Reconstructing protein interactions at enhancer-promoter regions in prostate cancer

Alexandros Armaos*[1] , François Serra*[2,3], Iker Núñez-Carpintero[2], Ji-Heui Seo[4], Sylvan C. Baca[4], Stefano Gustincich[1], Alfonso Valencia[2,5], Matthew L. Freedman[4], Davide Cirillo[#,2], Claudia Giambartolomei[#,1] , Gian Gaetano Tartaglia[#,1,5.6]

1 Istituto Italiano di Tecnologia, Via Enrico Melen 83, Building B, 7th floor, 16152 Genoa, Italy
2 Barcelona Supercomputing Center, Plaça Eusebi Güell, 1-3, 08034 Barcelona, Spain
3 Present: Josep Carreras Leukaemia Research Institute, Badalona, Barcelona, Spain
4 Department of Medical Oncology, The Center for Functional Cancer Epigenetics, Dana Farber Cancer Institute, Boston, MA 02215, USA.
5 ICREA - Institució Catalana de Recerca I Estudis Avançats, Pg. Lluís Companys 23, 08010 Barcelona, Spain
6 Sapienza University Rome, Biology and Biotechnologies Department C. Darwin, P.le Aldo Moro 5, 00185

** these authors contributed equally to this work*
# to whom correspondence should be addressed

Abstract

DNA-binding proteins (DBPs) and in particular transcription factors interact with enhancers and their target genes through enhancer-promoter (E-P) interactions. Technological advancements such as chromosome conformation capture allow to identify E-P interactions, but the protein networks involved have not yet been characterized. Most importantly, the role of nuclear protein networks in human diseases has been so far poorly investigated. Prostate cancer (PrCa) heritability is associated with variations in enhancers that affect specific gene expression. Here, we introduce a novel approach, called Promoter-ENhancer-GUided Interaction Networks (PENGUIN), to identify protein-protein interactions (PPI) in E-P interactions and apply it to our PrCa dataset. PENGUIN integrates chromatin interactions between a promoter and its enhancers defined by high-coverage H3K27ac-HiChIP data, with a tissue-specific PPI network inferred from DNA-binding motifs and refined with gene expression. Among a total of 4,314 E-P networks, PENGUIN performed unsupervised clustering. We functionally validated this clustering procedure by searching for enrichments of specific biological features. We first validated PENGUIN structural classification of E-P networks by showing a clear differential enrichment of the architectural protein CTCF. Next, and directly related to our PrCa case study, we observed that one of our 8 main clusters, containing 273 promoters, is particularly enriched for PrCA associated single nucleotide polymorphisms (SNPs) and oncogenes. Our approach proposes a mechanistic explanation for 208 PrCa SNPs falling either inside the binding sites of DNA-binding proteins (DBPs) or within genes encoding for intermediate proteins bridging E-P contacts. PENGUIN not only confirmed the relevance of key regulators in PrCa, but also identified new candidates for intervention, opening up new directions to identify molecular targets for disease treatment.

Introduction

Prostate cancer (PrCa) is the 2nd most common cancer in men (Rebello et al. 2021). Its distinct hormone-dependent nature is characterized by high expression and frequent genetic amplification of *AR*, regulator of homeostasis and proteases transcription, such as *KLK3* encoding PSA (Prostate-Specific Antigen), and principal therapeutically targeted oncogene in PrCa (Tan et al. 2015). Increased genetic instability resulting in chromosomal rearrangements and high frequency of mutations are deemed indicative of PrCa aggressiveness (Cancer Genome Atlas Research Network 2015) for which there is need of *ad hoc* treatments (de Bono et al. 2020). Recurrent mutations in *FOXA1*, involved in prostate organogenesis and regulator of *AR* transcription, have been observed in several populations (Adams et al. 2019; Parolia et al. 2019). Hundreds of PrCa-associated single nucleotide polymorphisms (SNPs) have been identified by genome-wide association studies (GWAS), including genomic regions within tumor suppressor genes and oncogenes, such as *MYC* (Ahmadiyeh et al. 2010). However, the functional relationship between most of these SNPs and PrCa pathophysiology is unknown. This missing part of the picture, together with the growing evidence of abnormal transcriptional programs driven by genetic instability, led us to investigate the role of chromatin architecture in PrCa. In particular we studied the nuclear proteins potentially involved in transcriptional regulation through the interaction of promoters and non-coding regulatory elements, enhancers.

Enhancer-promoter (E-P) interactions have an important role in gene regulation. DNA-binding proteins (DBPs), such as transcription factors (TFs), regulate gene expression by binding promoters to enhancers sometimes through intermediate proteins. This chromatin interaction between E-P can be mediated by co-activators (e.g., mediators), chromatin structural proteins (e.g., cohesin), noncoding RNA-binding proteins, and others. Disruption of these interactions centered in a single promoter (either protein-protein or protein-DNA interactions), which we collectively call E-P protein-protein Interaction Networks (EPIN), is increasingly linked to diseases such as cancer (Dekker and Misteli 2015; Norton and Phillips-Cremins 2017; Krumm and Duan 2019). Enhancers are often the target of sequence and structural variation in cancer, specifically deregulation of TFs and chromatin modifiers (Sur and Taipale 2016), and represent promising pharmacological targets in PrCa (Chen et al. 2020). To date, techniques such as HiC or, more specifically, its recent derivative HiChIP (Mumbach et al. 2016), HiC combined with ChIP-seq, are able to identify specific chromatin interactions between a promoter and its enhancers. For instance, H3K27ac-HiChIP is designed to specifically detect and amplify E-P interactions and it has been recently employed by our lab to uncover susceptibility genes in PrCa (Giambartolomei et al. 2021).

Although both protein-protein and protein-DNA interactions play central roles in E-P interactions, previous analyses focused on one or the other aspect. Indeed, a number of studies have focused on DBPs networks (Zhang et al. 2016; Wang et al. 2019). A more recent analysis exploited PPIs with chromosome conformation capture assays to facilitate prioritization of functional interactions (N. Liu et al. 2021). At present, the characterization of context specific intermediate protein-protein interactions (PPIs) involved in disease pathways and their association with DBPs remains largely unanswered (Deng and Blobel 2014).

To characterize protein interactions that take place at the E-P contacts in PrCa, we developed the Promoter-ENhancer-GUided Interaction Networks (PENGUIN) approach. For each promoter annotated in the genome and covered by at least one HiChIP interaction, PENGUIN builds an EPIN by integrating several sources of information: (1) high-resolution chromatin interaction maps enriched for a marker of active promoter-enhancer activity (H3K27ac-HiChiP); (2) gene expression (RNA-sequencing); (3) tissue-specific physical nuclear PPIs; (4) high-quality curated binding motifs of protein-DNA interactions. The PrCa specificity of this dataset is given by the H3K27ac-HiChiP of androgen-sensitive human prostate adenocarcinoma cells (LNCaP) and RNA-seq in the same cell line, and prostate-specific PPIs and DNA binding motifs extracted from publicly-available datasets (**Methods**).

PENGUIN is a software that can be also applied to other diseases. The method can identify clusters of PPI networks dependent on the potential binding proteins in promoters and enhancers, that are enriched for particular annotations like GWAS and CTCF. It offers a comprehensive framework to integrate disease data and provides insights into protein networks found in E-P interactions for further molecular validation. For instance, PENGUIN can be used to identify *trans*-acting factors (e.g., interaction cascades of TFs and chromatin regulators) that could be targeted by drugs, or *cis*-acting factors (e.g., DBPs with binding motifs in regulatory elements) whose DNA binding affinity could be modified through knock-outs via CRISPR for therapeutic intervention. Moreover, unlike traditional TF enrichment analysis which detects general enrichments of particular proteins, PENGUIN can help identify the specific protein cascade potentially disrupted at enhancer loci for the disease under study. This methodology identifies new directions in the molecular characterization of chromatin interactions as well as in the definition of potential targets for molecular screening towards disease treatment.

In this paper, we propose two main applications of PENGUIN. First, it facilitates the identification of key factors that may play a role in transcriptional regulation of PrCa. By clustering together promoters with similar EPINs, we identified 273 promoters whose genes are enriched in PrCa GWAS, known PrCa oncogenes, and ChIP-validated binding sites of transcriptional repressor CTCF. The proteins that populate such EPINs constitute putative PrCa-related bioentities, some of which have not been previously described to be associated with PrCa SNPs or oncogenes. Second, the EPINs detected by PENGUIN enable the characterization of distinct molecular cascades potentially affected by PrCa SNPs at E-P contacts. These represent potentially new molecular targets in PrCa that cannot be identified through conventional analytical procedures, such as E-P contacts and GWAS overlap. Finally, we provide a dedicated web server to explore the results at https://penguin.life.bsc.es/.

Results

PENGUIN identifies clusters of protein interactions based on chromatin contacts

We leveraged 24,547 E-P contacts (30,416 after refinement and prioritization, **Methods**) identified using H3K27ac-HiChIP data in LNCaP, 810 binding motifs from 639 DNA-binding proteins, and 31,944 prostate-specific, experimentally validated, physical and nuclear PPIs (filtering out proteins from unexpressed genes, **Methods**) to construct 4,314 EPINs using the PENGUIN pipeline outlined in **Figure 1** (**Methods**). Each EPIN is centered around one promoter that we found to be contacted by a median of 4 enhancers, with a maximum of 93 enhancers for the promoter of the gene *CRNDE* (**Table S1**). Altogether, the 4,314 EPINs contain a total of 8,215 interactions (edges) among a total of 885 proteins (nodes) that are expressed in LNCaP (**Methods**). A mean of 36% proteins found in these EPINs are encoded by differentially expressed genes in LNCaP versus LHSAR (**Methods** and **Table S1**).

Each EPIN is composed of three types of nodes (**Table S2**): proteins with DNA binding motifs in the promoter (promoter-bound nodes), proteins with DNA binding motifs in the enhancers (enhancer-bound nodes), and proteins interacting with promoter-bound or enhancer-bound nodes but without DNA binding motifs onto the promoter or the enhancers (intermediate nodes). Overall, 751 out of the 885 proteins represent intermediate nodes, with 127 of them acting both as intermediate and as DNA-bound nodes in different EPINs. 261 unique DNA-binding proteins have predicted binding sites in at least one of the anchors of enhancers and promoters. A mean of 32.8 (s.d. 11.5) distinct DBPs were identified per promoter anchor with SP1, EGR1, SP2 being the most represented; and a mean of 24.8 (s.d. 7.69) were predicted per enhancer anchor with SP1, IRF1 and TFAP2A being the most represented.

