Cis-regulatory Element Hijacking Overshadows Topological Changes in Prostate Cancer

James R. Hawley [1, 2, ‡], Stanley Zhou [1, 2,‡], Christopher Arlidge [1], Giacomo Grillo [1], Ken Kron [1], Rupert Hugh-White [8, 9, 10], Theodorus van der Kwast [3], Michael Fraser [1, 4], Paul C. Boutros [2, 6, 7, 8, 9, 10], Robert G. Bristow [1, 2, 11-14], Mathieu Lupien [1,2,5, *]

1. Princess Margaret Cancer Centre, University Health Network, Toronto, Canada
2. Department of Medical Biophysics, University of Toronto, Toronto, Canada
3. Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Canada
4. Department of Surgery (Urology), University of Toronto, Toronto, Canada
5. Ontario Institute for Cancer Research, Toronto, Ontario, Canada
6. Department of Pharmacology and Toxicology, University of Toronto, Toronto, Canada
7. Department of Human Genetics, University of California, Los Angeles, USA
8. Department of Urology, University of California, Los Angeles, USA
9. Institute for Precision Health, University of California, Los Angeles, USA
10. Jonsson Comprehensive Cancer Center, University of California, Los Angeles, USA
11. Department of Radiation Oncology, University of Toronto, Toronto, Canada
12. CRUK Manchester Institute and Manchester Cancer Research Centre, Manchester, UK
13. Division of Cancer Sciences, Faculty of Biology, Health and Medicine, University of Manchester, Manchester, UK
14. The Christie NHS Foundation Trust, Manchester, UK

‡ The authors co-led the study with equal contributions and can be interchangeably listed as first author.
* Corresponding author: Mathieu Lupien, mlupien@uhnresearch.ca

Keywords: prostate, cancer, topologically associating domains, genetic architecture, three-dimensional genome, CTCF, chromatin interaction, loop, structural variants, mutations, noncoding, cis-regulatory elements, H3K27ac, Hi-C, chromatin, epigenetics.

Abstract

Prostate cancer is a heterogeneous disease whose progression is linked to genome instability [1]. Despite large-scale tumour sequencing efforts, the impact of mutations on the genetic architecture in cancer remains ill-defined due to limited integration of genomics data across dimensions [2]. We addressed this limitation by assessing the impact of structural variants on the chromatin states and the three-dimensional organization across benign and malignant primary prostate genomes. We find high concordance in the three-dimensional genome organization between malignant and benign prostate tissues, arguing for constraints to the three-dimensional genome of prostate tumours. Moreover, we identify structural variants as effectors of changes to focal chromatin interactions, guiding *cis*-regulatory element hijacking [2,3] that imposes opposing expression changes on genes found at antipodes of a rearrangement. This leads to the repression of tumour suppressor gene expression and up-regulation of oncogenes, such as at the *TMPRSS2-ERG* and *PMEPA1-ZNF156* loci. Collectively, our results argue that *cis*-regulatory element hijacking by structural variants overshadows large-scale topological changes to alter gene regulation and promote oncogenesis.

Introduction

The human genome is organized into hubs of chromatin interactions within the nucleus, setting its three-dimensional topology [4]. These hubs of chromatin interactions, termed topologically associating domains (TADs), are rich in contacts between DNA sequences distant from each other in the linear scale, such as *cis*-regulatory elements (CREs) and their target gene promoters [5,6]. Insulating these hubs to prevent ectopic interactions are TAD boundaries that are maintained by CCCTC-binding Factor (CTCF) and the cohesin complex [7]. Disruption of TAD boundaries through genetic mutations or epigenetic alterations, such as aberrant DNA methylation, can activate oncogenes, as observed in medulloblastoma [3], acute myeloid leukemia [8], *IDH*-mutant gliomas [9] and salivary gland acinic cell carcinoma [10]. However, recent studies depleting CTCF or the cohesin complex produced little effect on gene expression despite global changes to the three-dimensional genome organization [2]. In contrast, CRE hijacking caused by genetic alterations results in large changes to gene expression in cancer, despite having little impact on the genome topology [11]. These contrasting observations raise questions about the interplay between components of the genetic architecture, namely, how genetic alterations, chromatin states, and the three-dimensional genome cooperate to regulate and misregulate genes. Hence, understanding the roles that chromatin organization and *cis*-regulatory interactions play in gene regulation is crucial for understanding how their disruption can promote oncogenesis.

Mutations can alter gene expression programs and protein function to drive cancer onset and progression [12]. Whilst coding mutations have been intensely investigated, recent studies reveal that noncoding mutations can similarly drive oncogenesis and disease progression by targeting CREs that are critical for gene regulation [11]. For instance, the *TERT* promoter harbours single-nucleotide variants (SNVs) driving *TERT* overexpression and producing

immortalized cancer cells [13–15]. Similarly, noncoding mutations have been found to target the CREs of the *ESR1* and *FOXA1* oncogenes in breast and prostate cancers, respectively [16,17]. In addition to SNVs, structural variants (SVs) have been reported to target CREs, resulting in aberrant gene expression programs. In prostate cancer (PCa), translocations, amplifications, and duplications of CREs for oncogenes such as *AR* [18], *FOXA1* [19,20] and *MYC* [20] lead to their overexpression. Importantly, while coding *FOXA1* mutations are found in ~10% of metastatic castration-resistant PCa patients, SVs that target the *FOXA1* CREs are found in over 25% of metastatic prostate tumours [20]. SVs driving cancer progression by targeting functional noncoding elements have also been reported in other cancers, such as the amplification of enhancers regulating *MYC* in lung adenocarcinoma and endometrial carcinoma [21], and *EGFR* in glioblastoma [22]. As such, we have historically underestimated the presence and driving role of noncoding mutations in cancer.

Despite large-scale tumour sequencing efforts identifying cancer drivers based on intra- and inter-tumour mutational frequencies [23], the genetic architecture in cancer remains ill-defined. This is partially due to limited integration of genetic alterations with chromatin states and the three-dimensional organization of the genome. To investigate this problem, we studied PCa, a disease where an estimated 97% of primary tumours contain SVs [1,24] and approximately 50% have overexpression of the *ERG* oncogene resulting from a fusion event on chromosome 21 (T2E fusion) [25,26]. In addition to oncogenic activation, SVs in prostate tumours disrupt and inactivate key tumour suppressor genes including *PTEN, BRCA2, CDK12* and *TP53* [19,26]. Furthermore, over 90% of prostate tumours contain complex SVs, including chromothripsis and chromplexy events [27]. Together, these results demonstrate the breadth of important roles SVs play in prostate tumours. In this work, we show that SVs in PCa repeatedly work through CRE hijacking to disrupt the expression of multiple genes with minimal impact to genome topology.

Results

The 3D genome is stable over oncogenesis

Chromatin conformation capture technologies enable the measurement of three-dimensional genome organization. These assays, however, are often limited to cell lines, animal models and liquid tumours due to the amount of input required [28]. Here, we conducted low-input Hi-C [29] on 10 μm thick cryosections from 12 primary prostate tumours and 5 primary benign prostate sections (see Methods, Supplementary Figure S1a-b). The 12 tumours were selected from the Canadian Prostate Cancer Genome Network (CPC-GENE) cohort previously assessed for whole-genome sequencing [1], RNA-seq [30] and H3K27ac ChIP-seq [31,32] (Supplementary Table 1). All 12 of these patients previously underwent radical prostatectomies. 6 of our 12 samples (50%) harbour the *TMPRSS2-ERG* genetic fusion (T2E) found in approximately half of the primary PCa patient population [1]. The total percent of genome altered ranges from 0.99%-18.78% (Supplementary Table 1) [1]. Our 12 tumour samples were histopathologically assessed to have ≥70% cellularity while the cellularity was ≥60% for our group of 5 normal prostate samples. Upon Hi-C library sequencing, we reached an average of 9.90×10^8 read pairs per sample (range 5.84×10^8 -1.49×10^9 read pairs) with minimal duplication rates (range 10.6% - 20.8%) (Supplementary Table 2). Pre-processing resulted in an average of 6.23×10^8 (96.13%) valid read pairs per sample (range 3.95×10^8 - 9.01×10^8, or 82.42 - 99.22%; Supplementary Table 2). Hence, we produced a high depth, high quality Hi-C library on 17 primary prostate tissues.

To characterize the chromatin organization of the primary prostate, we first identified TADs, hubs of preferential chromatin interactions that are postulated to control gene expression programs [33]. Across the 17 primary tissue samples, we observed an average of 2,305 TADs with a median size of 560 kbp (Figure 1a; Supplementary Tables 3-4). However, when

considering all hierarchical levels of TAD organization, we did not observe any global differences in the number of TADs identified across length scales (Figure 1a), nor in the strength of TAD boundaries (Figure 1b). This suggests few, if any, differences in three-dimensional genome organization at the TAD level between benign and tumour tissue. However, we observed differences in chromatin interactions around essential genes for PCa, previously profiled in cell lines. For example, chromatin interactions around the *AR* gene, previously found enriched in the 22Rv1 compared to RWPE1 prostate cell lines [34] were not recapitulated in either benign or tumour primary samples (Figure 1c). Moreover, when compared to other Hi-C datasets, the primary prostate samples clustered separately from cell lines, even after controlling for the restriction enzyme used and adjusting for sequencing depth (Supplementary Figure 1c). This suggests that primary tissues have disease-relevant differences in chromatin organization that are not recapitulated in cell line models. Cell lines derived from prostate cells (C4-2B, 22Rv1, and RWPE1) were most similar to the primary prostate samples (Supplementary Figure 1c). Median similarity scores between TADs in primary prostate tissues and cell lines was calculated at 72.1%, despite similar enrichment of CTCF binding sites near TAD boundaries (Supplementary Figure 1d). Comparably, the median similarity between prostate and non-prostate lines was calculated at 66.9%, and at 63.5% between primary prostate and non-prostate lines (Supplementary Figure 1c). Collectively, these results suggest that phenotypic differences between benign and tumour tissues cannot be explained by differences in large-scale three-dimensional genome organization, alone.

We next assessed whether there were differences in chromatin interactions connecting genomic features distal from each other on the linear genome, such as distal CREs and target gene promoters between benign and tumour tissues. We detected interactions for each sample, identifying 16,474 unique interactions across the entire cohort (n=17), of an estimated 20 602 interactions (~80% saturation) across primary tissues based on a nonlinear least squares

asymptotic regression (Supplementary Figure 1e). We detected a median of 4,416 interactions per sample (range 1,292 - 7,040; Supplementary Table 5). Amongst these detected interactions, we identified known contacts in PCa such as between two distal CREs on chromosome 14 and the *FOXA1* promoter [17] (Figure 1d), and CREs upstream of *MYC* on chromosome 8 that are frequently duplicated in metastatic disease [19] (Supplementary Table 5). We also identified two novel interactions in the primary prostate tumours to the *FOXA1* promoter missing in PCa cell lines. This suggests that the regulatory plexus of *FOXA1* consists of more CREs than previously reported (Figure 1d). We next compared the interactions between benign prostate and tumour samples that were detected in at least 3 samples, yielding 533 tumour- and 40 benign-specific interactions (Supplementary Figure 1f). However, upon visual inspection and aggregate peak analysis, these differences appear to be subtle (Supplementary Figure 1g). Differential interactions between benign and tumour samples may thus be artefacts of detection, or only marginally different between these two conditions. Together, these results suggest that chromatin interactions are globally stable across benign prostate and tumours.

