemHee Baek¹, Adam Sharp⁴, Alec Paschalis⁴, Mohammed Kanchwala⁸, Yunpeng Gao¹, Adam Aslam¹, Nida Safdar¹, Xiaowei Zhan⁹, Ganesh V. Raj¹⁰, Chao Xing⁸,⁹,¹¹, Paul C. Boutros⁵,⁶,⁷,¹²,¹³,¹⁴, Johann de Bono⁴, Michael Q. Zhang²,¹⁵, Ram S. Mani¹,¹⁰,¹⁶

¹Department of Pathology, UT Southwestern Medical Center, Dallas, TX 75390, USA
²Department of Biological Sciences, Center for Systems Biology, The University of Texas at Dallas, Richardson, TX 75080, USA
³Department of Molecular and Cellular Biosciences, Rowan University, Glassboro, NJ 08028, USA
⁴Prostate Cancer Targeted Therapy and Cancer Biomarkers Group, The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust, Sutton, UK
⁵Ontario Institute for Cancer Research, Toronto, ON M5G 0A3, Canada
⁶Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 1L7, Canada
⁷Vector Institute, Toronto, ON M5G 1M1, Canada
⁸Eugene McDermott Center for Human Growth & Development, UT Southwestern Medical Center, Dallas, TX 75390, USA
Department of Population and Data Sciences, UT Southwestern Medical Center, Dallas, TX 75390, USA

Department of Urology, UT Southwestern Medical Center, Dallas, TX 75390, USA

Department of Bioinformatics, UT Southwestern Medical Center, Dallas, TX 75390, USA

Department of Urology, University of California, Los Angeles, CA 90024, USA

Department of Human Genetics, University of California, Los Angeles, CA 90095, USA

Jonsson Comprehensive Cancer Centre, University of California, Los Angeles, CA 90024, USA

MOE Key Laboratory of Bioinformatics and Bioinformatics Division, Center for Synthetic and System Biology, TNLIST/Department Automation, Tsinghua University, Beijing 100084, China

Harold C. Simmons Comprehensive Cancer Center, UT Southwestern Medical Center, Dallas, TX 75390, USA

**Authorship note:** SGR, YC and JY are co-first authors.

*Correspondence:* ram.mani@utsouthwestern.edu, michael.zhang@utdallas.edu

**Conflict of interest:** AS, MBKL, SC, AP, WY and JDB are employees of The Institute of Cancer Research (ICR), which has a commercial interest in abiraterone. AS has received travel support from Sanofi and Roche-Genentech, and speakers' honorarium from Astellas Pharma. JDB has served on advisory boards and received fees from many companies including Astra Zeneca, Astellas, Bayer, Boehringer Ingelheim, Cellcentric, Daiichi, Genentech/Roche, Genmab, GSK, Janssen, Merck Serono, Merck
Sharp & Dohme, Menarini/Silicon Biosystems, Orion, Pfizer, Qiagen, Sanofi Aventis, Sierra Oncology, Taiho, Vertex Pharmaceuticals. JDB is an employee of The ICR, which have received funding or other support for his research work from AZ, Astellas, Bayer, Cellcentric, Daiichi, Genentech, Genmab, GSK, Janssen, Merck Serono, MSD, Menarini/Silicon Biosystems, Orion, Sanofi Aventis, Sierra Oncology, Taiho, Pfizer, Vertex, and which has a commercial interest in abiraterone, PARP inhibition in DNA repair defective cancers and PI3K/AKT pathway inhibitors (no personal income). He was named as an inventor, with no financial interest, for patent 8,822,438. He has been the CI/PI of many industry sponsored clinical trials. JDB is a National Institute for Health Research (NIHR) Senior Investigator. The views expressed in this article are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health. No relevant conflicts of interest were disclosed by other authors.

ABSTRACT

Transcriptional dysregulation is a hallmark of prostate cancer (PCa). We mapped the RNA Polymerase II (RNA Pol II) associated chromatin interactions in normal prostate cells and PCa cells. We discovered thousands of enhancer-promoter, enhancer-enhancer, as well as promoter-promoter chromatin interactions. These transcriptional hubs operate within the framework set by structural proteins—CTCF and cohesins, and are regulated by the cooperative action of master transcription factors, such as the Androgen Receptor (AR) and FOXA1. By combining analyses from metastatic castration resistant PCa (mCRPC) specimens, we show that AR locus amplification contributes to the transcriptional up-regulation of AR gene by increasing the total
number of chromatin interaction modules comprising of the AR gene and its distal enhancer. We deconvoluted the transcription control modules of several PCa genes, notably, the biomarker KLK3, lineage-restricted genes (KRT8, KRT18, HOXB13, FOXA1, ZBTB16), the drug target EZH2, and the oncogene MYC. By integrating clinical PCa data, we defined a novel germline-somatic interplay between the PCa risk allele rs684232 and the somatically acquired TMPRSS2-ERG gene fusion in the transcriptional regulation of multiple target genes—VPS53, FAM57A and GEMIN4. Our studies implicate changes in genome organization as a critical determinant of aberrant transcriptional regulation in PCa.

INTRODUCTION

Whole genome sequencing (WGS) and exome sequencing studies have revealed that PCa is generally characterized by lower mutation rates and higher rates of recurrent genomic rearrangements. The majority of PCa associated recurrent genomic rearrangements promote the up-regulation of transcription factor genes. For example, genomic rearrangements and gene fusions resulting in the up-regulation of ETS transcription factor family genes like ERG and ETV1 are observed in >50% of PCa (1). AR is a master transcription factor that is essential for the normal development of the prostate gland. Intriguingly, aberrant AR signaling drives many facets of PCa etiology, including lethal metastatic castration resistant prostate cancer (mCRPC) development (2). Copy number alterations and amplifications of the AR locus are observed in >60% of mCRPC (3, 4). AR and the pioneer factor FOXA1 co-occupy distant regulatory elements, such as enhancers, to regulate transcription (2, 5). The binding of ETS
transcription factors to these regulatory elements add an additional layer of complexity to the transcriptional output, ultimately resulting in the oncogenic phenotype (6, 7). Taken together, these studies indicate that transcriptional dysregulation is a distinguishing feature of PCa development.

Two-dimensional (2D) genomics approaches like ChIP-seq have contributed to the identification of thousands of enhancers in PCa (7, 8). However, the targets of most of these enhancers are unknown. It is unclear if individual enhancers regulate single or multiple genes, and conversely if individual genes are regulated by single or multiple enhancers in PCa. 2D genomics cannot reveal if multiple genes and enhancers are coordinately regulated in shared nuclear space. Methods like Hi-C have enabled the identification of topologically associating domains (TADs) that change in size between normal and PCa cells (9, 10). However, in general, Hi-C studies have limited utility in the precision mapping of enhancer-promoter contacts. The absence of this knowledge has stymied our understanding of the regulatory targets of somatic mutations and germline risk alleles residing in the intergenic regions.

We present, for the first time, the three-dimensional (3D) landscape of RNA Polymerase II (RNA Pol II) associated chromatin interactions in normal prostate cells and PCa cells. By pairing thousands of enhancers to their target genes, we identified thousands of transcriptional network hubs operating within the framework set by structural proteins like CTCF and cohesins. Integrative analyses of these data uncovered multiple, novel mechanisms of transcriptional regulation and its dysregulation in PCa. Finally, we demonstrate that many somatic and germline DNA alterations re-wire the landscape of RNA Pol II interactions in PCa.
RESULTS

RNA Pol II associated chromatin interactions in PCa

Chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) is a genome-wide integration of chromatin immunoprecipitation (ChIP) and chromosome conformation capture (3C) methods to map the genomic interaction landscape of any chromatin associated protein of interest (11, 12). We sought to identify global differences in transcriptional regulation by comparing RNA Pol II ChIA-PET, along with ChIP-seq and RNA-seq analyses in benign prostate cells (RWPE-1) and PCa cells (LNCaP, VCaP and DU145) (Tables S1 and S2). LNCaP and VCaP cells are AR positive, whereas RWPE-1 and DU145 cells are AR negative (Figure 1A).

Binding peaks represent the fundamental basis of ChIA-PET analysis; peaks that interact with other peaks are called anchor peaks, whereas non-interacting ones are termed non-anchor peaks. Thus, a single RNA Pol II ChIA-PET experiment can provide both 2D data in terms of RNA Pol II binding peaks, as well as 3D data in terms of RNA Pol II binding anchor peaks and non-anchor peaks. Paired-end-tags (PETs) are the units used for measuring the interactions between a pair of anchor peaks—more PETs between anchors signifying stronger interactions. We identified thousands of RNA Pol II associated chromatin interactions spanning all chromosomes in the cell lines analyzed. Saturation analysis indicated that we have sufficient sequencing depth to detect high confidence chromatin interactions supported by three or more PETs. For example, by using only 50% of the sequencing reads, we were able to re-discover 64%-80% of chromatin interactions supported by 3 or more PETs, and 76%-90% of chromatin...
interactions supported by 4 or more PETs (Figure S1). Importantly, the saturation analysis indicates that our data is comparable to the gold-standard ENCODE ChIA-PET datasets.

