Supplementary Materials for

The chromatin accessibility landscape of primary human cancers

M. Ryan Corces*, Jeffrey M. Granja*, Shadi Shams, Bryan H. Louie, Jose A. Seoane, Wanding Zhou, Tiago C. Silva, Clarice Groeneveld, Christopher K. Wong, Seung Woo Cho, Ansuman T. Satpathy, Maxwell R. Mumbach, Katherine A. Hoadley, A. Gordon Robertson, Nathan C. Sheffield, Ina Felau, Mauro A. A. Castro, Benjamin P. Berman, Louis M. Staudt, Jean C. Zenklusen, Peter W. Laird, Christina Curtis, The Cancer Genome Atlas Analysis Network, William J. Greenleaf†, Howard Y. Chang†

*These authors contributed equally to this work.
†Corresponding author. Email: howchang@stanford.edu (H.Y.C.); wjg@stanford.edu (W.J.G.)

Published 26 October 2018, *Science* **362**, eaav1898 (2018)
DOI: 10.1126/science.aav1898

This PDF file includes:

Other Supporting Online Material for this manuscript includes the following:
(available at www.sciencemag.org/content/362/6413/eaav1898/suppl/DC1)

Data S1 to S10 (Excel files)

The Cancer Genome Atlas Analysis Network Collaborators List

Rehan Akbani[19], Christopher C. Benz[20], Evan A. Boyle[21], Bradley M. Broom[19], Andrew D. Cherniack[22,23], Brian Craft[24], John A. Demchok[25], Ashley S. Doane[26], Olivier Elemento[26], Martin L. Ferguson[25], Mary J. Goldman[24], D. Neil Hayes[27], Jing He[28], Toshinori Hinoue[29], Marcin Imielinski[26], Steven J.M. Jones[30], Anab Kemal[25], Theo A. Knijnenburg[31], Anil Korkut[19], De-Chen Lin[32], Yuexin Liu[19], Michael K.A. Mensah[25], Gordon B. Mills[33], Vincent P. Reuter[34], Andre Schultz[19], Hui Shen[29], Jason P. Smith[34], Roy Tarnuzzer[25], Sheyla Trefflich[35], Zhining Wang[25], John N. Weinstein[19], Lindsay C. Westlake[22,23], Jin Xu[28], Liming Yang[25], Christina Yau[20,36], Yang Zhao[28], Jingchun Zhu[24]

[19]Department of Bioinformatics and Computational Biology, MD Anderson Cancer Center, Houston, TX 77030, USA.
[20]Buck Institute for Research on Aging, Novato, CA 94945, USA.
[21]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA.
[22]Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA.
[23]Department of Medical Oncology, Dana Farber Cancer Institute, Boston, MA 02215, USA
[24]Genomics Institute, University of California, Santa Cruz, CA 95064, USA.
[25]National Cancer Institute, Bethesda, MD 20892, USA.
[26]Caryl and Israel Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY 10021, USA.
[27]Department of Genetics and Genomics, University of Tennessee Health Science Center, Memphis, TN 38117, USA.
[28]Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA 94305, USA.
[29]Van Andel Research Institute, Grand Rapids, MI 49503, USA.
[30]Canada's Michael Smith Genome Sciences Center, BC Cancer Agency, Vancouver, V5Z 4S6, Canada.
[31]Institute for Systems Biology, Seattle, WA 98109, USA.
[32]Center for Bioinformatics and Functional Genomics, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA.
[33]Oregon Health Sciences University, Knight Cancer Institute, Portland, OR 97239, USA.
[34]Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22908, USA.
[35]Graduate Program in Bioinformatics, Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, 31270-901, Brazil.
[36]Department of Surgery, University of California, San Francisco, CA 94115, USA.

Materials and Methods:

Experimental methods

Tumor sample selection

Tumor samples were selected for this study based on (i) tissue availability (20 mg), (ii) low necrosis, (iii) the presence of other key TCGA data types (where possible), and (iv) to provide a variety of tumor types, representing the diversity found in the TCGA project.

Sample characteristics

In total, we processed 492 tumor samples representing 484 unique donors. Of these 492 samples, 83% (N=410) yielded ATAC-seq data that passed our stringent quality control thresholds based on enrichment of signal over background. Of the 410 passed-QC samples, we were able to isolate sufficient nuclei from 94% (N=386) to perform the ATAC-seq reaction in technical duplicate. In total, we sequenced 796 libraries representing 386 technical replicate pairs and 24 single replicate samples. The 17% (N=82) of samples that failed to pass our stringent data quality control thresholds failed for one of the following reasons: (i) unable to isolate nuclei from tumor sample (N=5), (ii) low enrichment in signal over background (N=77). We note that certain tissue types (CESC, ESCA, HNSC, LGG, and SKCM) had a higher rate of failure and that this failure could be caused by the quality of the tissue at the time of freezing or by inefficiencies in our nuclei isolation protocol for isolating nuclei from these tissues.

Isolation of nuclei and transposition for ATAC-seq

Each tumor sample weighed 20 mg and was cut from a larger tissue chunk on dry ice to prevent thawing. We developed a customized protocol by combining two previously published methods (*17*, *55*) (See Protocol S1 below). We find that this method has superior consistency across the diverse cancer types profiled in this study. Processing of samples was performed in randomized batches of 12 samples with cancer types randomly distributed across batches to avoid batch effects. All Tn5 transposase (*56*) and Tagment DNA (TD) Buffer (*17*, *56*) used was homemade but we note that purchased Tn5 transposase and TD buffer from Illumina (cat# FC-121-1030) can be used interchangeably at the same volumes.

ATAC-seq library preparation and high-throughput sequencing

Library preparation was performed as described previously (*17*, *57*). After library preparation, library concentration was checked by qPCR using the KAPA Library Quantification Kit. Libraries were sequenced to 100,000-200,000 reads on an Illumina MiSeq Sequencer to check library quality. Library quality was assessed primarily by the transcription start site (TSS) enrichment score (see below). A cutoff of "4" was used to determine whether a library was of sufficient quality to deep sequence. We note, however, that the absolute number of this score depends on the set of transcription start sites used and this metric may not necessarily be comparable across different studies. Here, a transcription start site enrichment score of 4 corresponds roughly to a fraction of

reads in peaks of 15%. However, calculation of the fraction of reads in peaks requires pre-determination of the peak set to be used and this makes the fraction of reads in peaks a less desirable method of quality control when studying novel samples. In this paper, we used the transcripts from "TxDb.Hsapiens.UCSC.hg38.knownGene" to define our transcription start sites and found that the enrichment scores from these transcripts were reproducible across genome builds hg19 and hg38. After quality control, 8 libraries were pooled together and this pool was purified on a homemade 6% poly-acrylamide TBE gel (similar to BioRad cat# 4565015). A gel slice (~1.5 cm long) was cut to retain all DNA above 125 bp. This gel slice was placed into a 0.5 ml tube, a small hole was made in the bottom of the tube with an 18 gauge needle, and the 0.5 ml tube was placed into a 1.5 ml tube. The gel slice was then crushed through the small hole via centrifugation at 21,000 RCF for 3 minutes. The crushed gel was resuspended in 300 ul of crush soak buffer (CSB - 500 mM NaCl, 1 mM EDTA, and 0.5% SDS in water) and incubated overnight at 55°C with 1400 RPM shaking in a thermoshaker. The next day, the gel slurry was transferred to a Spin-X tube (Sigma cat# CLS8162-96) using a wide-bore tip and centrifuged for 3 minutes at 9500 RCF. Five volumes of Zymo ChIP DNA Binding Buffer was added to the flow-through and the DNA was purified using a Zymo DNA ChIP Clean and Concentrator kit (cat# D5205). This gel purification is used to completely remove excess primers which interfere with downstream sequencing on certain Illumina machines, including the HiSeq 4000, through a process called index hopping. After purification, each library pool was quantified by Bioanalyzer and sequenced on a single lane of HiSeq 4000 using paired-end 75-bp reads. Libraries that did not attain at least 25 million aligned, filtered, deduplicated reads were sequenced again and the resultant reads were pooled prior to deduplication.

Manual phasing of the *FGD4* upstream locus

Genomic DNA from isolated nuclei was made using the QIAamp DNA mini kit (cat# 51304). Regions of the *FGD4* upstream locus were PCR amplified and TOPO cloned (Invitrogen cat# K280020). Individual clones, each representing an individual allele, were Sanger sequenced and genotyped at the specific SNPs of interest.

Generation of ATAC-seq data from human dendritic cell subsets

Plasmacytoid dendritic cells (CD123+) and myeloid dendritic cells (CD11c+, HLA-DR+) were isolated via fluorescence-activated cell sorting from human peripheral blood mononuclear cells and processed for ATAC-seq as described previously (*25*).

CRISPRi targeting of predicted peak-to-gene links

CRISPRi perturbations were performed similarly to previous reports (*58*). Briefly, 3 different guide RNA sequences were designed per target peak using GuideScan (*59*). In cases where 3 high-quality guide RNA target sequences could not be identified within the 501-bp peak, peak boundaries were extended by 500 bp on either side, consistent with the observation that CRISPRi is able to silence the 1-kbp region surrounding the target site. Each of the 3 guide RNA sequences

was cloned into a separate vector with either a mouse (Addgene 85996), human (Addgene 85997), or bovine (Addgene 85995) U6 promoter using the BstXI and BlpI restriction enzymes. U6 promoters from different species were used to avoid recombination during downstream cloning steps. After successfully inserting the guide RNAs into the U6 plasmids, each U6-guide RNA combination was amplified using primers designed with overlapping regions for Gibson assembly of the three U6-guide RNA combinations in tandem. Each of the three U6-guide RNA combinations were Gibson assembled into a modified version of pU6-sgGFP-NT1 (Addgene 46914) for lentiviral production as described previously (*42*). This final plasmid contains all three guide RNAs expressed from U6 promoters of different species origin and mCherry and puromycin expression for selection and tracking. Lentivirus was produced using 293T cells transfected with TransIT-LT1 reagent (Mirus cat# MIR2300) and psPAX2 (Addgene 12260) and pMD2.G (VSV-G, Addgene 12259) plasmids for viral packaging. After viral titration, cells were transduced with a target multiplicity of infection of 1. Two days after transduction, media containing puromycin (Thermo Fisher cat# A1113803) at 1 ug/ml for both MDA-MB-231 and MCF7 cells was added to the cells. Puromycin-containing media was refreshed after 2 days of treatment. After 4 total days of puromycin selection, puromycin-containing media was removed and replaced with regular media lacking puromycin. Cells were recovered for 2 days in the absence of puromycin. RNA was isolated from the cells using Trizol LS Reagent (Thermo Fisher cat# 10296-028) and chloroform followed by a QIAgen RNeasy mini purification (QIAgen cat# 74104). After isolation, RNA was then treated with TURBO DNase (Thermo Fisher cat# AM2238) and cleaned again using a QIAgen RNeasy mini cleanup. From this RNA, cDNA was made using the SuperScript III kit (Thermo Fisher cat# 18080051) and diluted 1 to 1 with water prior to qPCR. Quantitative PCR was performed with TaqMan Fast Advanced Master Mix (Thermo Fisher cat# 44-445-57) and the following TaqMan probes: *GAPDH* (Hs03929097_g1), *PDL1* (Hs00204257_m1), *BCL2* (Hs04986394_s1), and *SRC* (Hs01082246_m1). Cycle threshold (Ct) values were determined using a Roche 480 Lightcycler system. For each sample, a delta-delta Ct measurement was obtained by normalizing first to GAPDH and then to a non-targeting guide RNA control plasmid with guide RNAs targeting sequence from the enterobacteria phage lambda genome. All guide RNA sequences are available in Data S7.

Analytical methods

Unless otherwise specified, the term "BAM file" refers to aligned, filtered, deduplicated reads. All sequencing data was analyzed using the GRCh38/hg38 human reference genome. Unless otherwise stated, all genomic comparisons were performed on chr1-22 and chrX.

Terms

In this manuscript we use the following terminology:

Donor – An individual person. Multiple tumor samples can come from a single donor.

Sample – A piece of tumor tissue. Multiple technical replicates can exist per sample.

Technical Replicate – Tumor tissue was homogenized prior to ATAC-seq into a nuclei suspension. Each technical replicate represents an individual ATAC-seq reaction performed in a separate tube on different nuclei isolated from the same sample.
CPM – In our ATAC-seq analysis, we use the term CPM to mean [counts + scaled prior count using edgeR] per million mapped reads.
Insertion – An insertion refers to the precise single-base location where the Tn5 transposase accessed the chromatin. Each sequencing fragment has a forward and reverse read from paired-end sequencing. Each end of this fragment represents a unique Tn5 transposase insertion.

ATAC-seq data processing and alignment

ATAC-seq data processing and alignment was completed using the PEPATAC pipeline (http://code.databio.org/PEPATAC/). The hg38 genome build used for alignment was obtained using Refgenie (https://github.com/databio/refgenie). Briefly, all fastq files were first trimmed to remove Illumina Nextera adapter sequence using Skewer (*60*) with "-f sanger –t 20 –m pe –x" options. After trimming, FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc) was used to validate proper trimming and check overall sequence data quality. Bowtie2 (*61*) was then used for pre-alignments to remove reads that would map to chrM (revised Cambridge Reference Sequence), alpha satellite repeats, Alu repeats, ribosomal DNA repeats, and other repeat regions using "-k 1 -D 20 -R 3 -N 1 -L 20 -i S,1,0.50 -X 2000 –rg-id" options. After removal of these regions, Bowtie2 was used to align to the hg38 human reference genome using "--very-sensitive -X 2000 --rg-id" options. Samtools (*62*) was used to sort and isolate uniquely mapped reads using "-f 2 -q 10 -b -@ 20" options. Picard (http://broadinstitute.github.io/picard/) was then used to remove duplicates using the MarkDuplicates tool with "VALIDATION_STRINGENCY = LENIENT REMOVE_DUPLICATES = true" options. This resulted in the final aligned, de-duplicated BAM file that was used in all downstream analyses.

Peak calling

Peak calling for 796 ATAC-seq profiles and 23 cancer types was performed to ensure high quality fixed-width peaks. We chose to use fixed-width peaks because (i) it makes count based and motif focused analyses less biased to large peaks and (ii) with large datasets merging peak sets to obtain a union peak set can lead to many peaks being merged into one very large peak, limiting our ability to resolve independent peaks. Because each cancer type is not represented by an equal number of samples, we first determined a peak set for each cancer type individually. Initially, performing peak calling with MACS2, we found that peak calls were affected by changes in data quality (TSS enrichment scores ranged from 3.94 to 19 in our dataset) and read depth (range 26 million to 258 million per replicate). To overcome this issue, we designed a peak calling procedure that would produce a set of high confidence peaks. For each sample, peak calling was performed on the Tn5-corrected single-base insertions using the MACS2 callpeak command with parameters "--shift -75 --extsize 150 --nomodel --call-summits --nolambda --keep-dup all -p 0.01". The peak summits were then extended by 250 bp on either side to a final width of 501 bp, filtered by the ENCODE

hg38 blacklist (https://www.encodeproject.org/annotations/ENCSR636HFF/), and filtered to remove peaks that extend beyond the ends of chromosomes.

Overlapping peaks called within a single sample were handled using an iterative removal procedure. First, the most significant peak is kept and any peak that directly overlaps with that significant peak is removed. Then, this process iterates to the next most significant peak and so on until all peaks have either been kept or removed due to direct overlap with a more significant peak. This prevents the removal of peaks due to "daisy chaining" or indirect overlap and simultaneously maintains a compendium of fixed-width peaks. This resulted in a set of fixed-width peaks for each sample which we refer to here as a "sample peak set".

We found that when samples varied in read depth or quality, the MACS2 score varied proportionally due to the nature of the Poisson distribution test in MACS2. Typically, this type of inter-sample variation is handled using a reads-in-peaks normalization. However, this type of normalization is not possible during the process of calling peaks because one must know the peak set to perform a reads-in-peaks normalization. For this reason, we developed a robust method to normalize peak significance scores across samples and cancer types. To do this, the MACS2 peak scores (-log10(p-value)) for each sample were converted to a "score per million" by dividing each individual peak score by the sum of all of the peak scores in the given sample divided by 1 million. We carry out this procedure because as sample quality or read depth increases, the number of peaks called and the significance of those peak calls likewise increases. In this way, we are able to normalize peak calls for sample quality and total sequencing depth without performing a reads-in-peaks normalization (which is not possible at this stage of analysis). This normalization method allows for direct comparison of peaks across donors, enabling the generation of a merged peak set for each cancer type as described below.

We next wanted to compile a "cancer type-specific peak set" containing all of the reproducible peaks observed in an individual cancer type. First, all "sample peak sets" from a given cancer type were combined into a cumulative peak set and trimmed for overlap using the same iterative procedure mentioned above. Again, this procedure keeps the most significant (in this case, score per million) peak and discards any peak that overlaps directly with the most significant peak. To identify reproducible peaks from this merged peak set, the individual "sample peak sets" were overlapped with the merged peak set. Peaks from the merged peak set that were observed in at least two samples with a score per million value >=5 were labeled as reproducible. Lastly, we removed any peaks that spanned a genomic region containing "N" nucleotides and any peaks mapping to the Y chromosome. This resulted in a set of high quality, reproducible, fixed-width peaks for each cancer type (105,585 mean, 32,888 sd) that we refer to as a "cancer type-specific peak set".

Lastly, we wanted to obtain a "pan-cancer peak set" representing reproducible peaks from all cancer types that could then be used for cross-cancer comparisons. To start, we re-normalized the score per million scores for each "cancer type-specific peak set". This was performed to prevent cancer types with more samples or higher quality samples from comprising a disproportionate share of the union peak set. This re-normalization was performed for each cancer type. The

resulting peak sets were combined and we, once again, performed an iterative overlapping removal of this merged peak set based on the re-normalized score per million scores. This resulted in a final peak set of 562,709 peaks, each 501 bp in width. We provide these peaks with hg38 coordinates and the corresponding hg19 coordinates (from LiftOver) in Data S2.

