Published online 22 December 2022 *Nucleic Acids Research, 2023, Vol. 51, No. 3* ***e18***
https://doi.org/10.1093/nar/gkac1198

Optimized high-throughput screening of non-coding variants identified from genome-wide association studies

Tunc Morova[1], Yi Ding[2], Chia-Chi F. Huang[1], Funda Sar[1], Tommer Schwarz[2], Claudia Giambartolomei[3,4], Sylvan C. Baca[5], Dennis Grishin[5], Faraz Hach[1,6], Alexander Gusev[5,7], Matthew L. Freedman[5,8], Bogdan Pasaniuc[2,4,9,10] and Nathan A. Lack[1,6,11,12,*]

[1]Vancouver Prostate Centre, Vancouver, BC V6H 3Z6, Canada, [2]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA 90095, USA, [3]Central RNA Lab, Istituto Italiano di Tecnologia, Genova 16163, Italy, [4]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA, [5]Department of Medical Oncology, The Center for Functional Cancer Epigenetics, Dana Farber Cancer Institute, Boston, MA 02215, USA, [6]Department of Urologic Science, University of British Columbia, Vancouver, BC V5Z 1M9, Canada, [7]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA, [8]The Center for Cancer Genome Discovery, Dana Farber Cancer Institute, Boston, MA 02215, USA, [9]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA, [10]Department of Computational Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA, [11]School of Medicine, Koç University, Istanbul 34450, Turkey and [12]Koç University Research Centre for Translational Medicine (KUTTAM), Koç University, Rumelifeneri Yolu, Istanbul 34450, Turkey

Received September 26, 2022; Revised November 19, 2022; Editorial Decision November 26, 2022; Accepted December 06, 2022

ABSTRACT

The vast majority of disease-associated single nucleotide polymorphisms (SNP) identified from genome-wide association studies (GWAS) are localized in non-coding regions. A significant fraction of these variants impact transcription factors binding to enhancer elements and alter gene expression. To functionally interrogate the activity of such variants we developed snpSTARRseq, a high-throughput experimental method that can interrogate the functional impact of hundreds to thousands of non-coding variants on enhancer activity. snpSTARRseq dramatically improves signal-to-noise by utilizing a novel sequencing and bioinformatic approach that increases both insert size and the number of variants tested per loci. Using this strategy, we interrogated known prostate cancer (PCa) risk-associated loci and demonstrated that 35% of them harbor SNPs that significantly altered enhancer activity. Combining these results with chromosomal looping data we could identify interacting genes and provide a mechanism of action for 20 PCa GWAS risk regions. When benchmarked to orthogonal methods, snpSTARRseq showed a strong correlation with *in vivo* experimental allelic-imbalance studies whereas there was no correlation with predictive *in silico* approaches. Overall, snpSTARRseq provides an integrated experimental and computational framework to functionally test non-coding genetic variants.

INTRODUCTION

Germline genetic variants contribute to numerous diseases from COVID-19 (1) to cancer development (2–6). Disease-associated SNPs are primarily identified from GWAS (7). While those SNPs that occur in protein-coding regions have a predictable impact on protein sequence, the vast majority of disease-associated SNPs are located in non-coding regions (8,9). There is increasing evidence that these non-coding variants affect disease initiation and progression by altering critical *cis*-regulatory elements (CRE) that are involved in the spatiotemporal expression of target genes (10–14). These variants commonly occur at enhancers where they can alter transcription factors (TF) binding and gene transcription (15–19). For instance, a SNP in the PCAT19 locus disrupts NKX3.1 and YY1 binding which alters

[*]To whom correspondence should be addressed. Tel: +1 604 875 4411; Email: nlack@prostatecentre.com

enhancer activity causing dysregulation of oncogene expression and prostate cancer (PCa) progression (14,20,21). While most SNPs identified from GWAS studies are found within the non-coding region of the genome, it remains difficult to mechanistically characterize the impact of these variants (22,23).

Several *in silico* and experimental approaches are commonly used to characterize potential pathogenic non-coding variants. Current *in silico* methods apply machine learning that is trained on previously published data to predict activity without experimental input (23–26). Given their relative ease, these bioinformatic approaches are used to stratify candidates for validation. Yet, they do have several major limitations. Previous benchmarking studies demonstrated considerable variability between each computational method (27) with a very high error rate (28–30). In contrast, experimental approaches are much more robust. One promising method utilizes *in vivo* chromatin immunoprecipitation with sequencing (ChIPseq) to measure the impact of non-coding variants on allelic-imbalance of TF binding (31–34). Demonstrating the potential utility, recent work combined >7000 ChIPseq from 649 cell lines and identified 270 000 SNPs with altered TF binding affinity (25). While promising, this experimental approach is limited by the frequency of each variant in the tested population which requires a substantial number of clinical samples for accurate functional genomic testing. These challenges, therefore, limit allelic-imbalance studies to only a handful of specific experimental models, tissues, and disease states (35). To overcome these challenges, massively parallel reporter assay (MPRA) (36) including self-transcribing active regulatory region sequencing (STARRseq) (37) has been used to directly quantify the activity of thousands of non-coding sequences in a single experiment (38–41). Importantly, these methodologies do not require clinical samples and are amenable to functional perturbations. Several studies have adopted STARRseq to systematically screen the impact of SNPs on enhancer function (42–52). These methods can be broadly separated based on the source of the target non-coding sequence with variants. While a few obtained mutant and variant sequences from mixed DNA libraries (43,47), many used oligonucleotide synthesis to obtain enhancer sequences (42,44,45,49–53). While synthesis-based methods can generate sequences harboring targeted variants that allow fine-mapping of complex haplotypes (51,52), pooled methods incorporate longer fragments that provide additional sequence context of co-regulator binding (54) and produce more reproducible enhancer activity quantification (55). Despite the feasibility of the large-scale functional screens, their performance has been limited by library representation which causes poor signal-to-noise, false positives, or limited statistical robustness. Further, it is unclear how these plasmid-based methods compare to *in silico* predictions and other experimental approaches.

In this work, we developed a standardized experimental and computational STARRseq framework to identify disease-associated genetic variants that impact enhancer activity. To address the previous limitations due to short enhancer sequences, we developed asymmetrical Illumina sequencing and covered enhancer fragments up to 841 bp (mean: 543 bp). This can accurately identify critical non-coding SNPs and test hundreds to thousands of variants in a single experiment Using this approach we functionally tested 68 SNPs from known prostate cancer (PCa) risk-associated loci and demonstrated that 36 (51%) of them significantly altered enhancer activity. Combining these results with chromosomal looping we provided a mechanism of action for 20 PCa GWAS risk regions. Our methodology, snpSTARRseq, provides streamlined bioinformatic analysis and is amenable to different genomic regions, diseases, and sequencing approaches including PacBio Long Reads. Overall, snpSTARRseq functionally characterizes hits from GWAS studies and provides a mechanistic understanding of critical genetic variants.

MATERIALS AND METHODS

Detailed information can be found in Supplementary Methods.

snpSTARRseq capture library design and additional SNP expansion

We tested disease-associated 252 SNPs (Supplementary Table S1), which are located in enhancer regions that have H3K27Ac signal and chromosomal looping to a gene promoter (56) as well as 50 control regions (25 positive and 25 negative control). Positive control regions (strong enhancers) were identified from previously published whole genome STARRseq (57). Negative-control regions contained an androgen response element (ARE) motif but no AR binding or enhancer activity in published work (41). Chromosomal locations of all capture regions can be found in (Supplementary Table S2).

Generation of snpSTARRseq capture library

Pooled human genomic DNA (NA13421; consisted of 27 males and 27 females from CEPH Utah pedigrees) obtained from Coriell Institute for Medical Research (58) (Supplementary Table S3) was fragmented (500–800 bp), end-repaired and ligated with xGen stubby adaptors (IDT) containing random i7 3bp UMI. The captured regions were enriched using xGen biotinylated oligonucleotide probe pool (IDT) (59) and Dynabeads M-270 Streptavidin beads (IDT). Post-capture was PCR-amplified with STARR_in-fusion_F primer and STARR_in-fusion_R primer, and then cloned into AgeI-HF (NEB) and SalI-HF (NEB) digested hSTARR-ORI plasmid (Addgene plasmid #99296) with NEBuilder HiFi DNA Assembly Master Mix (NEB). The snpSTARRseq capture library was then transformed into MegaX DH10B T1R electrocompetent cells (Invitrogen) and the plasmid DNA was extracted using the Qiagen Plasmid Maxi Kit. The sequences of all the primers used for generating the snpSTARRseq capture library were listed (Supplementary Table S4).