We also estimated the length distribution of RNA Pol II chromatin interactions by setting multiple PET cutoffs for various interaction strengths (Figure S1B). Overall, the bulk of chromatin interactions ranged from 3 Kb to 1 Mb in all the cell lines analyzed. Interactions greater than 1 Mb were relatively rare, especially at higher PET cutoffs. This association was extended by comparing interactions between cell lines (Figure S2A). As expected, there was a greater overlap for interactions <1Mb between any two or more cell lines at all measured stringencies/PET cutoffs. Maximal overlap in RNA Pol II chromatin interactions was observed between LNCaP and VCaP cells. This is likely due to the fact that both LNCaP and VCaP cells are AR positive and are dependent on androgen signaling for survival. These results suggest that an overlap in master transcription factors are associated with a parallel overlap in RNA Pol II chromatin interactions. By partitioning the genome into three groups—promoters, gene bodies and other regions, we observed thousands of interactions within and between these groups (Figure S1C, top). We probed the chromatin interaction patterns of active enhancers by mapping acetylated Histone H3 at lysine 27 (H3K27ac) marked sites onto the RNA Pol II ChIA-PET anchor sites. In addition to interactions between enhancers and promoters, we also discovered thousands of enhancer-enhancer and promoter-promoter interactions (Figure S1C, bottom), implying that a complex network of chromatin interactions coordinates transcriptional regulation in these cells. We leveraged our ChIA-PET data to pair enhancers and their potential target genes in the four cell lines.
(Table S3). We observed that an increase in the RNA Pol II chromatin-interaction strength between enhancers and promoters is associated with a decrease in the 2D distances between the interacting regions (Figure S1D). In comparison to the benign RWPE-1 cells, all the PCa cell lines displayed an overall increase in enhancer usage per gene as well as an accompanying increase in gene targets per enhancer (Figure 1B and S1E-S1G). Next, we determined the relationship between enhancers per gene and gene-expression, and observed a very weak correlation (Figure S1H), implying that enhancers per gene is not a useful predictor of gene-expression.

We observed that RNA Pol II interaction strength, as measured by the number of PETs in gene promoters, positively correlated with mRNA abundance in all four cell lines (Figure 1C), and is a much better predictor of gene-expression than enhancers per gene. The interaction strength between anchors followed a negative binomial distribution (Figure S1I). By applying Difference of Negative Binomial distributions (DNB), a new statistical model for comparing two different negative binomial distributions, we have identified thousands of significantly gained and lost interactions across all the possible six pair-wise comparisons amongst the four cell lines (Figure 1D and Table S4). Next, we evaluated the relationship between changes in gene-expression and changes in promoter associated chromatin interactions between the benign RWPE-1 cells and the three PCa cell lines. We initiated the analyses by examining all genes that are up-regulated or down-regulated by two-fold or more, followed by additional analyses by filtering the top 80%, 60%, 40%, 20% and 10% up-regulated or down-regulated genes from the starting list. Remarkably, as we traversed from all genes to the top 10% genes, the correlation also increased (Figure 1E). This
trend was observed for all our comparisons. This result indicates that changes in RNA Pol II associated chromatin interactions are likely to be drivers of gene expression changes. This is particularly the case for the top dysregulated genes.

To extend this correlation further, we rank-ordered all the genes based on the number of PETs in their promoters (Figure 1F). Genes like FOS, JUN and KLF6 were highly expressed in all four cell lines and had >25 independent PETS in the promoter. The luminal like LNCaP and VCaP cells had more PETs in the luminal marker gene promoter KRT8, as compared to the basal marker gene promoter KRT5. The opposite trend was observed in the more basal-like RWPE-1 cells. Notably, the AR gene promoter had ~50 independent PETs in VCaP cells. The EZH2 gene promoter had a greater number of PETS in PCa cells as compared to benign RWPE-1 cells. Next, we examined PCa enriched pathways and gene ontology (GO) terms using gene lists obtained from differential transcript abundance and differential chromatin interactions. Although both these analyses indicated enrichment of common categories such as metabolic processes, cellular process and cell cycle related processes amongst others, we also identified categories unique to differential chromatin interactions such as chromatin organization and translation (Figures S2B and S2C).

We also investigated the DNA sequence features of promoters and enhancers (Figure S1J). In general, promoters exhibited lesser variability than enhancers. ETS, CCAAT and YY1 motifs were commonly observed in the promoters of all the cell lines. Enhancers exhibited considerable diversity in terms of DNA sequence features. ETS and CTCF motifs were commonly observed in the enhancers of all the cell lines. The FOXA1 forkhead motif and HOXB13 homeobox motif were unique to LNCaP and VCaP
cells. The leucine zipper motif bZIP was the most significantly enriched enhancer motif in RWPE-1 and DU145 cells.

**Integrative analysis of RNA Pol II associated chromatin interactions**

To better understand the regulatory concepts associated with transcriptional control, we integrated our RNA Pol II ChIA-PET data with multiple additional datasets (Table S1) (13). We examined a representative DNA region from LNCaP cells with both transcriptionally active and inactive genes by overlaying our RNA Pol II ChIA-PET data with RAD21 ChIA-PET, Hi-C, RNA-seq, DNase-seq and ChIP-seq data for CTCF (topologically associating domain (TAD) and sub-TAD boundary element mark), H3K4me3 (active promoter mark), and H3K27ac (active enhancer mark) (Tables S1 and S2). Our RNA Pol II ChIA-PET data were consistent with the architectural features of these independent datasets (Figure 2A). Both active and repressed regions exhibited significant chromatin contacts as visualized by the Hi-C track. However, the nature of the interactions within active and repressed regions showed characteristic differences. Consistent with RNA-seq, DNase-seq, H3K4me3, and H3K27ac signals, we observed that both RNA Pol II and RAD21 chromatin interactions were enriched in the active regions. The prominent RNA Pol II interactions almost always involved H3K4me3 and/or H3K27ac marks. The RAD21 interactions were primarily associated with CTCF bound regions. Next, we compared the RNA Pol II and RAD21 ChIA-PET datasets. The genome-wide total RNA Pol II peaks as well as interaction anchor RNA Pol II peaks overlapped with H3K27ac peaks with statistical significance (p-value<0.001, hypergeometric test), but not with CTCF or RAD21 peaks (Figure 2B and Table S5). The genome-wide total RAD21 peaks as well as interaction anchor RAD21 peaks
overlapped with CTCF peaks with statistical significance (p-value<0.001, hypergeometric test), but not with RNA Pol II or H3K27ac peaks (Figure 2C and Table S5). The length of RAD21 associated chromatin interactions were significantly longer than RNA Pol II associated interactions (Figure 2D). On the basis of these analyses, chromatin interactions can be hierarchically classified into two tiers—the outer RAD21 associated interactions at CTCF bound sites and the inner RNA Pol II associated interactions at H3K27ac marked sites. As RAD21 is a cohesin component, our results are consistent with the concept that RAD21 interactions represent boundaries of insulated neighborhoods or TADs. Functionally, the outer RAD21 interactions provide insulation for the inner RNA Pol II chromatin interactions. Conceptually, while lineage restricted enhancers trigger the transcriptional activation of their cognate target genes, this insulation draws the boundary and prevents these enhancers from activating genes outside the boundary. We present two examples to illustrate this point in the context of prostate epithelial lineage identity.

*KRT8-KRT18* gene cluster expression is limited to prostate luminal epithelial cells and luminal PCa represented by LNCaP and VCaP cells. The *KRT8* and *KRT18* genes display divergent transcription from a bi-directional promoter/enhancer element marked by FOXA1 and H3K27ac (Figure 2E). We observed that the inner, short-range, RNA Pol II associated chromatin interactions are flanked by outer, long-range, RAD21 associated chromatin interactions at sites marked by CTCF binding (Figures 2E and 2F). On the basis of these data, we hypothesize that while lineage restricted transcription factors, such as FOXA1, specify the activation of lineage restricted enhancers, the RAD21 interactions specify the gene targets of these enhancers by creating protective moats.
Although several type II keratin genes are located in this sub-chromosomal region, our study explains why the expression of \textit{KRT8} and \textit{KRT18} genes are unique to the prostate luminal epithelial lineage—because RAD21 interactions insulates these two genes and their enhancers. The same concept can be used to explain the lineage-specific regulation of \textit{HOXB13} and several other genes (Figure S3).