ATAC-seq data QC – Transcription start site enrichment, fragment length distribution, and fraction of reads in peaks

Enrichment of ATAC-seq accessibility at transcription start sites (TSSs) was used to robustly quantify ATAC-seq data quality without the need for a defined peak set (which is not available until all samples have been fully sequenced). First, BAM files were read into a Genomic Ranges object in R using Rsamtools "scanbam" and then corrected by a constant offset to the read start ("+" stranded +4 bp, "-" stranded -5 bp). To get the fragment length distribution, the width of each fragment/GRange was plotted. To get the TSS enrichment profile, each TSS from the R package "TxDb.Hsapiens.UCSC.hg38.knownGene" (accessed by transcripts(TxDb)) was extended 2000 bp in each direction and overlapped with the insertions (each end of a fragment) using "findOverlaps". Next, the distance between the insertions and the strand-corrected TSS was calculated and the number of insertions occurring in each single-base bin was summed. To normalize this value to the local background, the accessibility at each position +/- 2000 bp from the TSS was normalized to the mean of the accessibility at positions +/-1900-2000 bp from the TSS. The final TSS enrichment reported was the maximum enrichment value within +/- 50 bp of the TSS after smoothing with a rolling mean every 51 bp. To calculate the fraction of reads in peaks, we calculated the number of insertions that overlapped a peak using "countOverlaps" and divided by the total number of insertions (number of fragments x 2, as each paired-end read represents an individual Tn5 insertion).

ATAC-seq data QC – genotype correlation of ATAC-seq data with publicly available TCGA SNP array data

To validate that ATAC-seq data attributed to a specific TCGA donor actually originated from tissue taken from that TCGA donor, we performed genotyping analyses comparing our ATAC-seq data (N=926 individual sequencing experiments) to SNP calls from TCGA SNP array data using the Affymetrix SNP 6.0 array (N=11,127 TCGA donors). The ATAC-seq data used in these comparisons corresponds to all sequencing experiments prior to merging of any files, including resequencing of a library to achieve higher depth. The SNP array data was generated by the TCGA previously and serves as the ground truth in this assay. The genomic locations probed by the Affymetrix SNP 6.0 array (904,800 hg38-mappable probes) were overlapped with peak regions called in all ATAC-seq samples, resulting in 405,905 probes that fall within a called peak region. Genotype information for each ATAC-seq BAM file was collected at each of the 405,905 SNP locations and converted into a birdseed-style (*63*) format. A minimum depth of 6 reads was required to make a SNP call. If all reads mapped to either the A or B allele, the position was called as homozygous in the ATAC-seq data (birdseed call of 0 or 2). If the absolute value of the

difference between the counts for allele A and allele B was less than 50% of the total depth, then the position was called as heterozygous in the ATAC-seq data (birdseed value of 1). Otherwise, the position was called as homozygous due to excessive imbalance of the two alleles (birdseed value of 0 or 2). Each of these birdseed-style ATAC-seq genotyping lists were correlated with all available TCGA Affymetrix SNP 6.0 array data (11,127 individual donors). Of the 404 donors where high-quality ATAC-seq data was generated, 4 donors do not have available Affymetrix SNP 6.0 array data (Data S1) and could not be assessed in this analysis. Pearson correlations were performed for each individual ATAC-seq BAM file only at the genomic locations where a SNP call could be made in the ATAC-seq data (all locations with read depth greater than 6). Samples were considered to pass this QC if the correlation with the expected biological donor was greater than the correlation with all of the other 11,126 TCGA donors, indicating that the ATAC-seq data and the Affymetrix SNP 6.0 array data were generated from the same person.

Overlap of ATAC-seq peaks and Roadmap DNase peaks

To assess the degree of overlap between our peak calls and previously published Roadmap DNase-seq data we used reg2map (*18*) (https://personal.broadinstitute.org/meuleman/reg2map/). We first used LiftOver to convert the hg19 coordinates to hg38, retaining more than 95% of regions. We then compared the overlap of regulatory regions shared between the merged variable-width DNase-seq peak calls and our fixed-width union peak set. This comparison was done by using "overlapsAny" testing whether or not a Roadmap DNase-seq peak overlapped our pan-cancer peak-set, resulting in 64.8% of our peaks overlapping with Roadmap and 35.2% representing regulatory regions not observed in the Roadmap DNase-seq data. In addition, we wanted to test the overlap between Roadmap and the individual cancer type-specific peak sets. To do this, we first merged all Roadmap peaks corresponding to each "ANATOMY" subtype (ie. Lung or Cervical). Then, for each cancer type-specific peak set, we computed the percent of the peaks observed within each Roadmap subtype using "overlapsAny" in R.

Overlap of ATAC-seq peaks with chromHMM-defined regulatory states

To characterize the types of regulatory regions in which our ATAC-seq peaks occur, we used ChIP-seq-defined chromHMM states from the Roadmap Epigenomics Project. First chromHMM 15 state models were downloaded from the chromatin state learning site (https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html). We then determined how many regions of each chromHMM state were overlapped by an ATAC-seq peak summit. To determine the significance of these overlaps for each chromHMM state, we compared the proportion of ATAC-seq summits overlapping the given chromHMM state to the expected background determined by the total length covered by the chromHMM state and the length of the hg38 genome computed using a binomial test in R.

Saturation analysis to predict numbers of peaks

To estimate the number of peaks that would be discovered at saturation for each cancer type, we modeled the number of peaks contributed by each additional sample using Michaelis-Menten kinetics. For each cancer type, we used all samples including technical replicates to check for peak reproducibility. However, in this analysis to estimate the number of peaks each additional sample contributes we did not include technical replicates (we conservatively used the technical replicate with fewer called summits above a score per million of 5). First, we chose a random seed sample and determined the number of 501-bp peaks using the methods described above (filtered peaks overlapping the hg38 ENCODE blacklist, filtered peaks extending beyond chromosomes and removed overlaps with our iterative removal procedure). Next, samples were added randomly and the number of new peaks discovered was determined until all samples were added. This was repeated a total of 25 times. We then fit this saturation curve empirically using the "nls" function with the equation:

$$P(x) = \frac{P_{\max}}{1 + \left(\frac{1}{x-1}\right) S_{halfmax}}$$

where P_{max} represents the estimated number of peaks at saturation, $S_{halfmax}$ corresponds to the number of samples at half saturation, and x represents the number of samples. We subtracted the number of samples (x) by 1 to set the baseline at the number of peaks for the first sample, allowing us to better fit the number of peaks gained with each additional sample.

ATAC-seq data analysis – Constructing a counts matrix and normalization

To obtain the number of independent Tn5 insertions in each peak, first the BAM files were corrected for the Tn5 offset ("+" stranded +4 bp, "-" stranded -5 bp) (*16*) into a Genomic Ranges object in R using Rsamtools "scanbam". To get the number of Tn5 insertions per peak, each corrected insertion site (end of a fragment) was counted using "countOverlaps". This was done for all individual technical replicates and a 562,709 x 796 counts matrix was compiled. From this, a RangedSummarizedExperiment was constructed including peaks as GenomicRanges, a counts matrix, and metadata detailing information for each sample.

The counts matrix was then normalized by using edgeR's "cpm(matrix , log = TRUE, prior.count = 5)" followed by a quantile normalization using preprocessCore's "normalize.quantiles" in R. The prior count is used to lower the contribution of variance from elements with lower count values. We used a prior count of 5 for ATAC-seq because there are more features with fewer relative counts than RNA-seq. The reason this strategy was chosen over other normalization strategies was that it was (i) computationally inexpensive, (ii) reproducible, and (iii) directly interpretable. We tested DESeq2's variance-stabilized transform (VST) and regularized-log transform (Rlog), both of which gave results very similar to the above CPM approach but required much more computational effort. Lastly, we merged technical replicates using the log2-average for each ATAC-seq focused clustering analysis (t-SNE, PCA, Heatmaps)

and for integrative clustering analyses with existing data types merged by TCGA case ID to integrate with existing data.

This normalization procedure is analogous to a "reads-in-peaks" normalization. This is important to adjust for changes in data quality across this large cohort. This procedure assumes that global differences in chromatin accessibility (i.e. one cancer type is globally more accessible than another) are minor compared to the differences in chromatin accessibility observed in samples with differing data quality. As such, this normalization procedure prohibits the identification of potential differences in global chromatin accessibility.

RNA-seq data analysis – Constructing a counts matrix and normalization

Published RNA-seq data was downloaded from the NIH Genomic Data Commons portal using the gdc-client command line utility. To generate an RNA-seq matrix for integration analyses, we used the raw HTSeq counts for each "primary tumor" sample matching to the TCGA case IDs with matched ATAC-seq. For cases with multiple RNA-seq files (N=11) we summed the raw counts from each replicate together. We then calculated the exon lengths for each gene by downloading the "gencode.v22.annotation.gtf", read in the exon annotations using "import.gff" from rtracklayer and computed the exon lengths by summing the non-overlapping exons. We excluded all genes mapping to "chrM" and then length normalized the RNA-seq to transcripts per million (TPM). This was done first by length normalizing the read counts per gene by their respective exon lengths and then normalizing these values to a million within each sample. For gene names that were duplicated (mostly non-protein coding genes, N=2,096, 3.5%) we kept the genes that had the highest TPM variance.

Methylation data analysis – Constructing a matrix

Published Methylation 450K array data was downloaded from the NIH Genomic Data Commons portal using the gdc-client command line utility. To generate a matrix for integration analyses, we used the beta values for each "primary tumor" matching to the TCGA case IDs with matched ATAC-seq. For cases with multiple methylation data files (N=4) we took the mean beta value across the replicates. In addition, we logit transformed the beta values to make them normally distributed using "logit" from the car package in R. We then constructed a SummarizedExperiment that contained the beta values, GenomicRanges corresponding to the locations, and metadata to match up with the other data types.

ATAC-seq data Analysis – Reads-in-peaks-normalized bigwigs and sequencing tracks

To visualize our ATAC-seq data genome-wide we used ATAC-seq signal tracks that have been normalized by the number of reads in peaks. We do this because samples of varying quality will have varying percentages of reads in peaks and cannot be adequately compared using depth normalization. In the case of depth normalization, "background" reads are considered equal to reads falling within peaks. When depth normalization is applied, samples with higher background are artificially depressed. For reads-in-peaks normalization, we first constructed bigwigs based on

the Tn5 offset-corrected insertion sites. To do this, the genome was binned into 100-bp intervals using "tile" in GenomicRanges of the chromosome sizes in R. The insertion sites (GenomicRanges) were then converted into a coverage run-length encoding using "coverage". Then, to determine the number of Tn5 insertions within each bin we constructed a "Views" object and calculated the sum in each bin with "ViewSums". We then normalized the total number of reads by a scale factor that converted all samples to a constant 30 million reads within peaks. This approach simultaneously normalizes samples by their quality and read depth, analogous to the reads in peaks normalization within a counts matrix. This was then converted into a bigwig using rtracklayer "export.bw" in R. For plotting tracks, the bigwigs were read into R using rtracklayer "import.bw(as="Rle")" and plotted within R. All track figures in this paper show groups of tracks with matched y-axis scales.

ATAC-seq data analysis – ATAC-seq-centric clustering and visualization

It has previously been shown that distal elements (defined as non-promoter elements) are more cell type-specific than promoter elements (defined as occurring outside of the window -1000 to +100 bp from a TSS). From this, we split our data into promoter (N=45,782) and non-promoter regions (N=516,927) prior to clustering using the ChIPseeker "annotatePeak" function with "TxDb.Hg38.KnownGene". For clustering, we ranked the top 250,000 distal elements by row-variance using matrixStats "rowVars". PCA was then performed using "prcomp_irlba" (first 50 PCs) without scaling (the variance was already stabilized) and checked for any apparent biases in the data such as TSS enrichment, fraction of reads in peaks, and library complexity. We then plotted the variance explained across PCs and looked for the inflection point, which occurred at PC15, explaining cumulatively 85% of the variance. We then took the Euclidean distance across samples within these first 15 PCs using "dist" and performed density clustering (*27*) using "densityClust(distPCA, Gaussian=TRUE)" to get a decision plot (Figure S2A). We chose rho (2) and delta (200) values that determined the most distinct centers and extracted the cluster assignments with "findClusters". These cluster assignments were very robust across different iterations of the number of PCs and the number of peaks used. To visualize these cluster assignments, we performed t-Distributed Stochastic Neighbor Embedding (t-SNE) using "Rtsne(perplexity=20, max_iter=10000, pca=TRUE)" on the top 250,000 most variable peaks.

TumorMap representation

The layout method for the PanCanAtlas iCluster TumorMap was described previously (*28*, *64*). The ATAC-seq t-SNE cluster assignments were generated as described above.

Integrative data analysis – Comparing variation of information between platforms

The clustering information for five molecular platforms (mRNA, miRNA, DNA methylation, RPPA, and copy number analysis) and iCluster were obtained from the supplemental tables of the TCGA Cell-of-Origin study (*28*). ATAC-seq cluster membership was based on clustering using the distal elements peak scores (Figure S2B). The variation of information distance is a

distance measure characterizing the dissimilarity between two partitions of a set. It was applied to quantify the similarity of different TCGA clustering results from different molecular platforms. Variation of information distance was calculated using the implementation included in the mcclust R package. Different platforms were further hierarchically clustered based on the computed variation of information distance. It should be noted that each clustering is derived from a selected subset of features from the labeled molecular platform. For example, DNA methylation is selected based on a subset of CpG probes located in CpG Islands susceptible to hypermethylation. Copy number analysis is based on arm-level aneuploidy. ATAC-seq clustering is based only on distal elements.

ATAC-seq data analysis – Identification of cluster-specific peaks through distal binarization

Once we had determined clusters from density clustering of distal ATAC-seq peaks, we wanted to identify the peaks that were uniquely present in each cluster. Identification of these cluster-specific peaks is computationally challenging due to the number of samples and features present in our dataset. Use of traditional pairwise differential testing would require 153 (18 choose 2) separate tests. Additionally, traditional pairwise differential testing would not identify peaks that are unique to multiple clusters. To solve this problem, we developed a classification method that can be used to identify cluster-specific peaks, termed "distal binarization." Prior to classification, we calculate the intra-cluster mean and intra-cluster standard deviation across every peak for each cluster. Then, for each peak, we rank the groups by their intra-cluster mean. Then, we iterate from the second lowest cluster asking whether the mean of that cluster is greater than the maximum intra-cluster mean plus the intra-cluster standard deviation of the next-lowest sample (for cancer subtype-specific analyses we used 2 times the intra-cluster standard to account for the lower power in a subgroup due to fewer available samples for analysis). This iterative process proceeds until a cluster is identified that meets this criterion. This point is defined as the break point and all clusters with a higher intra-cluster mean are classified as positive for this peak and given a value of "1". All clusters below the break point are given a value of "0". If a peak does not have a break point it is discarded. This peak "binarization" procedure classifies all "1s" as being higher than every individual "0". This also captures the peaks that are unique to multiple groups. We kept all combinations that were unique to 4 or fewer groups. To facilitate multiple hypothesis testing, we computed a contrast matrix for all observed combinations and ran limma's eBayes test on the quantile-normalized log2(CPM) matrix. We then extracted all of the FDR-adjusted p-values from differential testing keeping those peaks that were below an FDR of 0.001. This resulted in the classification of 203,260 peaks. With the binary assignments and significance assessments, we were then able to sort the matrix to then create a heatmap of all of these elements using ComplexHeatmap in R.

ATAC-seq data analysis – Enrichment of motifs in cluster-specific peaks

To see which motifs were over-represented within each of the cluster-specific peak sets, we used the CIS-BP motif database (*65*) and the coordinates of our cluster-specific peak sets to calculate

motif occurrences using motifmatchr (*35*). This incorporates the GC bias observed within peaks for matches and returns a binary matrix denoting the presence or absence of a motif within the peak. For each cluster, we used all peaks classified as a "1" in distal binarization (see above) and then calculated the hypergeometric p-value testing the representation within the cluster-specific peaks compared to the pan-cancer peak set using phyper in R. Raw motif enrichment is insufficient to predict exactly which TF is mediating that activity. For this reason, we looked for the correlation of the motif's –log10(p-value) and the expression of the transcription factor (log2(TPM + 1) and identified those factors where $|R| > 0.4$ as potential mediators of the observed motif enrichment.

ATAC-seq data analysis – Methylation of cluster-specific peaks

To observe patterns in methylation in our cluster-specific peak sets, we first filtered positions on the Illumina 450K methylation array for those occurring within distal peaks. Importantly, the 450K array is mostly targeted at gene rich regions which are generally not enriched for regions of chromatin accessibility. To this end, we only observe 31,000 of the 450K array sites within distal peak regions. Since there was not high one-to-one peak to methylation site correspondence, we then calculated the average logit transformed methylation (*66*) for each sample within each group of cluster-specific peaks (18). For each cluster, we used all peaks classified as a "1" in our distal binarization (see above) and computed the average logit methylation.

Integrative data analysis – Comparing WGBS and ATAC-seq

One of the bladder carcinoma samples (TCGA-BL-A13J) profiled by ATAC-seq in this study had also been profiled for base-resolution DNA methylation level using whole genome bisulfite sequencing in a prior study (*5*). For each peak, we averaged the peak score from the two technical replicates for this sample and kept only peaks with normalized peak score greater than 1. Peaks were then classified into promoter peaks and distal element peaks. To visualize the relationship between ATAC-seq accessibility and DNA methylation, we aligned the peaks by their centers and ordered by a descending average peak score. We segmented each region of interest into 40 windows 50 bp in length and plotted average DNA methylation level and average bigwig signal intensities within the window.

ATAC-seq data analysis – Identification of cancer subtypes with k-means clustering

For profiling cancer subtypes, we used the cancer-specific peak sets and constructed a quantile-normalized log2(CPM) matrix analogous to the full cohort. We again used motifmatchr "motifMatches(out="matches")" with the CIS-BP motif set and found all motif occurrences in the cancer type-specific peak set. Next, we computed the row-wise z-scores of the top 25,000 most variable distal peaks and performed k-means clustering. For each cluster of peaks, we then calculated the hypergeometric p-value testing the representation within the k-means-defined peak sets compared to a background of all ATAC-seq peaks observed in the given cancer subtype using phyper in R. The nearest genes for each k-means-defined cluster of peaks were identified with ChIPseeker "annotatePeak".

Breast cancer known subtype calls

For breast cancer gene expression-based subtype classification, we used packages genefu (*67*) and iC10 (*68*). Variance stabilization transformation was used for normalization. Genefu was used for subtyping the SCMOD2 (Subtype Clustering Model) (*69*), ER, PR, and HER2 models from gene expression data using a mixture of Gaussians and PAM50 (*70*, *71*). Package iC10 was used for IC10 subgroup classification (*72*).