SV patterns differ across PCa subtypes

In prostate tumours, SVs populate the genome to aid disease onset and progression [1,19]. Advances in computational methods now enable the identification of SVs from Hi-C datasets [35,36]. Applying SV callers to our primary prostate tumour Hi-C dataset (See Methods), we found evidence of the *TMPRSS2-ERG* (T2E) genetic fusion spanning the 21q22.2-3 locus in 6/12 (50%) patients (CPCG0258, CPCG0324, CPCG0331, CPCG0336, CPCG0342, and CPCG0366) (Figure 2a), in accordance with previous whole-genome sequencing (WGS) findings [1]. We next computationally searched for SV breakpoints genome-wide [35], detecting a total of 317 unique breakpoints with a median of 15 unique breakpoints per tumour (range 3-95; Supplementary Figure 2a; Supplementary Table 6). Combining unique breakpoint pairs

into rearrangement events yielded 7.5 total events on average per patient (range 1 - 36, Figure 2b; Supplementary Figure 2b-c). These numbers are smaller than previously reported from matched WGS data [1]; however, the median distance between breakpoints on the same chromosome was much larger at 31.6 Mbp for Hi-C-identified breakpoints, compared to 1.47 Mbp from WGS-identified breakpoints (Supplementary Figure 2d). We also identified more inter-chromosomal breakpoint pairs with the Hi-C data in 11 of 12 tumours (Supplementary Figure 2a), including a novel translocation event that encompasses the deleted region between *TMPRSS2* and *ERG* into chromosome 14. This is consistent with the inherent nature and resolution of the Hi-C method to detect larger, inter-chromosomal events [35]. No SVs were detected in the 5 primary benign prostate tissue samples. While this does not rule out the presence of small rearrangements undetectable by Hi-C due to its low resolution, the absence of large and inter-chromosomal SVs supports a difference in genome stability between benign and tumour tissues [1,27,32,37]. Collectively, Hi-C defines a valid method to interrogate for the presence of SV in tumour samples, compatible with the detection of intra as well as inter-chromosomal interactions otherwise missed in WGS analyses.

By conducting a global assessment for SVs across our cohort of primary prostate tumours, we revealed a significant preference for SVs to populate T2E+ primary prostate tumours compared to T2E- tumours. T2E+ tumours had a median of 17 events whereas T2E- tumours had a median of 4.5 events (one-sided Mann-Whitney U test, p = 6.296e-3; Figure 2c). This significant difference between the T2E subtypes can also be seen upon subdividing the breakpoint pairs into inter- or intra-chromosomal events. T2E+ tumours had a median of 10.5 intra-chromosomal breakpoint pairs compared to a median of 1 for T2E- tumours (one-sided Mann-Whitney U test, p = 1.211e-2; Figure 2d). Similarly, T2E+ tumours had a median of 11 inter-chromosomal breakpoint pairs compared to 4 for T2E- tumours (one-sided Mann-Whitney U test, p = 1.388e-2; Figure 2e). This difference in abundance of breakpoints is not

restricted to intra-chromosomal events but can rather also be seen across all chromosomes between the two subclasses of tumours (Figure 2f). This corroborates WGS-based results, where T2E+ tumours harbour more SV breakpoints than T2E- tumours (Mann-Whitney U test, n = 130, W = 2814.5, p = 5.384e-4) [1]. Aside from the T2E fusion, no recurrent events were found in this cohort. Moreover, very few loci were recurrently altered in multiple tumours as 396/430 (92.1%) megabase bins contained an SV breakpoint from a single tumour (Supplementary Figure 2e). Together, these results show that SVs arise in primary prostate tumours and are 3 times more frequent, on average, in T2E+ patients than T2E- patients.

Amongst SVs detected in primary prostate tumours, we identified both simple and complex chains of breakpoints. While simple SVs correspond to fusion between two distal DNA sequences, complex chains are evidence of chromothripsis and chromoplexy [27]. These genomic aberrations affecting multiple regions of the genome are known to occur in both primary and metastatic PCa [1,24,27]. The chains can be pictured as paths connecting breakpoints in the contact matrix (Supplementary Figure 2c). 8 of the 12 (66.7%) tumour samples contained these chains, including one patient (CPCG0331) harbouring 11 complex events and three patients (CPCG0246, CPCG0345, and CPCG0365) each harbouring a single complex event. We observed a median of 1 complex event per patient (range 0-11) consisting of a median of 3 breakpoints (range 3-7) spanning a median of 2 chromosomes per event (range 1-4, Supplementary Table 7, Supplementary Figure 2f). In particular, patient CPCG0331 had 11 complex events, including a 6-breakpoint event spanning 3 chromosomes (Supplementary Figure 2b). A highly rearranged chromosome 3 was also found in the same patient (Figure 2g). The most common type of complex event involved 3 breakpoints and spanned 2 chromosomes, occurring 9 times across 5 of the 8 patients with complex events. While not significant, the T2E+ patients' trend towards having more complex SVs than the T2E- patients (one-sided Mann-Whitney U test, p = 0.1091), in accordance with previous findings [27]. Notably, complex

events involved significantly more chromosomes than simple events (Mann-Whitney U test, W = 2087, p = 8.086e-5). We did not identify any complex events in the benign primary prostate tissue samples since no breakpoints were identified. In summary, using Hi-C, we detected both simple and complex SVs in primary prostate tumours not previously identified using WGS-based methods. We were able to identify known observations, such as a highly mutated region on chromosome 3, as well as find novel inter-chromosomal events not previously reported [1].

TADs are principally immutable to SVs

Having delineated SVs from Hi-C data, we next systematically examined the impact of SVs on TAD structure. This led us to look at the intra-TAD and inter-TAD interactions around each breakpoint. We observed that only 18 of the 260 (6.9%) TADs containing SV breakpoints were associated with decreased intra-TAD or increased inter-TAD interactions (Figure 3a). 12 of 18 (66.7%) occurrences were within T2E+ tumours. We found no evidence that simple versus complex SVs were a factor in determining whether a TAD was altered (Pearson's chi-square test, $X^2 = 0.0166$, p = 0.8974, df = 1). Similarly, the type of SV (a deletion, inversion, duplication, or translocation) was not predictive of whether the TAD would be altered (Pearson's chi-square test, $X^2 = 4.7756$, p = 0.3111, df = 4). Overall, we find that SVs are associated with topological changes in a small percentage of cases, but the presence of an SV is not predictive of an altered TAD nearby.

Despite the evidence that SVs rarely impact large-scale chromatin topology, we evaluated whether SVs affected the expression of genes within the TADs surrounding the breakpoint using patient-matched RNA-seq data[30]. We found that 23 of 260 breakpoints (8.8%) are associated with significant changes to local gene expression programs (Figure 3b). Complex events can have different effects at each breakpoint. For example, the T2E fusion in

one patient (CPCG0366) leads to overexpression of *ERG* and under-expression of *TMPRSS2* [1,31]. However, the deleted locus between these two genes was found inserted into chromosome 14 as part of a complex translocation event (Figure 3c-f). The inserted locus positions *ERG* towards the 5' end of the *RALGAPA1* gene and *TMPRSS2* towards the 3' end (Figure 3c). This translocation is associated with a significant drop in intra-TAD contacts on chromosome 14 (two-sample unpaired t-test, t = 6.38, p=1.04e-9; Figure 3d). However, this insertion is not associated with any significant changes to expression for any gene in the TAD on chromosome 14 (Figure 3e). Thus, altered TADs are not sufficient to alter gene expression, as is seen in this case. Moreover, evaluating the impact of SVs requires considering genes around all breakpoints.

Conversely, TAD alterations are not required changes to gene expression. As part of a complex SV involving the *RIMBP2* gene (Figure 3g-j), both ends of the gene contain breakpoints (Figure 3g). This rearrangement is not associated with changes to intra-TAD contacts (two-sample unpaired t-test, t = 0.8101, p = 0.4183; Figure 3h). However, *RIMBP2* is over-expressed in this patient (Figure 3i). More generally, only a single breakpoint was observed with both TAD contact and gene expression changes, although we did not find evidence to suggest these are dependent events (Pearson's chi-square test, X^2 = 6.31e-3, p = 0.9367, df = 1). For TADs where at least one gene was differentially expressed, 19 (83%) of them had at least one gene with doubled or halved expression. Notably, we found that inter-chromosomal translocations are associated with altering the expression of genes nearby their breakpoints compared to intra-chromosomal breakpoints (Pearson's chi-square test, X^2 = 7.0088, p = 0.00811, df = 1; Supplementary Figure 3). Taken together, these results suggest that while SVs can alter TAD contacts, this is neither necessary nor sufficient to alter gene expression.

SVs hijack CRE to alter antipode genes

We next investigated other modes through which gene expression can change as a result of rearrangements. Focusing on oncogenes near SV breakpoints, we identified an inter-chromosomal SV connecting the q arm of chromosome 7 and the p arm of chromosome 19 centring on *BRAF* (Figure 4a). This breakpoint separates the last few exons of *BRAF* from its promoter and upstream enhancers while leaving the rest of the gene intact (Figure 4b). Concomitantly, the 3' end of *BRAF* has enriched SV-associated focal interactions at multiple active CREs on chromosome 19 (Figure 4b). Under the CRE hijacking model, this would predict that the most 3' exons of *BRAF* would be over-expressed in this patient and not the isoforms that exclude these last exons. Using matched RNA-seq data, this is in fact what we find, with an estimated 5 times overexpression compared to other tumours (fold-change = 4.976, FDR = 0.0181; Figure 4c). This suggests that the overexpression of disease-relevant oncogenes in PCa may result from CRE hijacking mediated by SVs.

To investigate the role of hijacking CREs in PCa more comprehensively, we considered all 22 SVs associated with altered gene expression near a breakpoint (Figure 4d). This resulted in 54 differentially expressed genes across the 22 SVs. 16 (72.7%) of these SVs are associated with altered expression of multiple genes. Notably, 15 of these 16 SVs (93.8%) are associated with both over- and under-expression of genes, instead of genes all being either over-expressed or under-expressed (Figure 4d-e). 12 of these 15 (80%) SVs are associated with expression changes at SV antipodes, opposite ends of a breakpoint pair (Figure 4f). Across 8 of 12 (66.7%) SVs we observed focal topological changes directly engaging with the body of differentially expressed genes at SV antipodes (Figure 4f). The remaining focal topological changes are indirectly linked to differentially expressed genes at SV antipodes (Figure 4f). These observations suggest that many SVs alter the expression of multiple genes, simultaneously, by bringing them into contact within the nucleus.

The T2E fusion is an example of this phenomenon, where the *TMPRSS2* promoter is hijacked by the *ERG* gene through the fusion of its promoter upstream of the *ERG* exons [25,31]. This CRE hijacking event results in overexpression of *ERG* and under-expression of *TMPRSS2* (Figure 4g-i) that coincides with H3K27ac histone hyperacetylation over the *ERG* gene body and histone hypoacetylation over the *TMPRSS2* gene body [31]. The *PMEPA1-ZNF156* fusion is another example of opposing expression changes at antipodes associated with acetylation changes extending beyond the breakpoint (Figure 4j-l). Specifically, the *PMEPA1-ZNF516* fusion leads to *PMEPA1* under-expression (Figure 4l) concomitant with histone hypoacetylation at its 3' region, despite its promoter region showing no reduction in acetylation (Figure 4k). Conversely, no changes in histone acetylation are detectable over the *ZNF516* promoter but its gene body shows histone hyperacetylation associated with its overexpression in the fusion positive tumour (Figure 4k, l). Overexpression of the oncogene *ERG* and suppression of the tumour suppressor *PMEPA1* showcases how SVs hijack CREs that result in opposing expression changes of genes at SV antipodes that contribute to disease onset.

Discussion

Genetic contributions to a given phenotype is entrusted to the chromatin. To capture the complexity of chromatin across dimensions, we explored the three-dimensional genome organization in primary prostate tumours of known genetic and chromatin identity. We found that oncogenesis has limited impact on genome topology. Instead, PCa development is paired with the acquisition of SVs hijacking CREs to alter the expression of their target genes, commonly resulting in opposing expression changes. Considering the contribution of SVs across human cancers [38], our collective work presents a framework inclusive of genetics, chromatin and three-dimensional genome organization to understand the genetic architecture across individual primary prostate tumours.