PENGUIN reconstructs an EPIN by grouping enhancers interacting with the same promoter based on the well-known partial redundancy of enhancers (Kvon et al. 2021). A mean of 1.43 (normalized) promoters (0.88 s.d.) are shared among enhancers, with a maximum of 15 promoters for the same enhancer. To identify communalities and differences among the 4,314 EPINs in PrCa, we performed an unsupervised, hierarchical clustering based on edge composition (Ward's linkage method, **Methods**). Using this approach, we identified 8 clusters of promoters with specific networks (**Table S1**,**Table S3, Figure 2 and Figure S1A**).

Protein networks of PrCa risk-associated regulatory elements

We characterized the 8 clusters using PrCa specific annotations. We used the 95% credible set of SNPs (henceforth referred to as PrCa SNPs) across 137 PrCa-associated regions fine-mapped from the largest publicly available GWAS summary statistics [N=79,148 cases and 61,106 controls (Schumacher et al. 2019)]. By comparing each cluster with all other clusters, we found a significant enrichment of PrCa SNPs in one specific cluster (cluster 8 or *GWAS+ cluster*; Fisher's exact test, Methods, Figures S3 B-C). Interestingly this enrichment is exclusively due to SNPs in enhancers (Table 1). Our results show that E-P interactions containing PrCa SNPs are clustered together (red branches in Figure 2A) indicating that they have similar characteristics in the way their PPI networks are wired.

We found that most edges (67.5%, or 5,550 out of 8,215 edges) are shared among all clusters. We identified the protein interactions that are enriched in each cluster and estimated the significance of overrepresentation of each edge in a cluster compared to all others (**Methods**). *GWAS+ cluster* (cluster 8 in **Figure 2**) exhibits the lowest number of promoters and distinctive network characteristics (**Table S3, Figure S1B**). But, per promoter, it displays the largest number of edges (p-value < 1e-16) and intermediate nodes (p-value < 1e-16), in line with its greater number of enhancers per promoter (p-value < 1e-16), see **Figure S2**. Moreover, the EPINs of the GWAS+ cluster have the lowest values of node-level centrality measures, namely betweenness and degree (**Figure S1C**). The degree of a node measures the amount of connections it has, while the betweenness centrality measures the amounts of shortest paths that pass through it. Low values of betweenness and degree indicate a lower amount of connections among different components and regions of the network. Betweenness and degree are significantly different across clusters (Kruskall-Wallis test p-value < 1e-16), but not with respect to the ensemble of all EPINs, which indicates that, despite the high number of shared pairwise interactions (67.5% of edges), the wiring of the cluster-specific EPINs are distinctive.

Since CTCF is a major actor in the formation and maintenance of transcriptionally productive E-P interactions (Zuin et al. 2014; Pugacheva et al. 2020), we tested the clusters identified by PENGUIN for enrichment in CTCF binding. For this analysis we used CTCF ChIP-seq peaks, from the same cell line (LNCaP), from the ENCODE project instead of predictions based on DNA-binding motifs (**Methods**). We found that the interactions with CTCF peaks, that we call CTCF+, cluster together (red branches in **Figure 2B, Figure S3A**) suggesting that the presence of CTCF in chromatin interactions results in the formation of characteristic PPI networks between the promoter and its enhancers.

CTCF+ clusters overlap the GWAS+ cluster (**Figure 2, Figures S3A to S3C**), also suggesting that CTCF-mediated interactions could be more functionally relevant to PrCa. In particular, GWAS+ cluster (representing 6% of the total number of promoters considered) is the only one presenting the unique and significant enrichment in CTCF binding, PrCa SNPs, and oncogenes (**Table 2, Table S3 , Figures S3D and S3E**). This cluster is enriched in the *Hippo signaling pathway* (KEGG:04390) (Bonferroni-corrected p-value=0.012) and the *Signal transduction pathway* (Reactome REAC:R-HSA-162582) with genes such as *FOXA1, MYC, FOS* (Bonferroni-corrected p-value = 0.047) (**Methods, Table S5**).

Interestingly GWAS+ cluster, or any other cluster did not significantly stand out in terms of overall expression level (**Figure S4A**) or, particularly in terms of fraction of differentially expressed genes (**Figure S4B**).

In conclusion, PENGUIN enabled the identification of a cluster of E-P contacts whose EPINs are uniquely enriched in PrCa SNPs, ChIP-seq CTCF binding sites, and oncogenes (a.k.a. GWAS+ cluster or cluster 8, **Figure 2** and **Table 2**). We note that using a different set of GWAS for PrCa, we also identified cluster 8 as enriched (**Methods, Table S6**).

Comparison with baseline analytical procedures

Among the 273 promoters belonging to the identified GWAS+ cluster (cluster 8 in **Figure 2A**), 11 belong to known oncogenes, *FOXA1, ZFHX3, CDKN1B, KDM6A, BRCA2, CDH1, CCND1, NKX3-1, BAG4, MYC, GATA2* (**Methods**). We compared enrichment of PrCa functional annotations in the reconstructed networks with and without inclusion of intermediate proteins. Including intermediates allows increasing the number of retrieved PrCa-related oncogenes in cluster GWAS+ from 6 to 11 and increasing significance of enrichment indicating improved specificity (**Table S4**). We then compared our results with the simple overlap of the genomic regions of E-P contacts and known oncogene promoters. **Table S1** also reports on the overlaps of E-P contacts with CTCF binding sites (in both enhancers and promoters, see **Methods**), and PrCa SNPs (in enhancers). Only 30 promoters (12 overlapping the GWAS+ cluster) would be identified that overlap both PrCa SNPs and CTCF binding sites. Of these, just 3 are promoters of known oncogenes (and only one, *ZMYM3*, is not in the GWAS+ cluster). We therefore conclude that PENGUIN, and the integration of intermediate PPI network, increases the number of promoters of candidate PrCa-related genes with the additional and unique information on their specific interactome.

Involvement of E-P protein interactomes in tumor-related functional processes

We analyzed the functional enrichment of the set of 885 proteins composing the universe of nodes used to create PPI networks in EPINs. 43 out of these 885 proteins are encoded by one of the 122 known PrCa oncogenes (32 intermediates, 7 DBPs among which MGA, ETV4, ETV1, GATA2, ETV3, ERF, NKX3-1, and 4 of both types among which TP53, MYC, FOXA1, AR. see **Methods** and **Table S2**). In total, 11 out of 885 have been targeted by PrCa-specific drugs (source: DrugBank; protein targets: ESR2, ESRRA, AR, PARP1, NFKB2, NFKB1, NCOA2, NCOA1, AKT1, TOP2A, TOP2B; drugs: Estramustine, Genistein, Flutamide, Nilutamide, Bicalutamide, Enzalutamide, Olaparib, Custirsen, Amonafide); and 190 out of 885 are targets of non-prostate drugs indicating the possibility of re-purposing.

Considering 500 out of 885 proteins with annotations for KEGG pathways retrieved using g:Profiler (Raudvere et al. 2019), 45 were enriched in the prostate cancer pathway (KEGG:05215) (adjusted p-value = 2.54e-27) (**Methods** and **Table S7**). We next studied specific protein enrichments in the nodes of the EPINs of each identified cluster (**Table S8)**. Although intermediates are ubiquitous and generally shared among all clusters, we could identify 22 significantly specific proteins enriched in the GWAS+ cluster (**Methods**). Functional enrichment analysis of these 22 proteins revealed significant relationships with tumorigenic processes (**Table S9). ** KEGG *Prostate cancer pathway* (KEGG:05215) appears highly enriched (adjusted p-value = 4.38e-5) together with other pathways related to tumors such as *Colorectal cancer* (KEGG:05210, adjusted p-value = 3.45e-9) *Pancreatic cancer* (KEGG:05212, adjusted p-value = 1.28e-6) and *Breast cancer* (KEGG:05224, adjusted p-value = 2.05e-6). KEGG pathway KEGG:04919 (*Thyroid hormone signaling pathway*) appears as the third most enriched pathway (adjusted p-value = 4.02^{-}e-8). Thyroid hormones have been previously described as modulators of prostate cancer risk (Mondul et al. 2012; Hsieh and Juang 2005; Lehrer et al.

2005; Hellevik et al. 2009). Pathway KEGG:05200 (called *Pathways in cancer*) appears as the fourth most enriched KEGG concept (adjusted p-value= 1.30e-7). Other classical tumorigenic pathways, such as *Wnt signaling pathway* (KEGG:04310, adjusted p-value = 3.48e-4) and *TGF-beta signaling pathway* (KEGG:04350, adjusted p-value = 3.30e-5) appear also enriched. In this regard, recent studies analyzed the involvement of Wnt signaling in the proliferation of prostate cancer cells (Ma et al. 2022; Wei et al. 2022), as well as the involvement and TGF-beta signaling (Natani et al. 2022; Xi et al. 2022).

SNPs path analysis in the E-P protein interactomes

Next, we sought to perform a SNPs analysis of the paths in an EPIN promoter (**Methods**). In this analysis, a path in a network is a sequence of edges joining a sequence of nodes and going from enhancer to promoter (**Figure 1A**). We distinguish between two possible scenarios: (1) PrCa SNPs fall in the DNA binding motifs found in enhancers, indicating a possible dysregulation of TFs binding and activity (**Figure 1B**); (2) PrCa SNPs in the genomic regions of the genes that encode for the intermediate nodes of the EPINs, indicating a possible alteration of the PPIs (**Figure 1C**). The first analysis aims to identify the location of enhancers that could be targeted by genetic perturbation techniques such as CRISPRi. The second analysis aims to identify the proteins that are potentially affected by mutations so as to enhance our understanding of prostate cancer biology. Overall, we characterized 188 PrCa SNPs falling within any path that connects enancers to promoters (rs4962419 is in both categories). In the following, we discuss the two scenarios and report on the *MYC* promoter as a unified illustrative example.