ATAC-seq data analysis – Inferring copy number amplification

To infer DNA copy number amplifications from ATAC-seq data, we first tiled the genome into 2-Mbp windows using "tile" of genomic ranges for chromosome sizes in R. These window positions were then filtered against regions with known artefactual mapping issues using the ENCODE blacklist with the "setdiff" function in R. Then, the number of insertions within each filtered window was determined using "countOverlaps". Next, the insertions per bp was determined within each filtered 2-Mbp window. Then, the percent GC content was computed for each filtered 2-Mbp window using the hg38 BSgenome in R. To estimate if a region is amplified, for each window we took the 100 nearest neighbors based on GC content and computed the average log2(fold change). If this was above 1.5 we considered this region as a candidate for amplification. This window size best captured smaller known amplifications, but added more false positives compared to 10-Mbp windows.

ATAC-seq data analysis – Transcription factor footprinting

We sequenced our ATAC-seq data to a median of 56.7 million aligned and deduplicated reads per technical replicate and wanted to use this deep data for transcription factor footprinting which measures the occupancy of a TF. One main challenge to TF footprinting is the insertion sequence bias of the Tn5 transposase (*73*) which can lead to misclassification of TF footprints. To account for Tn5 insertion bias we first identified the hexamer sequences surrounding each Tn5 insertion site (*34*). To do this, we identified single-base resolution Tn5 insertion sites (see above methods), resized these 1-bp sites to 6-bp windows (-3 and +2 bp from insertion), and then created a hexamer frequency table using the "oligonucleotidefrequency(w=6, simplify.as="collapse")" function from the Biostrings package. We then calculated the expected hexamers genome-wide using the same function with the hg38 BSGenome. To calculate the insertion bias for an aggregate footprinting plot, we first created a hexamer frequency matrix that is represented as the possible hexamers across a window +/- 250 bp from the motif center. Then, iterating over each motif site, we filled in the positioned hexamers into the hexamer frequency matrix. This was calculated 1 time for each CIS-BP motif genome-wide. Using the sample's hexamer frequency table, we could then compute the expected Tn5 insertions (shown in Figure S4A) by multiplying the hexamer position frequency table by the observed/expected Tn5 hexamer frequency. We would caution that this insertion correction model does not remove all potential bias signals, and a component of the footprint depth signal may still be driven by residual sequence bias. This unaccounted sequence bias may cause

weaker but still significant spurious negative or positive correlations between footprint depth and flanking accessibility. Thus correlations between footprint depth and flanking accessibility ought to be considered suggestive of binding modes, as discussed in the main text, not definitive evidence of such modes. We anticipate that more comprehensive bias models will further improve confidence in the measurement of footprint depth.

To calculate TF footprints, we first determined the location of TF motifs within peaks genome-wide. To do this, we used the pan-cancer peak set combined with the CIS-BP motifs to calculate the motif positions using motifmatchr "matchMotifs(positions = "out")". This returns a GenomicRangesList where each list object is a GenomicRanges of stranded motif positions genome-wide, which was then extended to +/- 250 bp centered at the motif. To calculate the insertions around these sites, first we converted the Tn5-corrected insertions GenomicRanges (see above) into a coverage run-length encoding using "coverage". For each individual motif, we iterated over the chromosomes, computing a "Views" object using "Views(coverage,positions)". This "Views" object was converted to a matrix using "as.matrix" which was then used to calculate the following:

$$\sum_{i=chr1}^{chrX} \left(colSums\left("+"Stranded\right) + rev\left(colSums\left("-"Stranded\right)\right)\right)$$

To better compare footprints across samples, we then calculated two values for each footprint: the flanking accessibility and the footprint depth. This is based on previous work from Baek et al. (*34*). We first defined 3 key relative positions: (i) the footprint base is the region encompassing the very center of the motif and is defined as length(position weight matrix)/2 + 5 bp from the motif center, (ii) the footprint flank is the region immediately adjacent to the transcription factor binding site and is defined as the region between the end of the footprint base and 50 bp away from the motif center, and (iii) the footprint background is defined as the region between 200 and 250 bp away from the motif center. The footprint depth was calculated as the log2(fold change) between the 10% trimmed mean within the footprint base and the mean accessibility within the footprint flank. A negative value for footprint depth indicates strong factor occupancy. The flanking accessibility was calculated as the log2(fold change) between the mean accessibility within the footprint flank and the mean accessibility of the footprint background. A positive value for flanking accessibility indicates motif-associated changes in local accessibility. These two values were then used for subsequent footprinting analyses.

Next, we determined which TF footprints were significantly correlated to the gene expression of the associated gene. We note that for transcription factor families comprising numerous proteins that bind to similar DNA motifs, this analysis will likely not generate a correlation between the expression of a single family member and chromatin features of the common motif – a significant limitation to global linkage of chromatin accessibility changes to changes in the gene expression of specific transcription factors. To determine which factors are correlated to their respective footprint, we compared both the flanking accessibility and the footprint depth to the expression of the paired gene (log2(TPM+1)). To test which are significantly

correlated, we generated a null mean and standard deviation for each motif by correlating each feature (flanking accessibility or footprint depth) to the expression of 250 random TFs (excluding its matched TF). This resulted in FDR-corrected p-values for all TFs for each footprint. In addition, we de-duplicated the TFs with multiple motifs for all subsequent analyses by keeping the motif that had the highest absolute correlation ($|cor_{flank}| + |cor_{depth}|$) to the motif's expression log2(TPM+1).

To compare methylation values within transcription factor binding sites, we calculated the mean methylation beta value in all sites that were within +/- 25 bp of the motif center (51-bp window). We chose this window size to maximize the occurrence of the 450K array data for more robust measures of methylation, while being as close to the motif center as possible. This was done by subsetting the methylation SummarizedExperiment using subsetByOverlaps with the 51-bp motif regions and then calculating the sample means using "colMeans(na.rm = T)".

ATAC-seq data analysis – chromVAR for transcription factor activity

In addition to TF footprinting, we wanted to measure global TF activity using chromVAR. We used as input the raw counts for all distal peaks and the CIS-BP motif matches within these peaks from motifmatchr. We then computed the GC bias-corrected deviations using the chromVAR "deviations" function. We then correlated the bias-corrected deviations to the log2(TPM+1) expression for the corresponding motif's TF (TFs with multiple motifs were de-duplicated as in the footprinting analysis). We then computed 5000 random correlations between transcription factor motifs and the RNA-seq gene expression of non-associated transcription factor genes to calculate an FDR for each correlation.

ATAC-seq data analysis – chromVAR for GWAS enrichment

To assess the variation of chromatin accessibility at known genetic risk loci, we gathered known GWAS SNPs from the European Bioinformatics Institute GWAS catalog (https://www.ebi.ac.uk/gwas/docs/file-downloads), and retrieved SNPs associated with 16 major cancers (breast cancer, lung cancer, colon cancer, non-melanoma skin cancer, melanoma, ovarian cancer, pancreatic cancer, brain cancer, lymphoma, bladder cancer, esophageal cancer, kidney cancer, prostate cancer, thyroid cancer, cervical cancer and uterine cancer). We expanded the SNP list by adding SNPs with Linkage Disequilibrium (LD) $r^2 > 0.8$ to the GWAS lead SNPs. This LD information was obtained from the haploreg website (http://archive.broadinstitute.org/mammals/haploreg/data/). We removed SNPs that were located in exons or UTR regions. This final SNP list was overlapped with our distal binarization peak set and we created a binary matches matrix; where each column represents a set of GWAS SNPs, grouped by cancer type, and each row represents a peak from our distal binarization peak set. The values of the matrix are a binary representation of the overlap of a SNP from the set of grouped GWAS SNPs with the given peak. We then computed the GC bias-corrected deviations using the chromVAR "deviations" function. We converted the "deviation scores" to a p-value using "pnorm" and adjusted using the Benjamini-Hochberg procedure.

ATAC-seq data analysis – Peak-to-gene linking predictions

To identify putative causal links between ATAC-seq peaks and gene expression we used a correlation-based approach. First, we removed the bottom 25% of both genes and peaks based on variance. Then, all possible interactions between ATAC-seq peaks and genes within 0.5 Mbp were identified. For each of these interactions, we computed the Pearson correlation between the ATAC-seq peak accessibility (log2(CPM)) and the gene's expression (log2(TPM+1)). To determine the significance of these correlations, we constructed a conservative null model meant to account for spurious association. For each chromosome, we correlated the ATAC-seq accessibility of 10,000 random peaks (not on the same chromosome and within 0.5 Mbp of a gene start) to the expression of every gene on the chromosome. We then calculated the mean and standard deviation for these "trans" correlations to represent nonspecific correlation. This enabled us to compute the p-value for each correlation and adjust for multiple hypothesis using the Benjamini-Hochberg procedure (FDR). We then selected all correlations with an FDR below 0.01.

Manual inspection of these peak-to-gene links, demonstrated that many of these links were present in regions of recurrent DNA copy number amplification (especially within a single cancer type). To remove links driven by this phenomenon, we leveraged published TCGA DNA copy number variation data. For each TCGA case, ATAC-seq peaks that occurred within known copy number amplified regions were marked as NA. We then recalculated all correlations (ignoring the NAs) and recomputed an FDR for each correlation. We again filtered for all CNA-adjusted correlations with an FDR below 0.01.

In addition, we wanted to construct a procedure that could account for diffuse correlations where large genomic regions are highly co-accessible. This occurs in regions where the expression of a gene is more strongly correlated to the overall chromatin accessibility of a large region than it is to the chromatin accessibility of any single peak. This situation is observed in regions such as the HOXB locus where gene expression is highly coordinated across a large genomic window. To do this, we tiled the genome into 100-kbp windows and calculated the number of Tn5 insertions for each sample. We then performed the same normalization as for the ATAC-seq peaks (quantile-normalized log2(CPM)) and merged samples by averaging samples with the same TCGA case ID. For every significantly linked ATAC-seq peak we determined within which 100-kbp window the summit resided and computed the correlation of the 100-kbp window's total accessibility to the expression of the linked gene. Predicted links with correlations that were lower than the absolute correlation of the gene's expression to the window's accessibility were then removed. Then, the remaining peak-to-gene links were filtered to remove links that overlapped the promoter region (defined as the ATAC-seq peak summit within -1000 to +100 bp from the TSS. Finally, for the remaining distal peak-to-gene links, we wanted to identify those that are driven by a single cluster identified in our analysis vs links that are more correlated across all cancer types. To do this, we recomputed all correlations removing all samples belonging to 1 cluster (for all 18 clusters individually) and keeping the lowest absolute correlation. We then labeled those that were no longer significant as "Cluster Driven" and those that remained significant as "Heterogeneous".

In this study, we use fixed-width 501-bp peaks to allow for motif searching and to create a union peak set that is not heavily length biased. However, this fixed-width windowing does not always conform to the size of every enhancer, some of which can be larger than 501 bp. In our significant links, there are many linked ATAC-seq peaks that are adjacent to other ATAC-seq peaks linked to the same gene. We found that these adjacent ATAC-seq peaks were more correlated to each other than to their corresponding linked gene, suggesting that these adjacent peaks likely correspond to a single "enhancer unit". To classify these adjacent linked peaks as so-called "enhancer units" we extended the peaks in increasing increments from 501 to 10 kbp, and if adjacent peaks linked to the same gene overlapped, they were denoted as a single "enhancer unit". Then, we identified the inflection point at 1500 bp where there were 58,092 unique enhancer-to-gene links corresponding to 81,323 peak-to-gene links. We then merged peaks denoted as an "enhancer unit" reporting the median correlation, median FDR (adjusted for CNA), and the closest peak summit distance to the gene start. We report both the individual peak-to-gene links and the merged enhancer-to-gene links in Data S7.

Integrating predicted ATAC-seq peak-to-gene links and regulon targets

Target genes for transcription factors in the METABRIC cohort 1 (N=997) have previously been identified (*74*). We used these METABRIC target gene sets in the TCGA breast cancer donors (N=74) for which ATAC-seq data was available.

For ESR1, we calculated regulon activity profiles with the two-tailed GSEA method (*45*) implemented in the RTN package, across the 74 donors. We then sorted this cohort by the activity profile (i.e. by dES), and generated a heatmap of chromatin accessibility z-scores for the sorted cohort for the subsets of negative and positive regulon target genes for which ATAC-seq distal peak-to-gene links had been predicted for the 74 BRCA donors. Peaks linked to multiple targets were used in the evaluation of each target. When more than one peak mapped to the same target, the average of all the linked peaks was used. Target genes that lacked a distal peak-to-gene link were dropped. Samples were averaged across technical replicates. Overall survival (OS) outcomes and tumor covariates were taken from Liu et al. (*75*). To complement the heatmap, we generated violin plots of ATAC-seq z-scores for negative and positive targets. In each plot, we assessed differences in distributions between samples with positive and negative dES with a t-test.

Survival analyses were performed using the Bioconductor RTNsurvival package on the full (N=1082) TCGA BRCA cohort, for which we re-calculated ESR1 regulon activity scores (dES) for each donor. For the Kaplan-Meier curve, we stratified the cohort into 3 groups – positive, undefined, and negative dES – and evaluated differences between the groups for 5-year OS, using a logrank test. Additionally, we fitted a Cox proportional hazards regression to further assess ESR1 regulon activity and survival, while considering age at initial diagnosis, tumor stage, ER status, and HER2 status as covariates. Finally, we tested the proportional hazards assumption of the Cox model using Schoenfeld's residual test.

Validation of predicted peak-to-gene links using ELMER

TCGA Illumina 450K DNA methylation array data was processed with SeSAMe (*76*), and matched to RNA-seq (FPKM-UQ) for 858 Breast Cancer samples. ELMER (*8*) supervised analysis was used to perform pairwise analysis of each PAM50 (*71*) subtype against all other subtypes, using only the default "distal" probe filter, which uses only 160,944 probes that are >2 kbp from a GENCODE 28 transcription start site. We used the following cutoffs for ELMER analysis: get.diff.meth(sig.diff)=0.15, get.diff.meth(p_value)=10^{-4}, get.pair(Pe)=10^{-9}, get.pair(raw.pvalue=10^{-9}), and get.pair(filter.probes)=false. ELMER version 2.5.4 (*43*) was used for all analyses.

We filtered the full set of BRCA-specific ATAC-seq peak-to-gene linkages described above to those overlapping one of the distal methylation probes by including any peak that overlapped a +/- 250-bp region around each probe. This yielded a set of 1,328 ATAC-seq peak-to-gene links. This set of ATAC-seq peak-to-gene links is represented by the rows of Figure S6A, and is the denominator for Figure 6D. Each ATAC-seq peak-to-gene link was associated with one or more ELMER probe-to-gene links if the gene matched by ENSEMBL ID, and if the ELMER probe was within a +/- 250-bp window of the ATAC-seq peak. If multiple ELMER probe-to-gene links were assigned to the same ATAC-seq peak-to-gene link, the beta values of all probes were averaged for the heatmap in Figure S6A. In total, 464 ATAC-seq peak-to-gene links (34.9%) were matched to one or more ELMER links, and the subtype-specific ELMER assignments for each of these were used to order rows in Figure S6A.

To evaluate the degree of overlap compared to chance, we generated "randomized" sets of ELMER links as follows. From the actual ELMER results, we compiled the full set of probe-to-gene links that were identified in any of the ELMER pairwise runs (N=108,587). For each real link, we generated a randomized link by selecting a random probe from among the 160,944 distal Illumina 450K methylation array probes used as input to ELMER. We then randomly assigned it to a nearby gene having the same relative position as the gene in the associated ELMER link (for instance, if the ELMER probe was linked to the 4th nearest gene upstream, we linked the randomized probe to the 4th nearest gene upstream). This randomized set thus contained exactly the same number of probe-to-gene links as the true ELMER output, and these links were overlapped and compared with the 1,328 ATAC-seq peak-to-gene links using the same procedure used for the true ELMER output. We produced 1000 such randomized sets, which had a mean of 48.3 overlaps (3.6%) and a standard deviation of 7.7 (0.6%). The number of randomized overlaps was approximately normal, with a minimum of 27 overlaps and a maximum of 69. The number of observed overlaps between ATAC-seq peak-to-gene links and ELMER probe-to-gene links (N=464) is thus extremely statistically significant, with a z-score of 53.8 used to estimate $p < 10^{-612}$. Because we only generated 1,000 randomized trials, we conservatively used "$p << 0.001$" in the text (Figure 6D).

All code to produce ELMER supervised analysis and ATAC-gene comparison figures is available as an HTML report at https://gdc.cancer.gov/about-data/publications/ATACseq-AWG.

Overlap of ATAC-seq peak-to-gene links with GTEx eQTLs

eQTLs from the Genotype-Tissue Expression project were used to support the ATAC-seq defined peak-to-gene links. GTEx eQTL data (version 7) was downloaded from https://gtexportal.org/home/datasets and the *.signif_variant_gene_pairs.txt.gz files were used. Next, the provided hg19 coordinates were converted to hg38 using "LiftOver" from rtracklayer with an hg19ToHg38 liftOver chain in R. All eQTLs located more than 500 kbp away from the predicted gene pair were removed to maintain consistency with the 500 kbp window used in predicting ATAC-seq peak-to-gene pairs. Next, the nearest genes for each eQTL were determined using "distanceToNearest" with the eQTL regions to all gene starts in gencode v22 in R. Then, all eQTLs that were paired with the nearest gene were removed to better test the predictive power of the non-nearest gene peak-to-gene link predictions. All peak-to-gene links were then overlapped with these filtered eQTLs using "findOverlaps" and then matched based on the predicted linked gene. To assess the significance of these overlaps, we created 250 random peak-to-gene link sets by taking all peaks from the peak-to-gene links (81,323 pan-cancer links and 9,711 BRCA links) and randomly assigning these peaks to any gene within 500 kbp of the peak summit. Then, we calculated the z-score and enrichment of our determined peak-to-gene links compared to the randomized peak sets. We then calculated the adjusted p-value using the Benjamini-Hochberg correction.

ATAC-seq data analysis – Immune-based analysis of peak-to-gene linking

To profile our predicted peak-to-gene links for links related to immune infiltration, we used published ATAC-seq data from the hematopoietic system (*25*) to characterize which of the peak-to-gene links are "immune-biased." We constructed a counts matrix from the Tn5 offset-corrected insertions for the hematopoietic subtypes in the union peak set and merged this matrix with the counts matrix of all samples profiled in this study. After log2-normalizing this matrix, we computed the median accessibility for every cancer subtype and the median accessibility across all cancer types to get a robust estimate for the average accessibility of each peak within our cancer cohort. Next, we computed the minimum log2 accessibility within each hematopoietic subtype. Then, we computed the median log2 accessibility across all cancer types by taking the median accessibility within each cancer type and then the median across all cancer types. We then computed the log2(fold change) between the ATAC-seq accessibility of the hematopoietic cell types and the cancer cohort's median accessibility for each peak. This gave a conservative estimate of immune "bias" for each linked ATAC-seq peak. To determine which of these links correlated with immune infiltration we first calculated the cytolytic activity (defined as the log-average of the geometric mean of granzyme A and perforin 1 expression in TPM (*49*)) from published TCGA RNA-seq data for each TCGA case ID. We then correlated the cytolytic activity to the accessibility of all linked ATAC-seq peaks. The correlation of ATAC-seq peak accessibility to cytolytic activity was plotted against the log2(fold change) for the most biased differentiated hematopoietic groups (Figure 6H, colored by immune function: B cells, CD4/8 T cells, NK cells, pDCs, Monocytes and mDCs). This analysis excluded links found to be most biased in hematopoietic progenitors.