Changes to the three-dimensional genome organization reported in disease onset or development are often inferred from alterations in TAD boundaries [2,39]. For instance, gains in DNA methylation at CTCF binding sites are linked to altered TAD structures in gliomas [9]. CTCF activity is also targeted by somatic mutations that enrich at its binding sites in colorectal, oesophageal, and liver cancers [40,41]. In primary PCa however, CTCF binding sites are not enriched with somatic mutations [32]. Furthermore, 97% of differentially methylated regions genome-wide in primary PCa are losses of DNA methylation [42,43] which have previously been shown to have limited impact on CTCF chromatin binding [44]. This suggests that altered CTCF binding at TAD boundaries may not underlie PCa development. In agreement, we find stable TAD structures between benign and primary prostate tumours as well as across T2E+ and T2E- tumours. This suggests that large disruptions to topology may not be necessary for transformation or divergent subtyping of prostate tumours, corroborating previous observations of conserved TADs across cell types [5,45]. These findings stand in contrast with how extrachromosomal circular DNA acquired during oncogenesis engage sequences that would

otherwise be constrained by the chromatin and topology [22,46,47]. However, the stable genome topology we observe is consistent with the conserved topologies seen between homologous regions in separate species [5,48], suggesting that TADs do not necessarily need to split or merge with neighbouring TADs to facilitate changes in gene regulation. Instead, SVs, chromatin states influenced by histone modifications or DNA methylation, better discriminate T2E+ and T2E- tumours than three-dimensional genome organization alone.

The genome of prostate tumours is populated with mutations that target CREs and promote oncogenesis by altering gene expression [17,19,20,32]. For instance, mutated CREs can alter the oncogenic expression of *ERG* [31], *FOXA1* [17–20] and *AR* [18,19,49]. Here, we observed similar findings whereby SVs hijack CREs to alter gene expression and focal chromatin interactions without interfering with TAD structures. We further demonstrate that CRE hijacking by SVs leads to opposing gene expression changes at SV antipodes, whereby genes on one flank of the breakpoint are upregulated while genes on the other flank are repressed. Gene expression changes are concomitant with histone hyperacetylation over the body of upregulated genes, in contrast to histone hypoacetylation over the body of repressed genes. Opposing changes in gene expression is not restricted to intra-chromosomal SVs such as seen at the *TMPRSS2-ERG* fusion event [31]. It also occurs in between different chromosomes, such as observed with the *ZNF516-PMEPA1* translocation event. It must be noted that it is common to identify SVs to be a part of chained, chromplexic rearrangement events as previously reported [27], suggesting that multiple instances of opposing gene expression changes may exist within the same chain of events. These insights stress the importance of investigating all breakpoints in SVs to assess the biological impact of these mutations on the *cis*-regulatory landscape, as opposed to focusing on CREs as single entities.

In conclusion, by bypassing technical limitations to characterize the three-dimensional genome organization across benign and tumour primary prostate tissue [29], our work reveals the

predominant stable nature of large-scale genome topology across oncogenesis. Instead, alterations to discrete CREs, reported as SV-mediated CRE hijacking events and reflected in focal topological changes, populate the PCa genome. Considering previous reports of CRE disruption by germline and somatic SNVs, our findings support the predominant contribution for noncoding genetic alterations to the genetic architecture of cancer.

Materials & Methods

Patient Selection Criteria

Patients were selected from the CPC-GENE cohort of Canadian men with indolent PCa, Gleason scores of 3+3, 3+4, and 4+3. The intersection of previously published data for whole genome sequencing [1], RNA abundance [30], and H3K27ac ChIP-seq [31] led to 25 samples having data for all assays. 11 of these tested positive for ETS gene family fusions (T2E status), and 14 without. To accurately represent the presence of this subtype of PCa in the disease generally, and to ensure minimum read depths required to perform accurate analysis on chromatin conformation data, we selected approximately half of these remaining samples (6 T2E+ and 6 T2E-).

Patient Tumour in situ low-input Hi-C Sequencing

We followed the general *in situ* low input Hi-C (Low-C) protocol from Díaz *et al.* [29], with our own re-optimization for solid tumour tissue sections. It is worth noting that throughout the protocol, the pellet would be hardly visible and would require careful pipetting. The specific modifications of the protocol are described below.

Tumour Tissue Preparation

Thirteen cryopreserved-frozen PCa tumour tissue specimens were obtained from primary PCa patients as part of the Canadian PCa Genome Network (CPC-GENE) effort [1]. Informed consent was obtained from all patients with REB approval (UHN 11-0024). These tumour specimens were sectioned into 10 µm sections. Sections before and after the sections used for Hi-C were stained with hematoxylin and eosin (H&E) and assessed pathologically for ≥ 70% PCa cellularity. The percentage of infiltrating lymphocytes was also estimated by pathological assessment to be ≤3%. Stratification into *TMPRSS2-ERG* (T2E)-positive or T2E-

negative was determined through either whole-genome sequencing detection of the rearrangement, immunohistochemistry or mRNA expression microarray data [1].

Normal Tissue Preparation

Five snap-frozen prostate tumour-adjacent normal tissue specimens were obtained. Informed consent was obtained from all patients with REB approval (UHN 11-0024). Tissue specimens were sectioned into 5, 10, and 20 µm sections. Sections used for Hi-C and RNA-seq were stained with H&E and assessed pathologically for ≥ 60% prostate glandular cellularity.

Fixation and Lysis

One or two sections (consecutive; depending on surface area) for each patient were thawed and fixed by adding 300 µL of 1% formaldehyde in PBS directly onto the tissue sample, followed by a 10-minute incubation at room temperature (RT) (Supplementary Figure 1b). The formaldehyde was quenched by adding 20 µL of 2.5M glycine to the sample reaching a final concentration of 0.2M followed by 5 minutes of incubation at RT. The samples were then washed three times with 500 µL cold PBS and scraped off the microscope slide with a scalpel into 1.5 mL centrifuge tube containing 250 µL of ice-cold Low-C lysis buffer (10 mM Tris-Cl pH 8.0, 10 mM NaCl, 0.2% IGEPAL CA-630 (Sigma-Aldrich)) supplemented with protease inhibitor. The samples were then mixed thoroughly by gentle pipetting and left on ice for 20 minutes with intermittent mixing. Upon lysis, the samples were snap-frozen with liquid nitrogen and stored at -80 °C until processing the next day. As a note, stagger fixation times when processing multiple samples to prevent needless rush and chance of under/over-fixation.

Enzyme Digestion and Overhang Fill-In

The samples stored at -80 °C were thawed on ice and spun down at 300 x g for 5 minutes at 4 °C. The samples were then resuspended in 125 µL of ice-cold 10X NEB2 Buffer (New

England Biolabs), and again spun down at 13,000 x g for 5 minutes at 4 °C. The pellet was then resuspended in 25 µL of 0.4% SDS and incubated at 65 °C for 10 minutes without agitation for permeabilization. To quench the SDS, 10% Triton X-100 in water (12.5 µL + 75 µL water) was then added to the samples and incubated at 37 °C for 45 minutes at 650 rpm. For enzymatic digestion, 35 µL of 10X NEB2.1 buffer (New England Biolabs) was added to each sample, follow by the addition of 50 U of MboI and 90 minutes incubation at 37 °C with gentle agitation (add 30 U first, incubate 45 minutes, followed by the addition of another 20 U and another 45 minutes of incubation). Upon digestion, the MboI enzyme was inactivated by incubating at 62 °C for 20 minutes. The overhangs generated by the MboI enzyme was then filled-in by adding a mix of dNTPs and DNA Polymerase I Klenow Fragment directly to each sample (10 µL of 0.4 mM biotin-14-dCTP, 0.5 µL of 10 mM dATP, 0.5 µL of 10 mM dGTP, 0.5 µL of 10 mM dTTP, 4 µL of 5U/µL DNA Polymerase I Klenow Fragment). The samples were then mixed by gentle pipetting followed by incubation at 37 °C for 90 minutes with gentle agitation.

Proximity Ligation and Decrosslinking

Upon overhang fill-in, each sample was subject to proximity ligation through the addition of 328.5 µL water, 60 µL of 10X T4 DNA Ligase Buffer (ThermoFisher Scientific), 50 µL of 10% Triton X-100, 6 µL of 20 mg/mL BSA (New England Biolabs) and 3.5 µL of 5 Weiss U/µL T4 DNA Ligase (ThermoFisher). The samples were mixed through gentle pipetting and incubated at RT (20-22 °C) with rotation for 4 hours. The samples were then spun down at 13,000 x g for 5 minutes at RT and resuspended in 250 µL of Extraction Buffer (50 mM Tris-Cl pH 8.0, 50 mM NaCl, 1 mM EDTA, 1% SDS) upon removal of supernatant. Next, 10 µL of 20 mg/mL Proteinase K (New England Biolabs) was added to each sample and incubated at 55 °C for 30 minutes at 1,000 rpm. Then 65 µL of 5 M NaCl was added to each sample and incubated at 65 °C at 1,000 rpm overnight.

DNA Extraction

Phenol-chloroform extraction columns were spun down at 17,000 x g for 1 minute at 4 °C to get gel down to the bottom of the tube. The samples incubated overnight were then added to the column. Next, an equal volume (~325 µL) of phenol-chloroform-isoamyl alcohol mixture (25:24:1) (Sigma) was also added to the column. The column was then inverted for thorough mixing and spun down at 17,000 x g for 5 minutes at 4 °C. The surface layer on top of the gel upon spinning contains the sample and is transferred to a clean 1.5 mL tube (~325 µL). Each sample was mixed with 31.5 µL of 3M sodium acetate, 2 µL of GlycoBlue (ThermoFisher Scientific), and 504 µL of 100% ethanol for DNA precipitation. The samples were inverted several times for mixing and incubated at -80 °C for 20 minutes, followed by a centrifuge spin at 17,000 x g for 45 minutes at 4 °C. The supernatant was carefully discarded and the pellet was washed with 800 µL of ice-cold 70% ethanol followed by a centrifuge spin at 17,000 x g for 5 minutes at 4 °C. The supernatant was then discarded and the tube was air-dried until no traces of ethanol was left prior to dissolving the DNA pellet with 30 µL of Elution Buffer (Qiagen PCR Clean-Up Kit). 1 µL of RNase A (ThermoFisher Scientific) was added to each sample followed by incubation at 37 °C for 15 minutes. A mix of 5 µL of 10X NEB2.1 buffer (New England Biolabs), 1.25 µL of 1 mM dATP, 1.25 µL of 1 mM dCTP, 1.25 µL of 1 mM dGTP, 1 mM of dTTP, 0.5 µL of 10 mg/mL BSA, 5 µL of water, 3.5 µL of 3 U/µL T4 DNA Polymerase (New England Biolabs) was added to each sample. The samples were mixed thoroughly by gentle pipetting, and then incubated at 20 °C for 4 hours.

Fragmentation and Biotin Pull-down

70 µL of water was added to each sample bringing total volume up to 120 µL, and the samples were transferred into Covaris sonication tubes. The samples were then sonicated using

Covaris M220 sonicator to attain 300-700 bp fragments. For biotin pull-down using a magnetic rack, 30 µL of Dynabeads MyOne Streptavidin C1 beads (Life Technologies) for each sample was washed once with 400 µL of 1X B&W buffer + 0.1% Triton X-100. The beads were then resuspended in 120 µL of 2X B&W buffer and transferred to the 120 µL of sample (1:1 ratio). The sample was then incubated with gentle rotation at RT for 20 minutes. The supernatant was discarded and the beads were resuspended with 400 µL of 1X B&W buffer + 0.1% Triton X-100 followed by a 2-minute incubation at 55 °C with mixing. The wash was repeated once more, then resuspended in 400 µL of 1X NEB2 buffer (New England Biolab).