Network paths with PrCa SNPs in enhancer binding motifs

We sought to detect SNPs located in the DNA binding motifs found in the enhancers of the EPINs. Based on previous evidence (Speedy et al. 2019; S. Zhou et al. 2020), our hypothesis is that SNPs in enhancers could disrupt the binding of proteins such as TFs having an impact on their interactome (**Figure 3B**).

In **Table S10** we list the 36 PrCa SNPs falling within 60 DBP motifs in enhancer regions linking 34 different promoters whose EPINs include 5,184 edges. Among these, we identified 17 PrCa SNPs falling within 16 promoter EPINs (1,894 edges) belonging to the *GWAS+ cluster* that had at least one PrCa SNP in their enhancers. Several of these EPINs promoters were also found differentially expressed (such as *DLL1, STOM and SEC11C* in the tumor/normal dataset; *ID2, RPS27, SEC11C, CASZ1, CRTC2, C5 and STOM* in the LNCaP/LHSAR dataset). Finally, at the level of intermediate proteins, we also found some encoded by genes reported to be differentially expressed. We observed that the mean proportion of intermediates that are differentially expressed is on average 40% (**Figure S4**). We tested whether promoters belonging to the GWAS+ cluster were significantly enriched for intermediate protein encoding for differentially expressed genes (**Methods**). Among the 16 EPINs belonging to the *GWAS+ cluster* that had at least one PrCa SNP in their enhancers, 11 contain expression data to study potential direct effects of the SNPs. In this subset we found 4 EPINs differentially expressed in promoters (3 also differentially expressed in intermediates: *CASZ1, ID2, SEC11C*), and 4 EPINs only differentially expressed in intermediates: *MIIP, MRPL14, MYC, TMEM63B* (**Table S1**). The differential expression of intermediates makes it easier to identify interesting and potentially novel cases. For instance, MYC is not differentially expressed but it has differentially expressed intermediates.

Network paths with PrCa SNPs in the genes coding for intermediate proteins

In this analysis, we identify EPINs with PrCa SNPs falling within genes that encode for intermediate nodes (**Table S11**), indicating a potential alteration of PPIs involved in E-P contacts. We found that the GWAS+ cluster has the highest proportion of PrCa SNPs in the intermediate nodes compared to all other clusters (mean = 53.2, SE = 18.0, p-value <= 0.01, **Table S12**). The EPINs of *STK40* and *GATA2* promoters GWAS+ cluster display the highest fraction of intermediate proteins with PrCa SNPs in their corresponding genes encoding them (**Table S1**).

We use the SNP paths to link 172 PrCa SNPs falling within the gene bodies of 26 genes of which 7 are known oncogenes (*MAP2K1, CHD3, AR, SETDB1, ATM, CDKN1B, USP28*). We identify edges that are most enriched in our GWAS+ cluster which could be pointing to essential links between the gene encoding for the intermediate and containing a PrCa predisposing SNP at a particular EPIN. For example, we identify the link between *MDM4* containing SNP rs35946963 (PrCa p-value 1e-24) and TP53 (Mejía-Hernández et al. 2022) and between *KDM2A* containing SNP rs12790261 (PrCa p-value 1e-7) and BCL6 (L. Liu, Liu, and Lin 2021) and *ARNT* continuing SNP rs139885151 (PrCa p-value 3e-13) and HIF1A (Mandl and Depping 2017).

We integrated information from pQTL associations between the 172 PrCa SNPs and protein levels (**Methods**). Two intermediate protein levels (CREB3L4, MAP2K1) were associated with PrCa SNPs falling within the gene encoding for them (p-value of association with protein levels were 7.75e-86 for CREB3L4 and 2.40e-5 for MAP2K1). We identified 3 out of 26 promoter EPINs (*TRIM26, MEIS1, POU2F2*) with suggestive evidence (p-value < 1e-5) of association between the PrCa SNP with the PENGUIN-linked promoter EPIN, pointing to the cancer promoting mechanistic action of these variants: gene with SNPs in *POU2F2* linked to the EPIN promoter of gene *PHGDH* (SNP with lowest p-value rs113631324 = 3.80e-8); gene with SNPs in *TRIM26* and EPIN promoter of gene *RRM2* (SNP with lowest p-value rs2517606 = 2.69e-7); gene with SNPs in *MEIS1* and EPIN promoter of gene *STOM* (SNP with lowest p-value rs116172829 = 8.19e-6).

A case study: the SNP paths connected to the MYC promoter

From HiChIP data, the *MYC* promoter (chr8:128747814-128748813) is in contact with 73 enhancer regions among which one holds the SNP rs10090154 (p-value of association with PrCa = 1.4e-188). This SNP is located in the binding motif of the transcription factor FOXA1. The integration of PrCa SNPs information highlights paths in the EPIN of *MYC* that are particularly compelling in the context of the disease (red line in **Figure 4A**). The promoter region of *MYC* binds 8 proteins TFAP2C, KLF5, RBPJ, SP1, ZBTB14, ATF6, ZBTB7A, PRDM1 and contains 17 protein interactors (dots in **Figure 4A**) that might be affected by the possible disruption of its binding motif, namely, FOXA1, HMGA1, RCC1, TFAP4, NFIC, PBX1, HOXB9, NFIX, NACC1, RARA, PIAS1, RPA2, H2AFY, RECQL, SATB2, CREB1, AR. The gene encoding for *FOXA1* is differentially expressed, along with others of its interactors (**Table S10**; **Methods**). Interestingly, 24 correlated PrCa SNPs fall within the genomic region of *AR* (marked by an asterisk next to the gene name), all with p-value of association with PrCa below 1e-11 (**Table S11**). AR is targeted by several drugs used in the treatment of prostatic neoplasms, such as apalutamide, bicalutamide, diethylstilbestrol, enzalutamide, flutamide, and nilutamide (triangle in the Figure 4A, source: DrugBank).

Notably, mutations in *FOXA1* enhancers were previously shown to alter TF bindings in primary prostate tumors (S. Zhou et al. 2020). And, also in line with our observations, *FOXA1* enhancer region has been previously reported to be coupled to *MYC* (Sur et al. 2013) and has been shown to have a strong binding of AR (Jia et al. 2009).

Discussion

Using PPI networks we uncovered a set of genes implicated in PrCa that could not be identified otherwise. Intermediate nodes of this PPI network carry the intrinsic properties to be used for the classification and characterization of E-P chromatin loops. We have shown that, without any prior information, PENGUIN was able to group genes according to their implication in our case study, PrCa. Our study opens a new path towards the understanding and identification of new biological markers in disease. Accordingly, the genes that we identified in the cluster most enriched in SNP associated to PrCa (GWAS+ cluster) can be regarded as candidate oncogenes or partners of oncogenes, for example they may be sharing “onco-enhancers” (enhancers participating in tumorigenic activity).

PENGUIN is based on the assumption that the PPI network structure between a promoter and its enhancers can be used as a signature to be associated to specific functional profile and to disease. This assumption is based on previous works showing the relation between loop 3D topology and chromatin state or expression (Galan, Serra, and Marti-Renom 2022). With PENGUIN here we propose a molecular explanation to the observed distinct structural features of loops, and directly relate them to disease. Indeed, we observed that interactions belonging in the GWAS+ cluster, but not carrying a GWAS SNP, can still be related to PrCa. The core of our method is agnostic to the presence of specific SNPs or oncogenes; these are used in the post-processing to label the defined clusters.

Previous methods have combined GWAS hits with PPI networks. For example, Ratnakumar and colleagues identify proteins enriched for PPI with GWAS hits (Ratnakumar et al. 2020). Recently, Dey and colleagues have demonstrated advantage in using strategies capturing both distal and proximal gene regulation to prioritize genes for disease (Dey et al. 2022). On the other hand, other methods have combined information from 3D chromatin interactions and GWAS SNPs to relate intergenic SNPs to gene regulation and to cancer (Javierre et al. 2016; López de Maturana et al. 2021; N. Liu et al. 2021). Our method is completely blind to the presence of SNPs, and combines information from PPI network and E-P information from enhancer-promoter conformation (H3K27ac-HiChIP) into one framework.

We have linked paths and identified examples where this could occur. We aimed to identify specific links that could be disrupted by a PrCa-predisposing variant, such as CTCF sites linking a promoter to its enhancers, or intermediate structural proteins involved in E-P network. More work is needed to understand the biology and mechanisms behind these links. To facilitate investigation of SNPs pathways involved in prostate cancer we provide a web interface at https://penguin.life.bsc.es/.

Interestingly, although PENGUIN finds clusters of EPINs significantly more related to cancer, our gene expression analysis did not reveal any significant trend. This observation is in apparent contradiction with our definition of EPIN clusters or even of our core definition of EPIN. In fact, given the level of evidence brought by our analysis, we believe that PENGUIN allows the detection of associations with cancer with a greater sensitivity than a differential expression analysis.

In this application, we have used intermediate nodes through PPI to inform clustering of E-P into networks (intermediate = 1 in **Table S4**). Yet, PENGUIN's clustering procedure allows the use of a variety of inputs. First, we can do without grouping the data by promoter into EPINs, considering only single, pairwise, E-P interactions. Alternatively, using our definition of EPINs, PENGUIN allows building networks either: providing the E-P interactions, without information from PPI (intermediate = 0 in **Table S4**); or only giving the intermediate PPI-network (not considering DBPs for the clustering) without the DNA-binding proteins bound to enhancer or to the promoter (data not shown); or increasing the number of intermediates (e.g. intermediates = 2, data not shown). In any of the above mentioned configurations PENGUIN is able to yield a clustering that significantly segregates CTCF-enriched and GWAS-enriched EPINs. In this report, we used PPI information (intermediates =1) and observed that the presence of intermediate PPI networks increases segregation significance compared to not using PPI (**Table S4**).

The PENGUIN approach used here to study PrCa in LNCaP cells can be applied to study any other human disease provided similar data, to study other scenarios (cell type/GWAS combination) of interest for future studies. For example, using an E-P set from another prostate cancer cell line would identify target genes regulated by enhancers from different cell-types. These can be prioritized using a genome-wide set of risk SNPs from a disease of interest.