HiChIP data analysis – MetaV4C

We wanted to validate our predicted links using previously published 3D chromosome conformation data. We made use of published histone H3 lysine 27 acetylation (H3K27ac) HiChIP data from the MDA-MB-231 basal-like breast cancer cell line and primary T cell subsets (Naive, Th17, and Treg) to serve as controls (*42*, *77*). H3K27ac HiChIP directly interrogates enhancer-centric chromatin interactions, enabling us to validate our predicted peak-to-gene links. First, we converted our peak-to-gene links to 10-kbp resolution by flooring each coordinate (gene start and peak center) to the nearest 10-kbp window. Next, we deduplicated all links at 10-kbp resolution. To make distance-scaled Meta-Virtual 4C plots, each chromosome was retrieved from the ".hic" interactions file using juicer dump at 10-kbp resolution and read into a "sparseMatrix" in R (each coordinate in the matrix corresponding to a 10-kbp interaction bin). Then, for each peak-to-gene link longer than 100 kbp, the upstream or downstream window (depending on the peak's location relative to the transcription start site) was identified and then interpolated linearly using the "approx" function to get the value at each 0.1%. This was then summed for each predicted interaction and divided by the total number of predicted interactions. Replicate reproducibility was visualized with the mean profile shown as a line and the shading surrounding the mean representing the standard deviation between replicates.

Whole genome sequencing and ATAC-seq genotyping

Whole genome sequencing somatic mutation calls were obtained from the TCGA through the NIH Genomic Data Commons. These mutations were called in hg19 so LiftOver was used to obtain the same positions in hg38. In total, 35 samples had available WGS somatic mutation calls and ATAC-seq data. Each sample was genotyped in the ATAC-seq reads at each position identified as a somatic mutation in the WGS data. Genotyping was performed using samtools mpileup with the options "--VCF --skip-indels --uncompressed --output-tags AD --positions --fasta-ref" to create a VCF file. Read counts at these SNV positions were compiled and compared to the "ground truth" observed in the published TCGA WGS data. SNVs were considered to affect chromatin accessibility if they met two criteria: (i) the SNV is observed at a higher variant allele frequency in the ATAC-seq data than in the WGS data (difference greater than 0.2), and (ii) the sample harboring the SNV has a higher overall accessibility at that position than other samples of the same cancer type (FDR less than 0.1). In this way, we were able to identify SNVs that are associated with an increase in accessibility in an allele-biased fashion.

Motif identification at positions of SNVs

To identify motifs whose binding affinity is predicted to be most perturbed by a given SNV, we constructed a REF and VAR string (+/- 10 bp from the SNV, 21 bp in length) using the hg38 BSGenome reference sequence. We then used motifmatchr with CIS-BP motifs to calculate the motif positions and motif scores (representing the significance of the motif match) using motifmatchr "matchMotifs(positions = "out", p.cutoff=0.01)" in both the REF and VAR string. Then, we intersected the motif positions with the SNV in both the REF and VAR string to ensure

the motifs overlap the SNV. For each motif present, the motif's score differential was computed if the motif was present at the same position in both the REF and VAR. If there was no motif present at the same position in both the REF and VAR, the differential was compared to a motif score of "0".

Nucleosomal and sub-nucleosomal sequencing tracks

To visualize the nucleosome-spanning and sub-nucleosome fragments in a track, we first separated the Tn5 offset-corrected fragments into sub-nucleosomal (<100 bp) and nucleosome-spanning (>160 bp) fragment sets. Then the insertions were converted to a run-length encoding in R using "coverage". Then these were scaled to 30 million total fragments and plotted in R.

Haplotype identification in the *FGD4* locus

Haplotypes were identified from the 1000 genomes project using LDlink (*78*).

Survival analysis

Clinical outcome data was obtained from the TCGA pan-cancer clinical curated data (*75*). Expression data was obtained from the NIH Genomic Data Commons. Package survcomp was used for drawing the Kaplan-Meier plots and defining the optimal threshold (function surv_cutpoint). Package rms was used for building the Cox Proportional Hazard model and plotting the hazard ratio plot.

For *FGD4* analysis in BLCA, the Cox Proportional Hazard model includes the binarized value of FGD4 expression, age, subtype, stage, and lymph node status as relevant covariates. The outcome is overall survival censored at 10 years. For *MECOM* analysis in KIRP, the Cox Proportional Hazard model includes the binarized value of *MECOM* expression, age, stage, and lymph node status as covariates. *MECOM* expression was divided into "overexpressed" and "normal" using the maxstat statistic. Outcome is overall survival censored at 10 years. P-values reported for the univariate model correspond to the logrank test, and the multivariate model corresponds to the anova test for the expression.

Supplementary Protocol #1 – Processing of frozen tissue fragments for ATAC-seq

Before you start the protocol:

1) All steps should be performed on ice or at 4°C. Pre-chill a swinging bucket centrifuge and a fixed angle centrifuge to 4°C.
2) Pre-chill all Dounces and pestles to 4°C in a fridge.
3) Pre-chill all tubes. For each sample you are processing, you will need:
    a. One 2 ml round-bottom LoBind tube for gradient separation
    b. One 1.5 ml LoBind tube for RNA homogenate
    c. One 2 ml Nunc Cryotube for extra nuclei
    d. One 50 ml conical for filtration step (often optional)
4) Prepare all buffers. For faster dissolution, crush protease inhibitor tablets prior to addition to 1x Homogenization Buffer Unstable Solution. DTT, Spermidine, Spermine, and digitonin are stored at -20°C. All other detergents, ATAC-RSB, and other buffers are stored at 4°C. Do not prepare transposition mix ahead of time.
    a. Remember that the catalog number provided for iodixanol from Sigma comes as a 60% solution (not 100%).
5) Fill up a 2 L beaker with 500 ml sterile water to soak the used Dounces and pestles.

Isolation of Nuclei via Dounce Homogenization and Density Gradient Centrifugation:

1) Remove samples from liquid nitrogen storage and keep on dry ice until use.
2) Place 20 mg frozen tissue into a pre-chilled 2 ml Dounce containing 1 ml cold 1x HB and let thaw for 5 minutes.
    a. For >30 mg tissue, use 2 ml 1x HB. For 10-20 mg tissue, use 1 ml 1x HB. For 50 um tissue sections, use 0.5 ml 1x HB.
3) If you would like to collect RNA from the same sample, add 10 ul RiboLock per ml of 1x HB and mix well.
4) Dounce with “A” loose pestle until resistance goes away (~10 strokes).
5) Place “A” pestle into beaker with sterile water to soak for cleaning later.
    a. Optional – If residual un-homogenized tissue makes it difficult to Dounce, filter homogenate through a pre-chilled 50 ml conical using a 70 um bucket-style cell strainer filter prior to using tight pestle “B”.
6) Dounce with “B” tight pestle for 20 strokes.
7) Place “B” pestle into beaker with sterile water to soak for cleaning later.
8) Filter during transfer using a 70 um Flowmi strainer and transfer homogenate to a pre-chilled 2 ml LoBind tube.
9) Place Dounce into beaker with sterile water to soak for cleaning later.
10) Pellet nuclei by spinning 5 min at 4°C at 350 RCF in a fixed angle centrifuge.
11) Remove all but 50 ul of supernatant (containing cytoplasmic RNAs) and transfer to a pre-chilled 1.5 ml LoBind tube. If the pellet is not clearly visible, you can leave more supernatant in the tube, up to 400 ul and add less of the 1x HB buffer in the next step.
12) Gently resuspend nuclei in a total volume of 400 ul 1x HB. If you only left 50 ul in the tube in the previous step, this means you should add 350 ul 1x HB. Make sure nuclei are fully resuspended without clumps.
13) Add 1 volume (400 ul) of 50% Iodixanol Solution and mix well by pipetting

14) Slowly layer 600 ul of 30% Iodixanol solution under the 25% mixture. To avoid mixing of layers, wipe the side of the pipette tip with a Kimwipe to remove excess Iodixanol solution from the external surfaces of the pipette tip.
15) Layer 600 ul of 40% Iodixanol solution under the 30% mixture. To avoid mixing of layers, wipe the side of the pipette tip with a Kimwipe to remove excess Iodixanol solution from the external surfaces of the pipette tip.
    a. During this step, you will need to gradually draw your pipette tip up to avoid overflowing the tube. However, the tip of your pipette must stay below the 30%-40% interface at all times.
16) In a pre-chilled swinging bucket centrifuge, spin for 20 min at 4°C at 3,000 RCF with the brake off. Handle tubes gently so as to not disturb the gradient.
    a. Iodixanol is meant to be used at higher speeds (10,000 RCF) but high-speed swinging bucket centrifuges are not always readily available so we perform this step at 3,000 g and have not had any issues.
17) Using a vacuum, aspirate the top layers down to within 200-300 ul of the nuclei band at the 30%-40% interface. Be careful not to get too close as you will disrupt the nuclei band.
18) Using a 200 ul volume, collect the nuclei band and transfer to a fresh tube. Do not aspirate more than 200 ul at this step as this can cause you to take too much of the 40% layer which sometimes contains debris.
19) Dilute nuclei by adding some volume of ATAC-RSB-Tween Buffer. Mix gently by pipetting. The precise volume of ATAC-RSB-Tween to add will depend on how many nuclei you have. If you don't dilute enough, it will be hard to get an accurate count. If you dilute too much, it will be similarly hard to get an accurate count. You should minimally add 200 ul of ATAC-RSB-Tween buffer to dilute the iodixanol as high concentrations of iodixanol can be too viscous for hemocytometers.

Transposition of Nuclei:

1) Count nuclei using Trypan blue staining (1:1 ratio of Trypan to sample) and a manual hemocytometer. We recommend using disposable hemocytometers for consistency but do not recommend automated cell counters.
2) We normally perform two technical replicates per sample. Each technical replicate should ideally have 50,000 nuclei, requiring 100,000 nuclei total. If you don't have at least 100,000 nuclei, follow this convention:
    a. More than 50,000 nuclei, still do 2 technical replicates using half of the volume for each replicate but reduce the volume of Tn5 transposase proportionately to the number of nuclei. Maintain all other transposition reaction volumes. For example, for 25,000 cells, use 1.25 ul of Tn5 transposase in a 50 ul total reaction volume. Replace omitted Tn5 volume with water.
    b. Less than 50,000 nuclei, only do 1 technical replicate and reduce the volume of Tn5 transposase proportionately to the number of nuclei. Maintain all other transposition reaction volumes. For example, for 25,000 cells, use 1.25 ul of Tn5 transposase in a 50 ul total reaction volume. Replace omitted Tn5 volume with water.
3) Label and chill 1.5 ml LoBind tubes according to how many tubes will be needed for transpositions.

4) Transfer 50,000 nuclei into a 1.5 ml LoBind tube containing 1000 ul of ATAC-RSB-Tween Buffer. If the total volume won't fit in a 1.5 ml tube, just reduce the amount of ATAC-RSB-Tween that you add to the tube to start.
5) Centrifuge nuclei for 10 minutes at 500 RCF at 4°C in a fixed angle centrifuge. At this point, the pellet should be clearly visible if 50,000 nuclei were used. Pellets of as few as 10,000 nuclei should be visible.
6) Using a p1000 pipette, remove all but the last 100 ul of supernatant. Remove last 100 ul with p200 pipette set to 200 ul using a single fluid pipetting motion. Place the tip of your pipette on the opposite side of the tube to where the nuclei pellet is located during this final aspiration step.
7) Add 50 ul ATAC-seq Reaction Mix to each tube and pipette up and down 6 times to resuspend nuclei pellet.
    a. Unlike the published ATAC-seq protocols, you do not need to do an individual lysis step in this protocol because the nuclei are exposed to NP40 throughout the Douncing portion of the protocol.
8) Incubate reactions at 37°C for 30 min in a thermoshaker with 1000 RPM constant shaking.
9) After incubation, add 250 ul (5 volumes) of Binding Buffer from the Zymo DNA Clean and Concentrator 5 kit. Mix well by vortexing and inverting to collect any condensate from the lid.
10) Pulse centrifuge to collect volume in the bottom of the tube.
11) Either finish the cleanup protocol using the Zymo DNA Clean and Concentrator 5 kit or transfer the binding buffer transposition mix to -20°C for short term storage for up to 1 week.
    a. If you store the binding buffer transposition mix at -20°C, allow it to equilibrate to room temperature and mix well before proceeding with the Zymo DNA Clean and Concentrator clean up protocol.

Cleanup and Freezing Down Tubes:

1) If you would like to save extra nuclei for other assays or to potentially use in additional ATAC-seq experiments downstream:
    a. Pellet remaining nuclei by centrifugation for 10 min at 500 RCF at 4°C
    b. Carefully aspirate supernatant using two pipetting steps (p1000 then p200) as above.
    c. Gently resuspend nuclei pellet in 100 ul of cold BAM Banker media and transfer to a pre-chilled 2 ml Nunc cryovial.
    d. Slow-freeze nuclei in a freezing container and move to -80°C or liquid nitrogen storage the next day.
2) If you would like to keep homogenate for making RNA (or potentially protein) downstream:
    a. Store homogenate at -80°C.
3) Cleaning Dounces and pestles:
    a. Rinse all Dounces and pestles thoroughly with sterile water (2x) followed by 70% ethanol (2x).
    b. Let Dounces and pestles dry on a kimwipe or paper towel for a few hours to overnight.

Processing Homogenate to Make RNA (optional):

1) Pre-chill a fixed-angle centrifuge to 4°C.
2) Thaw homogenate on ice.
3) Transfer 150 ul to a 2 ml LoBind tube containing 1500 ul Trizol and mix.
4) Add 400 ul chloroform to the Trizol and mix by vortexing for 15 seconds.
5) Immediately spin in a pre-chilled fixed-angle centrifuge at 21,000 RCF for 15 minutes at 4°C.
6) Pipette clear aqueous layer (~650 ul) into a clean 1.5 ml LoBind tube.
7) Add an equal volume (650 ul) of 100% ethanol and mix well.
8) Pass all volume through a QIAgen RNeasy column using two centrifugation steps.
9) Follow the QIAgen RNeasy protocol and elute in 27 ul of elution buffer
10) Add 3 ul of 10x Turbo DNase Buffer and 1 ul of Turbo DNase enzyme.
11) Incubate at 37°C for 30 minutes.
12) Add 70 ul of RNase free water and 350 ul of QIAgen RLT Buffer and mix well.
13) Add 250 ul 100% ethanol and mix well.
14) Apply to column. Wash 2x with RPE. Elute in 20 ul RNase-free water.

Stock Buffers

All stock solutions should be filtered using a 0.22 um PVDF filter system. All solutions except for the 50% Iodixanol solution are stable at 4°C for at least 6 months.

1.034x Homogenization Buffer Stable Solution		For 200 ml stock solution		
Stock	*Name*	*Final Conc.*	*Fold Dilution (x)*	*Total Vol. (ul)*
1	M Sucrose	0.26	3.87	51706.50
2	M KCl	0.03	77.36	2585.33
1	M MgCl2	0.01	193.40	1034.13
0.75	M Tricine-KOH pH 7.8	0.02	36.26	5515.36
-	Water	-	-	139158.69
			Total Vol. (ul)	200000.00

Diluent Buffer		For 100 ml stock solution		
Stock	*Name*	*Final Conc.*	*Fold Dilution (x)*	*Total Vol. (ul)*
2	M KCl	0.15	13.33	7500.00
1	M MgCl2	0.03	33.33	3000.00
0.75	M Tricine-KOH, pH 7.8	0.12	6.25	16000.00
-	Water	-	-	73500.00
			Total Vol. (ul)	100000.00

50% Iodixanol Solution		For 50 ml stock solution		
Stock	*Name*	*Final Conc.*	*Fold Dilution (x)*	*Total Vol. (ul)*
-	Diluent Buffer	1	-	8333.33
60	% Iodixanol	50	1.20	41666.67
**Remake monthly for stability			Total Vol. (ul)	50000.00

ATAC-RSB Buffer		For 500 ml stock solution		
Stock	*Name*	*Final Conc.*	*Fold Dilution (x)*	*Total Vol. (ul)*
1	M Tris-HCl pH 7.5	0.01	100.00	5000.00
5	M NaCl	0.01	500.00	1000.00
1	M MgCl2	0.003	333.33	1500.00
-	Water	-	-	492500.00
			Total Vol. (ul)	500000.00

1M Sucrose		For 300 ml stock solution		
Stock	*Name*	*Final Conc.*	*Fold Dilution (x)*	*Total*
-	Sucrose (Powder)	1000	-	102.69 g
	H2O			235.5 ml
			Total Vol. (ul)	300000.00

Same Day Buffers – should be prepared fresh each day

**Note – cOmplete Protease Inhibitors come as tablets. It is difficult to use less than 1/2 tablet so we prepare the 1x Homogenization Buffer Unstable Solution in batches of 12 as outlined below.