Experimental method for snpSTARRseq and sequencing

The cloned snpSTARRseq library (100 ug plasmid DNA/replica) was transiently transfected into LNCaP cells (5×10^7 cells/replica; three biological replicas) using the Neon Transfection System (Invitrogen). Cells were grown in Roswell Park Memorial Institute (RPMI) 1640 medium (Gibco) supplemented with 10% fetal bovine serum (FBS) and collected 48hrs post electroporation. These cells were lysed with Precellys CKMix Tissue Homogenizing Kit (Bertin Technologies) and total RNA was extracted using RNeasy Maxi Kit (Qiagen). mRNA was isolated with Oligo (dT) 25 Dynabeads (Thermo Fisher) and the reverse transcription was done with the plasmid-specific primer. The synthesized snpSTARRseq cDNA was treated with RNaseA (Thermo Fisher) and amplified by a junction PCR (15 cycles) with the RNA_jPCR_f primer and the jPCR_r primer. The snpSTARRseq capture library was PCR-amplified with DNA-specific junction PCR primer (DNA_jPCR_f primer) and jPCR_r primer. All primer sequences were listed in Supplementary Table S4. After purification with Ampure XP beads (Beckman Coulter), both the snpSTARRseq samples were PCR-amplified with TruSeq dual indexing primers (Illumina) to generate Illumina-compatible libraries. RNA samples were sequenced with a HiSeq4000 (150 bp; paired-end (PE)) while the DNA STARRseq capture library was asymmetrically sequenced with Illumina MiSeq PE reads. In the latter sequencing, we did two rounds of PE sequencing with round 1 being forward 75 cycles/reverse 425 cycles reverse and round 2 being forward 425 cycles/reverse 75 cycles reverse.

PacBio long-read sequencing

snpSTARRseq input DNA library was digested with NotI-HF (NEB) enzyme for linearization of the plasmid DNA. After that sequence library was sequenced by the PacBio SMRT link. Raw sequencing data (subreads.bam) was processed by the *ccs* function of SMRT tools (version 9.03). Output CCS file was processed by our pipeline (see the section below) to validate reconstructed enhancer fragments.

Reconstruction of enhancer sequences

We developed our computational framework to reconstruct enhancer sequencing by using ‘long’ pairs of asymmetrical reads that cover the full enhancer sequence. Briefly, UMI attached reads from ‘long-short’ and ‘short-long’ asymmetrical sequencing are first clustered with Calib's *cluster* (60) based on their UMI and sequence context (Figure 1B-Clustering). This step gathers reads belonging to the same enhancer fragments from each independent run. This step is followed up by consensus sequence generation to correct any random error due to sequencing (Figure 1B-Consensus). Having the high-quality reads generated for each enhancer fragment, in the next step, ‘long’ reads of ‘long-short’ and ‘short-long’ were matched using 12 bp sequences from 5' and 3' of the enhancer fragments. These short sequences represent enhancer fragments and are used as unique barcodes (Figure 1B-Match). Consequently, matched long reads then collapsed to reconstruct enhancer fragment (Figure 1B-Collapse) using *bbmerge* software (61). A detailed explanation of each step of the analysis pipeline can be found in the Supplementary Methods section. The asymmetrical processing pipeline can be found at https://github.com/mortunco/snp-starrseq.

Testing bi-allelic activity

To identify SNPs with allelic-specific enhancer activity, we conducted a Negative-Binomial Regression analysis to compare the expression of alternative allele-supporting fragments with reference allele-supporting fragments. Fragments overlapping at a SNP position were assigned as alternative or reference types based on the allele they carry. Only those SNPs with >15 unique plasmids for both alternative and reference type alleles were included for analysis. For each SNP, a negative Binomial regression was performed with the following model by using the *glm.nb( )* function in the MASS R package (version 7.3–54) (62):

$$log\,(E\,(Y_i)) = \beta_{oj} + \beta_{ij}\,X_{ij} + log\,(P_i)$$

where Y_i is the RNA read counts of fragment i, $X_{ij}$is the allele type of SNP j carried by fragment i, where X_{ij}=0 when the allele on fragment i is the reference allele-supporting type and X_{ij}=1 when the allele on fragment i is an alternative allele-supporting type, $\beta_{oj}$is the log expression per plasmid of the reference allele and β_{1j} is the log fold change of expression per plasmid comparing alternative type allele versus reference type allele, P_i is the plasmid DNA read count of the barcode serving as an offset term.

Empirical type-I error for NBR

As the fragment enhancer activity can be affected by not only the variants they carry but also the specific genomic region they cover, therefore, enough coverage of the SNP to be tested is required to reduce the impact of the position bias. We conducted an empirical analysis to investigate the relationship between the number of fragments and FDR. First, we selected 10 independent SNPs randomly with at least 30 fragments for each of the VAR and REF alleles, absolute alternative allele effect smaller than 0.1, and p-value larger than 0.5 to treat them as true null SNPs. For each SNP, we downsampled the fragments of each allele type to N (where N = 5, 10, 15, 20, 25, 30) and conducted the NBR to test the allelic-specific enhancer activity. We repeated the process 100 times to compute the proportion of tests with p-value < 0.05 as the empirical type-I error at a significance level of 0.05.

In silico method comparisons

We obtained five different prediction scores tables for each method (ncER (23), CADD (24), DVAR (63), LINSIGHT (64), deltaSVM (26)) and compared them with respect to snpSTARRseq absolute allelic effect abs(Log_2(ALT/REF)) and adjusted P-values (FDR). Detailed information about every step can be found in Supplementary Methods and the visualization code can be found at https://github.com/mortunco/snp-starrseq. We stored corresponding snpSTARRseq allelic-effect and in silico impact scores in supplementary table (Supplementary Table S5).

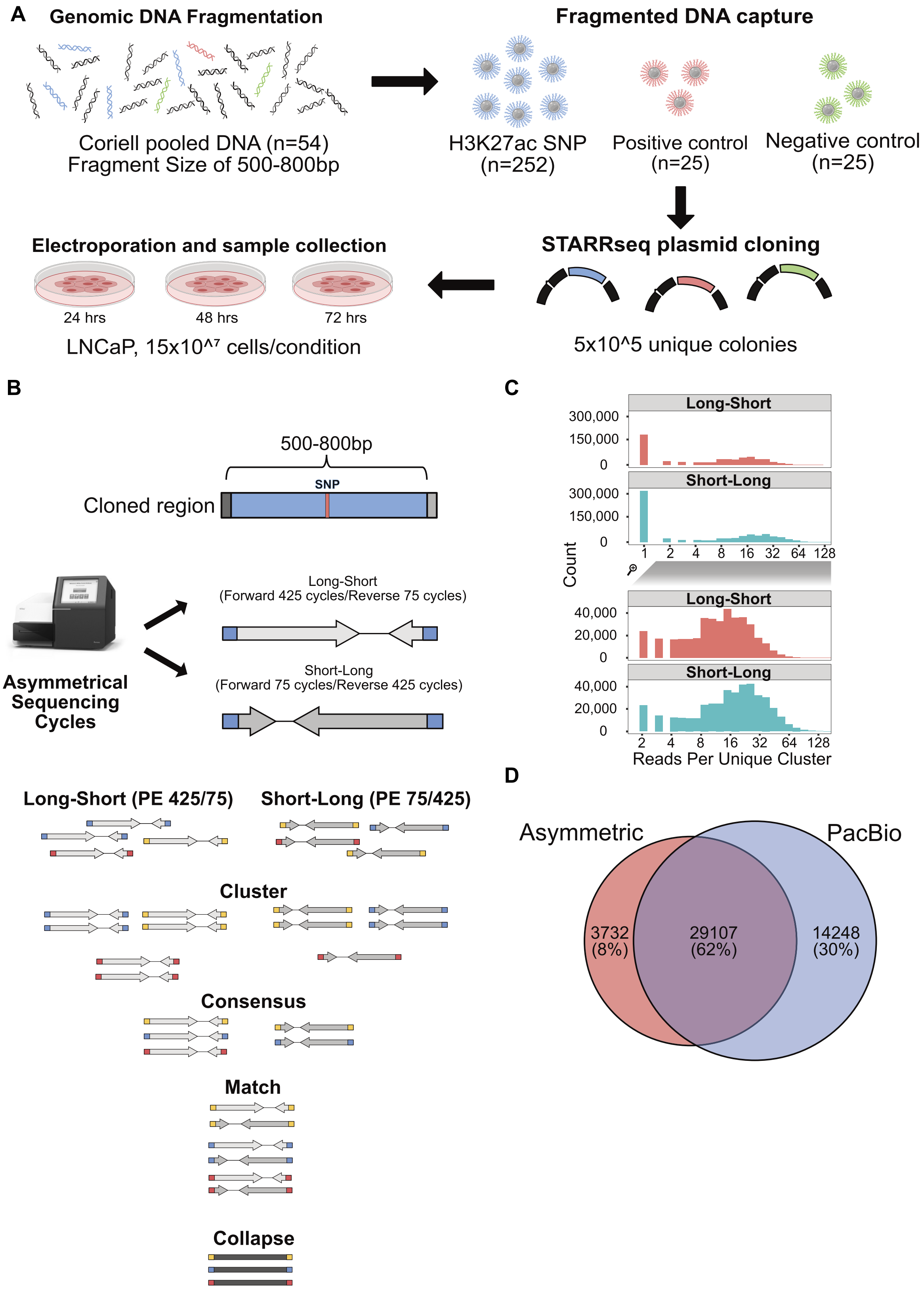

Figure 1. Schematic representation of enhancer fragment sequence reconstruction. (**A**) Experimental steps of snpSTARRseq method. (**B**) Computational analysis of asymmetrical reads. (**C**) Following the clustering step, the number of reads supporting each reconstructed fragment was investigated for each asymmetric run. The top histogram demonstrates the number of single reads supported fragments that are removed before the consensus step whereas the bottom histogram represents those reads included in the further analysis. (**D**) DNA input library was sequenced by PacBio CCS sequencing to validate the presence of reconstructed sequences

Comparison with *in vivo* methods

We obtained pre-processed allelic imbalance datasets from three different studies such as H3K27Ac ChIPseq (*n* = 200) (65), AR ChIPseq (*n* = 131) (65) or ATACseq (*n* = 26) (66). We matched all datasets (H3K27Ac and AR ChIPseq, ATACseq, and snpSTARRseq) using rsID therefore, we dropped 12 Coriell DNA SNPs which did not have the corresponding rsID in dbsnp150 common VCF. We shared corresponding AF for snpSTARRseq and *in vivo* methods as well as significance annotation (Supplementary Table S6).

