Inherited determinants of early recurrent somatic mutations in prostate cancer

Alessandro Romanel¹, Sonia Garritano¹, Blerta Stringa¹, Mirjam Blattner¹, Davide Dalfovo¹, Dimple Chakravarty², David Soong³, Kellie A. Cotter², Gianluca Petris¹, Priyanka Dhingra³, Paola Gasperini¹, Anna Cereseto¹, Olivier Elemento²,³, Andrea Sboner², Ekta Khurana²,³, Alberto Inga¹, Mark A. Rubin²,⁴,⁵ & Francesca Demichelis¹,²

Prostate cancer is a highly heritable molecularly and clinically heterogeneous disease. To discover germline events involved in prostate cancer predisposition, we develop a computational approach to nominate heritable facilitators of somatic genomic events in the context of the androgen receptor signaling. Here, we use a ranking score and benign prostate transcriptomes to identify a non-coding polymorphic regulatory element at 7p14.3 that associates with DNA repair and hormone-regulated transcript levels and with an early recurrent prostate cancer-specific somatic mutation in the Speckle-Type POZ protein (SPOP) gene. The locus shows allele-specific activity that is concomitantly modulated by androgen receptor and by CCAAT/enhancer-binding protein (C/EBP) beta (CEBPB). Deletion of this locus via CRISPR-Cas9 leads to deregulation of the genes predicted to interact with the 7p14.3 locus by Hi-C chromosome conformation capture data. This study suggests that a polymorphism at 7p14.3 may predispose to SPOP mutant prostate cancer subclass through a hormone-dependent DNA damage response.
Prostate cancer (PCa) is the second most frequent cancer in men causing each year more than 250,000 deaths worldwide. From a genomic perspective, PCa is a collection of molecular subclasses. Approximately 58% of risk for prostate cancer has been estimated to be due to inherited genetic factors. Genome-wide association studies have identified more than 100 common single-nucleotide polymorphisms (SNPs) associated with the risk of developing PCa. Most of these variants reside in non-coding regulatory regions and may affect the transcription factors (TFs)-binding affinity. Androgen receptor (AR) regulates genes expression in multiple tissues and diseases, by targeting binding elements in promoters and distant enhancers. A recent PCa whole-genome sequencing study revealed a significant correspondence between DNA breakpoints and AR-binding sites implicating an inter-play between hormone regulation and genomic events. These studies highlight an important role of androgens in the initiation and development of PCa. Indeed, at the earliest time point of clinical presentation, PCA already harbors a range of genomic lesions possibly due to DNA repair defects. We reasoned that, over a man’s lifetime, heritable variants could potentially predispose to genomic instability in the context of variable AR signaling leading to early PCa-specific somatic genomic events. To test this hypothesis, we interrogated the constellation of transcriptomic changes in benign prostate cells for clues as to how genetic variants could impact prostate cancer development through alterations in the expression of DNA repair genes and hormone-regulated genes. Here we report a link between an inherited non-coding variant and prostate cancer somatic mutations through the interrogation of large cohorts of human data and experimental support of the functional activity of the variant locus.

**Results**

**In silico selection of germline triggers of somatic mutations.** To quantitatively assess the predisposition to genomic changes in the context of AR signaling, we developed an approach to nominate potential heritable facilitators (referred hereafter as triggers) of somatic genomic events. We considered human variants within functionally active regions of the genome defined by the Encyclopedia of DNA Elements (ENCODE) histone mark ChIP-seq data, and established a ranking score, the trigger score, which quantifies the fraction of the transcriptome putatively modulated by each human variant leveraging individuals’ genotypes and transcript levels (Fig. 1a). The trigger score—unlike eQTL-based approach—only queries a predefined set of transcripts and ranks the variants for their likelihood to play a role in predisposition to cancer hallmarks. When applied to a RNA-seq data set comprising more than 200 samples including benign human prostate tissue from The Cancer Genome Atlas (TCGA) and samples from the 1000 Genomes project with known genotype at variants in transcriptionally active regulatory elements, the trigger score nominated 300 polymorphisms linked to DNA repair and hormone-regulated genes (Fig. 1b, Supplementary Data 1–3). Sixty-nine of those sites had a minimal trigger score in non-prostate samples (Supplementary Data 3).