1x Homogenization Buffer Unstable Solution				
Stock	*Name*	*Final Conc.*	*Fold Dilution (x)*	*Vol per 12 samp. (ul)*
1.0341	x HB Stable Solution	1	1.03	24175.00
1	M DTT	0.001	1000.00	25.00
500	mM Spermidine	0.5	1000.00	25.00
150	mM Spermine	0.15	1000.00	25.00
10	% NP40	0.3	33.33	750.00
-	cOmplete Protease Inhibitor	-	-	0.50 Tablets
			Total Volume (ul)	25000.00

30% Iodixanol Solution				
Stock	*Name*	*Final Conc.*	*Fold Dilution (x)*	*Vol per sample (ul)*
-	1x Homog. Buffer Unstable	-	-	240.00
50	% Iodixanol Solution	30	1.67	360.00
			Total Volume (ul)	600.00

40% Iodixanol Solution				
Stock	*Name*	*Final Conc.*	*Fold Dilution (x)*	*Vol per sample (ul)*
-	1x Homog. Buffer Unstable	-	-	120.00
50	% Iodixanol Solution	40	1.25	480.00
			Total Volume (ul)	600.00

ATAC-RSB-Tween Buffer				
Stock	*Name*	*Final Conc.*	*Fold Dilution (x)*	*Vol per sample (ul)*
-	ATAC-RSB	-	-	2970.00
10	% Tween-20	0.1	100.00	30.00
			Total Volume (ul)	3000.00

ATAC-seq Rxn Mix	
Reagent	*Vol per sample (ul)*
H2O	5
PBS	16.5
2x TD	25
1% Digitonin	0.5
10% Tween-20	0.5
Tn5	2.5

Order List

Item	Supplier	Cat Number
Eppendorf 2 ml Lo-Bind tubes	Sigma	Z666556-250EA
Eppendorf 1.5 ml Lo-Bind tubes	Sigma	Z666548-250EA
Nunc cryovials	Thermo	375418PK
Iodixanol (comes at 60%)	Sigma	D1556-250ML
Sucrose	Sigma	S7903-250G
NP40	Roche (Sigma)	11332473001
Tricine	Sigma	T0377-25G
Potassium Hydroxide (KOH)	Sigma	P5958-250G
cOmplete Protease Inhibitors	Roche	11697498001
MgCl2	Ambion (Thermo)	AM9530G
KCl	Ambion (Thermo)	AM9640G
DTT	Thermo	R0861
Spermidine	Sigma	S2501
Spermine	Sigma	S3256-1G
70 um Flowmi cell strainers	Fisher	03-421-228
70 um bucket-style cell strainers	BD Falcon	352350
Tris-HCl pH 7.5	Invitrogen	15567-027
NaCl	Ambion (Thermo)	AM9759
Tween 20	Roche (Sigma)	11332465001
H2O	Invitrogen	10977-015
Dounce Tissue Grinder Set	Sigma	D8938-1SET
INCYTO Disposable hemocytometers	Fisher	22-600-100
BAM Banker	Wako Chemicals	302-14681
RiboLock	Thermo	EO0384
0.22 um PVDF Filter Units (500 ml)	Millipore	SCGVU05RE
0.22 um PVDF Filter Units (50 ml)	Millipore	SE1M179M6

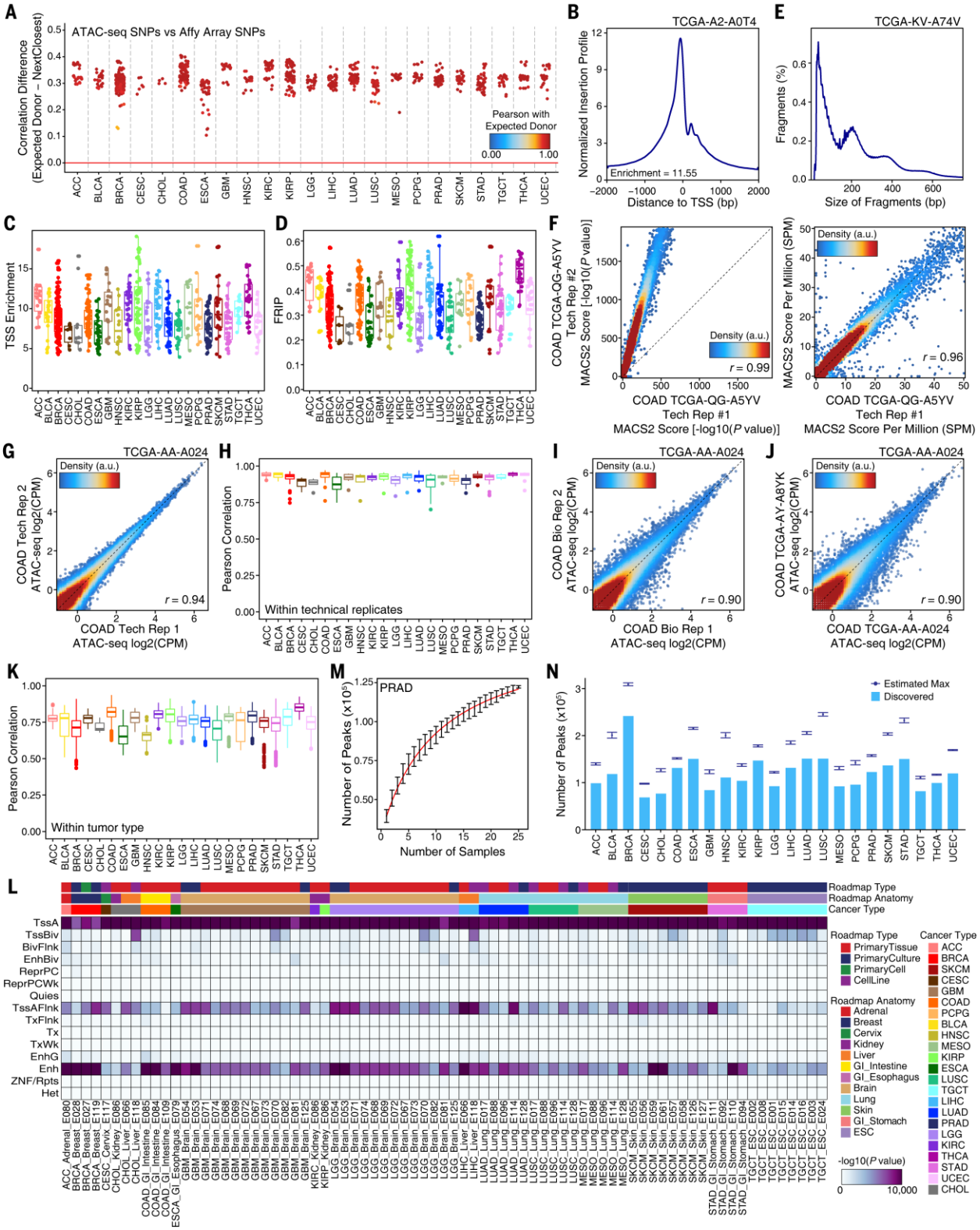

Fig. S1. Pan-cancer ATAC-seq data from frozen tissues is high quality and internally robust. (**A**) Genotype correlations between ATAC-seq-derived genotype and SNP array-derived genotype. Color represents the correlation of the sample with the expected donor. Correlation with next

closest match is derived from correlating with all other 11,126 donors profiled by SNP array by TCGA. Samples that match their expected donor better than all other donors have a correlation difference value above zero (red line). (**B**) Enrichment of ATAC-seq accessibility near transcription start sites in a representative sample (TCGA-A2-A0T4). The number of Tn5 insertions per base pair at regions +/- 2000 bp from the transcription start site is normalized to the number of insertions between +/- 1900-2000 bp from the transcription start site. (**C**) Box and whisker plots of the transcription start site enrichment values of all replicates of each cancer type. Each dot represents an individual technical replicate. The hinges represent the 25^{th} to 75^{th} percentile. (**D**) Box and whisker plots of the fraction of reads in peaks (FRIP) for all replicates of each cancer type. Each dot represents an individual technical replicate. The hinges represent the 25^{th} to 75^{th} percentile. (**E**) Distribution of fragment sizes for a representative sample (TCGA-KV-A7AV) from aligned, filtered and deduplicated reads. (**F**) Dot plot of the MACS2 peak call scores (-log10(p-value)) for all peaks across two technical replicates sequenced at unequal depth before (left) and after (right) "score-per-million" normalization. Each dot represents an individual peak. (**G**) Correlation plot of ATAC-seq accessibility in peaks from the pan-cancer peak set in two technical replicates of COAD sample TCGA-AA-A024. Color represents the density of points on the graph. Each dot represents an individual peak. (**H**) Box and whisker plots of the Pearson correlations between all technical replicates in each cancer type. The hinges represent the 25^{th} to 75^{th} percentile. Outliers are shown as individual dots. (**I**) Correlation plot of ATAC-seq accessibility in peaks from the pan-cancer peak set in two different tissue samples isolated from the same primary tumor from the same donor (TCGA-AA-A024). Color represents the density of points on the graph. Each dot represents an individual peak. (**J**) Correlation plot of ATAC-seq accessibility in peaks from the pan-cancer peak set in two biological donors (TCGA-AA-A024 and TCGA-AY-A8YK). Color represents the density of points on the graph. Each dot represents an individual peak. (**K**) Box and whisker plots of the Pearson correlations between all samples (N x N) from a given cancer type. The hinges represent the 25^{th} to 75^{th} percentile. Outliers are shown as individual dots. (**L**) Heatmap of the -log10(p-value) of the enrichment of pan-cancer ATAC-seq peaks within ChIP-seq-defined chromHMM states from the Roadmap Epigenomics Project. Roadmap tissue/cell type, Roadmap anatomy, and cancer type are all depicted by color across the top. Row abbreviations as follows: TssA – active TSS, TssBiv – bivalent/poised TSS, EnhBiv – bivalent enhancer, ReprPC – repressed polycomb, ReprPCWk – weak repressed polycomb, Quies – quiescent/low, TssAFlnk – flanking active TSS, TxFlnk – transcription at gene 5' and 3', Tx – strong transcription, TxWk – weak transcription, EnhG – genic enhancers, Enh – enhancers, ZNF/Rpts – ZNF genes and repeats, Het - heterochromatin. (**M**) In silico modeling of the number of peaks discovered with the addition of more samples (saturation curves) in prostate adenocarcinoma. Error bars represent the standard deviation of N=25 simulations. (**N**) Bar plot of the number of ATAC-seq peaks discovered in each cancer type in this study compared to the estimated number of peaks that would be discovered at saturation.

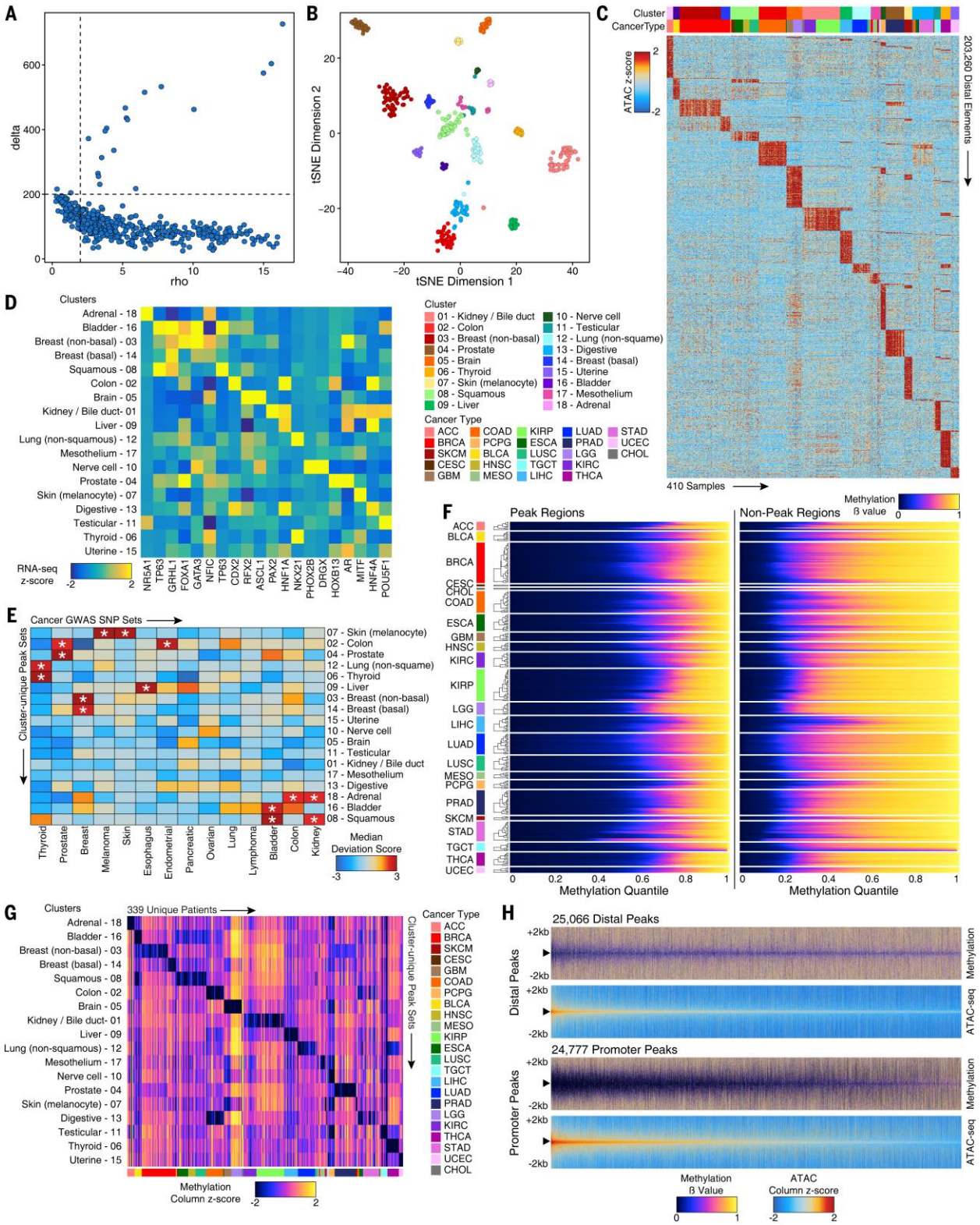

Fig. S2. Pan-cancer ATAC-seq identifies 18 distinct sample clusters. (**A**) Dot plot showing the decision metrics used for density clustering methodology. Rho refers to the number of points that are closer than the cutoff distance to a given point. Delta refers to the distance between a given

point and any other point with higher density. Points that are relatively high in both rho and delta represent putative cluster centers. The chosen values in this study were determined to be rho = 2 and delta = 200. (**B**) t-SNE as in Figure 2B but color represents the predicted cluster. (**C**) Heatmap showing the ATAC-seq accessibility at cluster-specific distal elements (N=203,260) across all 410 samples. Color represents the ATAC-seq z-score of each peak (rows). Cluster and Cancer Type colors shown in legend to the left. Each column represents an individual sample. (**D**) Heatmap showing the RNA-seq gene expression of transcription factors predicted to be drivers of cluster-specific identity. Transcription factors shown are the same as in Figure 3C. Color represents the RNA-seq z-score of log2(TPM+1) for each transcription factor corresponding to the motif (columns). (**E**) Heatmap showing the ATAC-seq variability at known GWAS polymorphisms within the 203,260 cluster-specific peaks. Color represents the median deviation score from chromVAR for the samples of each cluster. An asterisk denotes significant deviation from expectation with a median FDR < 0.1 within the cluster. (**F**) Heatmap of methylation quantiles across cancer types in peak regions (left) and non-peak regions (right). Color depicts the methylation beta value of each quantile. (**G**) Methylation beta value of peak sets identified in Figure 3B across 339 unique donors with matched 450K methylation array data. Each column represents an individual donor. Color represents the average methylation beta value in the given cluster-specific peak set for that donor. (**H**) Heatmap representation of methylation beta value from WGBS of a single BLCA sample (TCGA-BL-A13J) in distal (top) and promoter (bottom) peaks.

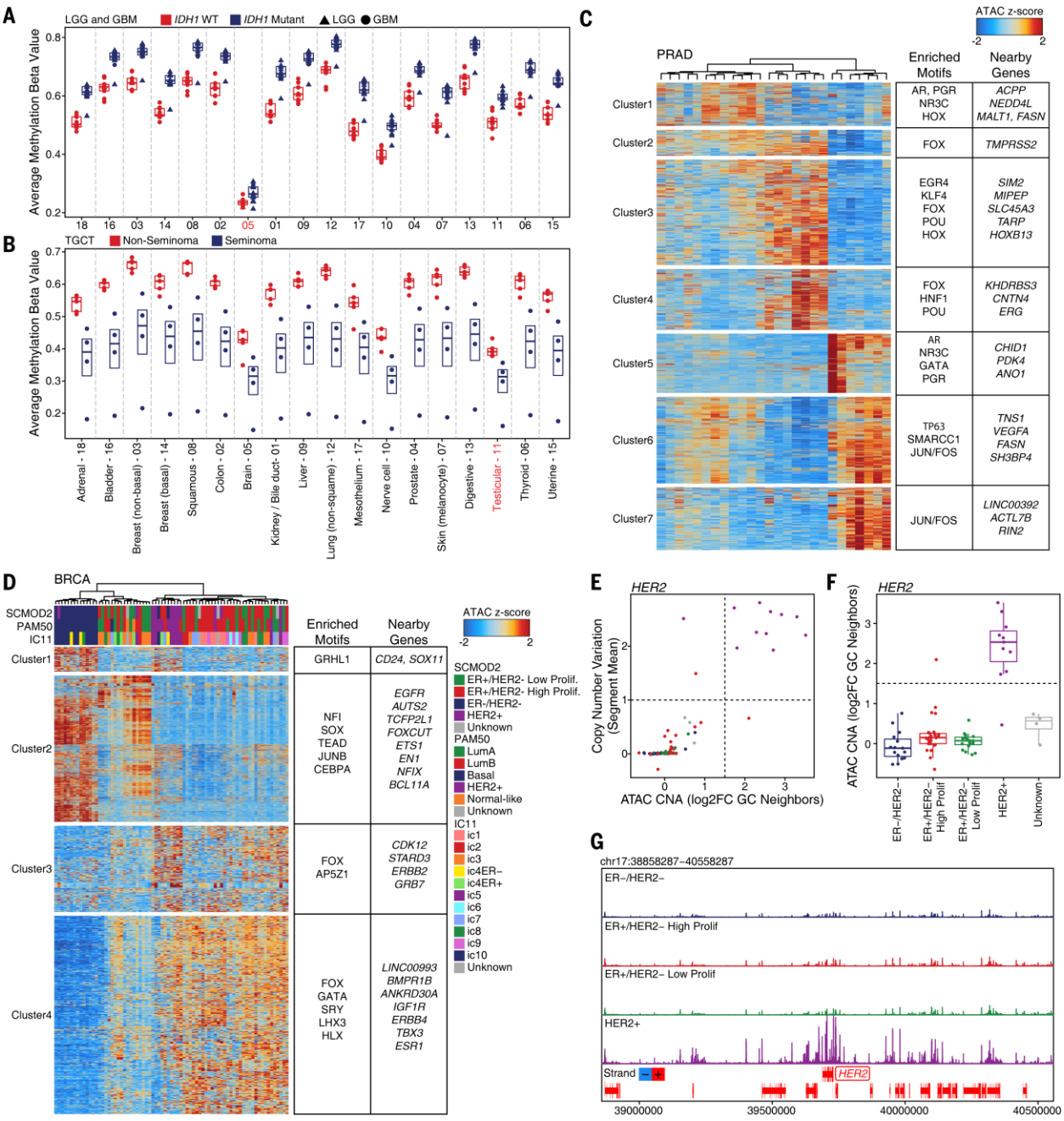

Fig. S3. Subtype-specific clusters identified from unsupervised analysis of ATAC-seq data. (**A**) Box and whisker plot of average methylation beta value for each of the cluster-specific peak sets identified in Figure 3B across all LGG and GBM samples. Color indicates the *IDH1* mutation status for each sample and shape indicates whether this sample is from the LGG or GBM cohort. Cluster names are shown below Figure S3B. Each dot represents an individual donor. The hinges represent the 25th to 75th percentile. (**B**) Box and whisker plot of average methylation beta value for each of the cluster-specific peak sets identified in Figure 3B across all TGCT samples. Color indicates if the sample is annotated as "seminoma" or "non-seminoma." Each dot represents an

individual donor. The hinges represent the 25th to 75th percentile. (**C**) Heatmap of ATAC-seq accessibility across all prostate cancer samples profiled in this study. Clusters of peaks are defined by k-means clustering followed by manual grouping. Enriched motifs in each cluster and nearby genes are shown to the right. Heatmap color represents the z-score of the given peak across each sample. (**D**) Same as Figure S3C but showing breast cancer. Various published classification schemes (SCMOD2, PAM50, IC11) are shown colorimetrically across the top. (**E**) Dot plot of DNA copy number amplification signal derived from ATAC-seq data (2-Mbp windows) and published DNA copy number array data. Dots are colored according to the SCMOD2 scheme as shown in Figure S3D. Each dot represents an individual donor. (**F**) Box and whisker plot showing the distribution of ATAC-seq-based copy number calls at the *HER2* locus (chr17:38000000-40000000) across all breast cancer samples profiled in this study. Boxes and dots are colored according to the SCMOD2 scheme as shown in Figure S3D. Each dot represents an individual donor. (**G**) Normalized ATAC-seq sequencing tracks at the *HER2* gene locus. Tracks represent the average of all samples from the given groups. Region represents chr17:38858287−40558287.