PCa risk loci analysis

We extracted 147 index SNPs from (5) manuscript's Supplementary Table S7 (European SNPs only) and Supplementary Table S8. Using rsID (rsXXX), we extracted SNP positions from dbsnp150 common VCF file (hg19). Later intersected with 252 custom capture probe locations. We accepted all interactions within a 150 kb distance (Supplementary Table S7).

SNP and gene overlap analysis

We obtained HiChIP-H3k27Ac chromosome interaction data paired-end BED file (BEDPE) from previous work (67) and used it for annotating our SNPs. In addition to this, from the same study, we also obtained COLOC annotation based TCGA (prostate) and (68) FUSION annotation based on TWAS genes. The final SNP-gene association table is deposited in (Supplementary Table S8).

RESULTS

Design considerations and enhancements

To develop a methodology that functionally characterizes the impact of genetic variants on enhancer activity we adopted STARRseq, given both the ease of use and demonstrated experimental feasibility (47–49). We designed our experimental approach with the following features: (i) to scale efficiently allowing hundreds or thousands of variants to be tested, (ii) to maintain a large insert size to ensure a high enhancer signal, (iii) to have a high experimental signal to noise ratio by increasing the number of tested plasmids, (iv) to reduce systemic false-positives associated with the earlier STARR plasmids. With this framework, we optimized several key parameters with the following improvements. To reduce false-positives we utilized the second-generation STARRseq plasmid and ensured that the experimental models had minimal IFN-gamma response, which can strongly influence STARRseq signal (69). Next, we utilized a DNA-capture to enrich the number of plasmids per variant and increase the signal to noise. Given that insert size influences enhancer signal, testing a low number of plasmids per variant is extremely error-prone as you cannot separate the impact of insert size from the impact of the genetic variant (69,70). Further, the use of DNA capture also solves the fragment size limitations that are intrinsic to oligonucleotide synthesis (150–200 bp) (71,72). Increasing the library insert size leads to an increase in relative signal strength, which reduces false positives and negatives (50,70,73). We also incorporated a unique molecular identifier (UMI) for each cloned fragment to increase statistical strength and limit amplification artifacts. Finally, we used pooled genomic material from a healthy population (Coriell DNA library; (58), *n* = 54) to introduce genetic diversity and increase the number of SNPs tested.

SNP selection criteria

With the modified design we chose to test 252 PCa risk-associated SNPs, which are located on 184 non-overlapping segments that include 52% (68/130) of the previously published PCa risk-associated loci (Supplementary Table S1; Supplementary Table S7) (5). These regions were selected as they are potential enhancers that have both H3K27Ac and chromosomal looping to a gene promoter (56). In addition to these risk-associated SNPs we designed the DNA capture to also target 25 strong enhancers (57) (positive control) and 25 inactive regions (negative control) as experimental controls (Figure 1.A; Supplementary Table S2; Methods). Using this capture approach, we enriched randomly fragmented DNA (400–600 bp) that was hybridized with adapters each containing flanking 3bp UMI (59). This was PCR amplified and then cloned into a second-generation STARRseq plasmid. Due to the relatively large size of the insert (median = 543 bp), it was not feasible to use conventional short-read Illumina paired-end sequencing as variants in these larger inserts potentially will not be sequenced. Therefore, to ensure that the entire insert is sequenced, we modified a 500-cycle Illumina sequencing protocol and did two rounds of asymmetric sequencing to include both a 'long (425)-short (75)' and 'short (75)-long (425)' reads (Figure 1B, top) (48). As we sequenced the same plasmid library twice, both asymmetric runs covered opposite ends of the potential enhancer fragments. We hypothesized that the full enhancer sequence could be reconstructed by matching the long sequences of the same fragment. To process this complex dataset, we developed a novel computational pipeline that clusters, collapses, and matches asymmetrical reads (Figure 1B, bottom). We first clustered the enhancer fragments using the UMI (6 bp) and sequence context of the enhancers to group the reads supporting each enhancer fragment. Next, we collapsed the short and long mates of asymmetric runs and took the consensus sequence (median 36 reads/enhancer; SEM = 0.03) (Figure 1C). By matching the unique fragment barcode from the UMI (6bp) and randomly captured insert (18 bp), we clustered each pair of asymmetric reads to obtain the full enhancer sequence. In total, we reconstructed 32 620 fragments that were located in the capture regions. These asymmetric sequencing results were confirmed with PacBio circular consensus sequencing (CCS) long reads, with 90% (29106/32620) of reconstructed fragments found with PacBio (Figure 1D). In addition to the PCa risk-associated germline variants, captured enhancer fragments also included variants from common population SNPs (dbSNP150 common VCF) (9) as well as SNPs that were specific to the Coriell DNA population (58). Overall, our library was represented by a median of 67 unique plasmids per SNP containing 41 reference (REF) alleles and 26 alternative alleles (ALT) respectively. We failed to capture 50 SNPs that were either not present in the Coriell DNA library (3/50) or were not sufficiently represented to pass variant filtration (47/50) (Supplementary Figure S1A). To limit the impact of the potential enhancer position and size bias we modelled the impact of unique plasmids thresholds on experimental data (Materials and Methods). We found that > 15 unique plasmids for both REF and ALT SNP minimizes the Type-I error to 0.05. Therefore, we focused our analysis on the 308 SNPs that have > 30 unique plasmids with REF and ALT allele supporting inserts (Supplementary Figure S1B; Table S9) and also a minimum of 1000 mRNA reads, which is 100× higher than previous work (47). These SNPs include 102 (of 252) PCa risk-associated SNPs (PCa), 194 SNPs commonly found in the 1000 Genome Project (G5), and 12 Coriell DNA library-specific SNPs (Supplementary Table S10). When we compared the variant-allele frequency (VAF) with these SNPs we observed a higher but not significant VAF of PCa risk-associated SNPs ($P = 0.56$) and G5 SNPs ($P = 0.18$) compared to Coriell SNP (Supplementary Figure S1C).

Characterization of genetic variants that impact enhancer activity

To test the impact of PCa-associated SNPs on enhancer activity, we electroporated the snpSTARRseq library into LNCaP, an androgen receptor (AR)-dependent prostate cancer cell line, and harvested at 24, 48 and 72 h ($n = 3$ biological replicas). With this, we quantified the insert self-transcription, a surrogate for enhancer activity, at all high-confidence REF and ALT fragments. Following normalization to the input library, we observed a clear increase in self-transcription at known enhancers (positive control) but not negative controls with a high correlation of allelic enhancer activity across each experimental time point (Figure 2A; Supplementary Figure S2A). Principal component analysis (PCA) of all samples demonstrated that 48- and 72-h samples had a closer enhancer profile compared to the 24-h samples (Supplementary Figure S2B). We next investigated how each SNP affected the activity of the enhancers. Using a differential allelic enhancer activity test based on a negative binomial regression model (NBR) we identified 78 unique SNPs across the three-time points that showed bi-allelic activity at the nominal significance level (P-value < 0.05) (60, 63 and 43 at 24, 48 and 72 h, respectively) (Figure 2B; Supplementary Figure S2C; Table S11). Of these a total of 31 (39%) nominally significant SNPs passed multiple hypothesis testing correction (FDR < 0.05) with 23 being PCa disease-associated SNPs identified from GWAS. Interestingly, the majority of SNPs that affected enhancer activity (36/78) were PCa risk-associated SNPs. Supporting our PCA analysis, we observed a higher correlation in bi-allelic activity across all significant SNPs in the 48 and 72 h samples (Pearson = 0.94) (Supplementary Figure S2D). Focusing on the 48 h samples we observed a similar frequency of activating ($n = 31$) and repressive ($n = 32$) events. Of the specific PCa-associated SNPs we observed that rs11083046 (chr18:51781019) alternative C allele had a 30% increased enhancer activity (P-value = 0.00367) compared to the reference T allele (Figure 2C). In contrast, the enhancer activity of rs13215402 (chr6:153447550) decreased by 50% (P-value = 0.00349) when the reference allele G was substituted with the alternative allele A (Figure 2C). Supporting these plasmid-based results, the SNPs with bi-allelic enhancer activity commonly affected target gene expression. Using previously published enhancer-promoter interactions from H3K27Ac-HiChIP in LNCaP cells (56), 20 of 36 'hit' PCa risk-associated SNPs correlated with altered expression of the target gene (Supplementary Table S8). We found that these 20 SNPs also overlapped with PCa-specific expression quantitative trait loci (eQTL) identified by both tumor-adjacent normal samples ($n = 471$) and multi-tissue transcriptome-wide association study (TWAS) (n-tissue = 45, n-individual = 4448), using COLOC (74) and FUSION (75) tools, respectively (Supplementary Table S8). Interestingly, we found two significant SNPs (rs13265330; 2.2-fold, FDR = 9.25×10^{-6}, rs11782388; 1.47-fold FDR = 0.07) that were located ~10 kb downstream of *NKX3-1,* a gene involved in early prostate tumorigenesis (76). While not significant, we also found supporting evidence of two risk loci enhancers that loop to the PCa-associated genes *CTBP2* and *PCAT19*. Similar to published work the two SNPs (rs11672691 and rs887391) near PCAT19 caused increased enhancer activity in all of the time points (14). At the CTBP2 loci, rs4962416 and rs12769019 caused repression and activation respectively in accordance with previous work (19). Taken together, these results demonstrate that snpSTARRseq can identify those SNPs that alter enhancer activity. When combined with chromosomal confirmation data these results can provide a mechanism of action for non-coding disease-associated SNPs.