The networks obtained with PENGUIN enable a finer characterization of the molecular associations at play in cancer chromatin and are suitable to train sophisticated machine learning models such as graph neural networks (GNN). We propose PrCa intermediates that interact in E-P networks in cancer cells (LNCaP) and are amenable to therapeutic intervention. High-throughput functional studies would validate the impact of genetic perturbation of thousands of enhancers at a time. For example CRISPR-Cas9 technology could allow the targeted editing of specific genomic regions.

Our analysis has several caveats. We considered E-P interactions from HiChIP technique, protein-DNA interactions from FIMO, and tissue-specific protein-protein interactions from the integrated interactions database (IID). The completeness of this information depends on the limits of the databases and methods used. Additionally, we consider networks that are bound to proteins (only proteins having edges are considered). Also, for sake of visualization purposes, we reduced the number of reported proteins and reported one intermediate only (expanded 1 edge away). However, we found that the clustering is even more evident when including three intermediate proteins (data not shown). Lastly, we considered E-P in a stable environment (LNCaP cells) representing a snapshot in time. Although this is an area of active research which requires further exploration, the literature to date supports minimal and quantitative small changes in E-P interactions.

Acknowledgments

The authors are grateful to José María Fernández González (Barcelona Supercomputing Center) for the crucial guidance with the PENGUIN web server. They also thank the Biola Javierre's lab at the Josep Carreras Leukaemia Research Institute for the support, the 'RNA initiative' at IIT and all the members of Tartaglia's lab at CRG, Sapienza and IIT.

FUNDING

The research leading to these results has been supported by the European Research Council [RIBOMYLOME_309545 and ASTRA_855923] and the H2020 projects [IASIS_727658 and INFORE_825080]. CG has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 754490 – MINDED project.

Methods

Conformation capture and E-P interactions

We used Hi-C followed by chromatin immunoprecipitation (**HiChIP**) **targeting H3K27Ac** in LNCaP cells (androgen-sensitive prostatic carcinoma cell line) across 5 biological replicates including 1 billion reads as previously described (Giambartolomei et al. 2021). HiChIP is an efficient protein-mediated chromatin-conformation assay (Mumbach et al. 2016). H3K27Ac is a **marker of active enhancers and promoters**. Briefly, we used HiC-Pro (Servant et al. 2015) to map HiCHiP reads and extract unique interactions; FitHiChIP (Bhattacharyya et al. 2019) was used to identify significant interactions with a predefined set of peaks from H3K27ac ChIP-seq in LNCaP to refine accurate anchor ranges. We used q-value < 0.01 and a 5 kb resolution and considered only interactions between 5 kb and 3 Mb. In this analysis, we restricted to a stringent global background estimation to reduce as much as possible the number of false-positive interactions. The corresponding FitHiChIP specifications used were “IntType=3” (the peak-to-all) for the foreground, meaning at least one anchor to be in the H3K27 peak, and “UseP2PBackgrnd=1” (the peak-to-peak (stringent)) for the global background estimation of expected counts and contact probabilities for each genomic distance for learning the background and spline fitting. We identified 49,565 significant interactions (FitHiChIP, FDR<0.01).

We categorized interactions by overlapping anchors with transcription start sites (TSS) and enhancers identified by H3K27ac ChIP-seq as previously described (Giambartolomei et al. 2021). Briefly, we first extended anchors by 5 kb on either side; we defined promoter regions around the TSS (+/- 500 bases) using RefSeq hg19 (see **Data Availability**); we defined enhancer regions using 49,638 regions from H3K27ac LNCaP in regular media (union of narrow and broad peaks). Out of the 49,565 significant interactions, we considered only the 24,547 E-P interactions. Specifically, we labeled the promoters and enhancer regions that overlap either right or left anchors, and considered E-P if only one anchor overlaps a promoter and the other an enhancer region. The enhancer anchors at this stage of the analysis are of length 15 kb (5 kb resolution of the HiChIP data analysis and additional 5 kb padding added to anchors on either side).

We further prioritized E-P interactions to 1 kb regions and discarded from enhancers the 1 kb bins with fewer HiChIP interactions with the promoter (see *E-P HiChIP prioritization* section). The 15 kb original E-P interactions dataset contained a mean of 1.6 (1.3 s.d.) promoter anchors per enhancer anchor (after prioritization of enhancer anchor to 1 kb region, mean of 1.4 (0.9 s.d.) promoters per enhancer). There were 11,127 (17,683 prioritized 1 kb regions) enhancer anchors in total; 7,341 (12,385 prioritized 1 kb regions) enhancer anchors are contacted by one promoter anchor with a maximum of 21 promoter anchors (15 using prioritized enhancer regions) sharing the same enhancer.

E-P HiChIP prioritization

In order to reduce experimental artifacts in the context of our EPINs, we developed a specific prioritization method. This prioritization start by normalizing the data assuming, as most used capture-C normalizations (ICE (Imakaev et al. 2012), Vanilla, or KR (Rao et al. 2014)) that all biases (e.g. GC content, number of restriction sites, mappability, or in the case of HiChIP, immunoprecipitation bias) can be corrected together. For this normalization step, we assume that there is a specific bias per any 1 kb genomic loci ($_x$ for loci x; see **Figures S5A and S5B**). This bias causes the difference between a theoretical expected number of interactions (E_{XY} between loci *X* and *Y*) and the observed number of interactions (O_{XY} between loci *X* and *Y*). In this representation we can define a system of 9 equations involving three 1 kb loci in the promoter and three 1 kb loci on the enhancer side. This system of equations is then solved using Sequential Quadratic Programming (SQP) (Virtanen et al. 2020). The procedure is repeated in an overlapping window manner along the 15 kb of the enhancer, alway against the target 1 kb of the promoter and its two 1 kb neighboring loci.

Before the normalization step, we observed a different interaction pattern for interactions below 10 kb (**Figure S5C**) due, in part, to the contiguity of restriction-enzyme fragments or chromatin persistence length. As these interactions may also be a source of bias in the construction of a PPI network, we removed them from our study.

We applied the normalization to the remaining interactions and observed a better correlation between genomic distance and interaction count (**Figures S5D and S5E**).

The normalized profile of interactions was finally used to prioritize most interacting 1 kb loci on the 15 kb enhancer (**Figure S5F**). The selected 1 kb regions are referred to as prioritized enhancer regions.

DNA binding motifs

DNA binding motifs were retrieved from JASPAR (Fornes et al. 2019), an open-access database of curated, non-redundant binding profiles of DBPs (a.k.a. motifs) stored as position frequency matrices (PFMs). To detect the binding motifs, we used FIMO from the MEME-suite software (Grant et al. 2011), with p-value <= 1e-4 and q-value <= 5e-2 cutoffs. JASPAR contains **810 DNA binding motifs of 640 proteins** that overlap the E-P contacts identified with HiChIP.

Gene expression data

We assayed RNA sequencing (RNA-seq) in the cell line LNCaP for two replicates using the VIPER pipeline as previously described (Giambartolomei et al. 2021), and fragments per kilobase of transcript per million mapped reads (FPKM) values were calculated for 20,114 RefSEQ genes. Genes with expression levels above the threshold of 0.003 in both replicates were considered in the entire analysis (**Figure S4C**).

Protein-protein interaction network

We obtained protein-protein interactions (PPIs) from the Integrated Interactions Database (IID) (Kotlyar et al. 2016). To better contextualize the interactome information, we combined the annotations of the PPIs from IID database with the LNCaP gene expression data. As for the IID annotations, we applied the following selection criteria. First we selected interactions annotated as "experimental" in the "*evidence type*" field and identified by at least two independent biological assays reported in the "*methods*" field. Then, we filtered only for interactions in the *prostate* or in *prostate cancer* cells and between *nuclear* proteins. Finally, we retain proteins whose gene expression levels were FPKM > 0.003 in both replicates (this cut-off removes ~30% of the genes). In total, 14,221 proteins from a pool of 20,111 human protein coding genes meet the gene expression criteria. The combination of the above filtering criteria (gene expression, using only nuclear, prostate cancer or prostate and experimentally by 2 methods) resulted in an unweighted network of **31,944 prostate-specific nuclear PPIs among 4,295 proteins**.(Kotlyar et al. 2016)

PENGUIN pipeline

We set up graph-based approach, called Promoter-ENhancer-GUided Interaction Networks (PENGUIN), to reconstruct individual networks of protein interactions that might occur between one promoter (P) and its contacting enhancers (E), that we call E-P protein-protein Interaction Networks (EPINs). To reconstruct the EPINs, PENGUIN integrates information about chromatin contacts, protein-DNA binding, and protein-protein interactions (PPIs). For the case under study in this work (prostate cancer, PrCa), chromatin contacts information comes from H3K27Ac HiChIP of LNCaP cells (4,314 promoters and 5,789 enhancer regions; see Methods, "Conformation capture and E-P interactions"), protein-DNA binding information (Rao et al. 2014; Virtanen et al. 2020) comes from the JASPAR database (810 DNA binding motifs of 640 proteins; see Methods, "DNA binding motifs"), and PPIs information comes from the IID database (31,944 prostate-specific nuclear PPIs among 4,295 proteins; see Methods, "Protein-protein interaction network").

The reconstruction of EPINs follows these steps: for each E-P contact, (1) DNA binding motifs are detected in the corresponding sequences of promoter and enhancer regions; (2) a subnetwork of PPIs is selected containing all promoter-bound proteins, all enhancer-bound proteins, and all their intermediate interactors, with a maximum of 1 intermediate node between enhancer and promoter bound DNA binding proteins; (3) intermediate interactors are discarded if they only connect promoter-bound proteins or enhancer-bound proteins. Using the provided PrCa information, PENGUIN reconstructed 4,314 EPINs consisting of a total of 9,141 PPIs among 885 proteins of which 751 are intermediate proteins linking promoter-bound and enhancer-bound proteins.

Node centrality measures

In several analyses we employed two measures of node centrality, namely betweenness and degree. **Betweenness** is a measure of centrality in a graph based on shortest paths. For every pair of nodes in a connected graph, there exists at least one shortest path between the vertices such that either the number of edges that the path passes through is minimized. The **degree** of a node in a network is the number of connections it has to other nodes and the degree distribution is the probability distribution of these degrees over the whole network.