Library Preparation and Size Selection

The beads containing the Hi-C samples were separated on a magnetic rack to remove the supernatant. The beads were then resuspended in a total volume of 10 µL for library preparation using the SMARTer ThruPLEX DNA-seq library preparation kit (Takara Biosciences) per manufacturer's protocol with an adjustment on the last step, a PCR reaction for library amplification. Upon reaching that step, the reaction was carried out on a regular PCR for two cycles to amplify the Hi-C samples off the streptavidin beads. Next, the samples were transferred onto a new tube where 20X SYBR was added. The samples were then subject to real-time qPCR and pulled out from the qPCR machine mid-exponential phase. Ultimately, this is done to reduce PCR duplication rates, a huge limitation for low-input Hi-C protocols. The Hi-C libraries were then double size-selected for 300-700 bp using Ampure XP beads and sent for BioAnalyzer analysis prior to sequencing.

Hi-C Sequencing and Data Pre-processing

Sequencing

The Hi-C libraries for each tumour sample were sent for shallow paired-end 150 bp sequencing (~10-15 million reads per sample) on NextSeq 500. Upon confirming library quality and low duplication rates (< 20%), samples were sent for deep paired-end 150 bp sequencing with the aim of ~1 billion raw reads per sample on NovaSeq 6000.

Sequence alignment and Hi-C artefact removal

Paired-end FASTQ files were pre-processed with HiCUP (v0.7.2). Reads were truncated at MboI ligation junction sites prior to alignment with `hicup_digester`. Each mate was independently aligned to the hg38 genome and were then paired and assigned to MboI restriction sites by `hicup_map`. `hicup_map` uses Bowtie2 (v2.3.4) as the underlying aligner which has the following parameters: `--very-sensitive --no-unal --reorder`. Reads that reflect technical artefacts were filtered out with `hicup_filter`. Duplicate reads were removed with `hicup_deduplicator`.

Reads that came from different sequencing batches were then aggregated for each tumour sample at this stage using `sambamba merge` (v0.6.9). This resulted in an average of 1.12×10^9 read per tumour sample (Supplementary Table 2).

Contact matrix generation and balancing

Aggregated binary alignment map (BAM) files were converted to the pairs format using pairtools (v0.2.2) and then the cooler format using the cooler package (v0.8.5). The pairs files were generated with the following command: `pairtools parse -c {genome} --assembly hg38 -o {output_pairs} {input_bam}`. The cooler files were generated at an initial matrix resolution

of 1000 bp with the following command: `cooler cload pairs --assembly hg38 -c1 2 -p1 3 -c2 4 -p2 5 {genome}:1000 {input_pairs} {output_cooler}`.

The raw contact matrices stored in the cooler file format were balanced using cooler's implementation of the ICE algorithm using the `cooler balance` command. Contact matrices at different resolutions were created with the `cooler zoomify` command.

Hi-C Data Analysis

TAD identification

Contact matrices were binned at a resolution of 40 kbp. To remove sequencing depth as a confounding factor, contact matrices for all samples were first downsampled to match the sequencing depth of the shallowest sample. For comparisons including cell lines, this was $120x10^6$ contacts (Figure 1a). For comparisons only involving primary samples, this was $300x10^6$ contacts (Figure 1b-c). This was achieved with Cooltools (v0.3.2) with the following command: `cooltools random-sample -c 120000000 {input}::/resolutions/40000 {output}`.

TADs were identified using TopDom on the downsampled, ICE-normalized contact matrices. To identify domains at multiple length scales, similar in concept to Artamus' gamma parameter , TopDom was run multiple times per sample, with the window size parameter set at values between 3 and 40, inclusive (corresponding to 120 kbp and 1.6 Mbp). The lower bound for the window size parameter allowed for the identification of domains multiple megabases in size at the upper end and domains < 100 kbp at the lower end without being dominated by false calls due to sparsity of the data.

Given the stochasticity of Hi-C sequencing, boundaries called at one window size may not correspond to the exact same location at a different window size. To attempt to resolve these different boundary calls and leverage power from multiple window sizes, boundaries for a given patient were considered at all window sizes. Boundaries within one bin (40 kbp) of

each other and called at different window sizes were marked as conflicting calls. If only two boundaries were in conflict and all the window sizes where the first boundary was called are smaller than the window sizes where the second boundary was called, the second boundary was selected since larger smoothing windows are less sensitive to small differences in contact counts. If only two boundaries were in conflict but there is no proper ordering of the window sizes, the boundary that was identified most often between the two was selected. If three boundaries are in conflict, the middle boundary was selected. If four or more boundaries were in conflict, the boundary that was identified most often was selected.

To determine the maximum window size for TAD calls, TAD calls were compared across window sizes for the same patient using the BPscore metric [50]. TAD calls are identical when the BPscore is 0, and divergent when 1. The cut-off window size for a single patient was determined when the difference between TAD calls at consecutive window sizes was < 0.005, twice in a row. The maximum window size was determined by the maximum window size cut-off across all samples in a comparison. For comparisons involving only primary samples, the maximum window size was determined to be w = 20 x 40 kbp. For comparisons involving cell lines, this was w = 32 x 40 kbp.

The persistence of a TAD boundary was calculated as the number of window sizes where this region was identified as a boundary.

Sample clustering by TADs

Using the TAD calls at the window size w = 32 x 40 kbp, the similarity between samples was calculated with BPscore. The resulting matrix, containing the similarity between any two samples, was used as the distance matrix for unsupervised hierarchical clustering with Ward.D2 linkage.

Identification of significant chromatin interactions

Chromatin interactions were identified in all 17 primary samples with Mustache. Using the Cooler files from above, Mustache was run on the ICE-normalized 10 kbp contact matrix for each chromosome with the following command: `mustache -f {input} -r 10000 -ch {chromosome} -p 8 -o {output}`. Interaction calls on each chromosome were merged for each sample to create a single table of interaction calls across the entire genome.

To account for variances in detection across samples and to identify similarly called interactions across samples, interaction anchors were aggregated across all samples to form a consensus set. Interaction anchors were merged if they overlapped by at least 1 bp. Interaction anchors for each sample were then mapped to the consensus set of anchors, and these new anchors were used in all subsequent analyses.

Chromatin interaction saturation analysis

To estimate the detection of all chromatin interactions across all samples, a nonlinear regression on an asymptotic model was performed. This is similar in method to peak saturation analysis used to assess peaks detected in ChIP-seq experiments from a collection of samples [31]. Bootstrapping the number of unique interactions detected in a random selection of *n* samples was calculated for *n* ranging from 1 to 17. 100 iterations of the bootstrapping process were performed. An exponential model was fit against the mean number of unique interactions detected in *n* samples using the `nls` and `SSaymp` functions from the stats R package (v3.6.3). The model was fit to the following equation:

$$\mu = \alpha + (R_0 - \alpha)\exp(kn)$$

Where μ is the mean number of chromatin interactions for a given number of samples, n, α is the asymptotic limit of the total number of mean detected interactions, R_0 is the response for $n = 0$, and k is the rate constant. The estimated fit was used to predict the number of samples

required to reach 50%, 90%, 95%, and 99% saturation of the asymptote (Supplementary Figure 1d).

Structural variant breakpoint pair detection

Breakpoint pairs for each patient were called on the merged BAM files using `hic_breakfinder` (commit 30a0dcc6d01859797d7c263df7335fd2f52df7b8) [35]. Pre-calculated expected observation files for the hg38 genome were downloaded from the `hic_breakfinder` GitHub repository on Jul 24, 2019, as per the instructions. Breakpoints were explicitly called with the following command: `hic_breakfinder --bam-file {BAM} --exp-file-inter inter_expect_1Mb.hg38.txt --exp-file-intra intra_expect_100kb.hg38.txt --name {Sample ID} --min-1kb`.

For the T2E fusion, only one patient had the deletion identified by hic_breakfinder with default parameters (CPCG0336). Difficulties identifying SVs with hic_breakfinder have been previously noted [36]. After adjusting the detection threshold, we were able to identify the fusion in other samples. To ensure the T2E+ tumours were effectively stratified for future analyses, the fusion was annotated using the same coordinates for the other T2E+ samples. No other additions to breakpoint calls were made. Certain breakpoints that appeared to be artefacts were removed, as described below.

Structural variant annotation and graph construction

The contact matrix spanning 5 Mbp upstream and downstream around the breakpoint pairs were plotted and annotated according to previously published heuristics (Supplementary Figure 4 for [35]). Breakpoint pairs that were nearby other breakpoints or did not match the heuristics in this figure were labelled as unknown. These annotations were matched against the

annotations identified from the previously published whole genome sequencing structural variants [1].

Breakpoint pairs matching the following criteria were considered as detection artefacts and were ignored.

1. At least one breakpoint was > 1 Mbp
2. At least one breakpoint was surrounded by empty regions of the contact matrix
3. At least one breakpoint corresponded to a TAD or compartment boundary shared across all samples that lacked a distinct sharp edge that is indicative of a chromosomal rearrangement

To identify unique breakpoints that were identified in multiple breakpoint pairs, breakpoints that were within 50 kbp of each other were considered as possibly redundant calls. This distance was considered as the resolution of the non-artefactual calls is 100 kbp. Plotting the contact matrix 5 Mbp around the breakpoint, breakpoints calls were considered the same breakpoint if the sharp edge of each breakpoint was equal to within 5 kbp.

Similar in concept to the ChainFinder algorithm [27], we consider each breakpoint as a node in a graph. Nodes are connected if they are detected as a pair of breakpoints by `hic_breakfinder`. Simple structural variants are connected components in the breakpoint graph containing only two nodes, and complex variants those with greater than two nodes. A visual representation of these graphs can be found in Supplementary Figure 2.

Determination of structural variant breakpoints altering TAD boundaries

Patients are assigned into one of two groups using hierarchical clustering (complete linkage) with the matrix of pairwise BPscore [50] values as a distance matrix. If the clustering equals the mutated samples from the non-mutated samples (i.e., the clustering matches the

mutation status in this locus), then the local topology was classified as “altered” as a result of the SV.

Virtual 4C

Two parts of the *BRAF* gene were used as anchors for virtual 4C data: the promoter region (1500 bp upstream, 500 bp downstream of the TSS) and the entire gene downstream of the breakpoint. Contact frequencies from the ICE-normalized, 10 kbp contact matrices were extracted, with the rows as the bins containing the anchor and the columns as the target regions (the x-axes in Figure 4k). The row means were calculated to produce a single vector where each element is the average normalized contact frequency between the anchor of interest and the distal 10 kbp bin. These vectors were plotted as lines in Figure 4k.

Patient Tumour Tissue H3K27ac ChIP-seq

Sequence alignment

ChIP-seq against H3K27ac was previously published for these matching samples in [31]. Sequencing data was processed similarly to the previous publication of this data[31]; however, the hg38 reference genome was used instead of hg19. FASTQ files from single-end sequencing were aligned to the hg38 genome using Bowtie2 with the following parameters: `-x {genome} -U {input} 2> {output_report} | samtools view -u > {output_bam}`. For FASTQ files from paired-end sequencing, only the first mate was considered and reads were aligned with the following parameters: `-x {genome} -U {input} -3 50 2> {output_report} | samtools view -u > {output_bam}`. This ensured that all H3K27ac ChIP-seq data had the same format (single-end) and length (52 bp) before alignment to mitigate possible differences in downstream analyses due to different sequencing methods. Duplicate reads were removed with `sambamba markdup -r` and were then sorted by position using `sambamba sort`.

Peak calling

Peak calling was performed using MACS2 (v2.1.2) with the following command: `macs2 callpeak -g hs -f BAM -q 0.005 -B -n {output_prefix} -t {seq_chip} -c {seq_input}`. ENCODE hg38 blacklist regions were then removed from the narrow peaks. Peaks calls are in Supplementary Table 8.