Comparison of snpSTARRseq to *in silico* methodologies

In silico based pathogenicity predictions are commonly used to stratify non-coding variants for functional characterization studies (44,77,78). To benchmark the performance of these methods to snpSTARRseq, we obtained the variant impact scores at the tested PCa disease-associated SNPs from ncER (23), CADD (24), DVAR (63), LINSIGHT (64) and deltaSVM (26). When comparing these *in silico* predictions to our experimental snpSTARRseq results we observed no positive correlation between the deleteriousness (CADD and DVAR), impact on chromatin accessibility (deltaSVM), or essentiality score (ncER, LINSIGHT) to the experimental (snpSTARRseq) bi-allelic effect ($\log_2$(ALT/REF)) or statistical significance (FDR) (Figure 3A, Supplementary Figure S3A, Supplementary Figure S3.B). In addition, we observed low negative correlation with ncER (Pearson = –0.12; P-value = 0.023) (Figure 3A). Further, when we separated SNPs into binary groups based on *in silico* annotations there were no significant changes in the enhancer activity (Supplementary Figure S3C). Our findings are consistent with the current literature (28,29) highlighting the challenges of *in silico* methods to accurately predict how variants predict enhancer activity (Supplementary Table S5).

snpSTARRseq correlates with clinical allelic-imbalance

We next compared our experimental snpSTARRseq results with *in vivo* allelic-imbalance from H3K27Ac ($n = 200$)

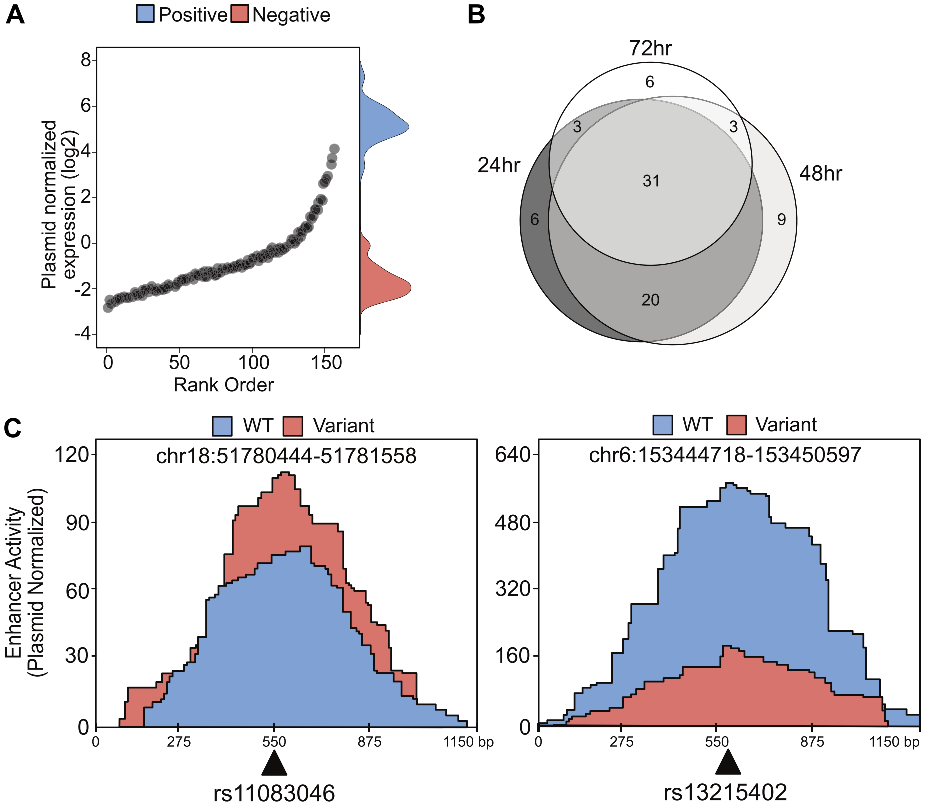

Figure 2. Characterisation of genetic variants that impact enhancer activity. (**A**) Quality control analysis of 48-hour time point demonstrates the SNP, positive control and negative control capture regions normalized count (mRNA/input DNA) distribution. Each black dot represents a single SNP capture region, whereas the density plot on the right-hand side represents the control capture regions (blue = positive control, red = negative control). (**B**) NBR model was used to determine SNPs with significant bi-allelic activity. As a result of the calculation, 78 unique SNPs were found among all time points. The number of overlapping SNPs from each time point is depicted by the Venn diagram. (**C**) Activating (rs11083046) and repressive (rs13215402) SNPs that cause bi-allelic enhancer activity were visualized.

and AR (n = 131) ChIPseq (65), as well as the assay for transposase-accessible chromatin using sequencing (ATAC-seq) (n = 26) in prostate tumors (66). This approach utilizes the endogenous heterogeneous allelic pool from clinical functional genomic studies to determine if there is an allelic preference for specific histone modifications, TFs, and chromosome accessibility. Of the 308 SNPs tested by snpSTARRseq, H3K27Ac ChIPseq had the highest coverage and captured 50% (148/308) of all SNPs while AR ChIPseq and ATACseq only captured 18% (57/308) of all SNPs tested. Based on our initial comparison without filtration, we observed low to moderate correlation between samples (Pearson; H3K27Ac = 0.17, AR = 0.44, ATAC-seq = 0.42) (Figure 3B, top; Supplementary Table S6). However, when non-significant SNPs were filtered out we observed a marked increase in correlation among all comparisons (Pearson; H3K27Ac = 0.31, AR = 0.75, ATAC-seq = 0.80) (Figure 3B, bottom; Supplementary Table S6). Based on 48 hr sample, 33% (26/78) of significant bi-allelic SNPs were supported by one, 24% (19/78) by two, and 2% (2/78) were supported by all of the allelic-imbalance *in vivo* methods. Those two SNPs that were captured by all methodologies (rs13215402 and rs11083046) demonstrated parallel repression (Figure 3C, left) or activation (Figure 3C, right) of enhancer activity and allelic imbalance. Overall, the *in vitro* snpSTARRseq results broadly correlate with *in vivo* allelic imbalance but not *in silico* predictions, highlighting the need for experimental validation of non-coding variants.

DISCUSSION

The impact of genetic variants on protein-coding amino acid sequences is generally well understood. However, the diversity of activity greatly limits large-scale testing, as each protein requires a specialized assay. Paradoxically, while non-coding variants are poorly understood, the common activity of enhancer CREs makes them extremely amenable to high-throughput screening. In a single experiment, hundreds to thousands of non-coding variants can be functionally tested. Further, when combined with chromosome conformation capture methods these massively multi-parallel assays offer a promising approach to systematically characterize the mechanism of disease-associated non-coding SNPs. However, previous adaptations were either designed to identify new variants (48,49) or were prone to false-positives (47). Therefore, we optimized snpSTARRseq to functionally test non-coding genetic variants.

In this work, we utilized a larger insert fragment (400–600 bp; ∼543 bp) to maximize TF and co-regulator interactions on CREs. This is the longest average fragment used in comparable MPRA methodologies (Supplementary Ta-