Several recent genomic studies now establish PCa as best being regarded as a collection of molecularly defined cancers—similar to breast and lung cancer—with major subclasses defined by either ETS gene fusions (most commonly TMPRSS2–ERG rearrangements), SPOP or FOXA1 single-nucleotide mutations, and are mainly prostate specific. To explore genotype/phenotype relationship for these common prostate cancer mutations, we assembled a data set comprising 539 prostate tumors from three recent studies, and observed 47.2, 12.1, and 5.4% incidence, respectively, (Supplementary Data 4). To test the relationship between the trigger candidates and the three somatic phenotypes, we used a computational in-silico cross-validation strategy that limits false positives results and implements multiple discovery and validation partitions from the entire cohort preserving somatic event incidence. No signal was detected for the FOXA1 phenotype and, surprisingly, no signal was observed for the largest genomic subclass defined by the ETS rearrangement phenotype (i.e., TMPRSS2–ERG). In contrast, a polymorphic site on 7p14.3 (rs1376350) was significantly associated with SPOP somatic mutations in 97.4% of the partitions with any positive signal (Fig. 1b, Supplementary Data 5 and Supplementary Table 1). When interrogating the whole prostate cancer cohort (N = 539 patients, Fig. 1c and Table 1), the association with the phenotype was highly significant (P = 6.7e−6, OR = 4.83 logistic regression analysis). Further analyses including SPOP wild-type samples of the prostate cancer Tyrol PSA screening cohort and of the 1000 Genomes Project confirmed the association (P = 9.2e−08, OR = 5.54 logistic regression analysis). Moreover, the genotype/phenotype relationship with SPOP mutation was confirmed in an independent non-Caucasian prostate cancer cohort of Korean men (P = 4e−02, OR = 5.84 logistic regression analysis), indicating that this phenomenon is not limited to Caucasian male populations but rather may extend across multi-ethnic populations (Table 1 and Fig. 1d). No association was detected between the genotype and the total number of somatic single-nucleotide variants (SNVs) in the tumor, but we observed increased somatic genomic burden in men with the minor allele associated with SPOP mutant prostate cancer (Supplementary Fig. 1). Linkage disequilibrium analysis on the 1000 Genomes Project data did not reveal variants in functional or coding regions linked to rs1376350 across populations (Supplementary Data 6).

**Regulatory impact of 7p14.3 variant.** Fourteen DNA repair genes and 15 hormone-regulated genes contributed to the high proportion test). Overall, this suggests a broad transcriptional deregulation analysis on the 1000 Genomes Project data did not reveal variants in functional or coding regions linked to rs1376350 across populations (Supplementary Data 6).

| DOI: 10.1038/s41467-017-00046-0 | www.nature.com/naturecommunications | NATURE COMMUNICATIONS | DOI: 10.1038/s41467-017-00046-0 | www.nature.com/naturecommunications |
an in vitro luciferase assay in two model systems, AR-negative (PC-3) and AR-positive (LNCaP) prostate cancer cells (Fig. 2a). In PC-3 cells, significantly increased activity was observed in the presence of the minor allele (adenine) associated with SPOP mutation compared to the ancestral one (guanine). In contrast, inhibitory activity was observed in LNCaP cells, suggesting differential effects of the variant with respect to AR status. TF DNA-binding site (TFBS) motifs analysis demonstrated an AR consensus motif at the variant locus with the minor but not with the ancestral allele (Supplementary Fig. 5a, Supplementary Data 8). In addition, we identified a consensus motif for the CEAP family (Supplementary Fig. 5b), which includes known AR co-repressors. RNA-seq data show high levels of CEAP transcripts in multiple prostate tissue cell lines and a marked anti-correlation with AR levels in human prostate cancers (N = 319, P = 8e-18 Pearson correlation, Supplementary Fig. 6a, b). A less stringent TFBS search in a wider genomic region revealed additional CEAP-specific consensus motifs in proximity of the variant locus. In addition, we found overlapping CEAPB and AR motifs ~70 bp downstream the variant and a CEAPB putative-binding site ~180 bp upstream the variant, along with motifs for MAFB and c-MYC TFs, known to co-localize with CEAPB (Supplementary Data 8). We, therefore, investigated the effect of AR over-expression on TFs binding at the 7p14.3 locus.

ChIP-qPCR experiments showed AR and CEAPB recruitment at the site of interest upon AR over-expression (KLK3 and IL-6 were used as positive controls, respectively) (Fig. 2b). Based on the wider TFBS search results, we also verified the recruitment of c-MYC at the 7p14.3 region upon AR over-expression and dihydrotestosterone (DHT) treatment (Supplementary Fig. 7). To address the potential functional consequences of AR and CEAPB binding to the polymorphic site, we examined the effects of modulating androgenic signaling. We observed decreased responsiveness of the reporter constructs in PC-3 cells upon AR over-expression and DHT treatment (Supplementary Fig. 8), whereas loss of repressive activity upon AR knock-down in LNCaP cells was seen (Supplementary Fig. 9a, b). PC-3 cells also showed a decreased enhancer activity upon CEAPB over-expression, which was even stronger upon AR over-expression (Fig. 2c), and an increase in enhancer transactivation upon CEAPB knock-down (Fig. 2d). The construct harboring the minor allele (A) showed the same behavior, while eliciting higher enhancer activity at all conditions (Supplementary Fig. 10). These data suggest that CEAPB may act as an AR co-repressor at the variant locus, potentially through its recruitment by AR and that the 7p14.3 locus may undergo early somatic event. The combination of the two variables demonstrate the nontrivial impact that DNA repair and hormone-regulated genes have on trigger score ranking.