Differential acetylation analysis

Unique peak calls and deduplicated pull-down and control BAM files from tumour samples were loaded into R with the DiffBind package (v2.14.0) using the DESeq2 (v1.26.0) as the differential analysis model. 3 of the 12 samples had low quality peak calls compared to the other 9 and were not considered when calculating differential acetylation. We considered each unique breakpoint one at a time in the remaining 9 samples. Samples were grouped by their mutation status (i.e., a design matrix where the mutation status is the only covariate) and DiffBind's differential binding analysis method was performed to identify all differentially acetylated regions between the two groups. Acetylation peaks outside of the TAD(s) overlapping the breakpoint were filtered out. Multiple test correction with the Benjamini-Hochberg FDR method was performed on all peaks after all breakpoints were considered, due to similar group stratifications depending on the breakpoint under consideration.

Primary Tissue RNA Data Analysis

Tumour sample RNA sequencing

Total RNA was extracted for the CPC-GENE tumour samples as previously described [30]. Briefly, total RNA was extracted with mirVana miRNA Isolation Kit (Life Technologies) according to the manufacturer's instructions. RNA samples were sent to BGI Americas where

it underwent QC and DNase treatment. For each sample, 200 ng of total RNA was used to construct a TruSeq strand-specific library with the Ribo Zero protocol (Illumina, Cat. #RS-122-2203). The libraries were sequenced on a HiSeq 2000 to a minimal target of 180 million, 2 x 100 bp paired-end reads.

RNA sequencing data pre-processing

RNA sequencing FASTQ files were pseudo-aligned to the hg38 genome using Kallisto (v0.46.1) with the following command: `kallisto quant --bootstrap-samples 100 --pseudobam --threads 8 --index /path/to/GRCh38.idx --output-dir {output_dir} {input_R1.fastq.gz} {input_R2.fastq.gz}`.

Differential gene expression analysis

To assess whether SVs were associated with local gene expression changes, we considered each unique breakpoint one at a time. For each breakpoint, we compared the gene expression between the mutated and non-mutated tumour samples using Sleuth (v0.30.0) with a linear model where the mutation status was the only covariate (ie. $d_{ti} = \mu_t + \mathbb{I}_{mut}\beta_{ti}$). To reduce the chance of falsely identifying genes as differentially expressed, only genes located within the TADs (window size 20) containing breakpoints were considered. Fold-change estimates of each transcript were assessed for significance using a Wald test. Transcript-level p-values are combined to create gene-level p-values using the Lancaster aggregation method provided by the Sleuth package. Correcting for multiple tests was then performed with the Benjamini-Hochberg FDR correction for all genes that were potentially altered in the mutated sample(s).

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Author Contributions

S.Z., J.R.H., and M.L. conceptualized the study. S.Z. designed and conducted all the experiments with help from C.A. G.G., AND K.K. J.R.H. implemented all the computational and statistical approaches and analyses. R.H.-W. pre-processed the RNA-seq data from the primary tumours. Figures were designed by S.Z. and J.R.H. The manuscript was written by S.Z., J.H., and M.L with assistance from all authors. T.v.d.K., M.F., P.C.B., R.G.B., and M.L. supervised the study. M.L. oversaw the study.

Acknowledgements

We thank all the Lupien lab members for their feedback, as well as Jesse Dixon for his support with hic_breakfinder and interpretation of structural variant calls. This work is supported by Prostate Cancer Foundation Canada, Ontario Institute for Cancer Research funded by the Government of Ontario, the Princess Margaret Cancer Foundation (R.G.B. and M.L.), Princess Margaret Cancer Centre Department of Surgical Oncology (M.F.), Princess Margaret Cancer Centre Genetics and Epigenetic Program (M.F. and M.L.), University of Toronto Department of Surgery Division of Urology (M.F.), Movember Foundation (RS2014-04 to M.L. and RS2014-01 to P.C.B.), the Radiation Medicine Program Academic Enrichment Fund (R.G.B.), Terry Fox Research Institute New Investigator Award (P.C.B.), Canadian Institute of Health Research (CIHR; FRN-153234 to M.L.) and New Investigator Award (P.C.B. and M.L.), Canadian Cancer Society Research Scientist Award (R.G.B.), Cancer

Society Impact Award (P.C.B), Investigator Award from the Ontario Institute for Cancer Research (M.L. and P.C.B), and Movember Rising Star Award from PCa Canada (M.L. and P.C.B).

References

1. Fraser, M. *et al.* Genomic hallmarks of localized, non-indolent prostate cancer. *Nature* **541**, 359–364 (2017).
2. Oudelaar, A. M. & Higgs, D. R. The relationship between genome structure and function. *Nat. Rev. Genet.* **571**, 489 (2020).
3. Northcott, P. A. *et al.* Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428–434 (2014).
4. Finn, E. H. & Misteli, T. Molecular basis and biological function of variability in spatial genome organization. *Science* **365**, eaaw9498 (2019).
5. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
6. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
7. Pombo, A. & Dillon, N. Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.* **16**, 245–257 (2015).
8. Gröschel, S. *et al.* A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell* **157**, 369–381 (2014).
9. Flavahan, W. A. *et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110–114 (2016).
10. Haller, F. *et al.* Enhancer hijacking activates oncogenic transcription factor NR4A3 in acinic cell carcinomas of the salivary glands. *Nat. Commun.* **10**, 368 (2019).
11. Zhou, S., Treloar, A. E. & Lupien, M. Emergence of the Noncoding Cancer Genome: A Target of Genetic and Epigenetic Alterations. *Cancer Discov.* **6**, 1215–1229 (2016).
12. Pleasance, E. D. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
13. Vinagre, J. *et al.* Frequency of TERT promoter mutations in human cancers. *Nat. Commun.* **4**, 2185 (2013).

14. Huang, F. W. *et al.* Highly Recurrent TERT Promoter Mutations in Human Melanoma. *Science* **339**, 957–959 (2013).

15. Stern, J. L., Theodorescu, D., Vogelstein, B., Papadopoulos, N. & Cech, T. R. Mutation of the TERT promoter, switch to active chromatin, and monoallelic TERT expression in multiple cancers. *Genes Dev.* **29**, 2219–2224 (2015).

16. Bailey, S. D. *et al.* Noncoding somatic and inherited single-nucleotide variants converge to promote ESR1 expression in breast cancer. *Nat. Genet.* **48**, 1260–1266 (2016).

17. Zhou, S. *et al.* Noncoding mutations target cis-regulatory elements of the FOXA1 plexus in prostate cancer. *Nat. Commun.* **11**, 441 (2020).

18. Takeda, D. Y. *et al.* A Somatically Acquired Enhancer of the Androgen Receptor Is a Noncoding Driver in Advanced Prostate Cancer. *Cell* doi:10.1016/j.cell.2018.05.037.

19. Quigley, D. A. *et al.* Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer. *Cell* **174**, 758–769.e9 (2018).

20. Parolia, A. *et al.* Distinct structural classes of activating FOXA1 alterations in advanced prostate cancer. *Nature* vol. 571 413–418 (2019).

21. Zhang, X. *et al.* Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. *Nat. Genet.* **48**, 176–182 (2016).

22. Morton, A. R. *et al.* Functional Enhancers Shape Extrachromosomal Oncogene Amplifications. *Cell* **179**, 1330–1341.e13 (2019).

23. Wedge, D. C. *et al.* Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets. *Nat. Genet.* **50**, 682–692 (2018).

24. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).

25. Tomlins, S. A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648 (2005).

26. Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163**, 1011–1025 (2015).

27. Baca, S. C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).

28. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).

29. Díaz, N. *et al.* Chromatin conformation analysis of primary patient tissue using a low input Hi-C method. *Nat. Commun.* **9**, 4938 (2018).

30. Chen, S. *et al.* Widespread and Functional RNA Circularization in Localized Prostate Cancer. *Cell* **176**, 831–843.e22 (2019).

31. Kron, K. J. *et al.* TMPRSS2-ERG fusion co-opts master transcription factors and activates NOTCH signaling in primary prostate cancer. *Nat. Genet.* (2017) doi:10.1038/ng.3930.

32. Mazrooei, P. *et al.* Cistrome Partitioning Reveals Convergence of Somatic Mutations and Risk Variants on Master Transcription Regulators in Primary Prostate Tumors. *Cancer Cell* **36**, 674–689.e6 (2019).

33. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin Domains: The Unit of Chromosome Organization. *Mol. Cell* **62**, 668–680 (2016).

34. Rhie, S. K. *et al.* A high-resolution 3D epigenomic map reveals insights into the creation of the prostate cancer transcriptome. *Nat. Commun.* **10**, 4154 (2019).

35. Dixon, J. R. *et al.* Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genet.* **50**, 1388–1398 (2018).

36. Ho, S. S., Urban, A. E. & Mills, R. E. Structural variation in the sequencing era. *Nat. Rev. Genet.* **21**, 171–189 (2020).

37. Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214–220 (2011).

38. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).

39. Akdemir, K. C. *et al.* Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nat. Genet.* (2020) doi:10.1038/s41588-019-0564-y.

40. Katainen, R. *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.* **47**, 818–821 (2015).

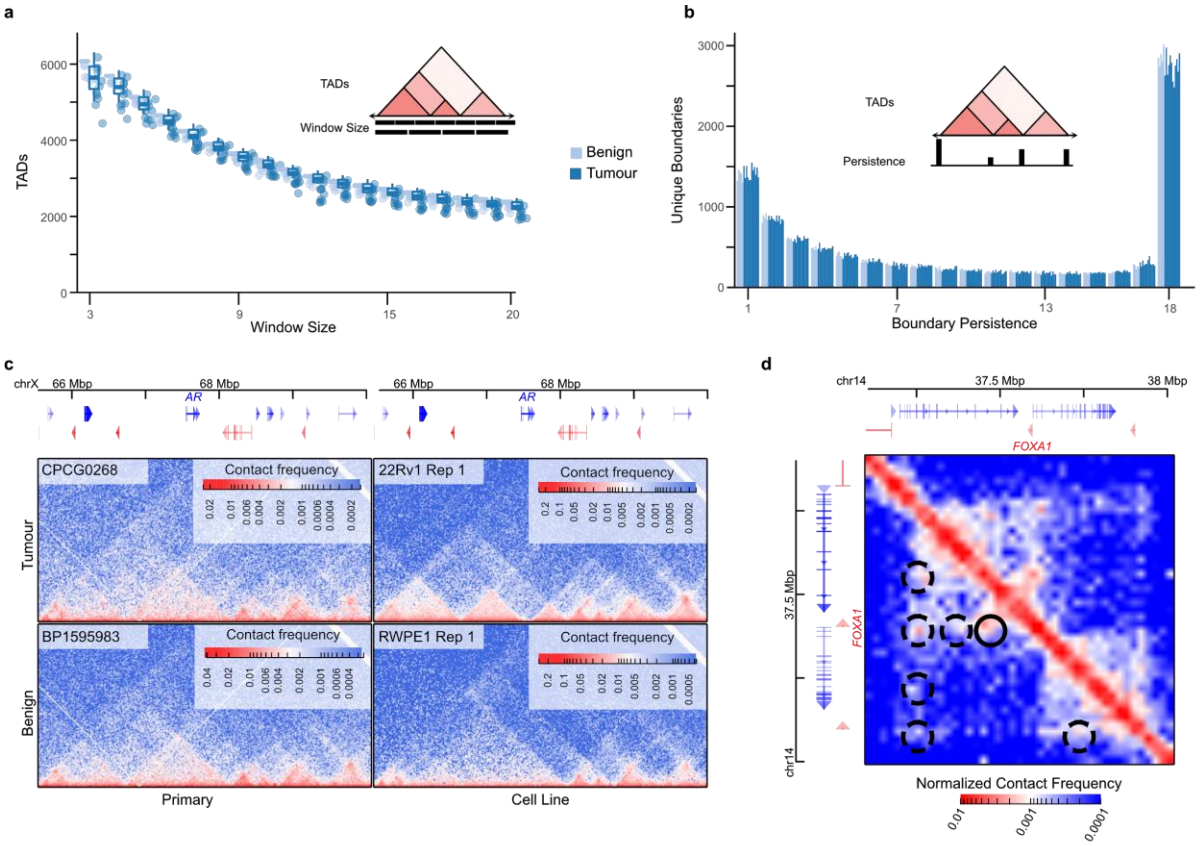
41. Guo, Y. A. *et al.* Mutation hotspots at CTCF binding sites coupled to chromosomal instability in

gastrointestinal cancers. *Nat. Commun.* **9**, 1–14 (2018).