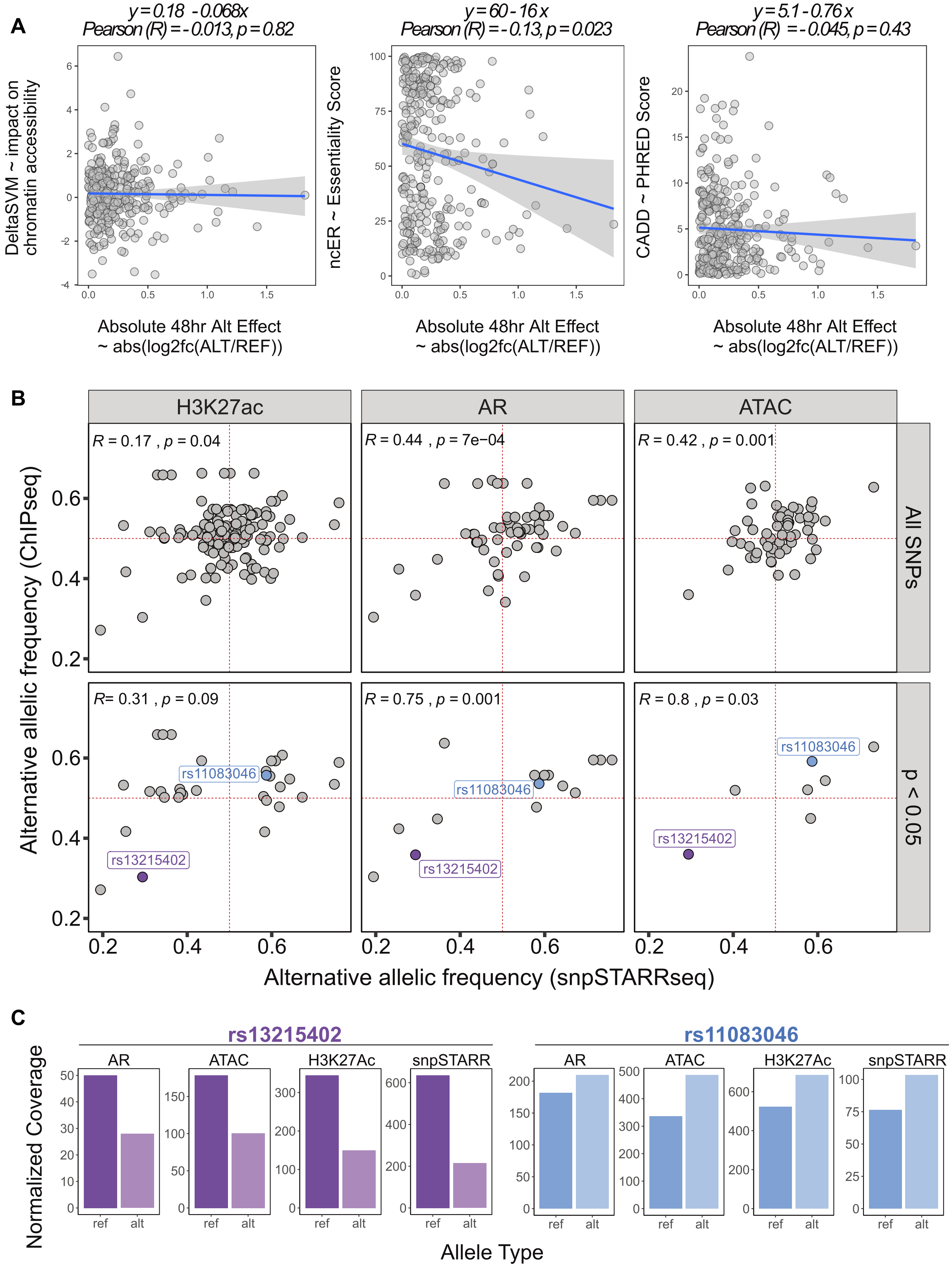

Figure 3. Comparison of snpSTARRseq to orthogonal methodologies. (**A**) snpSTARRseq allelic-effects (ALT/REF; Method; $\mathrm{Log_2}$FoldChange) was compared to DeltaSVM (Essentiality Score), ncER (Essentiality Score) and CADD (PHRED score) and no significant relation found. (**B**) snpSTARRseq allele frequency (AF) values were compared against in-vivo AF obtained by H3k27Ac and AR ChIPseq and ATACseq. The top row contains all SNP values without any filtration whereas the bottom row has only significant SNPs found by snpSTARRSeq (nominal p-value < 0.05). Two anecdotal examples of repressive (purple) and activating (blue) SNPs were colored to which were found by all 4 methodologies. (**C**) SNPs with activating and repressive effects found by previous *in vivo* work and snpSTARRseq were demonstrated. snpSTARRseq captures the bi-allelic effect of these SNPs accurately.

ble S12). While larger fragments (>600 bp) increase signal strength, they cannot be fully sequenced with standard short-read Illumina sequencing. To address this limitation, we modified existing Illumina technology to allow sequencing of DNA fragments lengths up to 850 bp using an asymmetric approach. Next, our design utilized a second-generation STARRseq plasmid, which provides a reduced background signal as compared to earlier reporter assays (69). Most importantly, by using a capture-based enrichment of specific genomic loci, we significantly increased the number of target fragments per genetic variant. This is critical as the unique number of plasmids can strongly influence Type-I error due to both position and insert size heterogeneity (Supplementary Figure S1B). To reduce these problems, we filtered all SNPs with less than 15 REF and ALT unique plasmids from our library. As a result, SNPs tested by our method had a minimum of 30 (REF + ALT) unique plasmids supported by >1000 mRNA reads. This is significantly higher than previous work (44,45,51,53), excluding one publication (45) (Supplementary Table S12; Figure S1A). This increased plasmid coverage dramatically reduces the overall noise caused by variable insert size (47). Overall, these modifications provide a robust platform for functional testing of GWAS hits.

With this approach, we targeted the bi-allelic enhancer activities of 252 PCa risk-associated SNPs from 184 non-overlapping regions that contain both a H3K27Ac mark and a chromatin loop to a gene promoter. Due to our very conservative threshold of REF + ALT plasmids (>30), we covered 35% (102/252) of PCa risk-associated variants. This could be improved with increased plasmid numbers during library generation. With this threshold, we observed that bi-allelic SNPs were highly concordant across multiple time points with minor differences between the earlier (24 h) and later (48 and 72 h) time points. Potentially, this may be due to cells reaching equilibrium in the later time points between snpSTARRseq mRNA transcription and degradation. From the PCa risk-associated SNPs tested we observed that 35% (36/102) had significantly altered enhancer activity. Many of these significant SNPs were located in 26 previously published PCa risk-associated loci (5). Moreover, 55% (20/36) of those with altered enhancer activity were associated with previously published eQTL (Supplementary Table S8). Consistent with the literature, we found supporting evidence for previously published SNPs that alter enhancer activity which impacts the expression of NKX3-1 (rs13215045 and rs11782388) (80–82), CTBP2 (rs4962416 and rs12769019), PCAT19 (rs11672691 and rs887391) and RGS17 (rs13215045, 6p25 RGS17 intron variant) genes (79).

We also compared the performance of snpSTARRseq to multiple *in silico* methodologies. While several of these approaches were designed to predict protein-coding mutations, we compared these results as we observed that *in silico* pathogenicity scores are commonly implemented for supporting enhancer variant annotation (80), GWAS prioritization (78,81,82) and driver gene calculation (83). This is particularly concerning as our experimental results did not strongly correlate with any *in silico* method. These differences could be potentially attributed to methodological differences (30) or a paucity of datasets that represent our experimental conditions. Specifically, there is an overall lack of representation of prostate models in public databases. For instance, ncER was trained on 38 databases that contain 9 targeted enhancer activity screens. However, none of these enhancer quantifications were based on PCa cell lines. Regardless of the cause, the low correlation between different *in silico* techniques (∼1%) suggests that there is a need for improved accuracy with these approaches (84). Recent studies have utilized semi-supervised methods to improve results by calibrating calculations and generating cell-type-specific predictions (85). For instance, MPRA datasets were incorporated to optimize feature weights for maximum tissue-specific separation (85,86). Contrasting these *in silico* methods, we observed a significant correlation between snpSTARRseq and clinical allelic imbalance of ATACseq, H3K27Ac and AR ChIPseq from tumor tissues. Overall, this supports the necessity of experimental validation of non-coding variants.

There are limitations to this methodology. Specifically, we missed 18% (47/252) of the targeted PCa risk-associated SNPs due to the low VAF of these variants in the DNA population. This can be easily overcome by increasing the number of plasmids during the cloning of the STARRseq library or genetic diversity of individuals in the DNA library (45). Further, as we are working with a pooled population, linkage disequilibrium (LD) makes GWAS traits to be harder to be finely mapped. Those events with low LD, conserved loci, and overlapping TF binding regions are more likely to be validated (87). To identify such complex events or multi-allelic events that are not present in the population, mutant enhancer sequences were previously generated either with saturation mutagenesis (29,88) or oligonucleotide synthesis (45,89). However, these methods lack control over the position of the mutations or lower enhancer activity due to the short fragment size. Shorter synthesized oligonucleotides can be used to characterize TF binding and co-regulator proteins on the target variants (90,91). For example, the binding affinity of TFs to oligonucleotides with reference and alternative alleles (26–40 bp) have been used to infer bi-allelic TF binding using SELEX (90,91). While there is little correlation between motifs and enhancer activity, these methods could be potentially incorporated to characterize the mechanism of variants identified from snpSTARRseq (41). Lastly, we did not characterize 36 INDELS as our snpSTARRseq computational pipeline exclusively focused on SNPs due to the ambiguity in INDELs calling (92).

Functionally characterization of non-coding variants is an emerging field, and we are now just beginning to learn the strengths and limitations of the various methodologies. For instance, in this work, 55% (20/36) of our PCa risk-associated variants were identified as eQTL. Given the availability of public databases (GTEx (93), eQTLgen (94)), these orthogonal results are important to validate our snpSTARRseq findings. However, eQTL-based studies also have limitations. For instance, eQTL studies only measure steady-state transcript levels. Consequently, the literature is now reporting that eQTLs explain only 11% of the heritability for an average trait (95,96) or up to 25% when transcription is profiled in disease-relevant tissue (97). Moreover, steady-state eQTLs are depleted near genes that are likely to contribute to complex phenotypes, including transcription factors, developmental genes, and highly conserved or essential genes (98). Finally, recent work demonstrated GWAS and eQTL studies are systematically biased toward different types of variants (99). Because of these limitations, there is a need for robust experimental approaches that can complement eQTL studies. In addition, the *in vitro* STARRseq can measure the activity of the enhancers generated by functional perturbations that can delineate complex regulatory mechanisms which is not possible in eQTL tissue-based approaches.