**Transcriptional regulation of AR and FOXA1 loci**

Having established genomic concepts that underlie transcriptional regulation, we next examined the mechanisms associated with dysregulated transcription in PCa. AR is the central effector of the androgen signaling pathway which is frequently dysregulated in PCa. Recent studies have reported the identification and amplification of an \textit{AR} enhancer in metastatic castration resistant prostate cancer (mCRPC) (4, 14, 15); however, unbiased maps of RNA Pol II chromatin interactions at the \textit{AR} locus are not available to date. We report the identification of an active \textit{AR} enhancer cluster located ~700 Kb upstream of the \textit{AR} promoter (Figures 3A and 3B). This H3K27ac marked enhancer cluster is occupied by AR and FOXA1 and interacts with the \textit{AR} promoter in LNCaP and VCaP cells. The \textit{AR} gene and its enhancers are amplified in VCaP cells, thereby resulting in its over-expression (16). Consistent with \textit{AR} locus amplification, VCaP cells displayed an increase in the total number of chromatin interaction modules comprising of the \textit{AR} gene and its distal enhancer, in comparison to LNCaP cells. These results suggest that \textit{AR} enhancers play a significant role in driving \textit{AR} expression, both in the absence or presence of \textit{AR} locus amplification. By integrating our RNA Pol II ChIA-PET data with the AR ChIA-PET dataset in VCaP cells (5), we observed that the complex chromatin interactions in the AR locus are, in part,
mediated by AR binding (Figure 3B). The VCaP cells express ERG, a transcription factor that is formed by genomic rearrangements involving TMPRSS2 and ERG genes (1, 17). Interestingly, ERG binding to the AR enhancer cluster was also observed in VCaP cells. Thus, the enhancer cluster functions as a landing site for master regulators, which govern AR transcription in a combinatorial manner. Consistent with their role in regulating enhancer driven gene expression (18), we observed that treatment with BET bromodomain inhibitors (BETi) reduced the expression of AR in a dose dependent manner (Figures S4A and S4B).

As targeting the AR signaling axis remains the mainstay of therapy for metastatic PCa, we addressed the relevance of AR enhancers in a clinical setting. We conducted high-density array comparative genomic hybridization (aCGH) to evaluate the copy number status of AR and its enhancers in metastatic tumor biopsies from 27 mCRPC patients (Figure 3C and Table S6). We observed (A) 20 patients with both AR gene amplification/copy-number gain as well as AR enhancer amplification/copy-number gain, (B) 1 patient with AR gene amplification/copy-number gain, but without accompanying changes in the AR enhancer, and (C) 4 patients with AR enhancer amplification/copy-number gain, but without accompanying changes in the AR gene. Two patients in this cohort did not have amplification or copy-number gain in the AR gene as well as the AR enhancer. In summary, AR enhancer alterations were more prevalent than AR gene alterations in our mCRPC cohort.

We validated these results by conducting droplet digital PCR (ddPCR) analysis in an independent, clinically well-annotated mCRPC cohort (n=46 mCRPC patients) (19) (Figure S4C). Again, we observed that amplification of AR enhancers correlated with
AR gene amplification (Figure S4D) and that copy-number gain of AR enhancers correlated with AR gene copy-number gain (Figure S4E). To establish the relationship between AR gene copy-number and mRNA abundance, we re-analyzed the data from 122 CRPC patients in the SU2C cohort (3). As hypothesized, AR gene copy number and mRNA abundance were significantly correlated (Figure S4F). Overall, our results point to an essential role for AR enhancers in promoting AR transcription in CRPC.

We suggest that the extra copies of genomic DNA from the AR locus, formed by the amplification process, are further subjected to iterative rounds of mutagenesis and epigenetic alterations, and are ultimately filtered by treatment-induced evolutionary selection. For example, WGS analysis of a representative mCRPC specimen with amplification of AR and its enhancers resulted in the discovery of a ~350 Kb subclonal deletion that brings the entire enhancer cluster closer to the AR gene (Figures 3D and S4G). This deletion is supported by 139 independent unique reads (out of 7775 reads) and is localized within the AR amplified (20 copy) region. Given the negative correlation between 2D distances separating the interacting regions and RNA Pol II chromatin-interaction strength (Figure S1D), as well as the positive correlation between RNA Pol II chromatin-interaction strength and transcript abundance (Figures 1C and 1E), we suggest that such deletions are likely to further influence AR transcription. Intriguingly, we observed enhanced CTCF binding to the 3'end of the AR gene in VCaP cells that have AR amplification, but not in LNCaP cells (Figures 3A and 3B). VCaP cells also express AR variant proteins that lack the C-terminus ligand binding domain (Figure S4B). CTCF binding to a subset of AR alleles in the context of AR amplification could activate neo-insulator elements that, would in turn, interfere with the RNA Pol II activity
to generate truncated AR variants. Thus, AR amplification can set the stage for multiple downstream regulatory processes to drive the lethal phenotype.

We also examined the transcriptional architecture of the gene encoding the pioneer factor FOXA1 which collaborates with AR to activate enhancers (2). We noticed that multiple enhancers located within the adjacent MIPOL1 gene interacted with the FOXA1 gene. These enhancers were co-bound by AR and FOXA1 (Figures 3E S5). ERG also bound to this enhancer cluster in VCaP cells. Thus, we conclude that master-regulators like AR and FOXA1 auto-regulate their own expression by binding to their enhancers. The gene fusion product, ERG, adds an additional layer of complexity to the regulation of and by the regulators.

**Transcriptional regulation by AR and FOXA1**

The overall expression of genes nearest to AR or FOXA1 binding peaks were not very different from randomly selected control genes (Figure 4A). Based on our observation that multiple genes and enhancers coordinately interact to regulate transcription, we hypothesized that chromatin bound AR/FOXA1 are likely to regulate genes that are located beyond their nearest neighbors. For example, the AR enhancers are closer to the EDA2R gene than the AR gene. By overlapping AR/FOXA1 binding peaks with RNA Pol II ChIA-PET anchor peaks, we created virtual chromatin-contact maps of AR and FOXA1 occupancy, and traced the target genes (Table S7). Importantly, the expression of target genes discovered by our new approach was significantly higher than the nearest neighbor genes as well as control genes in all our comparisons (Figure 4A). These studies highlight the utility of mapping RNA Pol II
associated chromatin interactions in deciphering the master transcription factor-enhancer-target gene regulatory networks.

To garner insights into the functional processes governed by the combinatorial binding of master transcription factors, we queried the RNA Pol II ChIA-PET data for chromatin interactions between AR-FOXA1 co-occupied regions and gene promoters (Figure 4B). This was followed by pathway analysis of these AR-FOXA1 chromatin-interaction-target genes. We were surprised to discover that the genes involved in the EIF2 signaling were maximally enriched in both the LNCaP and VCaP cells (Figure 4C). The EIF2 pathway regulates translation initiation. Our results are consistent with other studies reporting AR mediated transcriptional control of translation initiation (20). We now show that it is mediated by chromatin interactions involving AR and FOXA1. More generally, our results indicate that master transcription factors cooperatively regulate the 3D genome organization to control the transcriptome, and eventually the proteome to establish lineage identity.

**AR binding, histone acetylation and the regulation of the Kallikrein (KLK) gene cluster and ZBTB16**

Prostate-specific antigen (PSA), the protein product of the KLK3 gene, is one of the most commonly used PCa screening biomarkers (21). It is well established that AR up-regulates the expression of KLK3. The KLK3 gene resides in a gene cluster that also includes multiple members of the Kallikrein (KLK) family. We leveraged our ChIA-PET data to create a portrait of transcriptional regulation at the KLK gene cluster (Figure 5A). We identified multiple H3K27ac marked active enhancers surrounding KLK3 and
neighboring KLK genes in the AR positive LNCaP and VCaP cells which express KLK3. These enhancer locations were devoid of the H3K27ac mark in the AR negative RWPE-1 and DU145 cells which do not express KLK3. The active enhancers in the KLK cluster were occupied by AR and FOXA1 (Figures S6A and S6B). Remarkably, every H3K27ac marked enhancer in the KLK cluster interacted with an upstream H3K27ac marked distal enhancer located in the gene C19ORF48. The distal enhancer is not bound by AR or FOXA1 but is active in all the four prostate cell lines analyzed. In the AR negative cells, this enhancer interacted with other active enhancers away from the KLK3 gene. Analysis of this sub-chromosomal region by using H3K27ac ChIP-seq datasets from ~100 cell types indicated that the distal enhancer is constitutively active in all the cell lineages analyzed (Figure S7). However, the activity of enhancers flanking the distal constitutive enhancer is highly cell type specific and provides epigenetic fingerprints for every cell type. It is likely that the flanking enhancers are activated by binding of tissue specific transcription factors to their cognate recognition sites. Thus, we conclude that complex chromatin interactions between constitutive enhancers and tissue-specific enhancers specify the expression of lineage restricted genes like KLK3.