Clustering EPIN

We defined clusters by taking into account their edge content. We collected the full universe of edges using all existent edges between all promoter EPINs (the union graph) and we performed clustering using a binary representation that encodes this particular edgelist. Clustering was performed over the overlap index distance matrix, by calculating Euclidean distance and using Ward's linkage method. Each leaf in the obtained cluster is a promoter EPIN.

Identify transcription factors binding directly enhancer to promoter

The TFs binding in the Enhancer region (TFs.E) and the TFs binding in the Promoter region (TFs.P) were identified. The PPIs networks were then searched to see if there is a path linking the (TFs.E) and the (TFs.P).

Identifying enriched clusters using functional annotations

We performed fisher tests on the obtained clustering for every single branch of the dendrogram. We examined if there is an enrichment of any feature (CTCF, GWAS) for the leaves under the branch interest compared to those in the rest of the tree.

For the CTCF binding feature, we require a CTCF binding (see **CTCF ChiP-Seq peaks)** to Promoter and at least one of the Enhancer regions of the promoter EPIN.

For the GWAS feature, we require the presence/overlap of a GWAS SNP (see **Genome-wide association data**) in at least one of the Enhancers of the promoter EPIN.

Fisher tests were used to calculate the odds ratio (OR) and enrichment p-values for presence of PrCa annotations within the identified clusters.

Druggability information

We extracted information for target druggability from DrugBank (Wishart et al. 2018). The use of each drug was extracted from the Therapeutic Target Database (Y. Zhou et al. 2022). We annotated as PrCa druggable each protein node that is target for drugs that are assigned as Approved or under Clinical Trials (Phase 1, 2, 3) or Investigable for Prostate Cancer.

CTCF ChiP-Seq peaks

CTCF ChiP-seq peaks were retrieved from ENCODE project (https://www.encodeproject.org/) for the same Genome assembly, hg19 (Experiment ID: *ENCSR315NAC*, file ID: *ENCFF155SPQ*). These narrow peaks were mapped on the enhancer regions using the python package pyranges (see "E-P contacts" section). We found overlaps of the CTCT binding sites with enhancer and promoter anchors allowing a 10 kb gap between them.

PrCa SNPs

To explore enrichment of GWAS across the identified clusters, and to identify the SNP paths, we used the 95% credible set from fine-mapping of the largest PrCa genome-wide association studies (GWAS) collected from Schumacher et al. 2019 (N = 79,148 cases and 61,106 controls), which includes 20,370,946 SNPs, as previously described (Schumacher et al. 2018). Briefly, we fine-mapped 137 GWAS regions using PAINTOR (Kichaev et al. 2014), a Bayesian statistical method, with no functional annotations and specifying a maximum of 1 causal SNP. We then constructed a 95% credible set by taking the cumulative sum of the posterior probability until a cumulative 95% posterior probability was reached. This set was composed of 5412 distinct SNPs (rsid). We will refer to these 95% credible set SNPs as **PrCa SNPs** for brevity. Note that this set also includes SNPs that do not reach genome-wide-filters of p-value significance. We mapped the SNP location to prioritized enhancer regions anchor locations with a window of 10kb. 518 out of 5412 overlap our prioritized enhancer regions; 18 of them overlap our promoter regions. In total 218 prioritized enhancers and 14 promoters overlap a PrCa SNP.

SNP paths (PrCa SNPs in enhancer binding motifs)

A path in a network is a sequence of edges joining a sequence of nodes. We detected PrCa SNPs located in the DNA binding motifs in the enhancers, and identified the corresponding SNP paths (linked edges and nodes) for each EPIN promoter. For SNP paths analyses and the web-browser, we used all PrCa SNPs in the 95% credible set. There were 36 PrCa SNPs falling in enhancer binding motifs across clusters 3, 4, 5, 6, 7, 8. To report the most interesting cases in the Tables and Results, we used the subset of those passing genome-wide significance of p-value for PrCa association < 5e-8. There were 15 PrCa SNPs falling in enhancer binding motifs across clusters 3, 5, 6, 7, 8.

SNP paths (PrCa SNPs in intermediate proteins)

We detected PrCa SNPs falling within genes that encode for intermediate nodes, and identified the corresponding SNP paths (linked edges and nodes)for each EPIN promoter. For SNP paths analyses and the web-browser, we used all PrCa SNPs in the 95% credible set. To report the most interesting cases in the Tables and Results, we used the subset of those passing genome-wide significance of p-value for PrCa association < 5e-8.

PrCa GWAS enrichment using GWAS Catalog and comparison with other diseases

This analysis had two aims: 1) explore whether we could replicate our finding and identify the GWAS enriched cluster using a different source for the GWAS; 2) to compare the GWAS signal for different diseases. We estimated enrichment of SNPs overlapping the enhancers in each of the identified clusters by exploring the NHGRI GWAS Catalog associations (Buniello et al. 2019). First, we retrieved GWAS data and filtered the traits according to their “umlsSemanticTypeName” as defined in DisGeNet database (Piñero et al. 2020) to one of the following: "Mental or Behavioral Dysfunction", "Neoplastic Process", "Disease or Syndrome", "Congenital Abnormality; Disease or Syndrome", "Disease or Syndrome; Congenital Abnormality", "Disease or Syndrome; Anatomical Abnormality". We considered only traits with at least 10 genome-wide-significant SNPs (unadjusted p-value < 5e-8). We mapped the SNP location to prioritized enhancer anchor locations with a window of 10kb. 104 diseases had SNPs overlaps and 17 of them have more than 10 SNP overlapping (**Table S5).** For each cluster, we tested enrichment of disease-associated SNPs using Fisher tests and considered significant p-value < 0.01 and OR > 1.

Oncogenes Gene list

We used a previously identified list of 122 Genes ("PrCa_GeneList_Used.csv") known to be somatically mutated in PrCa oncogenesis (37 out of 4,314 promoters considered). As previously described (Giambartolomei et al. 2021), the 122 oncogenes are a set of prostate cancer–genes curated from three large-scale PrCa studies that show evidence of somatically acquired mutations, at both localized and advanced prostate cancer, known and recurrently altered in localized prostate cancer and metastatic prostate cancer.

Enriched edges within each cluster

Fisher tests were used to compute odds ratios and p-values of the edges and nodes in the eight different clusters. Specifically, each edge or node was tested for presence/absence in a cluster compared to all others. Therefore, one edge or node can be enriched in one or more than one cluster, it cannot be enriched in all clusters.

Enriched intermediate nodes within each cluster

We computed protein importance for each cluster in terms of two network centrality measures: betweenness and degree. For each protein we obtain both betweenness and degree specificity ratios in order to equitably quantify internal protein centrality differences between the clusters. For each of the found clusters we independently estimated the specificity of the observed protein centrality measures (“Betweenness” and “Degree”). For a given protein (P_i) in a particular cluster (C_j), we define the specificity as the ratio between the mean centrality value of P_i inside the fraction of networks belonging to C_j ; divided by the mean centrality value of P_i for the fraction of networks outside of the cluster C_j.

Specificity ratio (P_i, C_j) = (mean (Pi centrality in C_j networks) + 1) / (mean (P_i centrality in non C_j networks) + 1)

We assessed protein specificity ratio significance for each cluster upon random network cluster generation. Aiming to assess the significance of the different specificity ratios for the proteins within each cluster, we developed a significance analysis test based on random cluster subsamplings. In order to compute the significance of a given protein specificity ratio (P_i) within a particular cluster of analysis (C_j), we performed 1000 random network samplings to produce random network clusters containing the same amount of networks as the real cluster being analyzed (i.e. if the real cluster contains 100 networks, the random clusters generated will contain 100 random networks out of the 4,314 clustered networks). Within each of those 1000 random clusters, we compute the corresponding protein specificity ratios, with the p-value representing the probability of finding the protein specificity ratio to be higher or equal to the real value computed for the particular cluster of interest (C_j).

We also performed Fisher tests to assess enrichment for the presence of the node in the cluster (Fisher test p-value < 0.01). EP300 was excluded from the enrichment test as the presence of that node was not significantly enriched (Fisher test p-value < 0.01). 22 proteins (SMAD2, KAT5, NCOR2, MAPK8, SMAD4, CREBBP, CTNNB1, PGR, HDAC3, HDAC2, GSK3B, UBA52, UBE2I, JUND, PIAS1, XRCC5, CDK6, XRCC6, MAPK1, FOS, HIF1A and MAPK3) were found to be significantly specific for both betweenness and degree ratios (p-value < 0.01 for both centrality measures and over-represented in this cluster using Fisher tests).

Functional gene set enrichment analysis

Functional enrichment analysis was performed using the g:GOST module from g:Profiler, a web tool to perform simultaneous gene set enrichment analysis across multiple biomedical databases (Raudvere et al. 2019). g:GOST performs cumulative hypergeometric tests of an input geneset against preprocessed database-specific gene sets. We run the web service considering only annotated genes for the statistical domain scope. Reported adjusted p-values correspond to Benjamini-Hochberg correction for multiple testing, with adjusted p-values ≤ 0.05 considered to be significant. Gene set enrichment analysis results are provided for KEGG pathways, Reactome, Gene Ontology, Wikipathways, TRANSFAC, miRTarBase, Human Protein Atlas, CORUM and Human Phenotype Ontology.

For the enrichment analysis of significantly specific proteins of the GWAS+ cluster, we provided as input the 22 previously described proteins. For the enrichment analysis of the GWAS+ cluster, we provided as input all genes associated with the EPIN promoters in cluster GWAS+.

Differential Gene Expression

We integrated data from EPIN promoters with differential gene expression (DE) from two sources.

DE analysis on prostate cancer tumor versus normal was downloaded from GEPIA: http://gepia2.cancer-pku.cn/#degenes, which use the TCGA and GTEx projects databases to compare gene expression between tumor and normal tissues under Limma, both under and over expressed. We used the default thresholds of log2FC of 1 and qvalue cut-off of 0.01. These data covered 84 out of 885 genes encoding for intermediates in PENGUIN and 413 out of 4,314 promoter EPINs.