Herein, we developed snpSTARRSeq to improve the sensitivity and accuracy of large-scale non-coding enhancer assays. By increasing fragment length and reducing signal to noise, this approach can precisely identify functional variants. Potentially, the insert sizes could be further increased with long-read PacBio CCS sequencing (100,101). While not the goal of this work, genetic perturbations of snpSTARRseq systems could be used to identify how specific transcription factor binding is altered by SNPs. Further while focused on germline genetic variants, this same approach can be adopted to study somatic variants. Overall, snpSTARRseq offers an integrated experimental and computational approach to test the bi-allelic activity of hundreds to thousands of genetic variants in a single experiment.

DATA AVAILABILITY

We deposited the snpSTARRSeq computational framework at GitHub (https://github.com/mortunco/snp-starrseq). All visualization parameters and scripts could be found in publication-figures.rmd file in our repository. All datasets generated during this study along with other processed files are available at SRA under accession PRJNA791664.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Baraa Orabi and Fatih Karaoğlanoğlu for their contribution on the implementation of Calib and Snakemake, and Dr Yen-yi Lin for his useful discussions during manuscript preparation.

FUNDING

TUBITAK 1001 [119Z279]; Turkish Science Academy's Young Scientist Award Program (BAGEP). Funding for open access charge: Koç University School of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Pairo-Castineira,E., Clohisey,S., Klaric,L., Bretherick,A.D., Rawlik,K., Pasko,D., Walker,S., Parkinson,N., Fourman,M.H., Russell,C.D. *et al.* (2021) Genetic mechanisms of critical illness in COVID-19. *Nature*, **591**, 92–98.
2. Freedman,M.L., Monteiro,A.N.A., Gayther,S.A., Coetzee,G.A., Risch,A., Plass,C., Casey,G., De Biasi,M., Carlson,C., Duggan,D. *et al.* (2011) Principles for the post-GWAS functional characterization of cancer risk loci. *Nat. Genet.*, **43**, 513–518.
3. Schumacher,F.R., Berndt,S.I., Siddiq,A., Jacobs,K.B., Wang,Z., Lindstrom,S., Stevens,V.L., Chen,C., Mondul,A.M., Travis,R.C. *et al.* (2011) Genome-wide association study identifies new prostate cancer susceptibility loci. *Hum. Mol. Genet.*, **20**, 3867–3875.
4. Al Olama,A.A., Kote-Jarai,Z., Berndt,S.I., Conti,D.V., Schumacher,F., Han,Y., Benlloch,S., Hazelett,D.J., Wang,Z., Saunders,E. *et al.* (2014) A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nat. Genet.*, **46**, 1103–1109.
5. Schumacher,F.R., Al Olama,A.A., Berndt,S.I., Benlloch,S., Ahmed,M., Saunders,E.J., Dadaev,T., Leongamornlert,D., Anokian,E., Cieza-Borrella,C. *et al.* (2018) Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.*, **50**, 928–936.
6. Hazelett,D.J., Rhie,S.K., Gaddis,M., Yan,C., Lakeland,D.L., Coetzee,S.G., consortium,Ellipse/GAME-ON, consortium,Practical, Henderson,B.E., Noushmehr,H. *et al.* (2014) Comprehensive functional annotation of 77 prostate cancer risk loci. *PLoS Genet.*, **10**, e1004102.
7. Pomerantz,M.M. and Freedman,M.L. (2011) The genetics of cancer risk. *Cancer J.*, **17**, 416–422.
8. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 9362–9367.
9. Smigielski,E.M., Sirotkin,K., Ward,M. and Sherry,S.T. (2000) dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.*, **28**, 352–355.
10. Qian,Y., Zhang,L., Cai,M., Li,H., Xu,H., Yang,H., Zhao,Z., Rhie,S.K., Farnham,P.J., Shi,J. *et al.* (2019) The prostate cancer risk variant rs55958994 regulates multiple gene expression through extreme long-range chromatin interaction to control tumor progression. *Sci. Adv.*, **5**, eaaw6710.
11. Cong,Z., Li,Q., Yang,Y., Guo,X., Cui,L. and You,T. (2019) The SNP of rs6854845 suppresses transcription via the DNA looping structure alteration of super-enhancer in colon cells. *Biochem. Biophys. Res. Commun.*, **514**, 734–741.
12. Wasserman,N.F., Aneas,I. and Nobrega,M.A. (2010) An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer. *Genome Res.*, **20**, 1191–1197.
13. Kandaswamy,R., Sava,G.P., Speedy,H.E., Beà,S., Martín-Subero,J.I., Studd,J.B., Migliorini,G., Law,P.J., Puente,X.S., Martín-García,D. *et al.* (2016) Genetic predisposition to chronic lymphocytic leukemia is mediated by a BMF super-enhancer polymorphism. *Cell Rep.*, **16**, 2061–2067.
14. Hua,J.T., Ahmed,M., Guo,H. and Zhang,Y. (2018) Risk SNP-mediated promoter-enhancer switching drives prostate cancer through lncRNA PCAT19. *Cell*, **174**, 564–575.
15. Panigrahi,A. and O'Malley,B.W. (2021) Mechanisms of enhancer action: the known and the unknown. *Genome Biol.*, **22**, 108.
16. Morova,T., McNeill,D.R., Lallous,N., Gönen,M., Dalal,K., Wilson,D.M. 3rd, Gürsoy,A., Keskin,Ö. and Lack,N.A. (2020) Androgen receptor-binding sites are highly mutated in prostate cancer. *Nat. Commun.*, **11**, 832.
17. Zhou,S., Hawley,J.R., Soares,F., Grillo,G., Teng,M., Tonekaboni,S.A.M., Hua,J.T., Kron,K.J., Mazrooei,P., Ahmed,M. *et al.* (2020) Noncoding mutations target cis-regulatory elements of the FOXA1 plexus in prostate cancer. *Nat. Commun.*, **11**, 441.
18. Pomerantz,M.M., Ahmadiyeh,N., Jia,L., Herman,P., Verzi,M.P., Doddapaneni,H., Beckwith,C.A., Chan,J.A., Hills,A., Davis,M. *et al.* (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat. Genet.*, **41**, 882–884.
19. Takayama,K.-I., Suzuki,T., Fujimura,T., Urano,T., Takahashi,S., Homma,Y. and Inoue,S. (2014) CtBP2 modulates the androgen receptor to promote prostate cancer progression. *Cancer Res.*, **74**, 6542–6553.
20. Gao,P., Xia,J.-H., Sipeky,C., Dong,X.-M., Zhang,Q., Yang,Y., Zhang,P., Cruz,S.P., Zhang,K., Zhu,J. *et al.* (2018) Biology and clinical implications of the 19q13 aggressive prostate cancer susceptibility locus. *Cell*, **174**, 576–589.