Structural associations with 7p14.3 locus. To further demonstrate that the 7p14.3 region is functionally active, we deleted 731 bp of genomic sequence around the site via CRISPR-Cas9 (Supplementary Data 9) in PC-3 cells and sequenced their transcriptomes. Differential transcript expression analysis of edited vs. non-edited cells (control cells) showed edited vs. control and control vs. control (P = 1e-04 and P = 3e-05 Mann–Whitney test, respectively, Supplementary Fig. 11) as validated by real-time PCR in selected genes (Supplementary Fig. 12). The fraction of deregulated genes that showed upregulation resulted on average in 70%, 65%, and 62%,

Fig. 1 Genetic predisposition to SPOP mutant prostate cancer. a Schematic representation of the trigger score computation. The number of DNA repair (Dnarep) and hormone-regulated genes (HormReg) from healthy prostate cells that are modulated by a functional variant are combined into a ranking score that measures the likelihood to observe a prostate-specific early somatic event. The combination of the two variables demonstrate the nontrivial impact that DNA repair and hormone-regulated genes have on trigger score ranking. b Trigger score distribution (left) across all considered functional variants; top ranked variants are highlighted. Genotype/phenotype analysis (right) is performed on random partitions of the data set into discovery and validation sets for three early recurrent prostate cancer lesions (SPOP mutations, FOXA1 mutations, and TMPRSS2-ERG rearrangements). An 7p14.3 variant associated to SPOP was implicated in 97.4% of all collected associations (187 of the 192 partitions for which association signal was detected, red portion of the ring plot). No variants in the partition space for 7p14.3 variant highlighted, dominant test considered. d Hematoxylin and eosin stained prostate cancer frozen tissue sections and corresponding Sanger sequencing are shown for a patient carrying the 7p14.3 variant ancestral genotype and lacking SPOP mutation (left) and a patient carrying the 7p14.3 variant minor allele genotype and harboring SPOP F133L mutation (right)
respectively. Along chromosome 7 the analysis between edited vs. control cells, but not edited vs. edited or control vs. control, showed significant concordance with genes predicted to physically interact with the 7p14.3 locus by previously generated Hi-C chromosome conformation capture data from benign prostate cells18 (Fig. 3a, b, \( P = 1e^{-05} \) and \( P = 5e^{-09} \) Mann–Whitney test, respectively). Validation of ETV1, NT5C3A and IGFBP3 deregulation upon deletion of 7p14.3 locus is shown (Fig. 3b, c) further supporting medium-range interaction.

Discussion

Over the past 15 years, numerous non-coding SNPs have been linked to the susceptibility to developing prostate cancer3, the second most frequent cancer in men causing each year more than 250,000 deaths worldwide, with modest albeit highly significant effects. Evidence of predisposition to TMPRSS2-ERG prostate cancer subtype was previously tested in small familial or sporadic cases21. Recently, two out of 27 common prostate cancer risk variants were found associated with modest signal to the ETS subclass22. Here we tested a specific hypothesis on large collections of human prostate tissues and identified a strong association with an emerging class involving mutation in the SPOP gene. Its phenotype is related to DNA repair and AR dysfunction11, 23 and defines a distinct molecular class as confirmed by the TCGA prostate adenocarcinoma publication1. Our findings suggest that the genetic component of this common disease is linked, at least partially, to specifically molecularly defined sub-classes through the modulation of AR targets and DNA repair genes. SPOP mutations that predominately involve hotspots located in the MATH substrate binding domain of the protein are prostate cancer specific24. Using in vitro models, we recently linked the SPOP mutant prostate cancer to genomic instability due to defects in homologous repair23.

While the mechanism linking the 7p14.3 variant and SPOP mutation remains elusive and future studies should investigate the role of the allele in the emergence of SPOP somatic mutations, we propose a relevant role in cancer predisposition of non-coding variants that lead to allele-specific transactivation of central TF programs with age dependent and tissue-specific effect21, 25 manifesting in early somatic genomic events. This study has potential important implications for the aging male whose testosterone levels change with advancing age, where a subtle differential effect might become significant to the cell (Fig. 4, Supplementary Fig. 13) and facilitate or accelerate the initiation of tumorigenesis in hormone sensitive tissues.

Findings from this study may be generalizable to other cancer subtypes. The predisposition to neoplastic development manifests in a combination of genomic and epigenetic events over time. Unlike many of the model systems that require results to occur in short periods of time, neoplastic development is a fitness selection process. As suggested with the 7p14.3 allele in this study, association with disease-specific pathways potentially enables a cascade of events over time to provide an advantage for cells that are hormonally regulated, for example, to undergo more frequent DNA damaging events that are not repaired. As in the current study, this genetic predisposition may positively select for SPOP mutations as they accelerate the DNA damage phenotype with a complex interplay of multiple TFs at the site of interest involving AR, CEBPB, and possibly c-MYC. Future large scale studies should explore the role of germ cell trigger events to announce driver somatic mutations.