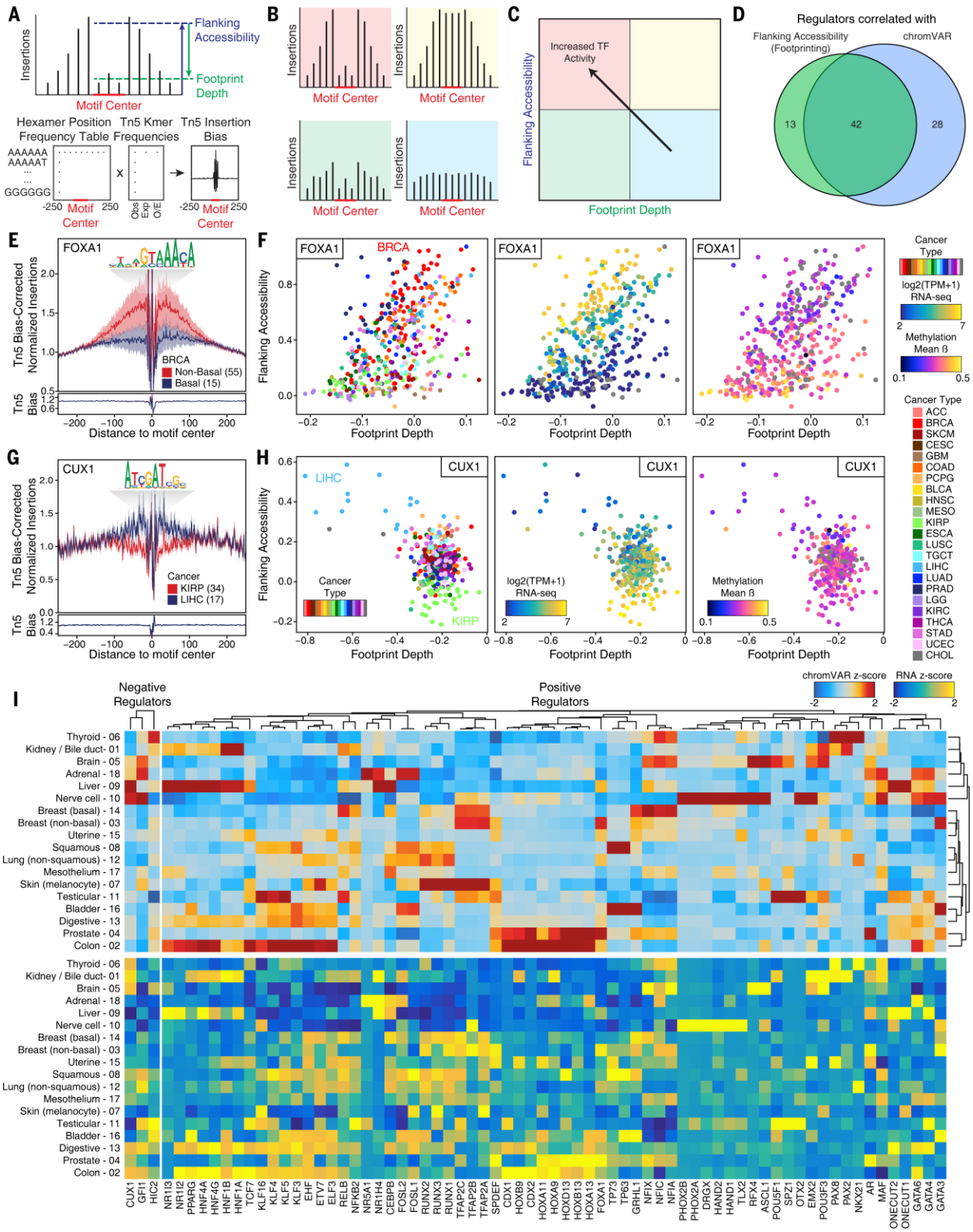

Fig. S4. Dynamics of transcription factor footprinting in pan-cancer ATAC-seq data. (**A**) Schematic illustrating how transcription factor footprints are determined. (**B**) Schematic illustrating the four predominant patterns observed in transcription factor footprints. (**C**) Schematic

illustrating the dynamic relationship between footprint depth and flanking accessibility for the four classes of footprints shown in Figure S4B. See methods for limitations. (**D**) Venn diagram showing the overlap of TFs whose expression is found to be correlated with flanking accessibility and TFs whose expression is found to be correlated with chromVAR GC bias-corrected deviation. (**E**) Transcription factor footprinting of the FOXA1 motif (CIS-BP M4566_1.02) in basal and non-basal breast cancer samples as defined by PAM50 classification. The Tn5 insertion bias track of FOXA1 motifs is shown below. (**F**) Dot plots showing the footprint depth and flanking accessibility of FOXA1 motifs across all samples studied. Each dot represents a unique sample. Color represents cancer type (left), RNA-seq gene expression (middle), or average methylation beta value (right). Samples without matching RNA or methylation data are shown in grey. (**G**) Transcription factor footprinting of the negative regulator CUX1 motif (CIS-BP M3021_1.02) in kidney renal papillary cell carcinoma and liver hepatocellular carcinoma samples. The Tn5 insertion bias track of CUX1 motifs is shown below. (**H**) Dot plots showing the footprint depth and flanking accessibility of CUX1 motifs across all samples studied. Each dot represents a unique sample. Color represents cancer type (left), RNA-seq gene expression (middle), or average methylation beta value (right). Samples without matching RNA or methylation data are shown in grey. (**I**) Heatmap representation of ATAC-seq chromVAR bias-corrected deviations (top) and RNA-seq gene expression (bottom) in all transcription factors predicted to have significant correlations (FDR < 0.1) of chromVAR bias-corrected deviations with gene expression. Color of each heatmap represents the z-score of chromVAR bias-corrected deviations for ATAC-seq or log2(TPM+1) for RNA-seq. Clustering order is dictated by the ATAC-seq heatmap.

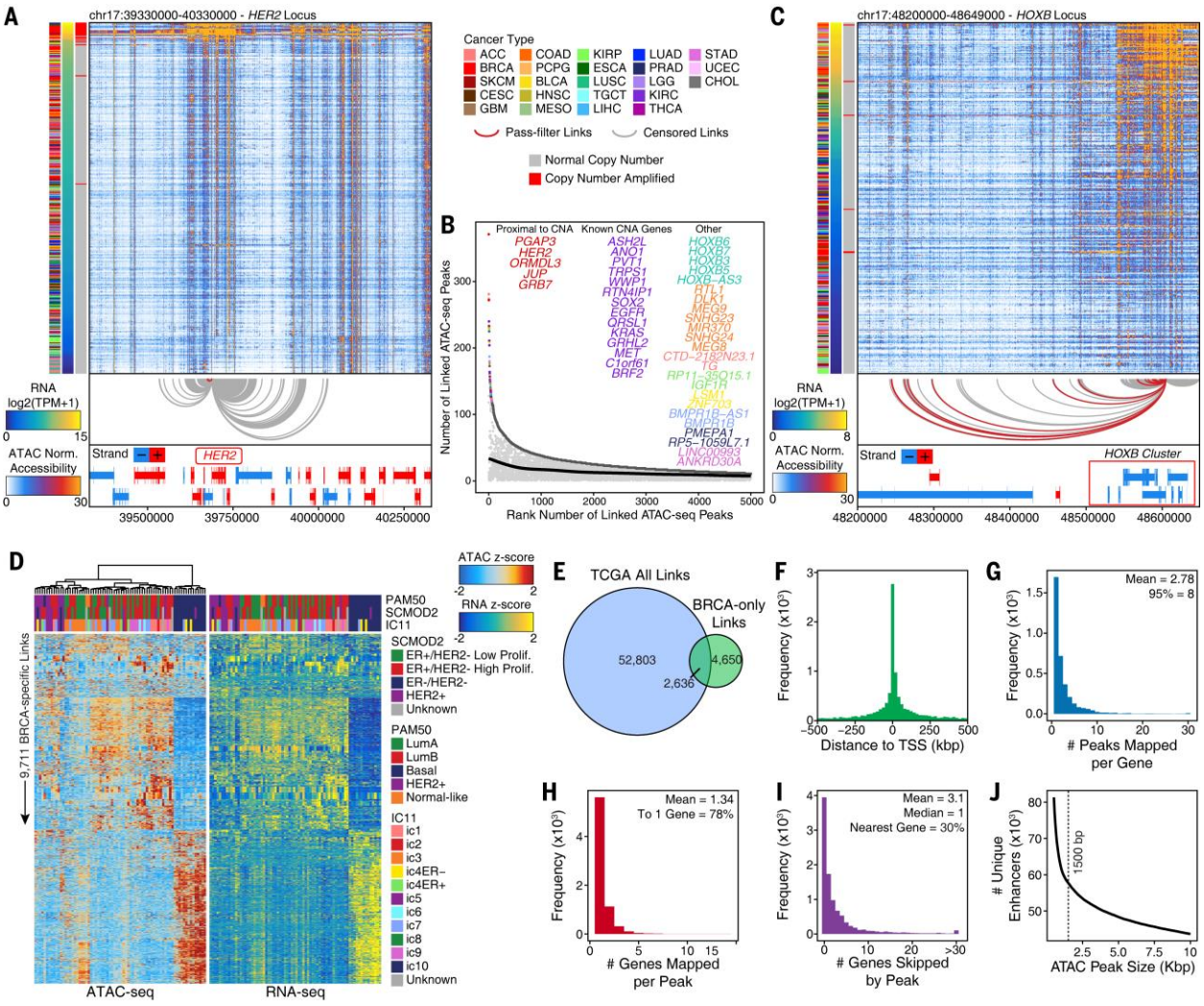

Fig. S5. Linking of ATAC-seq peaks with putative gene targets. (**A**) Heatmap representation of the ATAC-seq accessibility at the *HER2* locus across all samples profiled in this study. Arcs below the heatmap depict putative peak-to-gene links. Grey loops indicate spurious links that are filtered because they were found to be related to copy number amplification. Red loops indicate peak-to-gene links retained after filtering. Color bars to the left depict the cancer type, RNA-seq gene expression, and presence or absence of DNA copy number amplification of *HER2* for each sample. Colors described in the legend to the right. (**B**) Dot plot showing the number of ATAC-seq peaks linked to each gene before (colored dots) and after "CNA" and "diffuse" correction (light-grey dots). Only the top 5,000 genes are shown. Genes with many linked ATAC-seq peaks often are found in regions of recurrent copy number amplification or high local co-accessibility. Colors of gene names and dots represent groups of genes and are ordered according to the number of linked ATAC-seq peaks (y-axis). (**C**) Heatmap representation of the ATAC-seq accessibility at the *HOXB* locus across all samples profiled in this study. Arcs below the heatmap depict putative peak-to-gene links. Grey loops indicate links that are filtered out because they have high local correlation (aka "diffuse"). Red loops indicate peak-to-gene links retained after filtering. Color

bars to the left depict the cancer type, RNA-seq gene expression, and presence or absence of DNA copy number amplification for each sample. Colors described in the legend to the left of the figure. (**D**) Heatmap representation of the 9,711 unique peak-to-gene links predicted solely from breast cancer data. Each row represents an individual link between one ATAC-seq peak and one gene. Color represents the relative ATAC-seq accessibility (left) or RNA-seq gene expression (right) for each link. Published breast cancer classification schemes are shown above each plot. Clustering order is dictated by the ATAC-seq heatmap. (**E**) Venn diagram of the overlap of the peak-to-gene links predicted using all pan-cancer samples and the peak-to-gene links identified using only breast cancer data. Peak-to-gene links were de-duplicated at 5-kbp resolution. (**F**) Distribution of the distance of each peak to the transcription start site of the linked gene in breast cancer peak-to-gene links. (**G**) Distribution of the number of peaks mapped per gene in breast cancer peak-to-gene links. (**H**) Distribution of the number of genes mapped per peak in breast cancer peak-to-gene links. (**I**) Distribution of the number of genes "skipped" by a peak in order to reach its predicted linked gene in breast cancer ATAC-seq peak-to-gene links. (**J**) Line plot of the number of unique "enhancer units" identified (if multiple ATAC-seq peaks overlap they are denoted as the same enhancer unit) as the size of ATAC-seq peaks is extended. Based on this, we determined that extending ATAC-peaks to 1500bp enables us to classify enhancer units.

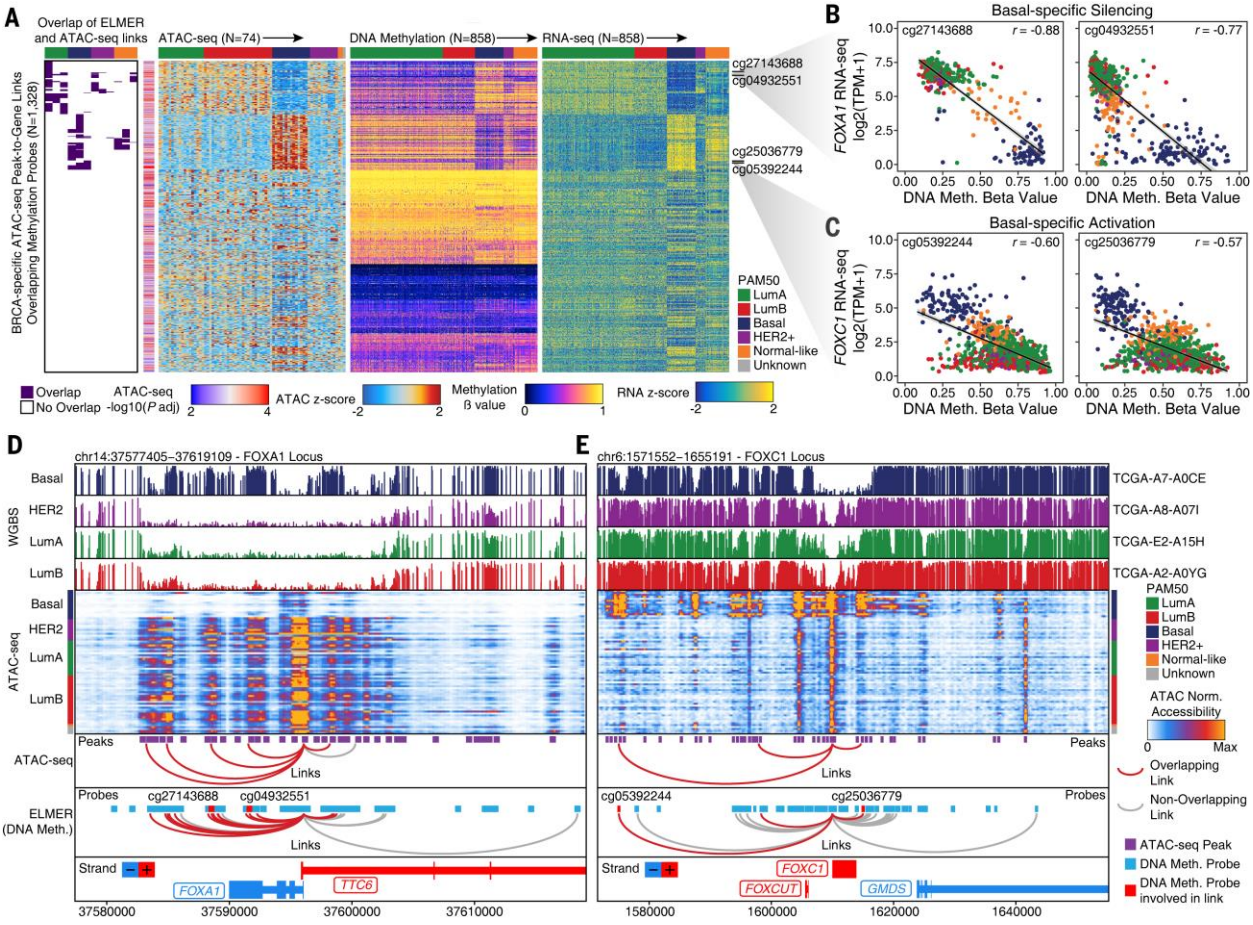

Fig. S6. Validation of predicted peak-to-gene links using ELMER. (**A**) Analysis of 1,328 BRCA ATAC-seq-based peak-to-gene links that are within a +/- 250-bp region of a distal Illumina 450K DNA methylation array probe. Each row represents an individual ATAC-seq-based peak-to-gene link. From left to right, heatmap showing the overlap between each ATAC-seq peak-to-gene link and PAM50 subtype-specific methylation probe-to-gene links, generated from all TCGA BRCA samples with methylation and expression data (N=858) using ELMER. Rows are ordered by the ELMER subtype for the ATAC-seq peak-to-gene links overlapping at least one ELMER link (N=464, 34.9%) and then by ATAC-seq accessibility (far left). Heatmap representation of ATAC-seq accessibility across all samples profiled in this study (N=74) (middle left). Side bar shows the –log10(adjusted p-value) of the ATAC-seq peak-to-gene link. Heatmap representation of the DNA methylation beta value for all TCGA BRCA samples used in the ELMER analysis (middle right). Heatmap representation of the RNA-seq gene expression for all TCGA BRCA samples used in the ELMER analysis (far right). (**B**) Scatter plot showing the correlation of DNA methylation at two probes (cg27143688 (left) and cg04932551 (right)) linked by ELMER to the basal-specific silencing of the *FOXA1* gene. Both links were also identified in the ATAC-seq BRCA-specific peak-to-gene links (labeled in Figure S6A). Each dot represents an individual donor. (**C**) Scatter plot showing the correlation of DNA methylation at two probes (cg05392244 (left) and cg25036779 (right)) linked by ELMER to the basal-specific activation of the *FOXC1*