ZBTB16 is another classic AR target gene that is mis-regulated in mCRPC (3, 6). ZBTB16 is expressed in the AR positive LNCaP cells, but is transcriptionally silent in DU145 cells, which do not express AR. In LNCaP cells, the binding of AR and FOXA1 is associated with H3K27ac marked intragenic enhancers. RNA Pol II ChIA-PET indicated interactions between these enhancers and the promoter of ZBTB16 (Figure 5B). These RNA Pol II interactions are insulated by outer chromatin interactions involving RAD21 at sites co-marked by CTCF binding. Enhancers having CTCF occupancy appear to draw
the RNA Pol II associated interactions towards the boundaries of the insulated neighborhood. In DU145 cells, the absence of H3K27ac mark indicated that the intragenic enhancers are inactive, presumably because these cells do not express AR (Figure 5C). As a consequence, there are no RNA Pol II interactions and the gene is silent. However, RAD21 interactions do exist at identical sites co-marked by CTCF binding. While AR/FOXA1 binding, enhancer activation and the associated RNA Pol II interactions up-regulate lineage restricted genes like \textit{ZBTB16}, cohesin-associated chromatin interactions appear to function as moats to prevent \textit{ZBTB16} enhancers from activating other genes outside the neighborhood.

**Transcriptional architecture of MYC and EZH2 loci**

\textit{MYC} is an oncogene that is mis-regulated in multiple cancers, including PCa. Tandem duplications involving the \textit{MYC} neighborhood are frequently observed in mCRPC (4, 15). We queried RNA Pol II ChIA-PET datasets to examine the landscape of chromatin interactions surrounding the \textit{MYC} gene. We observed extensive RNA Pol II associated chromatin interactions within the 1.8 Mb sub-chromosomal region harboring the \textit{MYC} oncogene (Figure 6A). The anchor peaks were characterized by the H3K27ac mark, indicative of transcriptional regulation by enhancers, and were highly cell-type specific. In RWPE-1, LNCaP and VCaP cells, the interactions spanned both the 5’ upstream and 3’ downstream regions of \textit{MYC}, whereas in DU145 cells, the interactions were comparatively restricted to the 3’ downstream region of \textit{MYC}. Interestingly, in LNCaP and VCaP cells, the RNA Pol II chromatin interactions surrounding \textit{MYC}, exhibited significant overlap and also displayed AR and FOXA1 occupancy (Figures S8A and S8B). Thus, it is conceivable that binding of AR and
FOXA1 facilitates the site-specific recruitment of coactivators, resulting in the formation of H3K27ac mark and enhancer activation. These activated enhancers in turn interact with the MYC promoter and transmit the signal to modulate RNA Pol II activity. Consistent with our study, recent HiChIP studies of H3K27ac in leukemia and lymphoma cells, and of cohesin in HCT-116 cells, reported that the enhancer landscape of the MYC locus is cell-type specific (22, 23). Analysis of the active enhancer mark H3K27ac in about 100 cell types revealed distinct patterns of enhancer activity on both sides of the MYC gene (Figure S8E). Due to the unavailability of RNA Pol II ChIA-PET data for most of these cell lines, we assigned the maximum enhancer peak score in the gene neighborhood as a surrogate for enhancer activity. We observed a positive correlation between enhancer activity and MYC transcript abundance (Figure S8F). As MYC is expressed in most human cancers, we propose a working model that depicts cell-type specific transcription factors specifying enhancer activation at their cognate binding sites in the neighboring sub-chromosomal region, and resulting in transcriptional regulation of MYC via chromatin interactions.

Another interesting feature of the MYC locus is that the region is devoid of protein coding genes but has an abundance of non-coding RNAs. We noticed that several MYC interacting anchor regions represent genes that encode non-coding RNAs. For example, the long intergenic non-coding RNA, PCAT-1, which is over-expressed in PCa (24), is associated with AR and FOXA1 binding, and is enriched with the H2K27ac mark (Figures S8A and S8B). PCAT-1 and its neighboring enhancers directly interact with the MYC promoter in VCaP cells. PCAT-1 and its neighboring enhancers show indirect interaction with the MYC promoter in LNCaP cells; these interactions converge
at an intervening DNA region which is flanked by germline PCa risk predisposing SNPs. The chromosome 8q24 harbors risk alleles for multiple cancers, including PCa (25). Several PCa risk SNPs from GWAS (n = 16) lie in the region upstream of MYC, seven of which are located within enhancer elements in one or more cell types analyzed (Figure 6B, left). Four of these seven SNP loci are also associated with chromatin interactions (Figure 6B, right). In summary, we conclude that many PCa risk loci in the MYC sub-chromosomal region are located on or near enhancers and are associated with RNA Pol II chromatin interactions.

The histone methyl transferase, EZH2, is over-expressed in several solid tumors, including PCa, however, the mechanisms underlying its transcriptional regulation are not completely understood (26). By integrative analysis of RNA Pol II ChIA-PET data, we identified chromatin interaction involving a H3K27ac marked enhancer and the EZH2 promoter in RWPE-1, LNCaP, VCaP, and DU145 cells (Figures 6C, 6D, S8C and S8D). The strength of this interaction was significantly higher in VCaP and DU145 cells. Conversely, the EZH2 promoter interacted with a single enhancer in benign RWPE-1 cells, but with multiple additional distant enhancers in the cancerous LNCaP, VCaP and DU145 cells. These enhancers in turn interact with additional modules of enhancers, promoters and gene bodies, resulting in complex transcriptional hubs. Thus, the strength of individual enhancer-promoter interactions and modular assembly of interacting enhancers govern EZH2 regulation in PCa cells. We analyzed the enhancer mark, H3K27ac, in the EZH2 neighborhood in a panel of cell lines (Figure S8G). Again, by assigning the maximum enhancer peak score in the gene neighborhood as a
surrogate for enhancer activity, we observed a positive correlation between enhancer activity and EZH2 transcript abundance (Figure S8H).

**Germline PCa risk variants and transcriptional regulation**

In most GWAS, disease-associated SNPs map to non-coding regions, creating bottlenecks in interpreting their functional role. We leveraged our RNA Pol II ChIA-PET datasets to interpret the function of PCa risk SNPs identified from GWAS. A curated list of 122 PCa risk SNPs from multiple GWAS was analyzed (27) to assess how risk alleles impact transcription control. We hypothesized that many risk alleles would be located in enhancers and/or promoters and hence influence transcriptional regulation of target genes by altering chromatin interactions. To understand the functional relevance of PCa risk loci in the context of genome architecture, we applied hypothesis testing using both simulation experiments and enrichment analysis.

We computationally partitioned the genome into several thousand bins and conducted simulations to test the observed versus expected values for (A) RNA Pol II peaks that overlap with PCa risk SNPs, and (B) chromatin interaction associated RNA Pol II peaks that overlap with PCa risk SNPs. The observed overlap between RNA Pol II peaks and PCa risk SNPs in RWPE-1, LNCaP, and VCaP cells was significantly greater than the expected values derived from 10,000 simulations (Figure 7A). The same trend was also observed for the subset of RNA Pol II peaks involved in chromatin interactions in RWPE-1, LNCaP and VCaP cells (Figure 7B). Interestingly, both RNA Pol II peaks, as well as chromatin interaction associated RNA Pol II peaks in AR non-expressing DU145 cells did not show a statistically significant enrichment for PCa risk SNPs. The
results obtained with computational simulations were largely recapitulated by conducting enrichment analysis using Fisher’s exact test (Figure 7C). These analyses indicate that RNA Pol II bound regions and chromatin interactions are enriched for PCa risk SNPs, both in the normal prostate epithelial cells as well as AR positive PCa cells. The absence of this enrichment in the AR negative DU145 cells provided both a valuable internal control and deeper biological insight. These results also imply that PCa risk alleles are not enriched in transcription associated cis elements for all cell types; rather this enrichment is restricted to certain lineages.