21. Spisák,S., Lawrenson,K., Fu,Y., Csabai,I., Cottman,R.T., Seo,J.-H., Haiman,C., Han,Y., Lenci,R., Li,Q. *et al.* (2015) CAUSEL: an epigenome- and genome-editing pipeline for establishing function of noncoding GWAS variants. *Nat. Med.*, **21**, 1357–1363.
22. Guo,Y.A., Chang,M.M. and Skanderup,A.J. (2020) MutSpot: detection of non-coding mutation hotspots in cancer genomes. *NPJ Genom Med*, **5**, 26.
23. Wells,A., Heckerman,D., Torkamani,A., Yin,L., Sebat,J., Ren,B., Telenti,A. and di Iulio,J. (2019) Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nat. Commun.*, **10**, 5241.
24. Kircher,M., Witten,D.M., Jain,P., O'Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
25. Abramov,S., Boytsov,A., Bykova,D., Penzar,D.D., Yevshin,I., Kolmykov,S.K., Fridman,M.V., Favorov,A.V., Vorontsov,I.E., Baulin,E. *et al.* (2021) Landscape of allele-specific transcription factor binding in the human genome. *Nat. Commun.*, **12**, 2751.
26. Lee,D., Gorkin,D.U., Baker,M., Strober,B.J., Asoni,A.L., McCallion,A.S. and Beer,M.A. (2015) A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.*, **47**, 955–961.
27. Drubay,D., Gautheret,D. and Michiels,S. (2018) A benchmark study of scoring methods for non-coding mutations. *Bioinformatics*, **34**, 1635–1641.
28. Liu,L., Sanderford,M.D., Patel,R., Chandrashekar,P., Gibson,G. and Kumar,S. (2019) Biological relevance of computationally predicted pathogenicity of noncoding variants. *Nat. Commun.*, **10**, 330.
29. Kircher,M., Xiong,C., Martin,B., Schubach,M., Inoue,F., Bell,R.J.A., Costello,J.F., Shendure,J. and Ahituv,N. (2019) Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.*, **10**, 3583.
30. Wang,Z., Zhao,G., Li,B., Fang,Z., Chen,Q., Wang,X., Luo,T., Wang,Y., Zhou,Q., Li,K. *et al.* (2022) Performance comparison of computational methods for the prediction of the function and pathogenicity of non-coding variants. *Genomics Proteomics Bioinformatics*, **7**, S1672-0229(22)00016-X.
31. Kasowski,M., Kyriazopoulou-Panagiotopoulou,S., Grubert,F., Zaugg,J.B., Kundaje,A., Liu,Y., Boyle,A.P., Zhang,Q.C., Zakharia,F., Spacek,D.V. *et al.* (2013) Extensive variation in chromatin states across humans. *Science*, **342**, 750–752.
32. McVicker,G., van de Geijn,B., Degner,J.F., Cain,C.E., Banovich,N.E., Raj,A., Lewellen,N., Myrthil,M., Gilad,Y. and Pritchard,J.K. (2013) Identification of genetic variants that affect histone modifications in human cells. *Science*, **342**, 747–749.
33. Cheng,Z., Vermeulen,M., Rollins-Green,M., DeVeale,B. and Babak,T. (2021) Cis-regulatory mutations with driver hallmarks in major cancers. *Iscience*, **24**, 102144.
34. Schizophrenia Working Group of the Psychiatric Genomics Consortium, Gusev,A., Mancuso,N., Won,H., Kousi,M., Finucane,H.K., Reshef,Y., Song,L., Safi,A., McCarroll,S. *et al.* (2018) Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat. Genet.*, **50**, 538–548.
35. Project Consortium,ENCODE, Moore,J.E., Purcaro,M.J., Pratt,H.E., Epstein,C.B., Shoresh,N., Adrian,J., Kawli,T., Davis,C.A., Dobin,A. *et al.* (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, **583**, 699–710.
36. Melnikov,A., Murugan,A., Zhang,X., Tesileanu,T., Wang,L., Rogov,P., Feizi,S., Gnirke,A., Callan,C.G.,Jr, Kinney,J.B. *et al.* (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.*, **30**, 271–277.
37. Arnold,C.D., Gerlach,D., Stelzer,C., Boryń,Ł.M., Rath,M. and Stark,A. (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, **339**, 1074–1077.
38. Zacher,B., Michel,M., Schwalb,B., Cramer,P., Tresch,A. and Gagneur,J. (2017) Accurate promoter and enhancer identification in 127 ENCODE and roadmap epigenomics cell types and tissues by GenoSTAN. *PLoS One*, **12**, e0169249.
39. Zhang,T., Zhang,Z., Dong,Q., Xiong,J. and Zhu,B. (2020) Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome Biol.*, **21**, 45.
40. Inoue,F., Kircher,M., Martin,B., Cooper,G.M., Witten,D.M., McManus,M.T., Ahituv,N. and Shendure,J. (2017) A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.*, **27**, 38–52.
41. Huang,C.-C.F., Lingadahalli,S., Morova,T., Ozturan,D., Hu,E., Yu,I.P.L., Linder,S., Hoogstraat,M., Stelloo,S., Sar,F. *et al.* (2021) Functional mapping of androgen receptor enhancer activity. *Genome Biol.*, **22**, 149.
42. Patwardhan,R.P., Lee,C., Litvin,O., Young,D.L., Pe'er,D. and Shendure,J. (2009) High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.*, **27**, 1173–1175.
43. Vockley,C.M., Guo,C., Majoros,W.H., Nodzenski,M., Scholtens,D.M., Hayes,M.G., Lowe,W.L. Jr and Reddy,T.E. (2015) Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome Res.*, **25**, 1206–1214.
44. Ulirsch,J.C., Nandakumar,S.K., Wang,L., Giani,F.C., Zhang,X., Rogov,P., Melnikov,A., McDonel,P., Do,R., Mikkelsen,T.S. *et al.* (2016) Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell*, **165**, 1530–1545.
45. Tewhey,R., Kotliar,D., Park,D.S., Liu,B., Winnicki,S., Reilly,S.K., Andersen,K.G., Mikkelsen,T.S., Lander,E.S., Schaffner,S.F. *et al.* (2016) Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*, **165**, 1519–1529.
46. Ernst,J., Melnikov,A., Zhang,X., Wang,L., Rogov,P., Mikkelsen,T.S. and Kellis,M. (2016) Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat. Biotechnol.*, **34**, 1180–1190.
47. Liu,S., Liu,Y., Zhang,Q., Wu,J., Liang,J., Yu,S., Wei,G.-H., White,K.P. and Wang,X. (2017) Systematic identification of regulatory variants associated with cancer risk. *Genome Biol.*, **18**, 194.
48. Wang,X., He,L., Goggin,S.M., Saadat,A., Wang,L., Sinnott-Armstrong,N., Claussnitzer,M. and Kellis,M. (2018) High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat. Commun.*, **9**, 5380.
49. Zhang,P., Xia,J.-H., Zhu,J., Gao,P., Tian,Y.-J., Du,M., Guo,Y.-C., Suleman,S., Zhang,Q., Kohli,M. *et al.* (2018) High-throughput screening of prostate cancer risk loci by single nucleotide polymorphisms sequencing. *Nat. Commun.*, **9**, 2022.
50. Klein,J.C., Keith,A., Rice,S.J., Shepherd,C., Agarwal,V., Loughlin,J. and Shendure,J. (2019) Functional testing of thousands of osteoarthritis-associated variants for regulatory activity. *Nat. Commun.*, **10**, 2434.
51. Choi,J., Zhang,T., Vu,A., Ablain,J., Makowski,M.M., Colli,L.M., Xu,M., Hennessey,R.C., Yin,J., Rothschild,H. *et al.* (2020) Massively parallel reporter assays of melanoma risk variants identify MX2 as a gene promoting melanoma. *Nat. Commun.*, **11**, 2718.
52. Abell,N.S., DeGorter,M.K., Gloudemans,M.J., Greenwald,E., Smith,K.S., He,Z. and Montgomery,S.B. (2022) Multiple causal variants underlie genetic associations in humans. *Science*, **375**, 1247–1254.
53. Weiss,C.V., Harshman,L., Inoue,F., Fraser,H.B., Petrov,D.A., Ahituv,N. and Gokhman,D. (2021) The cis-regulatory effects of modern human-specific variants. *Elife*, **10**, e63713.
54. Yáñez-Cuna,J.O., Kvon,E.Z. and Stark,A. (2013) Deciphering the transcriptional cis-regulatory code. *Trends Genet.*, **29**, 11–22.
55. Klein,J.C., Agarwal,V., Inoue,F., Keith,A., Martin,B., Kircher,M., Ahituv,N. and Shendure,J. (2020) A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat. Methods*, **17**, 1083–1091.
56. Giambartolomei,C., Seo,J.-H., Schwarz,T., Freund,M.K., Johnson,R.D., Spisak,S., Baca,S.C., Gusev,A., Mancuso,N., Pasaniuc,B. *et al.* (2021) H3K27ac HiChIP in prostate cell lines identifies risk genes for prostate cancer susceptibility. *Am. J. Hum. Genet.*, **108**, 2284–2300.
57. Liu,Y., Yu,S., Dhiman,V.K., Brunetti,T., Eckart,H. and White,K.P. (2017) Functional assessment of human enhancer activities using whole-genome STARR-sequencing. *Genome Biol.*, **18**, 219.