42. Zhao, S. G. *et al.* The DNA methylation landscape of advanced prostate cancer. *Nat. Genet.* **52**, 778–789 (2020).
43. Yu, Y. P. *et al.* Whole-genome methylation sequencing reveals distinct impact of differential methylations on gene transcription in prostate cancer. *Am. J. Pathol.* **183**, 1960–1970 (2013).
44. Role of DNA Methylation in Modulating Transcription Factor Occupancy. *Cell Rep.* **12**, 1184–1195 (2015).
45. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
46. Wu, S. *et al.* Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature* **575**, 699–703 (2019).
47. Kumar, P. *et al.* ATAC-seq identifies thousands of extrachromosomal circular DNA in cancer and cell lines. *Science Advances* **6**, eaba2489 (2020).
48. Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell Rep.* **10**, 1297–1309 (2015).
49. Viswanathan, S. R. *et al.* Structural Alterations Driving Castration-Resistant Prostate Cancer Revealed by Linked-Read Genome Sequencing. *Cell* (2018) doi:10.1016/j.cell.2018.05.036.
50. Zaborowski, R. & Wilczyński, B. BPscore: An Effective Metric for Meaningful Comparisons of Structural Chromosome Segmentations. *J. Comput. Biol.* **26**, 305–314 (2019).

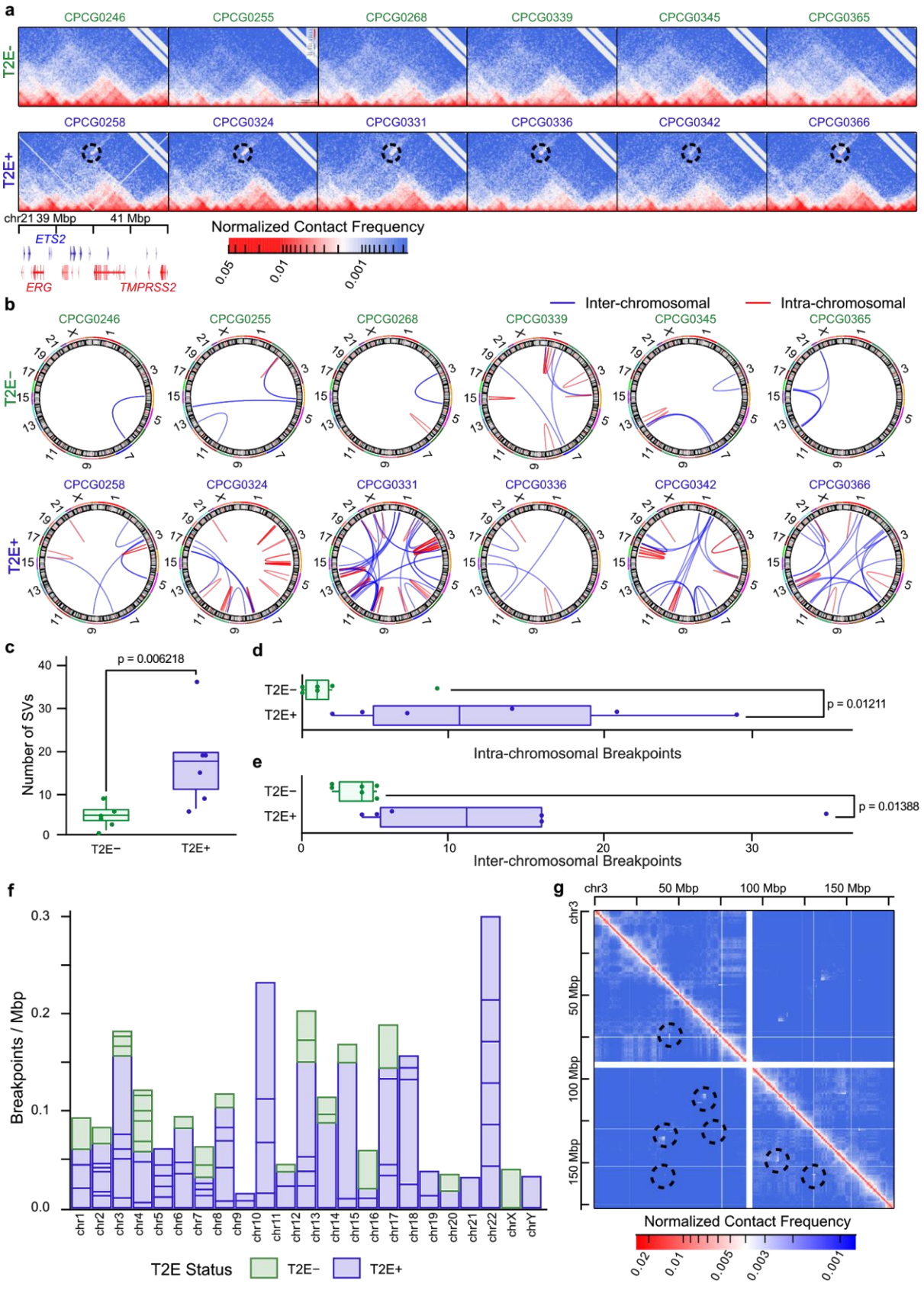
Figures

Figure 1



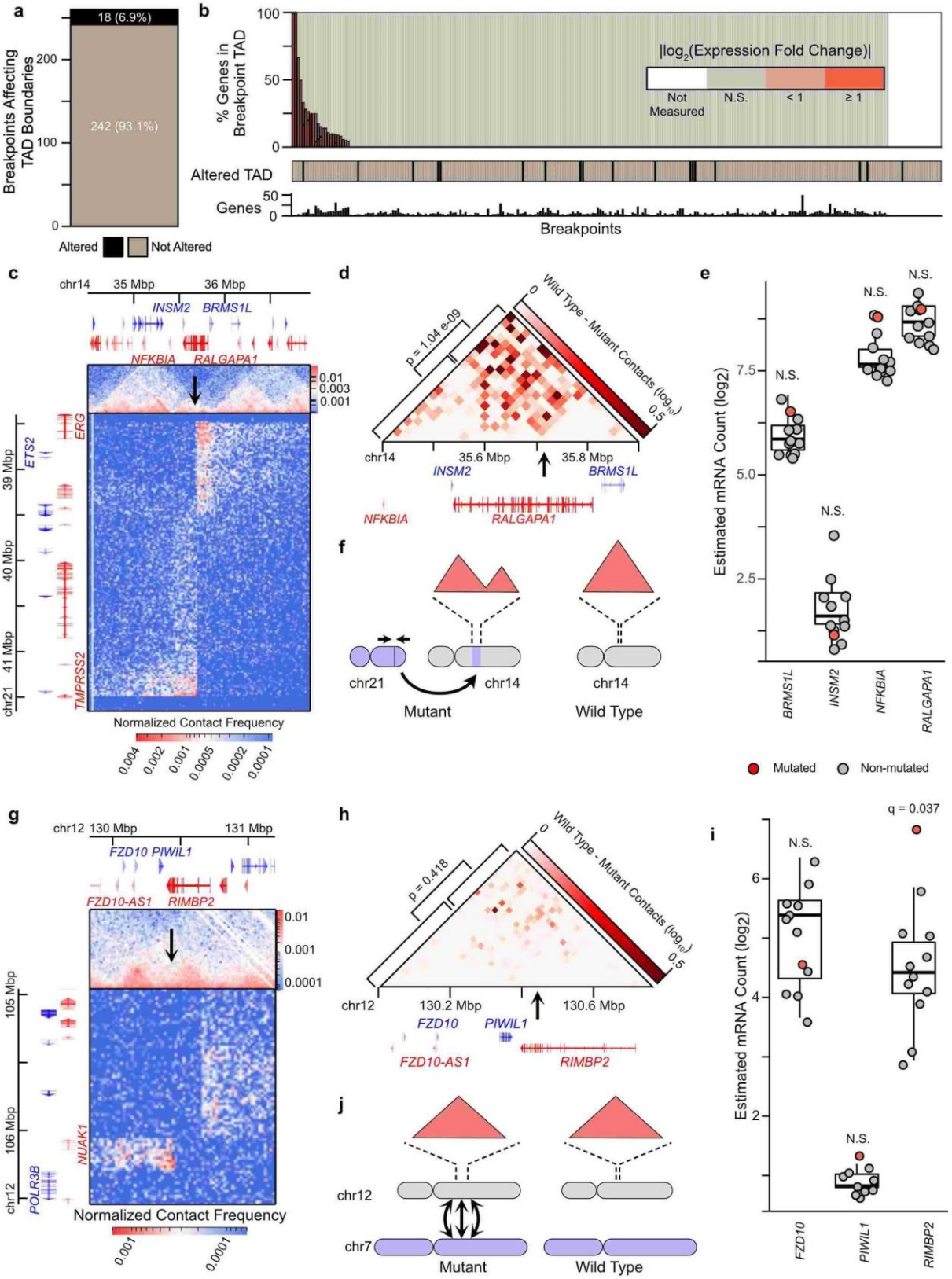
The 3D genome of primary prostate cancer. a-b. A comparison of the number of TADs detected at multiple window sizes (**a**) and boundary strength (**b**) in each patient sample, with inset schematics. **c.** Contact matrices around the *AR* gene demonstrate a difference in chromatin organization between primary samples and cell lines. Hi-C data for 22Rv1 and RWPE1 cell lines obtained from [34]. **d.** A snapshot of significant chromatin interactions called around the *FOXA1* gene. Identified interactions are highlighted as circles. The interaction marked by the solid border contains two CREs of *FOXA1* identified in [17] (listed in that publication as CRE1 and CRE2). The interactions marked by the dashed border indicate regions of increased contact that may contain more distal CREs of *FOXA1*.

Figure 2


a
CPCG0246
CPCG0255
CPCG0268
CPCG0339
CPCG0345
CPCG0365
T2E−
CPCG0258
CPCG0324
CPCG0331
CPCG0336
CPCG0342
CPCG0366
T2E+
chr21 39 Mbp
41 Mbp
ETS2
ERG
TMPRSS2
Normalized Contact Frequency
0.05
0.01
0.001
b
Inter-chromosomal
Intra-chromosomal
c
p = 0.006218
Number of SVs
T2E−
T2E+
d
p = 0.01211
Intra-chromosomal Breakpoints
e
p = 0.01388
Inter-chromosomal Breakpoints
f
Breakpoints / Mbp
T2E Status
T2E−
T2E+
g
chr3
50 Mbp
100 Mbp
150 Mbp
Normalized Contact Frequency
0.02
0.01
0.005
0.003
0.001