DE analysis of RNA-Seq on LHSAR (an immortalized prostate epithelial line overexpressing androgen receptor) versus LNCaP was performed as previously described. Briefly, RNA-seq data were processed using the VIPER pipeline (Cornwell et al. 2018). Reads were aligned to the hg19 human genome built with STAR. FPKM values were calculated with Cufflinks for 20,114 RefSEQ genes included in the VIPER repository. Differential expression analysis was performed with the DESeq2 R package (Love, Huber, and Anders 2014). 15,650 genes with DE data covered 884 of the 885 genes encoding for intermediates in PENGUIN and 3,286 genes out of 4,314 promoter EPINs.

We annotated whether the EPIN promoters themselves and the genes encoding the intermediate proteins in our data were DE using either of the two databases. We considered as DE those genes passing |log2 fold change| > 1 and adjusted p-value <= 0.01. For the LNCAP/LHSAR dataset, we could compute a Fisher test of enrichment of differentially expressed genes encoding for intermediate proteins within each EPIN promoter versus within the SNP paths (we could not compute this for the GEPIA since we did not have the full dataset of covered genes). The genes that were not passing these filters were considered non-DE and the genes not covered by the two datasets were excluded from the enrichment analysis described next. For each EPIN we calculated the fraction of DE intermediates within the SNP paths and we estimated the enrichment of those compared to the fraction of DE intermediates in the full EPIN network.

To find the enrichment of DE genes in SNP paths (PrCa SNPs in intermediate proteins) compared to those in the entire EPIN, we computed as enrichment the ratio of Fraction1 / Fraction2, where:

Fraction1 = (number of DE intermediates within SNP paths) / (number of covered intermediates within SNP paths), and

Fraction2 = (number of DE intermediates the EPIN) / (number of covered intermediates in the EPIN).

We report the EPIN genes passing enrichment (“**enrichment_DE_deseq_SNP.bs.TF.path**”) > 1.

pQTL look-up

We downloaded summary statistics with genome-wide association between SNPs and 4907 proteins reported in the deCODE study (Ferkingstad et al. 2021) and annotated with pQTL association the SNPs we identified falling in either in enhancer binding sites or in node genomic locations. The deCODE pQTL summary statistics data contained information on 4,907 proteins and 186 (201 PrCa SNPs out of the 213 PrCa SNPs we looked up were in the data and 186 also matched by alleles). 808 out of the 4,314 genes promoters ("Gene_network") and 278 out of the 885 gene intermediates (in total 997 out of 4,918 genes promoters and coding for intermediates in our networks) have information on associations with their respective coded proteins covered by the pQTL deCODE data.

Data Availability

RefSeq hg19 from UCSC Genome Browser is available at the following URL:
http://genome.ucsc.edu/cgi-bin/hgTables?hgsid=694977049_xUU5i1QkIJ50dj5miBt9wkAYuxN3&clade=mammal&org=&db=hg19&hgta_group=genes&hgta_track=knownGene&hgta_table=knownGene&hgta_regionType=genome&position=&hgta_outputType=selectedFields&hgta_outFileName=knownGene.gtf

Ensembl hg19 data for overlaps of SNPs with intermediates: biomart / ensembl from bioaRt package

All EPINs and related statistics can be downloaded through the PENGUIN web service at https://penguin.life.bsc.es/

References

Adams, Elizabeth J., Wouter R. Karthaus, Elizabeth Hoover, Deli Liu, Antoine Gruet, Zeda Zhang, Hyunwoo Cho, et al. 2019. “FOXA1 Mutations Alter Pioneering Activity, Differentiation and Prostate Cancer Phenotypes.” *Nature* 571 (7765): 408–12.

Ahmadiyeh, Nasim, Mark M. Pomerantz, Chiara Grisanzio, Paula Herman, Li Jia, Vanessa Almendro, Housheng Hansen He, et al. 2010. “8q24 Prostate, Breast, and Colon Cancer Risk Loci Show Tissue-Specific Long-Range Interaction with *MYC*.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (21): 9742–46.

Bhattacharyya, Sourya, Vivek Chandra, Pandurangan Vijayanand, and Ferhat Ay. 2019. “Identification of Significant Chromatin Contacts from HiChIP Data by FitHiChIP.” *Nature Communications* 10 (1): 4221.

Bono, Johann de, Joaquin Mateo, Karim Fizazi, Fred Saad, Neal Shore, Shahneen Sandhu, Kim N. Chi, et al. 2020. “Olaparib for Metastatic Castration-Resistant Prostate Cancer.” *The New England Journal of Medicine* 382 (22): 2091–2102.

Bosco-Lévy, Pauline, Caroline Foch, Angela Grelaud, Meritxell Sabidó, Clémentine Lacueille, Jérémy Jové, Emmanuelle Boutmy, and Patrick Blin. 2022. “Incidence and Risk of Cancer among Multiple Sclerosis Patients: A Matched Population-Based Cohort Study.” *European Journal of Neurology: The Official Journal of the European Federation of Neurological Societies* 29 (4): 1091–99.

Buniello, Annalisa, Jacqueline A. L. MacArthur, Maria Cerezo, Laura W. Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, et al. 2019. “The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019.” *Nucleic Acids Research* 47 (D1): D1005–12.

Cancer Genome Atlas Research Network. 2015. “The Molecular Taxonomy of Primary Prostate Cancer.” *Cell* 163 (4): 1011–25.

Chen, Xuanrong, Qianwang Ma, Zhiqun Shang, and Yuanjie Niu. 2020. “Super-Enhancer in Prostate Cancer: Transcriptional Disorders and Therapeutic Targets.” *NPJ Precision Oncology* 4 (1): 31.

Cornwell, Macintosh, Mahesh Vangala, Len Taing, Zachary Herbert, Johannes Köster, Bo Li, Hanfei Sun, et al. 2018. “VIPER: Visualization Pipeline for RNA-Seq, a Snakemake Workflow for Efficient and Complete RNA-Seq Analysis.” *BMC Bioinformatics* 19 (1): 135.

Dekker, Job, and Tom Misteli. 2015. “Long-Range Chromatin Interactions.” *Cold Spring Harbor Perspectives in Biology* 7 (10): a019356.

Deng, Wulan, and Gerd A. Blobel. 2014. “Manipulating Nuclear Architecture.” *Current Opinion in Genetics & Development* 25 (April): 1–7.

Dey, Kushal K., Steven Gazal, Bryce van de Geijn, Samuel Sungil Kim, Joseph Nasser, Jesse M. Engreitz, and Alkes L. Price. 2022. “SNP-to-Gene Linking Strategies Reveal Contributions of Enhancer-Related and Candidate Master-Regulator Genes to Autoimmune Disease.” *Cell Genomics* 2 (7). https://doi.org/10.1016/j.xgen.2022.100145.

Ferkingstad, Egil, Patrick Sulem, Bjarni A. Atlason, Gardar Sveinbjornsson, Magnus I. Magnusson, Edda L. Styrmisdottir, Kristbjorg Gunnarsdottir, et al. 2021. “Large-Scale Integration of the Plasma Proteome with Genetics and Disease.” *Nature Genetics* 53 (12): 1712–21.

Galan, Silvia, François Serra, and Marc A. Marti-Renom. 2022. “Identification of Chromatin Loops from Hi-C Interaction Matrices by CTCF-CTCF Topology Classification.” *NAR Genomics and Bioinformatics* 4 (1): lqac021.

Giambartolomei, Claudia, Ji-Heui Seo, Tommer Schwarz, Malika Kumar Freund, Ruth Dolly Johnson, Sandor Spisak, Sylvan C. Baca, et al. 2021. “H3K27ac HiChIP in Prostate Cell Lines Identifies Risk Genes for Prostate Cancer Susceptibility.” *American Journal of Human*

Genetics 108 (12): 2284–2300.

Hellevik, Alf Inge, Bjørn Olav Asvold, Trine Bjøro, Pål R. Romundstad, Tom Ivar L. Nilsen, and Lars J. Vatten. 2009. “Thyroid Function and Cancer Risk: A Prospective Population Study.” *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 18 (2): 570–74.

Hsieh, Ming-Li, and Horng-Heng Juang. 2005. “Cell Growth Effects of Triiodothyronine and Expression of Thyroid Hormone Receptor in Prostate Carcinoma Cells.” *Journal of Andrology* 26 (3): 422–28.

Huerta-Cepas, Jaime, François Serra, and Peer Bork. 2016. “ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data.” *Molecular Biology and Evolution* 33 (6): 1635–38.

Imakaev, Maxim, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R. Lajoie, Job Dekker, and Leonid A. Mirny. 2012. “Iterative Correction of Hi-C Data Reveals Hallmarks of Chromosome Organization.” *Nature Methods* 9 (10): 999–1003.

Javierre, Biola M., Oliver S. Burren, Steven P. Wilder, Roman Kreuzhuber, Steven M. Hill, Sven Sewitz, Jonathan Cairns, et al. 2016. “Lineage-Specific Genome Architecture Links Enhancers and Non-Coding Disease Variants to Target Gene Promoters.” *Cell* 167 (5): 1369–84.e19.

Jia, Li, Gilad Landan, Mark Pomerantz, Rami Jaschek, Paula Herman, David Reich, Chunli Yan, et al. 2009. “Functional Enhancers at the Gene-Poor 8q24 Cancer-Linked Locus.” *PLoS Genetics* 5 (8): e1000597.

Kichaev, Gleb, Wen-Yun Yang, Sara Lindstrom, Farhad Hormozdiari, Eleazar Eskin, Alkes L. Price, Peter Kraft, and Bogdan Pasaniuc. 2014. “Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies.” *PLoS Genetics* 10 (10): e1004722.

Kotlyar, Max, Chiara Pastrello, Nicholas Sheahan, and Igor Jurisica. 2016. “Integrated Interactions Database: Tissue-Specific View of the Human and Model Organism Interactomes.” *Nucleic Acids Research* 44 (D1): D536–41.

Krumm, Anton, and Zhijun Duan. 2019. “Understanding the 3D Genome: Emerging Impacts on Human Disease.” *Seminars in Cell & Developmental Biology* 90 (June): 62–77.

Kvon, Evgeny Z., Rachel Waymack, Mario Gad, and Zeba Wunderlich. 2021. “Enhancer Redundancy in Development and Disease.” *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-020-00311-x.