58. Carpen,J.D., Archer,S.N., Skene,D.J., Smits,M. and von Schantz,M. (2005) A single-nucleotide polymorphism in the 5'-untranslated region of the hPER2 gene is associated with diurnal preference. *J. Sleep Res.*, **14**, 293–297.
59. MacConaill,L.E., Burns,R.T., Nag,A., Coleman,H.A., Slevin,M.K., Giorda,K., Light,M., Lai,K., Jarosz,M., McNeill,M.S. *et al.* (2018) Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics*, **19**, 30.
60. Orabi,B., Erhan,E., McConeghy,B., Volik,S.V., Le Bihan,S., Bell,R., Collins,C.C., Chauve,C. and Hach,F. (2019) Alignment-free clustering of UMI tagged DNA molecules. *Bioinformatics*, **35**, 1829–1836.
61. Bushnell,B., Rood,J. and Singer,E. (2017) BBMerge – accurate paired shotgun read merging via overlap. *PLoS One*, **12**, e0185056.
62. Venables,W.N. and Ripley,B.D. (2002) In: *Modern applied statistics with S*. Springer, NY.
63. Yang,H., Chen,R., Wang,Q., Wei,Q., Ji,Y., Zheng,G., Zhong,X., Cox,N.J. and Li,B. (2019) De novo pattern discovery enables robust assessment of functional consequences of non-coding variants. *Bioinformatics*, **35**, 1453–1460.
64. Huang,Y.-F., Gulko,B. and Siepel,A. (2017) Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.*, **49**, 618–624.
65. Baca,S.C., Singler,C., Zacharia,S., Seo,J.-H., Morova,T., Hach,F., Ding,Y., Schwarz,T., Huang,C.-C.F., Anderson,J. *et al.* (2022) Genetic determinants of chromatin reveal prostate cancer risk mediated by context-dependent gene regulation. *Nat. Genet.*, **54**, 1364–1375.
66. Corces,M.R., Granja,J.M., Shams,S., Louie,B.H., Seoane,J.A., Zhou,W., Silva,T.C., Groeneveld,C., Wong,C.K., Cho,S.W. *et al.* (2018) The chromatin accessibility landscape of primary human cancers. *Science*, **362**, eaav1898.
67. Giambartolomei,C., Seo,J.-H., Schwarz,T., Freund,M.K., Johnson,R.D., Spisak,S., Baca,S.C., Gusev,A., Mancuso,N., Pasaniuc,B. *et al.* (2021) H3k27ac-HiChIP in prostate cell lines identifies risk genes for prostate cancer susceptibility. *Am. J. Hum. Genet.*, **108**, 2284–2300.
68. Thibodeau,S.N., French,A.J., McDonnell,S.K., Cheville,J., Middha,S., Tillmans,L., Riska,S., Baheti,S., Larson,M.C., Fogarty,Z. *et al.* (2015) Identification of candidate genes for prostate cancer-risk SNPs utilizing a normal prostate tissue eQTL data set. *Nat. Commun.*, **6**, 8653.
69. Muerdter,F., Boryń,Ł.M., Woodfin,A.R., Neumayr,C., Rath,M., Zabidi,M.A., Pagani,M., Haberle,V., Kazmar,T., Catarino,R.R. *et al.* (2018) Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods*, **15**, 141–149.
70. Lee,D., Shi,M., Moran,J., Wall,M., Zhang,J., Liu,J., Fitzgerald,D., Kyono,Y., Ma,L., White,K.P. *et al.* (2020) STARRPeaker: uniform processing and accurate identification of STARR-seq active regions. *Genome Biol.*, **21**, 298.
71. Song,L.-F., Deng,Z.-H., Gong,Z.-Y., Li,L.-L. and Li,B.-Z. (2021) Large-scale de novo oligonucleotide synthesis for whole-genome synthesis and data storage: challenges and opportunities. *Front. Bioeng. Biotechnol.*, **9**, 689797.
72. Palluk,S., Arlow,D.H., de Rond,T., Barthel,S., Kang,J.S., Bector,R., Baghdassarian,H.M., Truong,A.N., Kim,P.W., Singh,A.K. *et al.* (2018) De novo DNA synthesis using polymerase-nucleotide conjugates. *Nat. Biotechnol.*, **36**, 645–650.
73. Vockley,C.M., D'Ippolito,A.M., McDowell,I.C., Majoros,W.H., Safi,A., Song,L., Crawford,G.E. and Reddy,T.E. (2016) Direct GR binding sites potentiate clusters of TF binding across the human genome. *Cell*, **166**, 1269–1281.
74. Giambartolomei,C., Zhenli Liu,J., Zhang,W., Hauberg,M., Shi,H., Boocock,J., Pickrell,J., Jaffe,A.E., Consortium,CommonMind, Pasaniuc,B. *et al.* (2018) A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics*, **34**, 2538–2545.
75. Gusev,A., Ko,A., Shi,H., Bhatia,G., Chung,W., Penninx,B.W.J.H., Jansen,R., de Geus,E.J.C., Boomsma,D.I., Wright,F.A. *et al.* (2016) Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.*, **48**, 245–252.
76. Song,H., Zhang,B., Watson,M.A., Humphrey,P.A., Lim,H. and Milbrandt,J. (2009) Loss of Nkx3.1 leads to the activation of discrete downstream target genes during prostate tumorigenesis. *Oncogene*, **28**, 3307–3319.
77. Chalmers,Z.R., Connelly,C.F., Fabrizio,D., Gay,L., Ali,S.M., Ennis,R., Schrock,A., Campbell,B., Shlien,A., Chmielecki,J. *et al.* (2017) Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med.*, **9**, 34.
78. Jang,Y.J., LaBella,A.L., Feeney,T.P., Braverman,N., Tuchman,M., Morizono,H., Ah Mew,N. and Caldovic,L. (2018) Disease-causing mutations in the promoter and enhancer of the ornithine transcarbamylase gene. *Hum. Mutat.*, **39**, 527–536.
79. Han,Y., Hazelett,D.J., Wiklund,F., Schumacher,F.R., Stram,D.O., Berndt,S.I., Wang,Z., Rand,K.A., Hoover,R.N., Machiela,M.J. *et al.* (2015) Integration of multiethnic fine-mapping and genomic annotation to prioritize candidate functional SNPs at prostate cancer susceptibility regions. *Hum. Mol. Genet.*, **24**, 5603–5618.
80. Claringbould,A. and Zaugg,J.B. (2021) Enhancers in disease: molecular basis and emerging treatment strategies. *Trends Mol. Med.*, **27**, 1060–1073.
81. Lee,J., Suh,Y., Jeong,H., Kim,G.-H., Byeon,S.H., Han,J. and Lim,H.T. (2021) Aberrant expression of PAX6 gene associated with classical aniridia: identification and functional characterization of novel noncoding mutations. *J. Hum. Genet.*, **66**, 333–338.
82. Watanabe,K., Taskesen,E., van Bochoven,A. and Posthuma,D. (2017) Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.*, **8**, 1826.
83. Shuai,S., PCAWG Drivers and Functional Interpretation Working Group and PCAWG ConsortiumPCAWG Drivers and Functional Interpretation Working Group, Gallinger,S. and Stein,L.PCAWG Consortium (2020) Combined burden and functional impact tests for cancer driver discovery using DriverPower. *Nat. Commun.*, **11**, 734.
84. Li,J., Drubay,D., Michiels,S. and Gautheret,D. (2015) Mining the coding and non-coding genome for cancer drivers. *Cancer Lett.*, **369**, 307–315.
85. He,Z., Liu,L., Wang,K. and Ionita-Laza,I. (2018) A semi-supervised approach for predicting cell-type specific functional consequences of non-coding variation using MPRAs. *Nat. Commun.*, **9**, 5199.
86. Dong,S. and Boyle,A.P. (2022) Prioritization of regulatory variants with tissue-specific function in the non-coding regions of human genome. *Nucleic Acids Res.,***50**, e6.
87. Gorlova,O.Y., Xiao,X., Tsavachidis,S., Amos,C.I. and Gorlov,I.P. (2022) SNP characteristics and validation success in genome wide association studies. *Hum. Genet.*, **141**, 229–238.
88. Kvon,E.Z., Zhu,Y., Kelman,G., Novak,C.S., Plajzer-Frick,I., Kato,M., Garvin,T.H., Pham,Q., Harrington,A.N., Hunter,R.D. *et al.* (2020) Comprehensive in vivo interrogation reveals phenotypic impact of human enhancer variants. *Cell*, **180**, 1262–1271.
89. Schöne,S., Bothe,M., Einfeldt,E., Borschiwer,M., Benner,P., Vingron,M., Thomas-Chollier,M. and Meijsing,S.H. (2018) Synthetic STARR-seq reveals how DNA shape and sequence modulate transcriptional output and noise. *PLoS Genet.*, **14**, e1007793.
90. Yan,J., Qiu,Y., Ribeiro Dos Santos,A.M., Yin,Y., Li,Y.E., Vinckier,N., Nariai,N., Benaglio,P., Raman,A., Li,X. *et al.* (2021) Systematic analysis of binding of transcription factors to noncoding variants. *Nature*, **591**, 147–151.
91. Bray,D., Hook,H., Zhao,R., Keenan,J.L., Penvose,A., Osayame,Y., Mohaghegh,N., Chen,X., Parameswaran,S., Kottyan,L.C. *et al.* (2022) CASCADE: high-throughput characterization of regulatory complex binding altered by non-coding variants. *Cell Genom*, **2**, 100098.
92. Wang,N., Lysenkov,V., Orte,K., Kairisto,V., Aakko,J., Khan,S. and Elo,L.L. (2022) Tool evaluation for the detection of variably sized indels from next generation whole genome and targeted sequencing data. *PLoS Comput. Biol.*, **18**, e1009269.
93. The GTEx Consortium (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, **369**, 1318–1330.
94. Võsa,U., Claringbould,A., Westra,H.-J., Bonder,M.J., Deelen,P., Zeng,B., Kirsten,H., Saha,A., Kreuzhuber,R., Yazar,S. *et al.* (2021) Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.*, **53**, 1300–1310.

95. Umans,B.D., Battle,A. and Gilad,Y. (2021) Where are the disease-associated eQTLs?*Trends Genet.*, **37**, 109–124.
96. Yao,D.W., O'Connor,L.J., Price,A.L. and Gusev,A. (2020) Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.*, **52**, 626–633.
97. Chun,S., Casparino,A., Patsopoulos,N.A., Croteau-Chonka,D.C., Raby,B.A., De Jager,P.L., Sunyaev,S.R. and Cotsapas,C. (2017) Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.*, **49**, 600–605.
98. Wang,X. and Goldstein,D.B. (2020) Enhancer domains predict gene pathogenicity and inform gene discovery in complex disease. *Am. J. Hum. Genet.*, **106**, 215–233.
99. Mostafavi,H., Spence,J.P., Naqvi,S. and Pritchard,J.K. (2022) Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery. bioRxiv doi: https://doi.org/10.1101/2022.05.07.491045, 08 May 2022, preprint: not peer reviewed.
100. Ardui,S., Ameur,A., Vermeesch,J.R. and Hestand,M.S. (2018) Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res.*, **46**, 2159–2168.
101. Wenger,A.M., Peluso,P., Rowell,W.J., Chang,P.-C., Hall,R.J., Concepcion,G.T., Ebler,J., Fungtammasan,A., Kolesnikov,A., Olson,N.D. *et al.* (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.*, **37**, 1155–1162.