**Transcriptional regulation by the PCa risk SNP rs684232**

The PCa risk SNP rs684232 has been reported to function as an expression quantitative trait locus (eQTL) (28-31). We delved deeper into the mechanism by which this SNP impacts PCa risk by utilizing our RNA Pol II ChIA-PET data. The risk SNP rs684232 is located ~1 Kb upstream of the VPS53 transcriptional start site (TSS) in a proximal enhancer element. RNA Pol II ChIA-PET analysis revealed chromatin interactions between the VPS53 promoter and the adjacent genes, FAM57A and GEMIN4, indicating that these genes are co-regulated transcriptionally (Figure 8A). To assess the clinical significance of the chromatin interactions, we examined patient tumor data from three clinical cohorts—The Cancer Genome Atlas (TCGA), Canadian Prostate Cancer Genome (CPC-GENE), and the Porto cohort (8, 32, 33). By stratifying the tumors into AA, AB and BB genotypes, with A and B representing the wild-type and the risk allele, respectively, we observed that the presence of the risk allele was associated with decreased expression of VPS53, FAM57A and GEMIN4 genes in all the three clinical cohorts with statistical significance (Figure 8B). These data indicate that
the risk SNP rs684232 is likely to down-regulate these three neighboring genes. The locus harboring the risk SNP rs684232 has the H3K27ac enhancer mark and is co-occupied by AR, FOXA1 and ERG in VCaP cells (Figure 8C). We examined the epigenetic correlates of the germline risk SNP rs684232 in the Porto cohort by stratifying the patient tumors into AA, AB and BB genotypes (8). The presence of the risk allele was associated with decreased H3K27ac levels in the locus (Figure 8D). Furthermore, the presence of the risk allele was also associated with decreased H3K27ac levels in the \textit{FAM57A} and \textit{GEMIN4} genes (Figure 8E). These results suggest that the risk allele manifests transcriptional down-regulation by reducing enhancer activity. Next, we studied the effect of the wild-type and the risk allele in enhancing gene transcription by conducting dual-luciferase reporter assays in LNCaP and VCaP cells (Figures 8F and 8G). The wild-type allele markedly increased the reporter luciferase activity; the magnitude of the effect was significantly reduced in the presence of the risk allele in both the cell lines. Thus, by integrating cell-based assays and population genomics, we have unraveled novel mechanistic features of transcriptional regulation by the PCa risk SNP rs684232.

\textbf{The regulation and clinical significance of the transcriptional targets of the PCa risk SNP rs684232}

We next hypothesized that master transcription factors cooperatively act on enhancer elements to up-regulate the expression of the risk SNP rs684232 target genes—\textit{VPS53}, \textit{FAM57A} and \textit{GEMIN4}. SiRNA-based knock-down of AR resulted in the down-regulation of \textit{VPS53}, \textit{FAM57A} and \textit{GEMIN4} transcription in LNCaP cells (Figure 9A). To explore the relationship between the presence of the risk SNP rs684232 and
AR occupancy in the locus, we examined the Porto cohort by stratifying the patient
tumors into AA, AB and BB genotypes. The presence of the risk allele was associated
with decreased AR occupancy in the locus (Figure 9B). These results indicated that,
mechanistically, the risk SNP rs684232 diminished AR binding and enhancer activation,
resulting in the down-regulation of the three target genes.

We also addressed the role of ERG in the framework of transcription control by
the risk SNP rs684232. As ERG expression is a gene-fusion mediated acquired somatic
event, while the risk allele is inherited from the germline, the simultaneous presence of
these two properties may have opposing effects in terms of expression of the three
target genes. We further stratified the CPC-GENE cohort into ERG positive and
negative groups. Remarkably, the association between the presence of the risk allele
and the down-regulation of the target genes was more profound in ERG positive tumors
than in ERG negative tumors (Figure 9C). Thus, we conclude that the effect of the risk
allele rs684232 on its target genes is further modulated by acquired somatic events like
ERG gene fusions.

We observed a decrease in the transcript abundance of VPS53, FAM57A and
GEMIN4 genes with an increase in the PCa International Society of Urological
Pathology (ISUP) grade group—the effect was most prominent for GEMIN4 (Figure 9D).
We also stratified the PCa patient tumors on the basis of VPS53, FAM57A and GEMIN4
transcript abundance. Lower GEMIN4 expression in tumors was associated with shorter
biochemical recurrence (BCR)-free survival (Figure 9E). Although a weak trend was
also observed with VPS53 and FAM57A, this did not reach statistical significance.
Consistent with these findings, siRNA-based knock-down of VPS53, FAM57A and
GEMIN4 in LNCaP cells resulted in a modest increase in cell viability, both individually as well as in combination (Figure S9).

**DISCUSSION**

In this study, we describe the landscapes of transcription in PCa and have identified thousands of RNA Pol II associated long-range chromatin interactions—with implications ranging from basic biology to cancer etiology and risk. We have shown that RNA Pol II associated chromatin interactions strongly overlap with the H3K27ac marks, which are indicative of active enhancers. The enrichment of transcription factor binding motifs in enhancers provides a flexible mechanism for cells to change their state by altering their pool of transcription factors and/or their DNA binding sites—this feature is constantly exploited in cancers.

The physical nature and temporal dynamics of chromatin interactions are still elusive. These chromatin interactions can also be promoted by bromodomain and extraterminal (BET) proteins. Due to the presence of tandem acetyl lysine binding bromodomains, BET proteins can function as adaptors to connect enhancers with promoters; both enhancers and promoters are rich in acetylated proteins. In addition to cooperative interactions via specific structured bromodomains, it has not escaped our notice that the low complexity regions/disordered domains of BET proteins and perhaps other transcriptional regulators can also enhance chromatin interactions via liquid-liquid phase separation, and this could be further regulated by post-translational modifications (34-36). At present, it is not clear if the DNA regions in between enhancers and promoters loop out or if these are associated with weak, non-specific chromatin
interactions. Clearly, there is much complexity associated with such chromatin interactions. This complexity is necessary, as a single genotype gives rise to multiple phenotypes/cell-types at the organismal level. However, underlying this complexity is a simple design principle for transcriptional regulation conserved from bacteria to humans—genome-wide binding of trans-acting factors and cognate cis-regulatory elements to specify which genes are turned on and off by RNA Polymerase at any given time. This design and the genome organization is progressively altered during cancer development.

By integrative analysis, we have outlined the architectural features of transcription control. The RAD21 interactions associated with CTCF binding provide a structural framework for RNA Pol II interactions associated with the H3K27ac mark. RAD21 interactions are in general longer than RNA Pol II interactions. RAD21 interactions may form closed genomic neighborhoods (also called insulated neighborhoods, sub TADs or chromatin contact domains) to contain RNA Pol II interactions. A significant proportion of enhancers contain CTCF elements, but promoters are not enriched for these elements. Therefore, it could be speculated that enhancers with CTCF elements connect both with promoters and CTCF containing boundary elements; the promoters will get indirectly drawn towards boundary elements due to their interaction with enhancers. Thus, multiple layers of chromatin interactions specify the transcriptional output of individual cellular lineages: enhancer-promoter interactions direct specificity of transcriptional regulation whereas enhancer-boundary interactions and boundary-boundary interactions serve as protecting moats to reduce off-target chromatin interactions and control transcriptional noise. Genomic
rearrangements, gene fusions and CTCF mutations enable cancer cells to breach the boundaries of lineage dependent transcriptional regulation (1, 17, 37).

Mis-regulation of chromatin interactions is a distinguishing feature of most cancers. For example, the TMPRSS2-ERG gene fusions are observed in 50% of human PCa (1). As a result of the gene fusion, ERG is transcriptionally activated by the promoter and enhancer elements of the AR target gene, TMPRSS2. Most PCa treatment strategies target the AR signaling axis. Hence, it is critical to understand the transcriptional regulation of, and by, AR. We describe the identification of novel long-range AR enhancer clusters. AR locus amplification is associated with increased interaction between AR enhancers and promoter. Analysis of mCRPC specimens suggests that AR locus amplification frequently involves AR enhancer amplification. Thus, enhancers regulate AR function in cell line models and clinical PCa specimens.

We delineate general principles of oncogene activation. The genomic location of the MYC locus facilitates its activation by multiple enhancers in various cancers due to the availability of multiple transcription factor binding sites. Remarkably several PCa GWAS alleles lie on MYC enhancers suggestive of altered chromatin interactions and transcriptional mis-regulation. Cooperative action of multiple enhancers underlies EZH2 up-regulation in PCa. More generally, we suggest that RNA Pol II network hubs provide multiple nodes for receiving, processing and transmission of regulatory signals, thereby contributing to coordinated transcription as well as regulation of dynamic processes like transcriptional bursting (38, 39). Multi-nodal transcriptional hubs are likely to be associated with sub-clonal transcriptional variability in tumors, resulting in the selection of drug tolerant clones and the emergence of resistance to targeted therapies (40, 41).
We have shown that germline PCa risk alleles are enriched in the enhancers found in non-malignant basal like RWPE-1 prostate epithelial cells, as well as luminal AR positive cancer cells, but not AR negative cancer cells that have diverged transcriptionally. Unlike mutations in exons, which typically affect individual genes, PCa risk alleles like rs684232 can simultaneously alter the expression of multiple genes. Importantly, our discovery of a genetic interaction between the germline risk allele rs684232 and the somatically acquired TMPRSS2-ERG indicates a potential role of epistasis and modifier genes in modulating PCa risk. These results have implications for the development of intervention strategies to prevent or delay the onset of PCa.

In summary, we have described the genome architectural features of transcription control in multiple PCa cellular models, integrated the results with existing datasets, expanded our discoveries to clinical specimens, and finally made relevant connections to PCa germline susceptibility alleles. We anticipate that this work will create fertile avenues for future research in transcriptional regulation and cancer development.