Methods

Selection of human genome variants in functional regions. ChiP-Seq ENCODE datasets were selected based on availability of ETS1, H3K4me3, and H3K27ac regions of signal enrichment data (broadPeak format). The cell line set includes GM12878, H1-hESC, Hela-S3, HepG2, HMEC, HSMM, HSMMtube, HUVEC, K562, Monocytes-CD14+, NHA, NHDF-Ad, NHEK, NHLF, Osteobl, Du41. Consensus regions were determined for all three marker signals. Specifically, for each marker the consensus was generated as the overlap of all regions with enrichment signal above five detected in at least two cell lines as follow: (i) retain a region if it overlaps at least another cell line region for at least 50% of its length; (ii) concatenate the coordinates of regions from step (i) into a single coordinate list and perform BEDTools overlap in the order of regions for each chromosome. The list of hormone-regulated genes was obtained from27 (Table 1). The list of hormone-regulated genes was obtained from27 (Table 1). The list of hormone-regulated genes was obtained from27 (Table 1).

Selection of DNA repair genes and hormone-regulated genes. DNA repair genes list was obtained from the Human DNA Repair Genes database (http://sciencepark.mdanderson.org/labs/wood/dna_repair_genes.html) and an additional curated list (Supplementary Data 1) \( (N = 180) \). The list of hormone-regulated genes was obtained from27 (Table 1). Specifically, data for three biological replicates of LNCaP cells treated with small interfering RNA (siRNA) targeting AR and corresponding controls were considered (GM283829, GM288300, GM288301, GM288293, GM288294, GM288295) and for each replicate differentially expressed genes were selected as follows: (i) selection and quintile normalization of genes with reported detection \( P \)-values < 5% both in AR treated

Supplementary Table 1: Association signal of variant 7p14.3 with SPOP somatic phenotype

| Cohort               | AA+AG carriers | GG carriers | MAF | Comparison with SPOP.wt | Comparison with tyrol controls (\( N = 1,014 \)) | Comparison with tyrol extended controls # (\( N = 1,291 \)) | Comparison with 1000 Genomes Project controls ## (\( N = 2,504 \)) | Comparison with all controls ## (\( N = 3,795 \)) |
|----------------------|----------------|-------------|-----|------------------------|-----------------------------------------------|----------------------------------------------------------|----------------------------------------------------------|----------------------------------------------------------|
| Adenocarcinomas discovery | 24             | 217         | 0.052 | 5.75 | 3.0e-04 | 10.47 | 1.7e-07 | 10.2 | 9.1e-08 | 5.44 | 3.0e-05 | 6.38 | 4.5e-06 |
| Adenocarcinomas validation | 23             | 217         | 0.048 | 4.45 | 4.1e-03 | 8.22 | 8.8e-06 | 7.9 | 9.1e-06 | 4.04 | 1.5e-03 | 4.73 | 4.0e-04 |
| Adenocarcinomas complete | 47             | 434         | 0.050 | 4.83 | 6.7e-06 | 9.20 | 1.5e-10 | 8.9 | 7.4e-11 | 4.72 | 3.0e-07 | 5.54 | 1.2e-08 |
| Adenocarcinomas EUR only | 42             | 373         | 0.052 | 4.96 | 3.7e-05 | 10.07 | 1.3e-09 | 9.75 | 7.4e-10 | 7.86 | 1.1e-07** | 8.44 | 7.8e-10** |
| Validation Korean | 19             | 61          | 0.206 | 5.84 | 4.0e-02 | -    | -    | -    | -    | 4.78 | 4.3e-02*** | -    | -    |

Results refer to logistic regression analysis using dominant model corrected for age and prostate-specific antigen (PSA). First three rows show data from a random partition (discovery and validation) and the complete data set; columns include signal upon data set extension to controls from the Tyrol PSA Screening Cohort and the 1000 Genomes Project individuals collection. Data is also reported for EUR descent individuals only and for an independent cohort of Korean patients (EAS from 1000 Genomes Project collection included as controls). #ETS positive/SPOP.wt tumors and controls, ##analysis not corrected for age and PSA, **EUR individuals only included (\( N = 503 \)), ***EUR individuals only included (\( N = 504 \)).
and control cells. (ii) Selection of genes with absolute-change, i.e., log2(treated/control), equal or greater than 1. The final hormone-regulated gene list (Supplementary Data 2) is obtained by merging the genes differentially expressed in at least one of the three replicates.

Somatic phenotype data sets. Whole-exome or whole-genome sequencing data from prostate cancer tissue samples was queried for early somatic lesions\[1,9,11\]. Patients with relevant clinical annotations (age, PSA), functional variant genotypes and lesion status for SPOP (N = 539, 12.1% mutated), TMPRSS2-ERG (N = 451, 47.2% rearranged) and FOXA1 (N = 520, 5.4% mutated) were included in the study (N total = 539, Supplementary Data 4). Variants genotypes were determined using standard APT tools 1.16.1 pipeline from Affymetrix SNP 6.0. As all data sets used clinically localized prostate cancer cases, none of these data sets have meaningful clinical follow up data, which would require ten or more years.