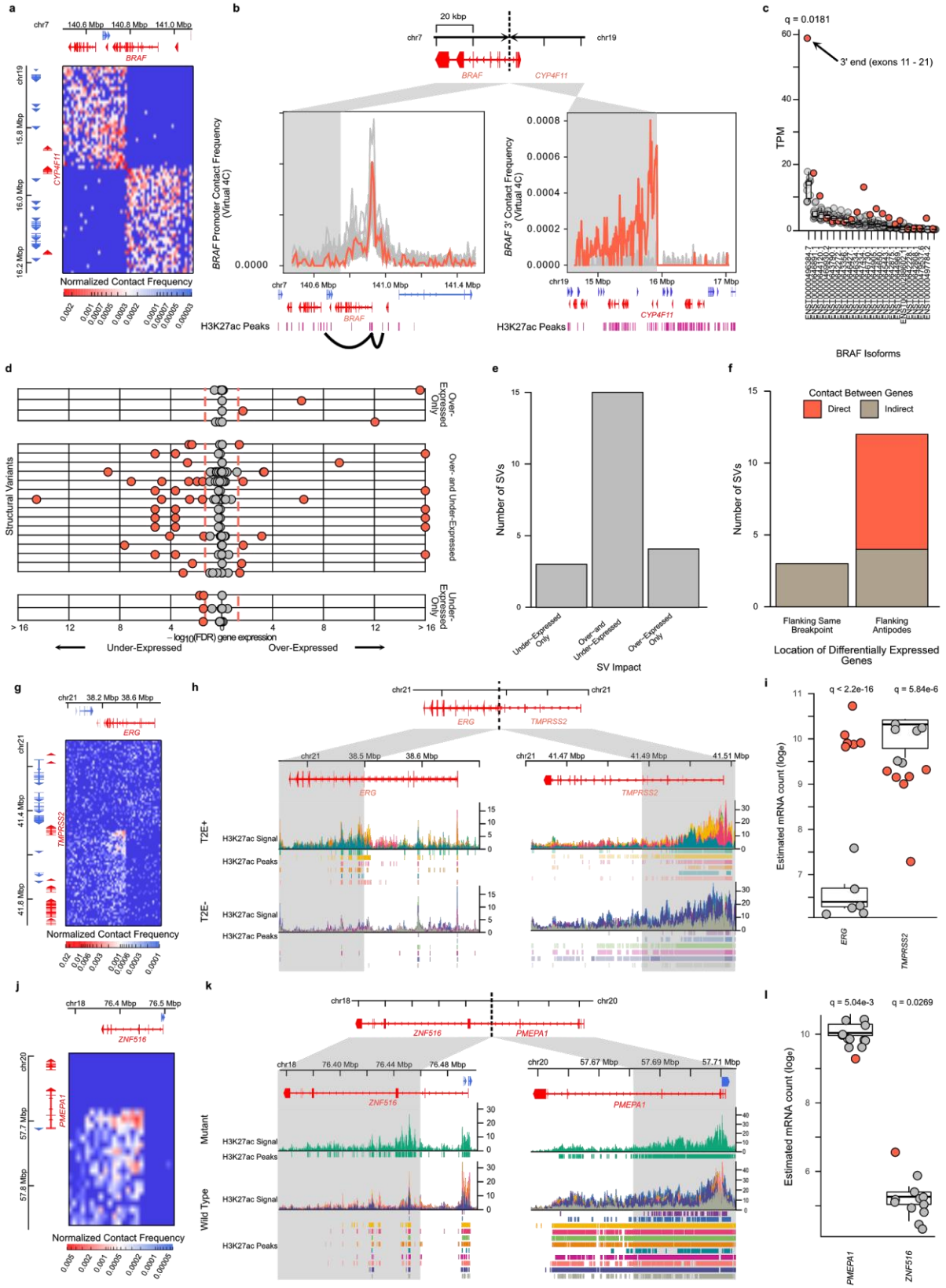
SVs are identified across the 12 primary prostate tumours through chromatin conformation capture. a. Hi-C contact matrices of the chr21:37-42 Mbp locus harbouring the *TMPRSS2* and *ERG* genes. Circles indicate increased contact between *TMPRSS2* and *ERG* in the T2E+ tumours. **b.** Circos plots of structural variants identified in the 12 primary prostate tumours. **c.** Boxplot comparing the number of structural variants between the T2E+ and T2E- tumours. **d-e.** Boxplots comparing the number of intra-chromosomal (**d)** and inter-chromosomal (**e**) SV breakpoints between the T2E+ and T2E- tumours. **f.** Chromosome location and frequency of the structural variants of the 12 primary prostate tumours. **g.** An example of a complex set of rearrangement across both arms of chromosome 3 in a patient. One-sided Mann-Whitney U tests were performed in panels **c-e.**

Figure 3


a
18 (6.9%)
242 (93.1%)
Breakpoints Affecting TAD Boundaries
Altered
Not Altered
b
% Genes in Breakpoint TAD
|log2(Expression Fold Change)|
Not Measured
N.S.
< 1
≥ 1
Altered TAD
Genes
Breakpoints
c
chr14
35 Mbp
36 Mbp
INSM2
BRMS1L
NFKBIA
RALGAPA1
ERG
ETS2
TMPRSS2
39 Mbp
40 Mbp
41 Mbp
chr21
Normalized Contact Frequency
d
p = 1.04 e-09
Wild Type - Mutant Contacts (log10)
chr14
35.6 Mbp
35.8 Mbp
INSM2
BRMS1L
NFKBIA
RALGAPA1
e
Estimated mRNA Count (log2)
N.S.
BRMS1L
INSM2
NFKBIA
RALGAPA1
Mutated
Non-mutated
f
chr21
chr14
Mutant
Wild Type
g
chr12
130 Mbp
131 Mbp
FZD10
PIWIL1
FZD10-AS1
RIMBP2
NUAK1
POLR3B
105 Mbp
106 Mbp
Normalized Contact Frequency
h
p = 0.418
Wild Type - Mutant Contacts (log10)
chr12
130.2 Mbp
130.6 Mbp
FZD10
PIWIL1
FZD10-AS1
RIMBP2
i
q = 0.037
Estimated mRNA Count (log2)
N.S.
FZD10
PIWIL1
RIMBP2
j
chr12
chr7
Mutant
Wild Type

SVs can alter TADs or gene expression around breakpoints, but rarely alters both. a. A count of the number of SV breakpoints associated with altered TAD boundaries. **b.** Bar plot showing the number of genes differentially expressed around SV breakpoints. **c-f.** An example of an SV that alters TAD boundaries without significantly affecting gene expression of the nearby genes. **c.** The contact matrix showing a translocation of the *TMPRSS2-ERG* locus into chr14 in the *RALGAPA1* gene. **d.** The differential contact matrix between the tumour containing this translocation and another tumour without it to show the decreased contacts between sites upstream and downstream of the insertion site. **e.** Expression of the genes within the broken TAD show no significant changes to their expression. **f.** A schematic representation of the translocation. **g-j.** An example of an SV that does not alter TAD boundaries but does alter the expression of a nearby gene. **g.** The contact matrix showing a complex rearrangement around the *RIMBP2* gene. **h.** The differential contact matrix between the tumour containing this translocation and another tumour without it to show the decreased contacts between sites upstream and downstream of the insertion site. **i.** Expression of the genes within the broken TAD show no significant changes to their expression. **j.** A schematic representation of the translocation.

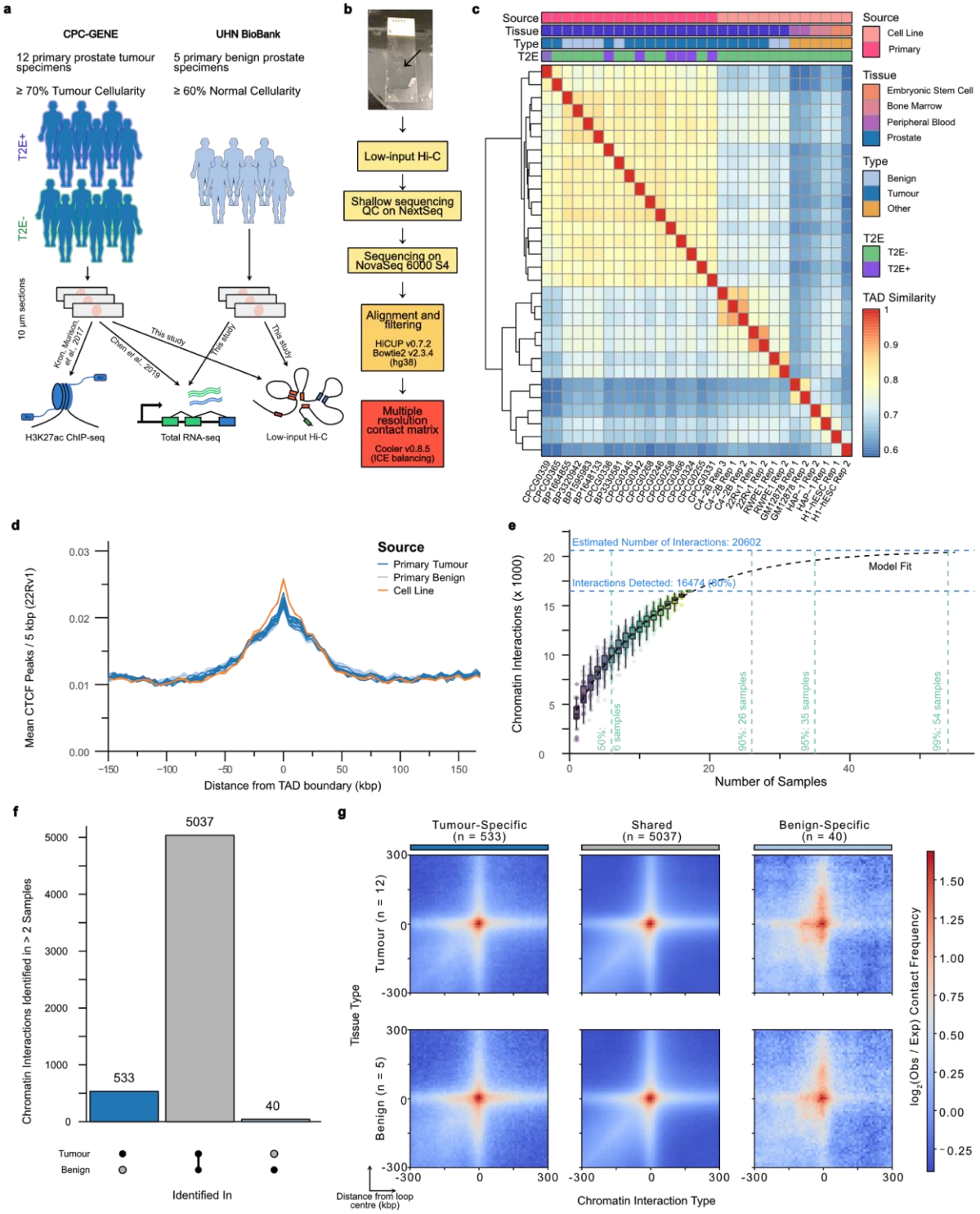
Figure 4


a
chr7
140.6 Mbp
140.8 Mbp
141.0 Mbp
BRAF
chr19
15.8 Mbp
16.0 Mbp
16.2 Mbp
CYP4F11
Normalized Contact Frequency
b
20 kbp
chr7
chr19
BRAF
CYP4F11
BRAF Promoter Contact Frequency (Virtual 4C)
BRAF 3' Contact Frequency (Virtual 4C)
0.0000
0.0008
0.0006
0.0004
0.0002
140.6 Mbp
141.0 Mbp
141.4 Mbp
15 Mbp
16 Mbp
17 Mbp
H3K27ac Peaks
c
q = 0.0181
3' end (exons 11 - 21)
TPM
BRAF Isoforms
d
Structural Variants
Over-Expressed Only
Over- and Under-Expressed
Under-Expressed Only
> 16
12
8
4
0
− log10(FDR) gene expression
Under-Expressed
Over-Expressed
e
Number of SVs
Under-Expressed Only
Over- and Under-Expressed
Over-Expressed Only
SV Impact
f
Contact Between Genes
Direct
Indirect
Number of SVs
Flanking Same Breakpoint
Flanking Antipodes
Location of Differentially Expressed Genes
g
chr21
38.2 Mbp
38.6 Mbp
ERG
chr21
41.4 Mbp
41.8 Mbp
TMPRSS2
Normalized Contact Frequency
h
chr21
ERG
TMPRSS2
38.5 Mbp
38.6 Mbp
41.47 Mbp
41.49 Mbp
41.51 Mbp
T2E+
T2E-
H3K27ac Signal
H3K27ac Peaks
i
q < 2.2e-16
q = 5.84e-6
Estimated mRNA count (log2)
ERG
TMPRSS2
j
chr18
76.4 Mbp
76.5 Mbp
ZNF516
chr20
57.7 Mbp
57.8 Mbp
PMEPA1
Normalized Contact Frequency
k
chr18
chr20
ZNF516
PMEPA1
76.40 Mbp
76.44 Mbp
76.48 Mbp
57.67 Mbp
57.69 Mbp
57.71 Mbp
Mutant
Wild Type
H3K27ac Signal
H3K27ac Peaks
l
q = 5.04e-3
q = 0.0269
Estimated mRNA count (log2)
PMEPA1
ZNF516