METHODS

Cell Lines. The cell lines used in this study (RWPE-1, LNCaP, VCaP and DU145,) were procured from American Type Culture Collection (ATCC) and grown as per their recommendations. All cell lines were verified by genotyping.

Chromatin interaction analysis by paired-end tag sequencing (ChIA-PET). ChIA-PET was performed using published methods with a few modifications (42). Briefly, about 200 million cells per cell line were cross-linked using 1.5mM EGS (35 minutes), and 1%
Formaldehyde (8 minutes), and then quenched using 0.125M Glycine (5 minutes) at room temperature. The cells were lysed and the chromatin was sheared using a HighCell ChIP kit protein G (Diagenode) and used for Chromatin Immunoprecipitation (ChIP). The anti-PolII monoclonal antibody 8WG16 (Biolegend) was used. A quality check for the ChIP product was carried out by qPCR using the primers GAPDH Promoter F: 5' TACTAGCGGTTTTACGGGCG, GAPDH Promoter R: 5' TCGAACAGGAGGAGCAGAGCGA as a positive control and GAPDH Exon8 F: 5' CCATCACTGCCACCCAGAAG, GAPDH Exon 8 R: 5' AGCTTCCCGTTCAGCTCAGG as the negative control. The ChIP DNA was quantified by flurometry (Qubit, Invitrogen), and end blunted, A-tailed and proximity ligated overnight with a biotinylated bridge linker. The DNA was reverse cross-linked, fragmented using transposase (Nextera DNA Library Preparation Kit, Illumina) and purified (Genomic DNA Clean & Concentrator, Zymo Research). Libraries were constructed by using a minimal number of cycles for PCR amplification, and the PCR products were purified, size selected (Agencourt Ampure XP, Beckman Coulter) and quality checked using Tape Station (High sensitivity D1000 Screentape, Agilent). Paired-End sequencing (2X150bp) in Nextseq 500 (Illumina, CA) was carried out to yield approximately 70-80 million reads per sample.

**ChIA-PET data processing and analysis.** Data processing of ChIA-PET reads was performed by using ChIA-PET2 software (43) to obtain the binding peaks and the interactions among peaks. Pair-end read (PET) sequences were scanned to filter the bridge linker sequence ‘CGCGATATCTTATCTGACT’. The filtered reads were mapped to the human reference genome (hg19) and only uniquely aligned (MAPQ ≥ 30) PETs were retained. By evaluating the genomic locations of the two ends of a PET, each PET
was categorized as either intra-chromosomal PET (two ends in the same chromosome) or inter-chromosomal PET (two ends in different chromosomes). All intra-chromosomal PETs were further separated as a self-ligation PETs (two ends in same peak region) or regular intra-chromosomal PETs (between two different peak regions). This data was used to define the interaction strength between two peak regions. The RNA Pol II binding peaks was then compared between different cell lines to identify commonalities and the R package Vennerable was used to generate the Venn diagram.

**Difference of Negative Binomial distributions (DNB) model.** For a pair of interacting peaks, the number of PETs between them is defined as their interaction strength. We found that the interaction strengths of a cell line follow negative binomial (NB) distribution, which is commonly observed in analysis of RNA-seq and ChIP-seq data (44, 45). Since there is a lack of computational methods for directly comparing two samples of ChIA-PET data that are fitted as NB distributions, we introduced a statistical model, named DNB model (Difference of Negative Binomial distributions), to compare two general NB distributions whose parameters are real positive numbers. Suppose $X$, $Y$ are the random variables of two NB distributions, where $X \sim \text{NB}(r_1, p_1)$, $Y \sim \text{NB}(r_2, p_2)$. We calculated the probability function of their difference $Z = X - Y$. By using the convolution formula, we have

$$P(Z = k) = \sum_{y=0}^{\infty} P(Y = y, X = k + y) = \sum_{y=0}^{\infty} P(Y = y)P(X = k + y).$$

We then calculated the probability mass function (PMF) of $Z$ as
For a peak pair $i$, we can calculate their difference of interaction strength $X_i$ and $Y_i$ for two samples. We then calculate the probability of $P(Z > X_i - Y_i)$ to see if the difference is significant or not. Clearly, smaller the p-value, the more significant is the difference of the interaction strengths between two samples. With a significance level of 0.05 as threshold, we obtained large number of peaks pairs whose interactions are significantly changed by comparing LNCaP, VCap, DU145 with RWPE-1 cells. For the peak pair with significant different interactions, their nearest genes were extracted for functional analysis of GO terms and pathways by using PANTHER database (46).

**Data and software availability.** The accession numbers for the ChIA-PET, ChIP-seq and RNA-seq data are GEO: GSE121020, GSE121021, and GSE121022, respectively. The accession numbers for the aCGH and WGS data are ArrayExpress: E-MTAB-7326 and European Nucleotide Archive (ENA): ERS2773662, respectively.

**Statistics.** All quantitative reverse transcription polymerase chain reaction (qRT-PCR) experiments were performed in triplicate, and quantitative data are presented as the mean ± SD. All reporter luciferase assays were performed in six replicates, and quantitative data are presented as the mean ± SD. Two-tailed Student’s t-tests were used to determine significance for qRT-PCR and reporter luciferase assays, and P values of less than 0.05 were considered statistically significant.

$$P(Z = k) = p_1^k(1 - p_1)^{r_1}(1 - p_2)^{r_2} \sum_{y=0}^{y} \frac{(p_1 p_2)^y}{y! (k + y)!} \frac{(y + r_2 - 1)! (k + y + r_1 - 1)!}{(r_1 - 1)! (r_2 - 1)!}.$$
Study Approval. All procedures involving human subjects were approved by the Royal Marsden NHS Foundation Trust Hospital (London, UK) ethics review committee (reference no. 04/Q0801/60).

AUTHOR CONTRIBUTIONS


ACKNOWLEDGEMENTS

We thank James Malter, Diego Castrillon and James Brugarolas for insightful comments and discussion; Weibo Luo for assistance with luciferase assays; Xiangyi Li for technical assistance. This work was supported by the NIH Pathway to Independence (PI) Award (R00CA160640), the NIH/NCI R01 grant (R01CA245294), the CPRIT Individual Investigator Research Award (RP190454), the US Department of Defense Prostate Cancer Research Program (PCRP) - Impact Award (W81XWH-17-1-0675), Kidney Cancer Specialized Programs of Research Excellence (SPORE) Career Enhancement Program (CEP) award and UT Southwestern startup funds to R.S.M. M.Q.Z. thanks the
support from the UTD startup fund, Cecil H. and Ida Green Endowment, the Cancer Prevention and Research Institute of Texas (CPRIT) grant (RP180826), and NSFC grants (31671384,91329000). Work in the J.D.B. laboratory was supported by funding from the Movember Foundation, Prostate Cancer UK, US Department of Defense, the Prostate Cancer Foundation, Stand Up To Cancer, Cancer Research UK, and the UK Department of Health through an Experimental Cancer Medicine Centre grant. A.S. is supported by the Medical Research Council, the Academy of Medical Sciences, Prostate Cancer UK and Prostate Cancer Foundation. P.C.B. was supported by Prostate Cancer Canada, a Movember Prostate Cancer Canada Rising Star Award and by CIHR and TFRI New Investigator Awards. G.V.R. acknowledges support from the US Department of Defense, the Wilson Foundation and the Mimi and John Cole Foundation. The results described here are in part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov.

Address correspondence to: Ram S. Mani, Department of Pathology, UT Southwestern Medical Center, 5323 Harry Hines Blvd NB6.444, Dallas, TX 75390, USA. Phone: 214-645-7007. Email: ram.mani@utsouthwestern.edu. Or to: Michael Q. Zhang, Department of Biological Sciences, Center for Systems Biology, The University of Texas at Dallas, 800 W. Campbell Road, Richardson, TX 75080, USA. Phone: 972-883-2523. Email: michael.zhang@utdallas.edu.

REFERENCES


Figure 1. Analysis of RNA Pol II associated chromatin interactions.

(A) The pipeline for ChIA-PET data processing and identification of chromatin interactions. (B) Enhancers per gene (left panel) and genes per enhancer (right panel) for each cell line. For enhancers per gene, all expressed genes (FPKM>1) were included; for genes per enhancer, all enhancers located in RNA Pol II peak anchor regions were included. Boxplot represents median, 25% and 75% quantiles with lines at 1.5x interquartile range. The significance for each pair comparison was tested by Kolmogorov–Smirnov test, and *** represents p-value < 2.2E-16. The analysis was normalized by adjusting for sequencing depth. (C) Correlation between RNA expression and chromatin interaction across the four cell lines. The expression level was measured by FPKM transformed by log2, and chromatin interactions were measured by the number of promoter PETs. The correlation efficient was calculated by spearman correlation and p-value is labeled. (D) Significant gain and loss of RNA Pol II associated chromatin interactions in the four cell lines. The three solid lines show the distributions of the three PCa cell lines compared with benign cells (RWPE-1), while the three broken lines show the distributions among the three PCa cell lines. The significant gain/loss interactions were obtained by using DNB model. The total number of interactions that are significantly altered between cell lines are listed accordingly. (E) The scatterplot shows the correlation between changes in gene-expression and changes in promoter associated chromatin interactions (left panel; VCaP vs RWPE-1). The spearman correlation coefficient for top 10%, 20%, 40%, 60% and 80% differentially expressed genes (right panel). (F) Total number of PETs at promoter regions in the four cell lines. Genes are ranked by increasing number of PETs. Eight representative genes are labeled as 1 to 8 in the plots.
This page contains a variety of diagrams and tables related to gene expression and chromatin interactions.