Ethnicity analysis. Ethnicity of all individual’s samples was inferred using an approach based on inspection of differential germline variants genotype. First, by combining genotype data of individuals with known ethnicity a reference model is built; genotype data by the International HapMap Project was used. A target model is then created using genotype data from all 539 individuals in the somatic data set. Principal component analysis (PCA) is then performed by means of smart pca module\[28\] on aggregated target and reference models genotype data. Euclidean space defined by the first two PCA components is then inspected to, first, generate smallest convex sets identifying main ethnic groups (EUR, AFR, EAS, AMR, and

Fig. 2 Functional characterization of 7p14.3 variant. a Luciferase assays were performed on PC-3 and LNCaP cells transfected with pGL4.26 vectors containing 7p14.3 (A or G allele, represented in light grey and dark grey, respectively) or empty vector (white); mean ± s.d. of three biological replicates. b PC-3 cell were transfected with pCMV_Empty (solid bars) or pCMV_AR (dashed bars) vectors; AR (left) or CEBPB (right) chromatin binding at 7p14.3 locus in PC-3 cells were evaluated by ChIP-qPCR. Occupancy level at KLK3 enhancer and IL-6 promoter was used as positive control of AR and CEBPB, respectively. Data are represented as mean ± s.d. of two biological replicates. c Luciferase assays on PC-3 cells co-transfected with pCMV6_CEBPB and/or CMV_AR (dashed bars) along with the different pGL4.26 reporter vectors described above. The enhancer activity is inhibited upon CEBPB overexpression. The inhibition becomes stronger upon AR overexpression. Data are represented as mean ± s.d. of two biological replicates. d Luciferase assays on PC-3 cells transfected with siRNA against CEBPB or scrambled siRNA. Then, cells were co-transfected with pCMV_Empty or pCMV_AR vectors along with the pGL4.26 reporter vectors described above. Data are represented as mean ± s.d. of two biological replicates. Where indicated, cells were treated for 16 h with ETOH or DHT. *P < 0.05, **P < 0.01, ***P < 0.005, Student’s t-test.

Ethnicity analysis. Ethnicity of all individual’s samples was inferred using an approach based on inspection of differential germline variants genotype. First, by combining genotype data of individuals with known ethnicity a reference model is built; genotype data by the International HapMap Project was used. A target model is then created using genotype data from all 539 individuals in the somatic data set. Principal component analysis (PCA) is then performed by means of smart pca module\[28\] on aggregated target and reference models genotype data. Euclidean space defined by the first two PCA components is then inspected to, first, generate smallest convex sets identifying main ethnic groups (EUR, AFR, EAS, AMR, and
Tyrol Early Prostate Cancer Detection Program cohort (N = 80) were queried. This set included 1036 healthy controls and additional 492 individuals considered as SPOP wild type (Supplementary Data 4). No statistically significant genotype/phenotype association was found when testing the 7p14.3 variant in the Tyrol cohort against prostate cancer risk (P = 0.47, logistic regression analysis), TMPRSS2-ERG rearrangement (P = 0.11, logistic regression analysis) or aggressive PCa (P = 1, logistic regression analysis).

Korean cohort. Prostate cancer patients of Korean descent were previously annotated for SPOP mutations. Individuals with relevant clinical annotations (age and PSA) and SPOP mutation status (N = 80, 8.7% SPOP mutated) were retained. The rs1376350 variant genotype was assessed by TaqMan assay (Supplementary Data 4).

Control cohorts. Genotype data for the rs1376350 study variant was retrieved for 1903 individuals from the Korean cohort (Supplementary Data 4). Prostate cancer patients of Korean descent were previously genotyped and matched genotype data for the rs1376350 study variant was retrieved for 20130502). Genotype data and clinical information for 1903 individuals from the Tyrol Early Prostate Cancer Detection Program cohort were queried. This set includes 1036 healthy controls and additional 492 individuals considered as SPOP wild type (Supplementary Data 4). No statistically significant genotype/phenotype association was found when testing the 7p14.3 variant in the Tyrol cohort against prostate cancer risk (P = 0.47, logistic regression analysis), TMPRSS2-ERG rearrangement (P = 0.11, logistic regression analysis) or aggressive PCa (P = 1, logistic regression analysis).

Transcriptome analysis and trigger score assessment. Benign (N = 63) and tumor (N = 319) prostate tissues RNA-Seq data with available FASTA files were used. Functional annotations were aligned to the reference genome hg19 using STAR aligner and logarithm transformed (two based) RPKM+1 of each gene (UCSC knownGenes) were computed using mrfQuanti. Real-time PCR validation of selected genes nominated by RNA-seq, ETV1, NT5C3A, and IGFBP3; dotted lines represent fold thresholds in RNA-Seq analysis to identify deregulated transcripts (Supplementary Fig. 12). Data are represented as mean ± s.d. of three technical replicates.
Among the 21,364 functional variants considered in the study only 881 (~4%) had a positive trigger score (Supplementary Data 3). Top ranked variants in the highest tertile of positive trigger score distribution were retained for further analysis (N = 300). The relationship between the variants minor allele frequencies (MAF) and trigger scores was investigated (Supplementary Fig. 14); no association detected.