SVs altering gene expression by rewiring focal chromatin interactions. a. The Contact matrix of an inter-chromosomal break between chromosome 7 and chromosome 19. **b**. Contact frequencies of the *BRAF* promoter on chromosome 7 (left) and the 3' end of *BRAF* on chromosome 19 (right). Grey regions are loci brought into contact. SV-associated contacts between the 3' end of *BRAF* on chromosome 19 (right) are focally enriched at H3K27ac peaks downstream of *CYF4P11*. **c**. *BRAF* isoforms in mutant (red) and wild type patients (grey). **d**. Scatterplot of gene expression changes flanking SV breakpoints. Red dots are differentially expressed genes (FDR < 0.05), grey dots are genes not differentially expressed. **e**. Bar plot of SVs categorized by how differentially expressed genes altered. **f**. bagplot of all 15 SVs associated with both over- and under-expression, categorized by which breakpoints are flanked by the differentially expressed genes. Red SVs contain differentially expressed genes whose gene bodies are in direct contact with each other, i.e., immediately flank the breakpoint. **g**. Contact matrix of the deletion between *TMPRSS2* and *ERG*. **h.** Genome tracks of H3K27ac ChIP-seq data. **i**. Gene expression of *TMPRSS2* and *ERG*. Boxplots represent the distribution of T2E- patients (grey dots). T2E+ patients are represented by red dots. **j**. Contact matrix of the deletion between *PMEPA1* and *ZNF516*. **k**. Genome tracks of H3K27ac ChIP-seq data. **l**. Gene expression of *PMEPA1* and *ZNF516*. Boxplots represent the distribution of wild type patients (grey dots) and red dots are the mutated patient.

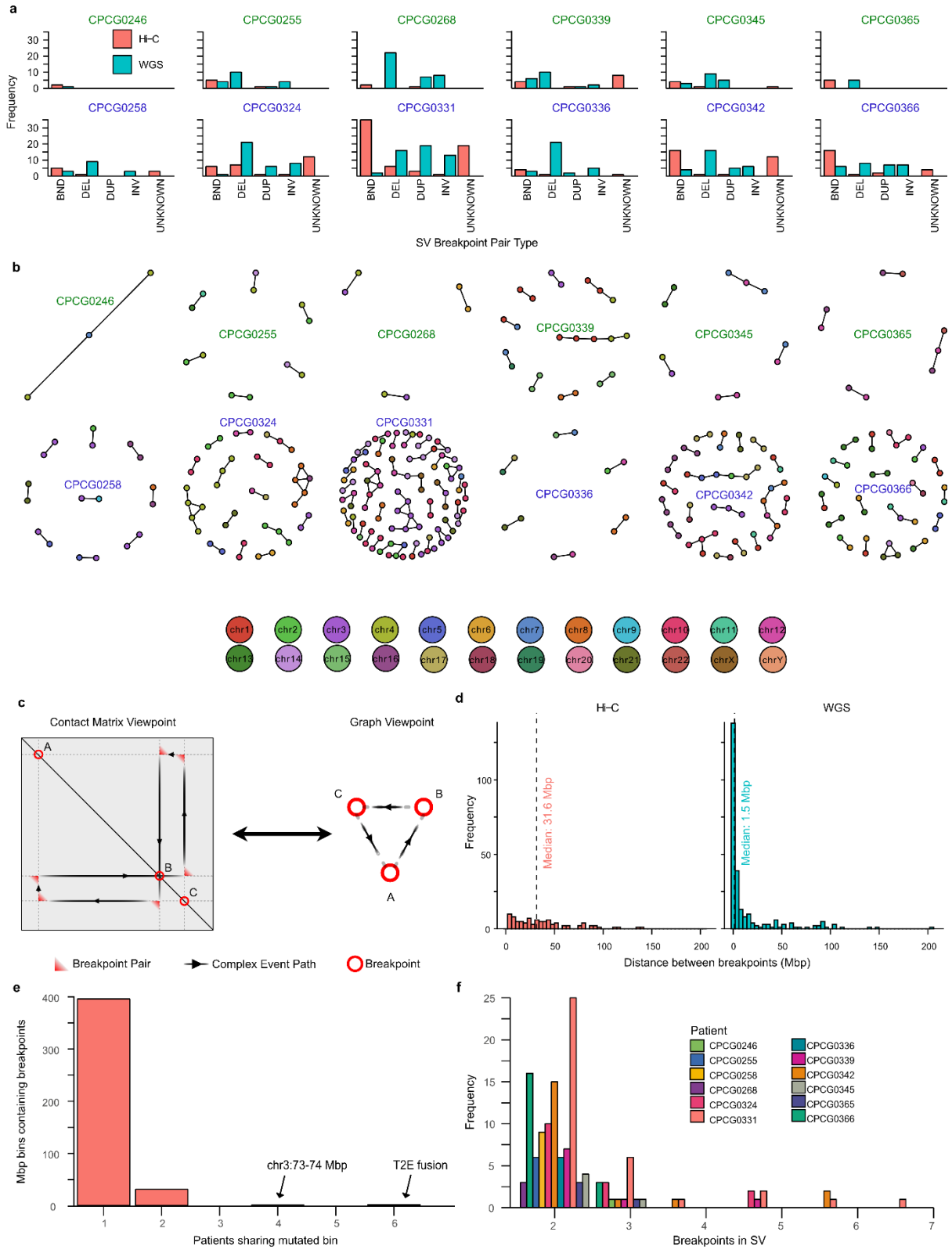
Supplementary Figures

Supplementary Figure 1



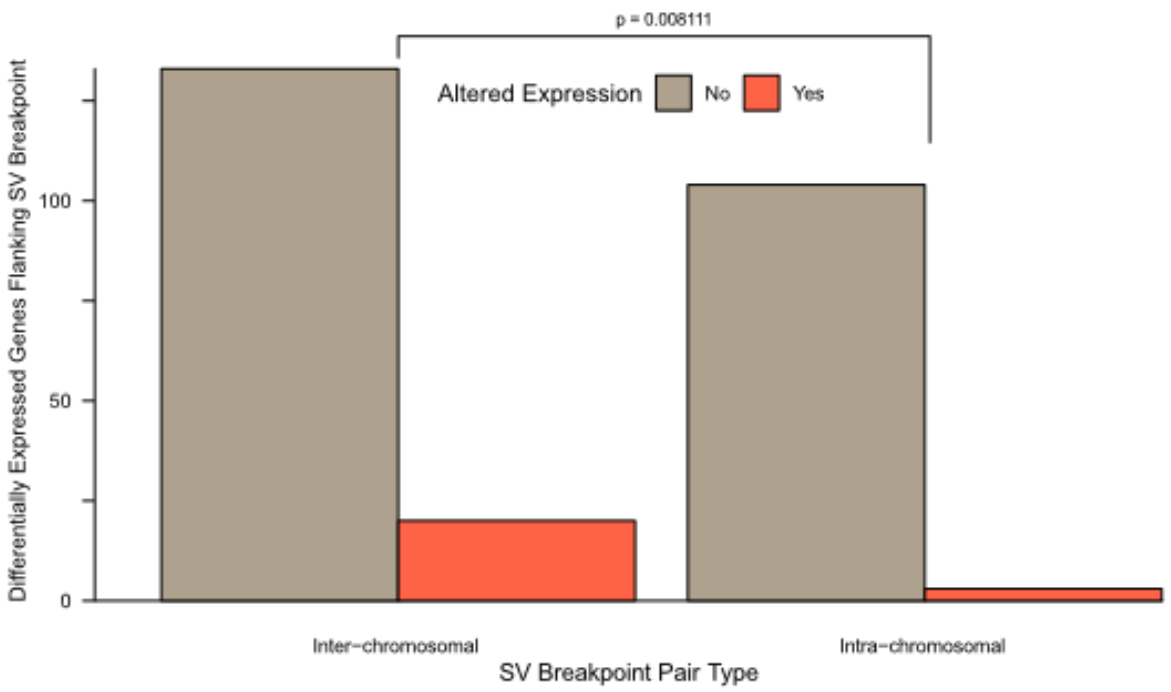
a. The sample collection and data usage of primary prostate samples in this study. 10 µm sections from 6 tumours previously identified as T2E+ and 6 T2E- were used for Hi-C sequencing. 5 additional 10 µm sections were collected from benign prostate specimens in the UHN BioBank. **b.** Schematic representation of the protocol and data pre-processing pipeline used in this study to obtain Hi-C sequencing data. **c.** Heatmap of TAD similarities between primary prostate samples, prostate cell lines, and non-prostate cell lines. **d**. Local enrichment of CTCF binding sites from the 22Rv1 PCa cell line around TAD boundaries identified in the primary samples. **e.** Chromatin interactions detected in our cohort of prostate samples versus the theoretical estimation obtained through asymptotic estimation from bootstraps. **f**. Upset plot of significant chromatin interactions identified in at least 2 patient samples. Chromatin interactions in grey are detected in both tumour and benign samples, dark blue are detected in at least 2 tumour samples and no benign samples, and light blue are detected in at least 2 benign samples and no tumour samples. **g**. Aggregate peak analysis of detected chromatin interactions. The top row is the aggregation of contact matrices over all tumour samples, and the bottom row over all benign samples. The columns correspond to tumour-specific, shared, and benign-specific chromatin interactions, respectively.

Supplementary Figure 2



a. Bar plot of SV breakpoint pairs identified by Hi-C and WGS [1] on matched samples. BND = inter-chromosomal translocation, DEL = deletion, DUP = duplication, INV = inversion, UNKNOWN = breakpoint pair of unknown type. **b.** Graph reconstructions of the SV breakpoints in all 12 tumours. The node colour corresponds to the chromosome of origin. The nodes are spaced by a spring-force layout which is then adjusted using the Kambda Kawai optimization. **c.** Correspondence between the breakpoint representation in the contact matrices and a graph representation. Each node represents a breakpoint and each edge determines whether the breakpoints were directly in contact, as identified by the Hi-C contact matrix. **d.** Histogram showing the distance between breakpoints on the same chromosome detected by Hi-C (left) versus WGS [1] (right). **e.** Bar plot showing the lack of recurrence of SV breakpoints between patients. Almost all breakpoints belong to a unique megabase-sized bin. No SV, other than the T2E fusion, is identified as common between any of the PCa patients. **f.** Bar plot of the number of SVs and the number of breakpoints involved, for each tumour. Most SVs are simple events (2 breakpoints), but many complex events (> 2 breakpoints) are found.

Supplementary Figure 3



Differentially expressed genes flanking SV breakpoints are significantly more associated with inter-chromosomal translocations than intra-chromosomal rearrangements.

Supplementary Tables

Supplementary Table 1 - Clinical information of samples involved in this study.

Supplementary Table 2 - Sequencing metrics as calculated by HiCUP for all Hi-C libraries generated in this study.

Supplementary Table 3 - Summary statistics for TAD counts in all 12 tumour and 5 benign samples, across multiple window sizes.

Supplementary Table 4 - Individual TAD calls in all 12 tumour and 5 benign samples.

Supplementary Table 5 - Detected chromatin interactions in all 12 tumour and 5 benign samples.

Supplementary Table 6 - Detected SV breakpoints in each tumour sample.

Supplementary Table 7 - Simple and complex SVs reconstructed from SV breakpoints.

Supplementary Table 8 - H3K27ac peaks identified in each of the 12 primary PCa patients. Raw sequencing data as previously published in [31] was remapped to the hg38 reference genome.