**Figure 2**

### Panel A
- Heat map showing RNA Pol II and H3K27ac peaks across different cell types (LNCaP, VCaP, RWPE-1).
- Interaction peaks are highlighted in different colors for each cell type.

### Panel B
- Heat map showing Dnase-Seq and H3K4me3 peaks for LNCaP and VCaP.
- Interaction peaks are shown for specific regions.

### Panel C
- Heat map showing CTCF and RAD21 ChIA-PET peaks for LNCaP and DU145.
- Interaction peaks are highlighted in different colors.

### Panel D
- Bar graphs showing the length of PETs (kb) for RAD21 and RNA Pol II in LNCaP.
- p-value < 2.2e-16 indicated.

### Panel E
- Heat map showing RNA-Seq, ChIP-Seq, and RNA Pol II ChIA-PET peaks for LNCaP.
- RefSeq genes are highlighted.

### Panel F
- Heat maps showing RNA Pol II ChIA-PET and RAD21 ChIA-PET for Chr12.
- Data resolution: 10kb.
Figure 2. Integrative analysis of chromatin interactions.

(A) Integrated genome view representing DNase-seq, ChIP-seq for CTCF, H3K4me3 and H3K27ac, RNA-seq, ChIA-PET for RNA Pol II and RAD21, and HiC data from the LNCaP cell line for a representative region with both active and repressed genes. (B) Overlap analysis between H3K27ac ChIP-seq and RNA Pol II ChIA-PET data in four cell lines. Left: overlap in the number of total peaks; right: overlap in the number of peaks with intra-chromosomal interactions. (C) Overlap analysis between CTCF ChIP-seq and RAD21 ChIA-PET data in LNCaP and DU145 cells. Left: overlap in the number of total peaks; right: overlap in the number of anchor peaks with intra-chromosomal interactions. (D) Comparison of PETs length between RNA Pol II ChIA-PET and RAD21 ChIA-PET data from LNCaP and DU145 cells. Boxplot represents median, 25% and 75% quantiles with lines at 1.5x interquartile range. The p-value was calculated by Kolmogorov–Smirnov test. (E) Integrated genome view of gene KRT8-KRT18 gene cluster and its neighborhood in LNCaP cells. The data tracks represent RNA-seq, ChIP-seq for CTCF, FOXA1, AR, H3K27ac and RNA Pol II, ChIA-PET of RNAPol II and RAD21, and HiC data in the LNCaP cell line. (F) ChIA-PET contact heatmap representing RNA Pol II (left) and RAD21 (right) associated chromatin interactions for the KRT8-KRT18 gene cluster and neighborhood regions. The KRT8-KRT18 gene cluster is shown in light green color.
Figure 3
Figure 3. Transcriptional regulation of AR and FOXA1 loci.

(A-B) Integrated genome view of AR gene and its adjacent regions from -1400 Kb to +400 Kb in LNCaP and VCaP cells. RNA Pol II ChIA-PET, RNA-seq and CTCF, FOXA1, AR, H3K27ac and RNA Pol II ChIP-seq data are described; in addition, AR ChIA-PET, phospho RNA Pol II and ERG ChIP-seq is described for VCaP cells. The AR gene and up-stream regions are highlighted in light-blue color. (C) Summary of the copy number aberrations associated with the AR and its enhancers. A heatmap of the high dense probes array comparative genomic hybridisation (aCGH) of 27 mCRPC patients is shown. Gains are depicted in pink, losses in light blue whereas amplifications are in red and deep deletion in dark blue. Each row is a probe in the aCGH platform and each column represents a sample. Probes that cover the EDA2R and AR and regions of the enhancer peaks are shown on the left of the heatmap. (D) Schematic representation of a deletion between AR gene and its enhancers. (E) Comparison of RNA Pol II associated chromatin interactions at the FOXA1 locus and its adjacent regions in the four cell lines.
Figure 4

A

B

C

LCaP

\(-\log_{10}(p\text{-value})\)

\begin{itemize}
  \item \text{LCaP AR}
  \item \text{LCaP FOXA1}
  \item \text{LCaP AR+FOXA1}
  \item \text{VCaP AR}
  \item \text{VCaP FOXA1}
  \item \text{VCaP AR+FOXA1}
\end{itemize}

Pathway Enrichment Analysis

LCaP

\begin{itemize}
  \item EIF2 Signaling
  \item ILK Signaling
  \item ATM Signaling
  \item p53 Signaling
  \item mTOR Signaling
  \item Regulation of eIF4 and p70S6K Signaling
  \item Wnt/β-catennin Signaling
  \item Role of NFAT in Regulation of Immune Response
  \item Cyclins and Cell Cycle Regulation
  \item GADD45 Signaling
  \item Hereditary Breast Cancer Signaling
  \item Prostate Cancer Signaling
  \item Corticotropin Releasing Hormone Signaling
  \item Sumoylation Pathway
  \item Sirtuin Signaling Pathway
  \item Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis
  \item Cardiac Hypertrophy Signaling
  \item p38 MAPK Signaling
  \item Protein Kinase A Signaling
  \item Colorectal Cancer Metastasis Signaling
\end{itemize}

VCaP

\begin{itemize}
  \item EIF2 Signaling
  \item mTOR Signaling
  \item IGF-1 Signaling
  \item Calcium Transport I
  \item ILK Signaling
  \item HIPPO Signaling
  \item NAD Phosphorylation and Dephosphorylation
  \item Regulation of eIF4 and p70S6K Signaling
  \item p53 Signaling
  \item Regulation of eIF4 and p70S6K Signaling
  \item Dopamine-DARPP32 Feedback in cAMP Signaling
  \item Protein Kinase A Signaling
  \item Endoplasmic Reticulum Stress Pathway
\end{itemize}
Figure 4. Chromatin interaction associated transcriptional targets of AR and FOXA1.

(A) The expression of AR/FOXA1 target genes discovered by integrating RNA Pol II ChIA-PET with AR/FOXA1 ChIP-seq are compared to genes that are nearest to AR/FOXA1 binding peaks from ChIP-seq data and randomly selected control genes in VCaP (top panel) and LNCaP (bottom panel) cells. The y-axis represents expression level, measured as FPKM transformed by log2. Boxplot represents median, 25% and 75% quantiles with lines at 1.5x interquartile range. The p-value was calculated by Kolmogorov–Smirnov test (*** represents p-value is less than 2.2E-16). (B) Gene promoters that interact with AR, FOXA1, and AR-FOXA1 co-occupied regions in the RNA Pol II ChIA-PET datasets are shown in yellow. Gene promoters that interact with AR, FOXA1, and AR-FOXA1 co-occupied enhancers in the RNA Pol II ChIA-PET datasets are shown in orange. (C) Pathway analysis for gene promoters that interact with AR-FOXA1 co-occupied regions in the RNA Pol II ChIA-PET datasets. The left and right panels represent LNCaP and VCaP cells, respectively.
Figure 5. Transcriptional regulation the Kallikrein (KLK) gene cluster and ZBTB16.

(A) Comparison of RNA Pol II associated chromatin interactions and H3K27ac ChIP-seq signals in KLK3 gene and its neighborhood regions from -350 Kb to +350 Kb in four cell lines. (B-C) Transcriptional regulation of ZBTB16 gene in (B) LNCaP and (C) DU145 cells.
Figure 6

A

Chr8

RWPE-1: RNA-Seq
RWPE-1: H3K27ac
RWPE-1: ChIA-PET

LNCaP: RNA-Seq
LNCaP: H3K27ac
LNCaP: ChIA-PET

VCaP: RNA-Seq
VCaP: H3K27ac
VCaP: ChIA-PET

DU145: RNA-Seq
DU145: H3K27ac
DU145: ChIA-PET

RefSeq Genes

SNPs

SNPs overlapped with peaks
SNPs overlapped with interactions peaks

B

C

Chr7

RWPE-1: RNA-Seq
RWPE-1: H3K27ac
RWPE-1: ChIA-PET

LNCaP: RNA-Seq
LNCaP: H3K27ac
LNCaP: ChIA-PET

VCaP: RNA-Seq
VCaP: H3K27ac
VCaP: ChIA-PET

DU145: RNA-Seq
DU145: H3K27ac
DU145: ChIA-PET

RefSeq Genes

SNPs ID
1: rs12543663
2: rs10086908
3: rs1016343
4: rs1456315
5: rs13254378
6: rs6983561
7: rs16901979
8: rs10505843
9: rs16902094
10: rs445114
11: rs620661
12: rs6983267
13: rs1447295
14: rs4242382
15: rs4242384
16: rs7837688

D

EZH2

Actin
Figure 6. Transcriptional regulation of the MYC and EZH2 genes.