gene. Both links were also identified in the ATAC-seq BRCA-specific peak-to-gene links (labeled in Figure S6A). Each dot represents an individual donor. (**D**) Normalized sequencing tracks and predicted links showing overlapping and non-overlapping link calls from ELMER and ATAC-seq in the *FOXA1* locus. WGBS tracks are shown at the top for a single representative sample of each of the four main PAM50 subtypes. The heatmap shows the ATAC-seq accessibility at this locus across all BRCA donors (N=74) profiled in this study, ordered by PAM50 subtype. Below the heatmap are ATAC-seq peaks (purple), DNA methylation probes (light blue) and links (shown as arcs). Red arcs represent links that are predicted both by ATAC-seq and DNA methylation. Grey arcs represent links that are only predicted by one data type. Specific links shown in Figure S6A and S6B (cg27143688 and cg04932551) are labeled. (**E**) Normalized sequencing tracks and predicted links showing overlapping and non-overlapping link calls from ELMER and ATAC-seq in the *FOXC1* locus. WGBS tracks are shown at the top for a single representative sample of each of the four main PAM50 subtypes. The heatmap shows the normalized ATAC-seq accessibility at this locus across all BRCA donors (N=74) profiled in this study, ordered by PAM50 subtype. Below the heatmap are ATAC-seq peaks (purple), DNA methylation probes (light blue) and links (shown as arcs). Red arcs represent links that are predicted both by ATAC-seq and DNA methylation. Grey arcs represent links that are only predicted by one data type. Specific links shown in Figure S6A and S6C (cg05392244 and cg25036779) are labeled.

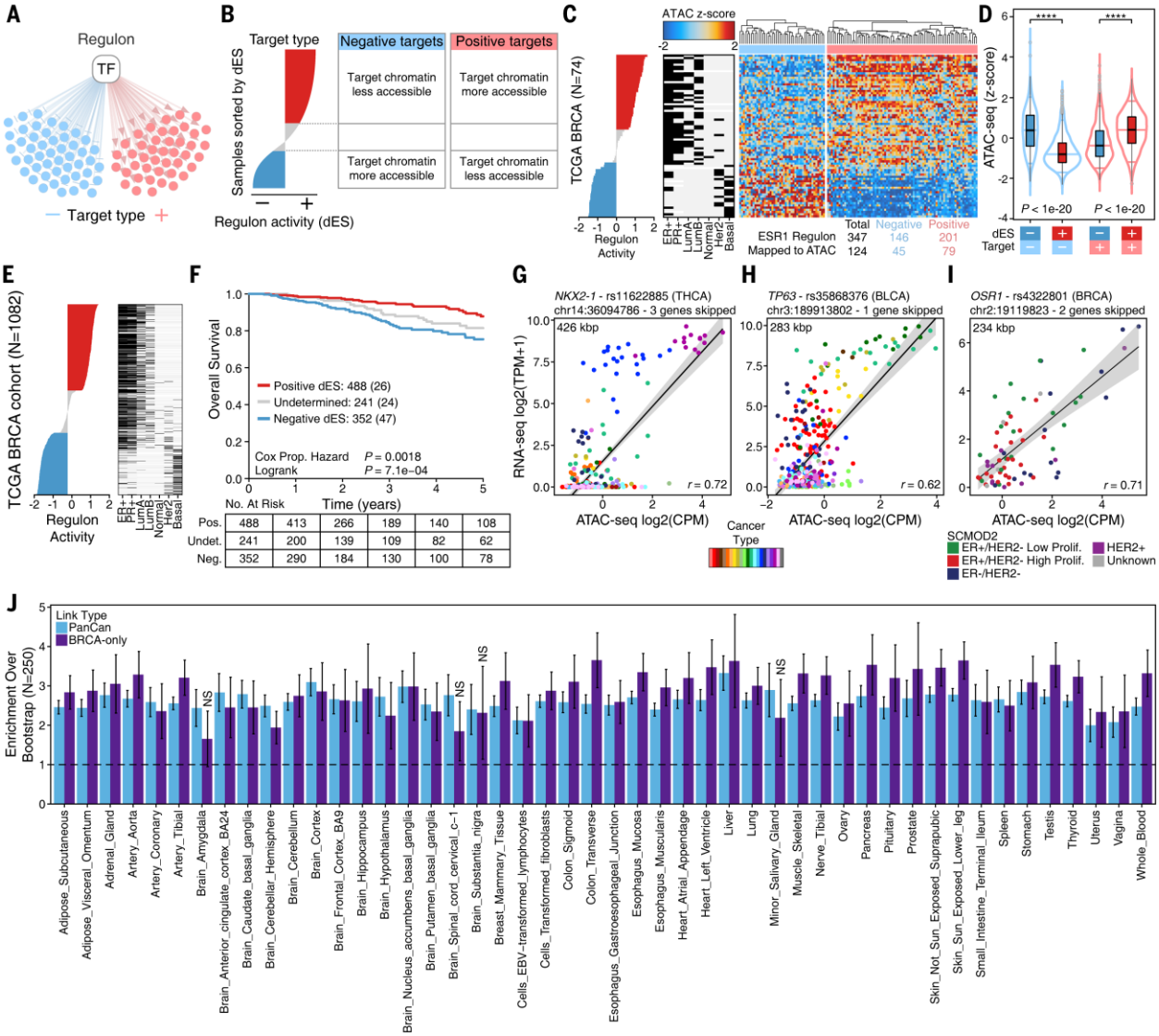

Fig. S7. Validation of predicted peak-to-gene links via regulon analysis and eQTLs. (**A**) Schematic of a regulon showing a transcription factor regulator and sets of negative and positive target genes. (**B**) Schematic of expectations for chromatin accessibility changes for a regulator's negative and positive target genes. A red (vs. blue) color indicates that we expect target genes to be associated with chromatin that is more (vs. less) accessible. (**C**) From left to right: sorted regulon activity profile for the ESR1 regulator in all BRCA donors profiled in this study (N=74); sorted ER, PR, and PAM50 subtypes; sorted chromatin accessibility heatmap, with color representing the ATAC-seq z-score for each BRCA-specific distal peak that is linked to a regulon target gene (column). (**D**) Distribution of chromatin accessibility z-scores for BRCA-specific distal peaks linked to negative and positive target genes and samples with dES greater than or less than 0. P-value determined by t-test. (**E**) Sorted regulon activity profile for the ESR1 regulator in all TCGA donors (N=1082). Covariates as in Figure S7C. (**F**) Kaplan-Meier plot for each dES group shown in Figure S7C. (**G**) Dot plot showing the correlation of ATAC-seq accessibility and RNA-seq gene expression for all samples profiled in this study for the peak-to-gene link predicted

between SNP rs11622885 and the *NKX2-1* gene. Each dot represents an individual donor. (**H**) Dot plot showing the correlation of ATAC-seq accessibility and RNA-seq gene expression for all samples profiled in this study for the peak-to-gene link predicted between SNP rs35868376 and the *TP63* gene. Each dot represents an individual donor. (**I**) Dot plot showing the correlation of ATAC-seq accessibility and RNA-seq gene expression for all BRCA samples profiled in this study for the peak-to-gene link predicted between SNP rs4322801 and the *OSR1* gene. Each dot represents an individual sample. (**J**) Bar plot showing the enrichment of overlap between ATAC-seq-defined peak-to-gene links and GTEx eQTLs. Error bars represent the standard deviation of the enrichment based on 250 randomly permuted peak-to-gene link sets. Unless marked as not significant ("NS"), all tests showed significance as defined by FDR < 0.001.

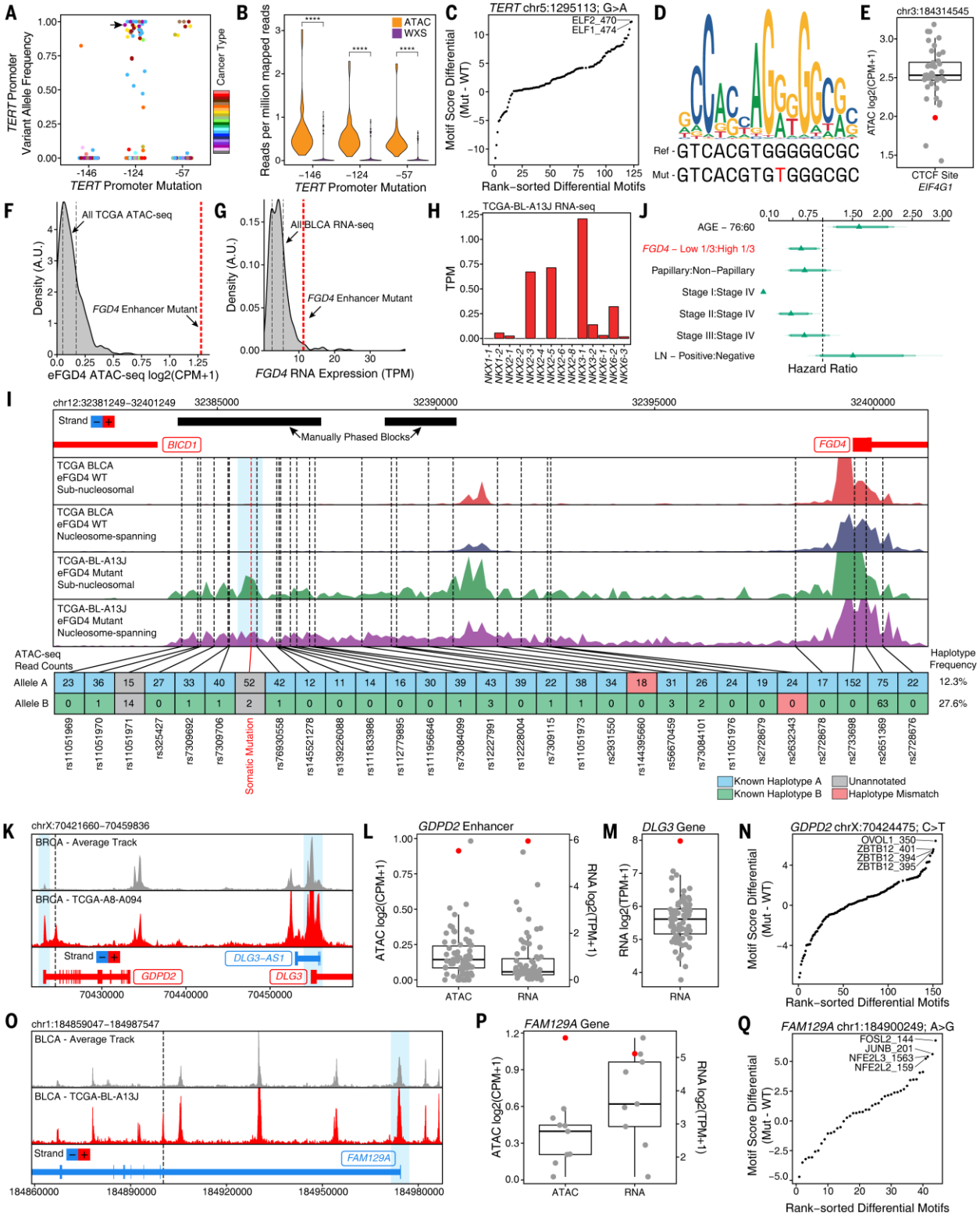

Fig. S8. Intersection of WGS and ATAC-seq data. (**A**) Variant allele frequency from ATAC-seq data at positions in the *TERT* gene promoter that have previously been shown to harbor activating mutations. Arrow indicates the same sample shown in Figure 7B (blue dot). Each dot represents an individual sample and the color indicates the cancer type using the same colors as

used throughout. (**B**) Reads per million mapped reads in ATAC-seq and whole exome sequencing (WXS) at positions in the *TERT* gene promoter that have previously been shown to harbor activating mutations. ****$P < 10^{-7}$ by two-tailed t-test. (**C**) Difference in motif score in the wildtype and mutant *TERT* promoter region associated with the -146 mutation (chr5:1295135). Motif score represents the degree of similarity between the sequence of interest and the relevant motif. (**D**) Overlay of the CTCF motif (CIS-BP M4502_1.02) and the wildtype and mutant sequences of the CTCF motif site near the *EIF4G1* gene. (**E**) Box plot showing the distribution of ATAC-seq accessibility at the mutated position shown in Figure S8D. Gray dots represent samples that are wildtype for this position. Red dot represents the single colon cancer sample with a mutation in this CTCF site. (**F**) Distribution of ATAC-seq accessibility across all samples profiled in this study (N=410) at the *FGD4* enhancer mutant locus (chr12:32385775). Red dotted line indicates the ATAC-seq accessibility observed in sample TCGA-BL-A13J which has a mutation at this position. Black dotted lines indicate the 25th (left) and 75th (right) percentile of ATAC-seq accessibility from all samples. (**G**) Distribution of RNA-seq gene expression of *FGD4* across all TCGA bladder cancer samples with available RNA-seq data (N=407). Red dotted line indicates the gene expression observed in sample TCGA-BL-A13J which has a mutation upstream of the FGD4 gene. Black dotted lines indicate the 25th (left) and 75th (right) percentile of RNA-seq gene expression from all TCGA BLCA samples. (**H**) Expression of NKX factors in sample TCGA-BL-A13J. (**I**) ATAC-seq sequencing tracks of the *FGD4* enhancer locus (chr12:32381249−32401249) in wildtype bladder cancer samples and the eFGD4 mutant bladder cancer (TCGA-BL-A13J). The eFGD4 mutation is denoted as "Somatic Mutation" below the plot. ATAC-seq fragments were split into sub-nucleosomal (less than 100 bp) and nucleosome spanning (greater than 160 bp) fragments and then scaled for equal representation (to account for fragment distribution differences) to show putative protein binding sites in the sub-nucleosomal fraction (blue highlighted box). All other nearby germline SNPs observed in sample TCGA-BL-A13J by ATAC-seq and WGS are shown by black dotted lines. ATAC-seq allele counts at each of these locations are shown in the boxes below the plot. Allele counts are divided into Allele A and Allele B based on manually phased blocks (shown as black bars) as well as known haplotypes from the 1000 genomes project (indicated by color). Polymorphisms that do not match the known haplotypes are shown as red boxes. (**J**) Hazard plot of risk of dying from BLCA based on multiple covariates including *FGD4* gene expression (HR=0.64, 95% confidence interval = 0.44 – 0.94). Lines represent 95% confidence intervals. (**K**) Normalized ATAC-seq sequencing tracks of the *GDPD2* locus in breast cancer samples including the one sample with a mutation in the intronic peak region (TCGA-A8-A094). The top track shows the average ATAC-seq accessibility across all samples excluding TCGA-A8-A094 which is shown in the bottom track. Locus shown represents chrX:70421660−70459836. The mutation position is indicated by a black dotted line. The promoter regions of the *GDPD2* and *DLG3* genes are highlighted by blue boxes. (**L**) ATAC-seq accessibility at the *GDPD2* intronic mutation location and RNA-seq expression of the *GDPD2* gene in all breast cancer samples profiled in this study. Red dot represents sample TCGA-A8-A094 which has a mutation in the *GDPD2* intronic locus. (**M**) RNA-seq expression of the *DLG3*

gene in the same samples shown in Figure S8L. (**N**) Difference in motif score in the wildtype and mutant *GDPD2* intronic region (chrX:70424475). Motif score represents the degree of similarity between the sequence of interest and the relevant motif. (**O**) Normalized ATAC-seq sequencing tracks of the *FAM129A* locus in bladder cancer samples including the one sample with a mutation in the intronic peak region (TCGA-BL-A13J). The top track shows the average ATAC-seq accessibility across all samples excluding TCGA-BL-A13J which is shown in the bottom track. Locus shown represents chr1:184859047−184987547. The mutation position is indicated by a black dotted line. The promoter region of the *FAM129A* gene is highlighted by a blue box. (**P**) ATAC-seq accessibility at the *FAM129A* intronic mutation location and RNA-seq expression of the *FAM129A* gene in all bladder cancer samples profiled in this study. Red dot represents sample TCGA-BL-A13J which has a mutation in the *FAM129A* intronic locus. (**Q**) Difference in motif score in the wildtype and mutant *FAM129A* intronic region (chr1:184900249). Motif score represents the degree of similarity between the sequence of interest and the relevant motif.

Captions for Data S1 to S10:

Data S1. (separate file)
Cancer types studied, donor characteristics, and sequencing statistics.

Data S2. (separate file)
Pan-cancer and breast cancer peak calls.

Data S3. (separate file)
Overlap of peaks with Roadmap DNase-seq, peak saturation analysis, and t-SNE positions of all samples.

Data S4. (separate file)
Distal binarization analysis and enrichment of motifs in cluster-specific peak sets.

Data S5. (separate file)
GWAS and eQTL analyses and overlap with peak-to-gene links.

Data S6. (separate file)
TF footprinting analyses and correlation to gene expression.

Data S7. (separate file)
Pan-cancer and breast cancer-specific peak-to-gene links and enhancer-to-gene links.

Data S8. (separate file)
ELMER and Regulon analyses.

Data S9. (separate file)
Peak-to-gene links related to immune response in cancer.

Data S10. (separate file)
Integration of ATAC-seq and WGS to identify noncoding mutations.