For the 7p14.3 variant we then also performed genome-wide association analysis considering 18,758 sequenced transcripts (all transcripts with normalized RPKM greater or equal to 1 in at least one individual). Linear regression of RPKMs across genotype classes using dosage model was performed applying a FDR threshold of 5%. The variant was found genome-wide associated with other 1515 genes of which 723 (48%) show downregulation in presence of the minor allele while 792 (52%) present upregulation.

DNA repair and hormone-regulated genes associated with 7p14.3 variant (Supplementary Data 7) were tested for differential expression across prostate tissue samples and the score computed from the 1000 Genomes Project samples set (to mimic the prostate tissue sample size) and the score computed using mrfQuanti er32 and were quintile normalized.

For each of the top selected trigger score variants (N = 300), we measured the trigger score prostate specificity by comparing the score computed from the benign prostate tissue samples and the score computed from the 1000 Genomes Project samples. We performed 100 random sampling of 63 individuals from the 1000 Genomes Project samples set (to mimic the prostate tissue sample size) and computed the trigger score for all top 300 variants. We then annotated a variant as non-global, if no positive score was observed across the 100 experiments; a global trigger score was annotated if at least one experiment provided a positive score. A total of 69 (23%) variants showed non-global scores, where the score was positive only in the prostate tissue dataset (see Supplementary Data 3). No association between variants MAF and global or non-global annotations was detected.

**Genotype/phenotype association analysis.** Genotype/phenotype association analysis was performed on the top selected trigger score variants (N = 300), after excluding variants with genotype calling rate <85% (N = 423) (Supplementary Data 5). Logistic regression analysis was used to test genotype/phenotype associations and was performed using PLINK 1.073 considering allele, dominant and recessive models. Dominant and recessive models were tested for the minor allele. Association analyses were performed applying age and PSA correction as available. In order to minimize genotype/phenotype FDR, we computed multiple rounds of discovery and validation by partitioning the whole data set in two subsets two hundred times for each somatic phenotype (lesions in SPOP, TMPRSS2-ERG, and FOXA1) by preserving the lesion incidence in each subset. Specifically, in each partition the genotype/phenotype association was tested for the top selected variants applying corrected P-value (FDR) cutoff at 20% and P-value cutoff at 5% for discovery and validation, respectively. A genotype/phenotype association was considered only if both thresholds were met.

Variant 7p14.3 genotypes were tested for association with total number of tumor SNVs and with tumor genomic burden in 427 and 474 patients, respectively, (data availability as per cBioPortal), using Mann–Whitney test statistics. The tumor genomic burden was measured as the fraction of the genome with absolute log2(ratio) of tumor over normal above conventionally used threshold (0.15) (Supplementary Fig. 1).

**Protein-protein interaction analysis.** To characterize the PPI network of the transcriptome fraction modulated by the study variant, we first built a reference PPI network by merging information of five databases: BioGRID http://thebiogrid. org/ (Release 3.2); HPRD http://www.hprd.org/ (Release 9 20100413); IntAct http://www.ebi.ac.uk/intact/ (Release 20150120); MINT http://mint.bio.uniroma2.it/mint/ (Release 20130326); STRING http://string-db.org/ (Release 9).

Interactions between nodes that represent human proteins and with a confidence score greater than 0.7 were retained (all HPRD interactions were included because no score measure is associated to protein–protein interactions in that database). The resulting network contains 263,369 interactions and 16,002 human proteins. The PPI network involving 7p14.3 variant-associated genes was built from the reference PPI network and composed of 953 genes and 1755 interactions.

To determine how likely is that the fraction of the transcriptome modulated by a variant reflects in a PPI network with a connected component comparable to the 7p14.3 variant network component, which is made of 552 genes and 1717 interactions (Supplementary Fig. 3), we built three distributions: (i) for each functional variant considered in the study (i.e., functional variants in active enhancers), the relative proportion of the biggest connected component present in the corresponding induced PPI network was calculated and a reference distribution was built; (ii) in 10,000 random variants along the genome were selected (among all variants available in the Affymetrix SNP 6.0 platform) and the relative proportion of the biggest connected component present in corresponding induced PPI network was calculated for each variant to build a reference distribution; and (iii) using all genes from the reference PPI network (N = 16,002), 10,000 random sets of size 953 were generated and the reference distribution of the relative proportions of the biggest connected component present in the induced network was built. P-values were then computed for 7p14.3 variant-induced network using the three computed reference distributions (see Supplementary Fig. 4). Graphical visualization of PPI networks was performed using both igraph library33 of R programming language and Cytoscape tool35.