(A) Comparison of chromatin interactions, H3K27ac ChIP-seq and RNA-seq signals of MYC gene and its adjacent regions from -1000kb to +600 kb in four cell lines. The MYC gene is highlighted in light-blue color. PCa risk SNP loci located in the MYC neighborhood are shown. (B) Left panel: venn diagram describes the number of PCa risk SNP loci located within the coordinates of RNA Pol II peaks for three cell lines. Right panel: venn diagram shows the number of PCa risk SNP loci located within the coordinates of anchor regions of RNA Pol II associated chromatin interactions for three cell lines. (C) Transcriptional regulation of EZH2 gene and its neighborhood genes. Comparison of RNA Pol II ChIA-PET, H3K27ac ChIP-seq and RNA-seq signals at EZH2 locus and its adjacent regions from -150kb to +800 kb in four cell lines. (D) Immunoblot representing the endogenous EZH2 expression in RWPE-1, LNCaP, VCaP and DU145 cells. The EZH2 and actin blots were obtained from separate gels which were run contemporaneously.
Figure 7

A

B

C

Non-peak Regions
Peak Regions
Non-interaction Regions
Interaction Regions

% Containing SNPs

0 0.005 0.010 0.015 0.020 0.025 0.030

RWPE-1 LNCaP VCaP DU145

Number of Germline SNPs (122) Number of Germline SNPs (122) Number of Germline SNPs (122) Number of Germline SNPs (122)

Simulations
Figure 7. Evaluation of RNA Pol II associated peaks and interaction with 122 prostate-specific germline SNP locations.

(A) Peak analysis. The red lines indicate the observed number of peaks containing SNPs for each cell line. The histograms illustrate the results from 10,000 simulations that assessed the expected number of peaks containing SNPs. The mean of the simulations is the black line. Significant differences between the expected and observed values are indicated by stars. Three stars indicate a p-value less than 0.001. RWPE-1, LNCaP, VCaP, and DU145 have 17, 13, 14, and 4 peaks that overlap SNPs, respectively. The black dotted lines indicate the expected number of overlapping peaks (the mean of all the simulations). The expected values for RWPE-1, LNCaP, VCaP, and DU145 are 4.88, 4.09, 4.03, 3.44. (B) Interaction analysis. The same procedure was repeated except using only the peaks involved in interactions. Red lines indicate the observed number of SNPs, and black lines show expected values. Stars indicate a significant difference between the expected and observed values. One star indicates a p-value less than 0.05, and three stars indicate a p-value less than 0.001. RWPE-1, LNCaP, VCaP, and DU145 have an observed value 10, 9, 5, and 1, as indicated by the red dotted lines. The expected values for RWPE-1, LNCaP, VCaP, and DU145 are indicated by the black dotted lines and is 1.81, 2.17, 2.60, and 1.30, respectively. (C) Enrichment analysis. Fisher enrichment analysis was performed to compare the number of SNP-positive peaks to the rest of the genome as well as to compare the number of SNP-positive, interaction-positive peaks to the rest of the genome. Three stars indicate significant enrichment with a p-value less than 0.001.
Figure 8

A. Chr17

B. CPC-GENE

C. FOXA1

D. VCaP ChIP-Seq

E. FAM57A

F. rs684232

G. LNCaP, VCaP

H. Genotype

I. mRNA abundance

J. SNP

K. rs684232

L. FOXA1

M. VPS53

N. H3K27ac

O. RefSeq Genes

P. ChIP-Seq

Q. FOXA1

R. AR

S. H3K27ac

T. ERG

U. RNAPII

V. RNAPI

W. phospho

X. SNP

Y. rs684232

Z. Patient

aa. Genotype

bb. mRNA abundance

cc. SNP
Figure 8. Transcriptional regulation by the PCa risk SNP rs684232.

(A) Integrated genome view of RNA Pol II associated chromatin interactions in the genomic region harboring the PCa-risk SNP rs684232. (B) rs684232 is significantly associated with mRNA abundance of FAM57A, VPS53 and GEMIN4 in the CPC-GENE, TCGA and Porto cohorts. Boxplot represents median, 0.25 and 0.75 quantiles with whiskers at 1.5x interquartile range. mRNA abundance in FPKM. Numbers next to genotypes reflect number of samples in each group. P-value and effect size are from linear model. (C) Epigenetic features of the PCa-risk SNP rs684232 locus in VCaP cells are described using ChIP-seq analysis. (D) rs684232 falls in an active enhancer region and the alternative allele is significantly associated with decreased H3K27ac binding. Heatmap shows H3K27ac ChIP-seq signal within chr17:614900-622900 (x-axis) for 92 patients (y-axis). Colour indicates ChIP-seq signal intensity while black bar in covariate along the top indicates location of rs684232. Boxplot shows H3K27ac signal intensity stratified by genotype in the Porto cohort (Mann-Whitney test of recessive model). Y-axis represents the number of H3K27ac ChIP-seq read counts mapped to the SNP rs684232 region, and is normalized by the TMM method. (E) Boxplots show H3K27ac signal intensity in the promoter regions of FAM57A and GEMIN4 stratified by genotype in the Porto cohort (Mann-Whitney test of recessive model). P-value and effect size are from linear model. (F) Sequence analysis to confirm the cloning of WT and risk allele (rs684232) in the pGL2 luciferase reporter plasmid. (G) Luciferase reporter assays in LNCaP (left panel) and VCaP cells (right panel). Cells were co-transfected with pSV-Renilla and the luciferase reporter encoding WT or risk allele (rs684232) and processed 48 hr post transfection. The Firefly Luc/Renilla Luc activity was determined (mean ± SD, n = 6; ****P < 0.0001 by 2-tailed Student's t test).
Figure 9

A

Relative Fold Change

VPS53  FAM57A  GEMIN4  AR

**  ***  ***

rs684232 | VPS53

ERG negative  ERG positive

rs684232 | FAM57A

ERG negative  ERG positive

rs684232 | GEMIN4

ERG negative  ERG positive

C

rs684232 | VPS53

rs684232 | FAM57A

rs684232 | GEMIN4

D

VPS53

FAM57A

GEMIN4

E

VPS53

FAM57A

GEMIN4

**  ***  ***

ρ = -0.12  ρ = -0.21  ρ = -0.33

P = 1.21E-2  P = 2.60E-5  P = 1.36E-11

Low  98  82  79  62  49  25  7  0

High  98  92  87  64  44  20  8  0

0.6  1.0

0.8  0.4

0.2  0.0

P: 0.35  P: 0.076  P: 0.015
Figure 9. Transcriptional regulation and clinical correlates of the chromatin interaction targets of the PCa risk SNP rs684232.

(A) qRT-PCR validation of AR knockdown and the expression of VPS53, GEMIN4 and FAM57A genes upon treatment of LNCaP cells with AR siRNA. (* P<0.05, ** P<0.01, *** P<0.001 by two-tailed Student's t-test; Error bars, standard deviation of 3 technical replicates). (B) Boxplot shows AR ChIP-seq signal intensity stratified by genotype in the Porto cohort (Mann-Whitney test of recessive model). Y-axis is the number of AR ChIP-seq read counts mapped to SNP rs684232 region, and is normalized by the TMM method. Boxplot represents median, 0.25 and 0.75 quantiles with whiskers at 1.5x interquartile range. (C) Regulatory impact of rs684232 is enhanced in the presence of the TMPRSS2-ERG fusion. Boxplot shows mRNA abundance (FPKM) of each gene stratified by genotype and further split by ERG status in the CPC-GENE cohort. FC_{genotype} (fold-change) and P_{genotype} quantifies difference in mRNA abundance between ERG positive and negative individuals with same genotype (Mann-Whitney test). β_{positive} and β_{negative}, and associated p-values, quantifies eQTL within ERG positive and negative patients, respectively (linear model). (D) Boxplots show the mRNA abundance of FAM57A, GEMIN4 and VPS53 genes in PCa specimens from the TCGA cohort. Tumors are classified into various ISUP grade groups. P-value and effect size are from linear model. (E) BCR-free survival curves for PCa patient groups defined by transcript abundance for FAM57A, GEMIN4 and VPS53 genes in the CPC-GENE cohort. Logrank test was constructed to obtain the p-values.