Lehrer, Steven, Edward J. Diamond, Nelson N. Stone, and Richard G. Stock. 2005. “Serum Thyroid-Stimulating Hormone Is Elevated in Men with Gleason 8 Prostate Cancer.” *BJU International* 96 (3): 328–29.

Liu, Lisheng, Jiangnan Liu, and Qinghai Lin. 2021. “Histone Demethylase KDM2A: Biological Functions and Clinical Values (Review).” *Experimental and Therapeutic Medicine* 22 (1): 723.

Liu, Ning, Wai Yee Low, Hamid Alinejad-Rokny, Stephen Pederson, Timothy Sadlon, Simon Barry, and James Breen. 2021. “Seeing the Forest through the Trees: Prioritising Potentially Functional Interactions from Hi-C.” *Epigenetics & Chromatin* 14 (1): 41.

López de Maturana, Evangelina, Juan Antonio Rodríguez, Lola Alonso, Oscar Lao, Esther Molina-Montes, Isabel Adoración Martín-Antoniano, Paulina Gómez-Rubio, et al. 2021. “A Multilayered Post-GWAS Assessment on Genetic Susceptibility to Pancreatic Cancer.” *Genome Medicine* 13 (1): 15.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550.

Ma, Fen, Seiji Arai, Keshan Wang, Carla Calagua, Amanda R. Yuan, Larysa Poluben, Zhongkai Gu, et al. 2022. “Autocrine Canonical Wnt Signaling Primes Noncanonical Signaling through ROR1 in Metastatic Castration-Resistant Prostate Cancer.” *Cancer Research*, February. https://doi.org/10.1158/0008-5472.CAN-21-1807.

Mandl, Markus, and Reinhard Depping. 2017. “ARNT Is a Potential Direct HIF-1 Target Gene in Human Hep3B Hepatocellular Carcinoma Cells.” *Cancer Cell International* 17 (August): 77.

Mejía-Hernández, Javier Octavio, Dinesh Raghu, Franco Caramia, Nicholas Clemons, Kenji Fujihara, Thomas Riseborough, Amina Teunisse, et al. 2022. “Targeting MDM4 as a Novel Therapeutic Approach in Prostate Cancer Independent of p53 Status.” *Cancers* 14 (16). https://doi.org/10.3390/cancers14163947.

Mondul, Alison M., Stephanie J. Weinstein, Tracey Bosworth, Alan T. Remaley, Jarmo Virtamo, and Demetrius Albanes. 2012. “Circulating Thyroxine, Thyroid-Stimulating Hormone, and Hypothyroid Status and the Risk of Prostate Cancer.” *PloS One* 7 (10): e47730.

Mumbach, Maxwell R., Adam J. Rubin, Ryan A. Flynn, Chao Dai, Paul A. Khavari, William J. Greenleaf, and Howard Y. Chang. 2016. “HiChIP: Efficient and Sensitive Analysis of Protein-Directed Genome Architecture.” *Nature Methods* 13 (11): 919–22.

Natani, Sirisha, K. K. Sruthi, Sakkarai Mohamed Asha, Priyanka Khilar, Pampana Sandhya Venkata Lakshmi, and Ramesh Ummanni. 2022. “Activation of TGF-β - SMAD2 Signaling by IL-6 Drives Neuroendocrine Differentiation of Prostate Cancer through p38MAPK.” *Cellular Signalling* 91 (March): 110240.

Norton, Heidi K., and Jennifer E. Phillips-Cremins. 2017. “Crossed Wires: 3D Genome Misfolding in Human Disease.” *The Journal of Cell Biology* 216 (11): 3441–52.

Parolia, Abhijit, Marcin Cieslik, Shih-Chun Chu, Lanbo Xiao, Takahiro Ouchi, Yuping Zhang, Xiaoju Wang, et al. 2019. “Distinct Structural Classes of Activating FOXA1 Alterations in Advanced Prostate Cancer.” *Nature* 571 (7765): 413–18.

Piñero, Janet, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I. Furlong. 2020. “The DisGeNET Knowledge Platform for Disease Genomics: 2019 Update.” *Nucleic Acids Research* 48 (D1): D845–55.

Pugacheva, Elena M., Naoki Kubo, Dmitri Loukinov, Md Tajmul, Sungyun Kang, Alexander L. Kovalchuk, Alexander V. Strunnikov, Gabriel E. Zentner, Bing Ren, and Victor V. Lobanenkov. 2020. “CTCF Mediates Chromatin Looping via N-Terminal Domain-Dependent Cohesin Retention.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (4): 2020–31.

Rao, Suhas S. P., Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, et al. 2014. “A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.” *Cell* 159 (7): 1665–80.

Ratnakumar, Abhirami, Nils Weinhold, Jessica C. Mar, and Nadeem Riaz. 2020. “Protein-Protein Interactions Uncover Candidate ‘Core Genes’ within Omnigenic Disease Networks.” *PLoS Genetics* 16 (7): e1008903.

Raudvere, Uku, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. 2019. “g:Profiler: A Web Server for Functional Enrichment Analysis and Conversions of Gene Lists (2019 Update).” *Nucleic Acids Research* 47 (W1): W191–98.

Rebello, Richard J., Christoph Oing, Karen E. Knudsen, Stacy Loeb, David C. Johnson, Robert E. Reiter, Silke Gillessen, Theodorus Van der Kwast, and Robert G. Bristow. 2021. “Prostate Cancer.” *Nature Reviews. Disease Primers* 7 (1): 9.

Schumacher, Fredrick R., Ali Amin Al Olama, Sonja I. Berndt, Sara Benlloch, Mahbubl Ahmed, Edward J. Saunders, Tokhir Dadaev, et al. 2018. “Association Analyses of More than 140,000 Men Identify 63 New Prostate Cancer Susceptibility Loci.” *Nature Genetics* 50 (7): 928–36.

Schumacher, Fredrick R., Ali Amin Al Olama, Sonja I. Berndt, Sara Benlloch, Mahbubl Ahmed, Edward J. Saunders, Tokhir Dadaev, et al. 2019. “Author Correction: Association Analyses of More than 140,000 Men Identify 63 New Prostate Cancer Susceptibility Loci.” *Nature Genetics* 51 (2): 363.

Schwartz, G. G. 1992. “Multiple Sclerosis and Prostate Cancer: What Do Their Similar

Geographies Suggest?” *Neuroepidemiology* 11 (4-6): 244–54.

Servant, Nicolas, Nelle Varoquaux, Bryan R. Lajoie, Eric Viara, Chong-Jian Chen, Jean-Philippe Vert, Edith Heard, Job Dekker, and Emmanuel Barillot. 2015. “HiC-Pro: An Optimized and Flexible Pipeline for Hi-C Data Processing.” *Genome Biology* 16 (December): 259.

Speedy, Helen E., Renée Beekman, Vicente Chapaprieta, Giulia Orlando, Philip J. Law, David Martín-García, Jesús Gutiérrez-Abril, et al. 2019. “Insight into Genetic Predisposition to Chronic Lymphocytic Leukemia from Integrative Epigenomics.” *Nature Communications* 10 (1): 3615.

Sur, Inderpreet, and Jussi Taipale. 2016. “The Role of Enhancers in Cancer.” *Nature Reviews. Cancer* 16 (8): 483–93.

Sur, Inderpreet, Sari Tuupanen, Thomas Whitington, Lauri A. Aaltonen, and Jussi Taipale. 2013. “Lessons from Functional Analysis of Genome-Wide Association Studies.” *Cancer Research* 73 (14): 4180–84.

Tan, M. H. Eileen, Jun Li, H. Eric Xu, Karsten Melcher, and Eu-Leong Yong. 2015. “Androgen Receptor: Structure, Role in Prostate Cancer and Drug Discovery.” *Acta Pharmacologica Sinica* 36 (1): 3–23.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, et al. 2020. “Author Correction: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python.” *Nature Methods* 17 (3): 352.

Wang, Ruimin, Yunlong Wang, Xueying Zhang, Yaliang Zhang, Xiaoyong Du, Yaping Fang, and Guoliang Li. 2019. “Hierarchical Cooperation of Transcription Factors from Integration Analysis of DNA Sequences, ChIP-Seq and ChIA-PET Data.” *BMC Genomics* 20 (Suppl 3): 296.

Wei, Xing, Martine P. Roudier, Oh-Joon Kwon, Justin Daho Lee, Kevin Kong, Ruth Dumpit, Lawrence True, et al. 2022. “Paracrine Wnt Signaling Is Necessary for Prostate Epithelial Proliferation.” *The Prostate* 82 (5): 517–30.

Wishart, David S., Yannick D. Feunang, An C. Guo, Elvis J. Lo, Ana Marcu, Jason R. Grant, Tanvir Sajed, et al. 2018. “DrugBank 5.0: A Major Update to the DrugBank Database for 2018.” *Nucleic Acids Research* 46 (D1): D1074–82.

Xi, Xinhua, Zhengbo Hu, Qiang Wu, Konghe Hu, Zhengguo Cao, Jun Zhou, Junjian Liao, et al. 2022. “High Expression of Small Nucleolar RNA Host Gene 3 Predicts Poor Prognosis and Promotes Bone Metastasis in Prostate Cancer by Activating Transforming Growth Factor-Beta Signaling.” *Bioengineered* 13 (1): 1895–1907.

Zhang, Kai, Nan Li, Richard I. Ainsworth, and Wei Wang. 2016. “Systematic Identification of Protein Combinations Mediating Chromatin Looping.” *Nature Communications* 7 (July): 12249.

Zhou, Stanley, James R. Hawley, Fraser Soares, Giacomo Grillo, Mona Teng, Seyed Ali Madani Tonekaboni, Junjie Tony Hua, et al. 2020. “Noncoding Mutations Target Cis-Regulatory Elements of the FOXA1 Plexus in Prostate Cancer.” *Nature Communications* 11 (1): 441.

Zhou, Ying, Yintao Zhang, Xichen Lian, Fengcheng Li, Chaoxin Wang, Feng Zhu, Yunqing Qiu, and Yuzong Chen. 2022. “Therapeutic Target Database Update 2022: Facilitating Drug Discovery with Enriched Comparative Data of Targeted Agents.” *Nucleic Acids Research* 50 (D1): D1398–1407.