Pathway enrichment analysis was performed for the genes in the 7p14.3 variant PPI connected component (N = 552) on the REACTOME pathway database36.
ChIP assay. PC-3 cells were maintained into 150 mm Petri dishes in RPMI medium without phenol red, supplemented with 10% charcoal/dextran treated FBS. Two days after, as PC-3 cells do not express AR, they were transfected either with control pCAGGS(2X) expression vector or with the pcCMV6-CEBPB plasmid vector. Then, cells were treated with EtOH or DHT (100 nM) and after 16 h of treatment, cells were subjected to ChIP with an AR-anti AR antibody (3 µg, 17-10489 ChIP+ and androgen receptor Antibody Assay Kit, Millipore), anti-CEBPB antibody (3 µg, 18F8 Abcam), anti-c-Myc antibody (3 µg, N262 Santa Cruz Biotechnology) or a normal IgG (3 µg, CS200581), using the MagnaChIP HiSens Chromatin IP Kit (17-10461 Upstate, Millipore) as previously described.

Briefly, the experiment procedure includes chromatin crosslinking with formaldehyde, chromatin shearing for 45 cycles of 30 s ON/30 s OFF with the BioRuptor Pico (Diagenode), protein–DNA complex immunoprecipitation and reverse crosslinking with protease K. Pre-implicated DNA was analyzed by real-time qPCR with KAPA SYBR FAST Universal 2X qPCR Master Mix (Kapa Biosystems, Resnova) using the CFX384 or CFX96 Detection Systems (BioRad). KLK3 enhancer region, IL-6 promoter region and NPM1 intron 1 were used as positive control of AR, CEBPB and c-Myc, respectively; a desertic region in chr12 as per ENCODE annotations (hg19, chr12:17456963-17457066) was used as negative control. The region of interest surrounding the chr14.3 variant was amplified as well (Supplementary Data 10). AR, CEBPB or c-Myc specific recruitment was calculated as enrichment respect to the IgG according to the ΔCt method.

CRISPR-Cas9. 7p14.3 deleted locus in PC-3 cells. Single guide RNA oligos (sgRNAs) to induce 7p14.3 locus deletion were selected using the GPP Web Portal (http://portals.broadinstitute.org/gpp/public/) that ranks candidates according to their predicted on-target and off-target activity (Supplementary Data 9). Selected sequences were ligated into pSpCas9(1.1)-2A Puromycin vector, derived from px330 (Addgene 42230). Plasmid eSpCas9(1.1)-2A Puromycin, derived from Addgene 71814, were generated through the addition of nucleotides encoding 2 A peptide and Puromycin resistance. To induce genomic deletions four pairs of sgRNA (listed in Supplementary Data 9) vectors were cotransfected with 25 nM sgRNA expression vector. On 6 wks plated, 25% of PC-3 cells were seeded and after 24 h were transfected with 1.5 µg of eSpCas9(1.1)-2A Puromycin plasmid and 250 ng of sgRNA plasmid using FuGENE HD (Promega E2311). Three days after transfecion, cells were selected for 7 days with 2 µg/ml Puromycin (Sigma-Aldrich P8833) in order to prioritize transfected cells only. A representative amount of cells was used for DNA extraction and the remaining cells were replated and cultured for RNAseq experiments. Genomic DNA was extracted following the procedures of Nucleospin tissue kit (Macherey-Nagel). To evaluate the editing, 25 µl PCR reaction was performed using 12.5 µl Platinum SuperFi Green PCR Master Mix (2X), 50 ng DNA template and 0.5 µM forward and reverse primers (Supplementary Data 10). Predicted PCR bands (Supplementary Fig. 16), across the genomic deletion, are listed in Supplementary Data 9. Based on editing efficiency, combinations A and B were selected for downstream experiments (deleted segments of length 625 bp and 731 bp, respectively).

RNAseq experiments. PC-3 cells (7p14.3 deleted and not deleted cells) were seeded in 24-well plate, transfected with pCMV-AR24Q expression vector using FuGENE HD (Promega E2311) and then treated with DHT for at least 16 h. CEBPB silencing was performed by transfection of PC-3 cells with siRNA against CEBPB (20 nM) (FlexiTubeGeneSolution, Qigen) and Lipofectamine RNAiMAX Transfection Reagent (Invitrogen, Carlsbad, CA). 18F8 cells were transfected with the RNeasy kit (Qagen) according to the manufacturer’s instructions. The RNA integrity number (RIN) was quantified on the Agilent 2100 bioanalyzer. cDNA libraries were prepared with Trueseq stranded mRNA library prep Kit (RS-122-2101, Illumina) using 500 ng of total RNA. Single end (100 bp) sequencing was performed on a HiSeq 2500 Illumina, FASTA files were aligned to the reference genome hg19 using STAR aligner and log2 ratio greater or equal than 1. Concordance of deregulation in cells edited with their predicted on-target and off-target activity (Supplementary Data 9). Selected sequences were ligated into pSpCas9(1.1)-2A Puromycin vector, derived from px330 (Addgene 42230). Plasmid eSpCas9(1.1)-2A Puromycin, derived from Addgene 71814, were generated through the addition of nucleotides encoding 2 A peptide and Puromycin resistance. To induce genomic deletions four pairs of sgRNA (listed in Supplementary Data 9) vectors were cotransfected with 25 nM sgRNA expression vector. On 6 wks plated, 25% of PC-3 cells were seeded and after 24 h were transfected with 1.5 µg of eSpCas9(1.1)-2A Puromycin plasmid and 250 ng of sgRNA plasmid using FuGENE HD (Promega E2311). Three days after transfecion, cells were selected for 7 days with 2 µg/ml Puromycin (Sigma-Aldrich P8833) in order to prioritize transfected cells only. A representative amount of cells was used for DNA extraction and the remaining cells were replated and cultured for RNAseq experiments. Genomic DNA was extracted following the procedures of Nucleospin tissue kit (Macherey-Nagel). To evaluate the editing, 25 µl PCR reaction was performed using 12.5 µl Platinum SuperFi Green PCR Master Mix (2X), 50 ng DNA template and 0.5 µM forward and reverse primers (Supplementary Data 10). Predicted PCR bands (Supplementary Fig. 16), across the genomic deletion, are listed in Supplementary Data 9. Based on editing efficiency, combinations A and B were selected for downstream experiments (deleted segments of length 625 bp and 731 bp, respectively).
Hi-C signal above the 90th percentile of the overall intra-chromosomal 7p14.3 normalized Hi-C signal distribution (Supplementary Fig. 18).