Zuin, Jessica, Jesse R. Dixon, Michael I. J. A. van der Reijden, Zhen Ye, Petros Kolovos, Rutger W. W. Brouwer, Mariëtte P. C. van de Corput, et al. 2014. “Cohesin and CTCF Differentially Affect Chromatin Architecture and Gene Expression in Human Cells.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (3): 996–1001.

Figure 1. **A.** Schematic representation of an enhancer-promoter protein-protein interaction network (EPIN) reconstructed with PENGUIN for a given E-P contact detected by H3K27ac-HiChIP. Promoter and enhancer DNA binding motifs found in HiChIP regions after enhancer prioritization and the corresponding bound proteins are indicated in orange; their physical interactions with other factors of the EPIN (in gray) are represented as gray lines. **B.** Workflow data processing and reconstruction of the EPINs. We considered 24,547 HiChIP E-P interactions and built EPINs centered around 4,314 promoters.

Figure 2: Clustering of the promoters originating the PENGUIN reconstructed EPINs (figure generated using ETE3 (Huerta-Cepas, Serra, and Bork 2016)). In both panels, clustering is based on edge composition of the EPINs. Leaf radius is proportional to network size. Color code (Fisher's exact test): red, enriched; blue, depleted; **A**: Enrichment of PrCa SNPs in enhancers. We identified one PrCa SNP enriched cluster (GWAS+; cluster 8), and multiple PrCa SNP depleted (GWAS-; clusters 1, 2) and neutral (GWAS=; clusters 3, 4, 5, 6, 7) clusters. **B**: Enrichment of CTCF ChIP-seq binding sites. We identified multiple CTCF enriched (CTCF+; clusters 3, 7, 8), depleted (CTCF-; clusters 1, 2, 6) and neutral (CTCF=; clusters 4, 5) clusters.

Figure 3. Schematic representation of different types of network paths found in the EPINs reconstructed by PENGUIN. In general, a network path is defined by an intermediate protein (gray circle), encoded by a gene (dark red line; $Gene_i$), that interacts with DBPs (orange circles) with binding motifs (orange lines) on the enhancer (green line) and the promoter (red line) of another gene (dark red line; $Gene_j$) (**A**). If a PrCa SNP (asterisk) falls in the enhancer binding motif, the interaction between the DBP and the enhancer is disrupted and possibly its interactions (**B**). If a PrCa SNP (asterisk) falls in the gene that encodes for the intermediate protein, the gene product is affected and possibly its interactions **(C)**. Colors are consistent with **Figure 1**.

Figure 4. Reconstructed protein interactions between MYC promoter and its enhancers. DBPs with binding motifs on the promoter region are aligned on the left, while those with binding motifs on the enhancers are aligned on the right. In the middle, proteins that connect DBPs through a shortest path. Each dot is a protein. The **color of lines** represent SNP.bs.TF.path = red; SNP.intermediate.path = green; other protein interactions are hidden. **Color of dot**: Fisher test OR for strength of enrichments in GWAS+ cluster (red is strongest odds ratio of enrichment and blue is less specific to the cluster). **Size**: the degree of enrichment for intermediates or just the degree for DNA-binding proteins (i.e., big nodes are very connected and worst for specificity). **Shape** is triangle if the protein is a druggable target from Drugbank. The **asterisk** indicates a PrCa credible SNP falling within the genomic region of the gene encoding for the intermediate proteins in the network. Names of proteins are specified if the node connects to a PrCa SNP within the enhancer region, or the node contains a PrCa SNP within the gene encoding for the intermediate protein. **rsID** is listed only for PrCa SNP overlapping a TF binding site. **Bold text** indicates the 22 enriched proteins identified in cluster GWAS+. The filters for the images and corresponding tables are the following: no filter for enrichmed nodes; no filter for enriched edges; no filter on expression; panel A, plot only the tf paths; panel B, plot only the intermediate paths. The user on the web-server https://penguin-analytics.herokuapp.com/ can choose among other options, including selecting for both SNP paths, or for only enriched edges in cluster GWAS+ (OR>1 and p-value <= 0.01).

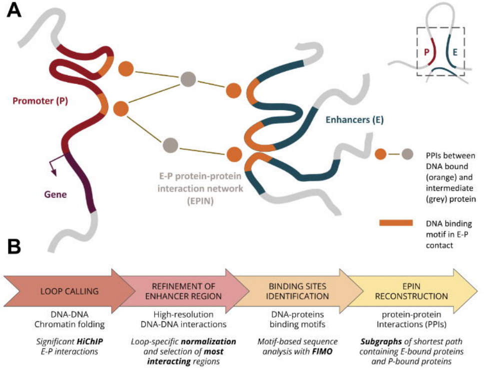
Table 1. Enrichment of PrCa SNPs in cluster 8 (GWAS+) when considering SNPs overlapping enhancers, promoters, either or both.

Table 2. Enrichment of PrCa SNPs, CTCF ChIP-seq binding sites ("CTCF" in the header), and other PrCa annotations (oncogene promoters and PrCa SNPs from GWAS Catalog) across the eight clusters

identified by PENGUIN. Cluster 8 is enriched in CTCF binding, PrCa SNPs, and oncogenes. Symbols code: +, enriched; -, depleted; =, neutral. OR: Fisher’s exact test Odds Ratio.

PrCa SNPs overlaps	Odds Ratio (OR)	p-value
Only enhancers	11.329	1.80e-12
Only promoters	1.139	0.6
Either enhancers or promoter	8.551	2.68e-11
Both enhancers and promoter	0	1

Cluster	Number of genes	CTCF	OR CTCF	P-value CTCF	PrCa SNPs	OR PrCa SNPs	P-value PrCa SNPs	Number of oncogene promoters	OR oncogenes	P-value oncogenes
1	825	-	0.617	1.91e-9	-	0.28	2.46e-2	8	1.17	0.67
2	399	-	0.613	9.65e-6	-	0.00	2.00e-2	5	1.54	0.38
3	544	+	1.348	1.35e-3	=	0.80	8.27e-1	2	0.39	0.31
4	491	=	1.084	4.09e-1	=	0.51	3.60e-1	4	0.94	1.00
5	465	=	0.841	9.12e-2	=	0.75	8.14e-1	1	0.23	0.17
6	641	-	0.664	4.24e-6	=	0.38	1.03e-1	1	0.16	0.03
7	676	+	1.655	2.12e-9	=	1.42	3.18e-1	5	0.84	1.00
8	273	+	3.287	3.64e-20	+	11.33	1.80e-12	11	6.48	1.04e-5



A
P
E
Promoter (P)
Enhancers (E)
Gene
E-P protein-protein interaction network (EPIN)
PPIs between DNA bound (orange) and intermediate (grey) protein
DNA binding motif in E-P contact
B

LOOP CALLING

DNA-DNA Chromatin folding

*Significant **HiChIP** E-P interactions*

REFINEMENT OF ENHANCER REGION

High-resolution DNA-DNA interactions

*Loop-specific **normalization** and selection of **most interacting** regions*

BINDING SITES IDENTIFICATION

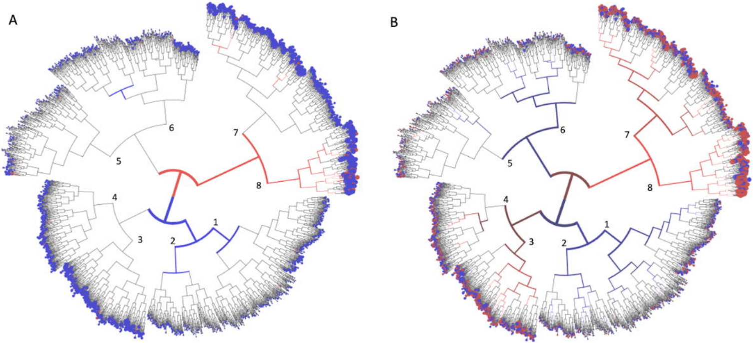
DNA-proteins binding motifs

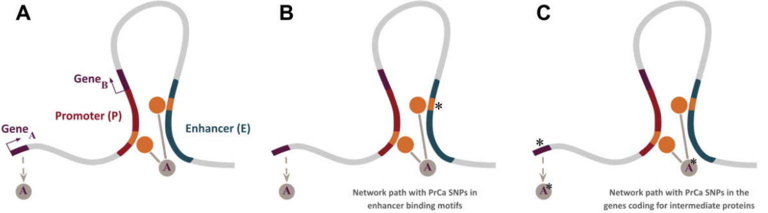
*Motif-based sequence analysis with **FIMO***

EPIN RECONSTRUCTION

protein-protein Interactions (PPIs)

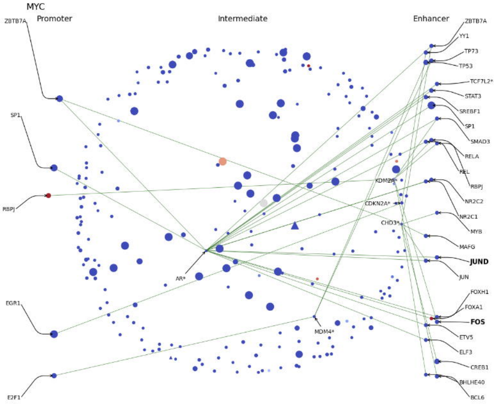
***Subgraphs** of shortest path containing E-bound proteins and P-bound proteins*


A
6
7
5
8
4
3
2
1
B
6
7
5
8
4
3
2
1



Network path with PrCa SNPs in enhancer binding motifs

Network path with PrCa SNPs in the genes coding for intermediate proteins


MYC
Promoter
Intermediate
Enhancer
ZBTB7A
SP1
RBPJ
EGR1
E2F1
AR*
MDM4*
KDM2A*
CDKN2A*
CHD3*
ZBTB7A
YY1
TP73
TP53
TCF7L2*
STAT3
SREBF1
SP1
SMAD3
RELA
REL
RBPJ
NR2C2
NR2C1
MYB
MAFG
JUND
JUN
FOXH1
FOXA1
FOS
ETV5
ELF3
CREB1
BHLHE40
BCL6