Real-time qPCR. PC-3 cells were seeded in 6-well plate. Total RNA was extracted using the RNeasy kit (Qiagen) according to the manufacturer’s instructions. Two-hundred nanogram of total RNA was retro-transcribed into cDNA using the Revert Aid First Strand cDNA Synthesis Kit (ThermoFisher Scientific). Then, qPCR reactions in real-time were performed using KAPA SYBR FAST Universal 2X qPCR Master Mix (Kapa Biosystems, Resnova) using the CFX384 or CFX96 Detection Systems (BioRad). Analysis of relative mRNA expression was performed using the ΔΔCt method with GAPDH (glyceraldehyde 3-phosphate dehydrogenase) as reference genes (primers sequences in Supplementary Data 10). RNA-seq validation of selected transcripts (Supplementary Data10) was performed with qPCR including RAB1B, BAP1, and BCAP31 as negative controls. As control (Fig. 6B), Human cDNA was used for a parallel qPCR, normalized with 36B4, and GAPDH as reference.

Western blot. Proteins were extracted from PC-3 cells with ice-cold RIPA (Radio Immuno-Precipitation Assay) lysis buffer supplemented with protease inhibitor cocktail (11873580001, Roche), then homogenized in a dounce homogenizer for 1 h and centrifuged at 13,400 × g at 4 °C for 10 min. The supernatants were collected and boiled with 6X sample buffer at 95 °C for 5 min. The samples were separated by sodium dodecyl sulfate polyacrylamide gel electrophoresis and transferred to nitrocellulose membranes using wet transfer or the semi-dry blot Transfer System (Invtrogen, Life Technologies). The membranes were blocked with 5% non-fat dry milk in PBS-T overnight at 4 °C. Membranes were then incubated with secondary goat anti-mouse (Santa Cruz Biotechnology A0944, 1:10000) or goat anti-rabbit antibodies (Sigma-Aldrich A9165, 1:12000) for 1 h at room temperature. Detection was achieved using the ECL Select detection reagent (Amersham, GE Health Care) with the ChemiDoc XRS + System (BioRad) (Supplementary Figs.19 and 20).

Data availability. RNA-seq data of control and edited PC-3 cells have been deposited in the BioProject database under the accession code PRJNA831797. All other remaining data are available within the article and Supplementary Files, or available from the authors upon request.

Received: 8 September 2016 Accepted: 28 April 2017
Published online: 29 June 2017

References
1. Cancer Genome Atlas Research N. The molecular taxonomy of primary prostate. Cancer Cell 163, 1011–1025 (2015).
2. Takai, K. et al. Recurrent deletion of an androgen receptor enhancer in prostate cancer genomes. Proc. Natl. Acad. Sci. USA 107, 9095–9100 (2010).
3. Bregman, A. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 58, 1255–1266 (2015).

Acknowledgements
This work was supported by the Associazione Italiana per la Ricerca sul Cancro (AIRC, MIUR 13562 and IG 19221 to F.D.) and the U.S. Department of Defense (PCP94536 to F.D.). We would like to thank Nima Sharif and Francesca Lorenzo for constructive discussions on the manuscript, Myriam Kossai for technical work, and the LaBSAAH-CIBIO Next Generation Sequencing Facility of the University of Trento for sequencing samples.
Author contributions
A.R., S.G., M.A.R., and F.D. designed the study. S.G., B.S., and P.G. performed the experiments with input from D.C., K.A.C., A.I., G.P., A.C., K.A.C., and M.B. characterized the tumor samples. A.R., D.D., and F.D. performed the computational and statistical analysis with input from D.S., P.D., O.E., A.S., and E.K. A.R., S.G., and F.D. wrote the manuscript. F.D. oversaw the project. All authors read and approved the final manuscript.

Additional information
Supplementary Information accompanies this paper at doi:10.1038/s41467-017-00046-0.

Competing interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© The Author(s) 2017