References and Notes

1. C. Hutter, J. C. Zenklusen, The Cancer Genome Atlas: Creating lasting value beyond its data. *Cell* **173**, 283–285 (2018). doi:10.1016/j.cell.2018.03.042 Medline

2. W. A. Flavahan, E. Gaskell, B. E. Bernstein, Epigenetic plasticity and the hallmarks of cancer. *Science* **357**, eaal2380 (2017). doi:10.1126/science.aal2380 Medline

3. D. Hanahan, R. A. Weinberg, Hallmarks of cancer: The next generation. *Cell* **144**, 646–674 (2011). doi:10.1016/j.cell.2011.02.013 Medline

4. M. Egeblad, E. S. Nakasone, Z. Werb, Tumors as organs: Complex tissues that interface with the entire organism. *Dev. Cell* **18**, 884–901 (2010). doi:10.1016/j.devcel.2010.05.012 Medline

5. W. Zhou, H. Q. Dinh, Z. Ramjan, D. J. Weisenberger, C. M. Nicolet, H. Shen, P. W. Laird, B. P. Berman, DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat. Genet.* **50**, 591–602 (2018). doi:10.1038/s41588-018-0073-4 Medline

6. M. Almamun, B. T. Levinson, A. C. van Swaay, N. T. Johnson, S. D. McKay, G. L. Arthur, J. W. Davis, K. H. Taylor, Integrated methylome and transcriptome analysis reveals novel regulatory elements in pediatric acute lymphoblastic leukemia. *Epigenetics* **10**, 882–890 (2015). doi:10.1080/15592294.2015.1078050 Medline

7. Y. He, D. U. Gorkin, D. E. Dickel, J. R. Nery, R. G. Castanon, A. Y. Lee, Y. Shen, A. Visel, L. A. Pennacchio, B. Ren, J. R. Ecker, Improved regulatory element prediction based on tissue-specific local epigenomic signatures. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E1633–E1640 (2017). doi:10.1073/pnas.1618353114 Medline

8. L. Yao, H. Shen, P. W. Laird, P. J. Farnham, B. P. Berman, Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Genome Biol.* **16**, 105 (2015). doi:10.1186/s13059-015-0668-3 Medline

9. M. Ceccarelli, F. P. Barthel, T. M. Malta, T. S. Sabedot, S. R. Salama, B. A. Murray, O. Morozova, Y. Newton, A. Radenbaugh, S. M. Pagnotta, S. Anjum, J. Wang, G. Manyam, P. Zoppoli, S. Ling, A. A. Rao, M. Grifford, A. D. Cherniack, H. Zhang, L. Poisson, C. G. Carlotti Jr., D. P. C. Tirapelli, A. Rao, T. Mikkelsen, C. C. Lau, W. K. A. Yung, R. Rabadan, J. Huse, D. J. Brat, N. L. Lehman, J. S. Barnholtz-Sloan, S. Zheng, K. Hess, G. Rao, M. Meyerson, R. Beroukhim, L. Cooper, R. Akbani, M. Wrensch, D. Haussler, K. D. Aldape, P. W. Laird, D. H. Gutmann, H. Noushmehr, A. Iavarone, R. G. W. Verhaak; TCGA Research Network, Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* **164**, 550–563 (2016). doi:10.1016/j.cell.2015.12.028 Medline

10. H. Noushmehr, D. J. Weisenberger, K. Diefes, H. S. Phillips, K. Pujara, B. P. Berman, F. Pan, C. E. Pelloski, E. P. Sulman, K. P. Bhat, R. G. W. Verhaak, K. A. Hoadley, D. N. Hayes, C. M. Perou, H. K. Schmidt, L. Ding, R. K. Wilson, D. Van Den Berg, H. Shen, H. Bengtsson, P. Neuvial, L. M. Cope, J. Buckley, J. G. Herman, S. B. Baylin, P. W. Laird, K. Aldape; Cancer Genome Atlas Research Network, Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* **17**, 510–522 (2010). doi:10.1016/j.ccr.2010.03.017 Medline

11. T. Hinoue, D. J. Weisenberger, C. P. E. Lange, H. Shen, H.-M. Byun, D. Van Den Berg, S. Malik, F. Pan, H. Noushmehr, C. M. van Dijk, R. A. E. M. Tollenaar, P. W. Laird, Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res.* **22**, 271–282 (2012). doi:10.1101/gr.117523.110 Medline

12. P. A. Northcott, I. Buchhalter, A. S. Morrissy, V. Hovestadt, J. Weischenfeldt, T. Ehrenberger, S. Gröbner, M. Segura-Wang, T. Zichner, V. A. Rudneva, H.-J. Warnatz, N. Sidiropoulos, A. H. Phillips, S. Schumacher, K. Kleinheinz, S. M. Waszak, S. Erkek, D. T. W. Jones, B. C. Worst, M. Kool, M. Zapatka, N. Jäger, L. Chavez, B. Hutter, M. Bieg, N. Paramasivam, M. Heinold, Z. Gu, N. Ishaque, C. Jäger-Schmidt, C. D. Imbusch, A. Jugold, D. Hübschmann, T. Risch, V. Amstislavskiy, F. G. R. Gonzalez, U. D. Weber, S. Wolf, G. W. Robinson, X. Zhou, G. Wu, D. Finkelstein, Y. Liu, F. M. G. Cavalli, B. Luu, V. Ramaswamy, X. Wu, J. Koster, M. Ryzhova, Y.-J. Cho, S. L. Pomeroy, C. Herold-Mende, M. Schuhmann, M. Ebinger, L. M. Liau, J. Mora, R. E. McLendon, N. Jabado, T. Kumabe, E. Chuah, Y. Ma, R. A. Moore, A. J. Mungall, K. L. Mungall, N. Thiessen, K. Tse, T. Wong, S. J. M. Jones, O. Witt, T. Milde, A. Von Deimling, D. Capper, A. Korshunov, M.-L. Yaspo, R. Kriwacki, A. Gajjar, J. Zhang, R. Beroukhim, E. Fraenkel, J. O. Korbel, B. Brors, M. Schlesner, R. Eils, M. A. Marra, S. M. Pfister, M. D. Taylor, P. Lichter, The whole-genome landscape of medulloblastoma subtypes. *Nature* **547**, 311–317 (2017). doi:10.1038/nature22973 Medline

13.The Cancer Genome Atlas Research Network, Comprehensive molecular characterization of papillary renal-cell carcinoma. *N. Engl. J. Med.* **374**, 135–145 (2016). doi:10.1056/NEJMoa1505917 Medline

14. B. Akhtar-Zaidi, R. Cowper-Sal·lari, O. Corradin, A. Saiakhova, C. F. Bartels, D. Balasubramanian, L. Myeroff, J. Lutterbaugh, A. Jarrar, M. F. Kalady, J. Willis, J. H. Moore, P. J. Tesar, T. Laframboise, S. Markowitz, M. Lupien, P. C. Scacheri, Epigenomic enhancer profiling defines a signature of colon cancer. *Science* **336**, 736–739 (2012). doi:10.1126/science.1217277 Medline

15. H. Chen, C. Li, X. Peng, Z. Zhou, J. N. Weinstein, H. Liang,; The Cancer Genome Atlas Research Network, A pan-cancer analysis of enhancer expression in nearly 9000 patient samples. *Cell* **173**, 386–399.e12 (2018). doi:10.1016/j.cell.2018.03.027 Medline

16. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013). doi:10.1038/nmeth.2688 Medline

17. M. R. Corces, A. E. Trevino, E. G. Hamilton, P. G. Greenside, N. A. Sinnott-Armstrong, S. Vesuna, A. T. Satpathy, A. J. Rubin, K. S. Montine, B. Wu, A. Kathiria, S. W. Cho, M. R. Mumbach, A. C. Carter, M. Kasowski, L. A. Orloff, V. I. Risca, A. Kundaje, P. A. Khavari, T. J. Montine, W. J. Greenleaf, H. Y. Chang, An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017). doi:10.1038/nmeth.4396 Medline

18. A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. R. Pfenning,

X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, M. Kellis; Roadmap Epigenomics Consortium, Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015). doi:10.1038/nature14248 Medline

19. J. Schuijers, J. C. Manteiga, A. S. Weintraub, D. S. Day, A. V. Zamudio, D. Hnisz, T. I. Lee, R. A. Young, Transcriptional dysregulation of *MYC* reveals common enhancer-docking mechanism. *Cell Reports* **23**, 349–360 (2018). doi:10.1016/j.celrep.2018.03.056 Medline

20. G. Andrey, T. Montavon, B. Mascrez, F. Gonzalez, D. Noordermeer, M. Leleu, D. Trono, F. Spitz, D. Duboule, A switch between topological domains underlies *HoxD* genes collinearity in mouse limbs. *Science* **340**, 1234167 (2013). doi:10.1126/science.1234167 Medline

21. M. Yeager, N. Orr, R. B. Hayes, K. B. Jacobs, P. Kraft, S. Wacholder, M. J. Minichiello, P. Fearnhead, K. Yu, N. Chatterjee, Z. Wang, R. Welch, B. J. Staats, E. E. Calle, H. S. Feigelson, M. J. Thun, C. Rodriguez, D. Albanes, J. Virtamo, S. Weinstein, F. R. Schumacher, E. Giovannucci, W. C. Willett, G. Cancel-Tassin, O. Cussenot, A. Valeri, G. L. Andriole, E. P. Gelmann, M. Tucker, D. S. Gerhard, J. F. Fraumeni Jr., R. Hoover, D. J. Hunter, S. J. Chanock, G. Thomas, Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39**, 645–649 (2007). doi:10.1038/ng2022 Medline

22. I. P. M. Tomlinson, E. Webb, L. Carvajal-Carmona, P. Broderick, K. Howarth, A. M. Pittman, S. Spain, S. Lubbe, A. Walther, K. Sullivan, E. Jaeger, S. Fielding, A. Rowan, J. Vijayakrishnan, E. Domingo, I. Chandler, Z. Kemp, M. Qureshi, S. M. Farrington, A. Tenesa, J. G. D. Prendergast, R. A. Barnetson, S. Penegar, E. Barclay, W. Wood, L. Martin, M. Gorman, H. Thomas, J. Peto, D. T. Bishop, R. Gray, E. R. Maher, A. Lucassen, D. Kerr, D. G. R. Evans, C. Schafmayer, S. Buch, H. Völzke, J. Hampe, S. Schreiber, U. John, T. Koessler, P. Pharoah, T. van Wezel, H. Morreau, J. T. Wijnen, J. L. Hopper, M. C. Southey, G. G. Giles, G. Severi, S. Castellví-Bel, C. Ruiz-Ponte, A. Carracedo, A. Castells, A. Försti, K. Hemminki, P. Vodicka, A. Naccarati, L. Lipton, J. W. C. Ho, K. K. Cheng, P. C. Sham, J. Luk, J. A. G. Agúndez, J. M. Ladero, M. de la Hoya, T. Caldés, I. Niittymäki, S. Tuupanen, A. Karhu, L. Aaltonen, J.-B. Cazier, H. Campbell, M. G. Dunlop, R. S. Houlston; CORGI Consortium; EPICOLON Consortium, A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat. Genet.* **40**, 623–630 (2008). doi:10.1038/ng.111 Medline

23. I. K. Sur, O. Hallikas, A. Vähärautio, J. Yan, M. Turunen, M. Enge, M. Taipale, A. Karhu, L.

A. Aaltonen, J. Taipale, Mice lacking a *Myc* enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science* **338**, 1360–1363 (2012). doi:10.1126/science.1228606 Medline

24. R. E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B.-K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford, J. A. Stamatoyannopoulos, The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012). doi:10.1038/nature11232 Medline

25. M. R. Corces, J. D. Buenrostro, B. Wu, P. G. Greenside, S. M. Chan, J. L. Koenig, M. P. Snyder, J. K. Pritchard, A. Kundaje, W. J. Greenleaf, R. Majeti, H. Y. Chang, Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016). doi:10.1038/ng.3646 Medline

26. L. J. P. van der Maaten, G. E. Hinton, Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

27. A. Rodriguez, A. Laio, Clustering by fast search and find of desity peaks. *Science* **344**, 1492–1496 (2014). doi:10.1126/science.1242072 Medline

28. K. A. Hoadley, C. Yau, T. Hinoue, D. M. Wolf, A. J. Lazar, E. Drill, R. Shen, A. M. Taylor, A. D. Cherniack, V. Thorsson, R. Akbani, R. Bowlby, C. K. Wong, M. Wiznerowicz, F. Sanchez-Vega, A. G. Robertson, B. G. Schneider, M. S. Lawrence, H. Noushmehr, T. M. Malta, J. M. Stuart, C. C. Benz, P. W. Laird; The Cancer Genome Atlas Network, Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173**, 291–304.e6 (2018). doi:10.1016/j.cell.2018.03.022 Medline

29. M. Kulis, A. Merkel, S. Heath, A. C. Queirós, R. P. Schuyler, G. Castellano, R. Beekman, E. Raineri, A. Esteve, G. Clot, N. Verdaguer-Dot, M. Duran-Ferrer, N. Russiñol, R. Vilarrasa-Blasi, S. Ecker, V. Pancaldi, D. Rico, L. Agueda, J. Blanc, D. Richardson, L. Clarke, A. Datta, M. Pascual, X. Agirre, F. Prosper, D. Alignani, B. Paiva, G. Caron, T. Fest, M. O. Muench, M. E. Fomin, S.-T. Lee, J. L. Wiemels, A. Valencia, M. Gut, P. Flicek, H. G. Stunnenberg, R. Siebert, R. Küppers, I. G. Gut, E. Campo, J. I. Martín-Subero, Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nat. Genet.* **47**, 746–756 (2015). doi:10.1038/ng.3291 Medline

30. H. S. Kim, Y. Tan, W. Ma, D. Merkurjev, E. Destici, Q. Ma, T. Suter, K. Ohgi, M. Friedman, D. Skowronska-Krawczyk, M. G. Rosenfeld, Pluripotency factors functionally premark cell-type-restricted enhancers in ES cells. *Nature* **556**, 510–514 (2018). doi:10.1038/s41586-018-0048-8 Medline

31. H. Shen, J. Shih, D. P. Hollern, L. Wang, R. Bowlby, S. K. Tickoo, V. Thorsson, A. J. Mungall, Y. Newton, A. M. Hegde, J. Armenia, F. Sánchez-Vega, J. Pluta, L. C. Pyle, R. Mehra, V. E. Reuter, G. Godoy, J. Jones, C. S. Shelley, D. R. Feldman, D. O. Vidal, D.

Lessel, T. Kulis, F. M. Cárcano, K. M. Leraas, T. M. Lichtenberg, D. Brooks, A. D. Cherniack, J. Cho, D. I. Heiman, K. Kasaian, M. Liu, M. S. Noble, L. Xi, H. Zhang, W. Zhou, J. C. ZenKlusen, C. M. Hutter, I. Felau, J. Zhang, N. Schultz, G. Getz, M. Meyerson, J. M. Stuart, R. Akbani, D. A. Wheeler, P. W. Laird, K. L. Nathanson, V. K. Cortessis, K. A. Hoadley; The Cancer Genome Atlas Research Network, Integrated molecular characterization of testicular germ cell tumors. *Cell Reports* **23**, 3392–3406 (2018). doi:10.1016/j.celrep.2018.05.039 Medline

32. S. Werner, S. Frey, S. Riethdorf, C. Schulze, M. Alawi, L. Kling, V. Vafaizadeh, G. Sauter, L. Terracciano, U. Schumacher, K. Pantel, V. Assmann, Dual roles of the transcription factor grainyhead-like 2 (GRHL2) in breast cancer. *J. Biol. Chem.* **288**, 22993–23008 (2013). doi:10.1074/jbc.M113.456293 Medline

33. S. K. Denny, D. Yang, C.-H. Chuang, J. J. Brady, J. S. Lim, B. M. Grüner, S.-H. Chiou, A. N. Schep, J. Baral, C. Hamard, M. Antoine, M. Wislez, C. S. Kong, A. J. Connolly, K.-S. Park, J. Sage, W. J. Greenleaf, M. M. Winslow, Nfib promotes metastasis through a widespread increase in chromatin accessibility. *Cell* **166**, 328–342 (2016). doi:10.1016/j.cell.2016.05.052 Medline

34. S. Baek, I. Goldstein, G. L. Hager, Bivariate genomic footprinting detects changes in transcription factor activity. *Cell Reports* **19**, 1710–1722 (2017). doi:10.1016/j.celrep.2017.05.003 Medline

35. A. N. Schep, B. Wu, J. D. Buenrostro, W. J. Greenleaf, ChromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017). doi:10.1038/nmeth.4401 Medline

36. Y. Yin, E. Morgunova, A. Jolma, E. Kaasinen, B. Sahu, S. Khund-Sayeed, P. K. Das, T. Kivioja, K. Dave, F. Zhong, K. R. Nitta, M. Taipale, A. Popov, P. A. Ginno, S. Domcke, J. Yan, D. Schübeler, C. Vinson, J. Taipale, Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239 (2017). doi:10.1126/science.aaj2239 Medline

37. T. Ellis, L. Gambardella, M. Horcher, S. Tschanz, J. Capol, P. Bertram, W. Jochum, Y. Barrandon, M. Busslinger, The transcriptional repressor CDP (Cutl1) is essential for epithelial cell differentiation of the lung and the hair follicle. *Genes Dev.* **15**, 2307–2319 (2001). doi:10.1101/gad.200101 Medline

38. B. M. Javierre, O. S. Burren, S. P. Wilder, R. Kreuzhuber, S. M. Hill, S. Sewitz, J. Cairns, S. W. Wingett, C. Várnai, M. J. Thiecke, F. Burden, S. Farrow, A. J. Cutler, K. Rehnström, K. Downes, L. Grassi, M. Kostadima, P. Freire-Pritchett, F. Wang, H. G. Stunnenberg, J. A. Todd, D. R. Zerbino, O. Stegle, W. H. Ouwehand, M. Frontini, C. Wallace, M. Spivakov, P. Fraser; BLUEPRINT Consortium, Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* **167**, 1369–1384.e19 (2016). doi:10.1016/j.cell.2016.09.037 Medline

39. C. P. Fulco, M. Munschauer, R. Anyoha, G. Munson, S. R. Grossman, E. M. Perez, M. Kane, B. Cleary, E. S. Lander, J. M. Engreitz, Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**, 769–773 (2016). doi:10.1126/science.aag2445 Medline

40. L. S. Qi, M. H. Larson, L. A. Gilbert, J. A. Doudna, J. S. Weissman, A. P. Arkin, W. A. Lim,

Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013). doi:10.1016/j.cell.2013.02.022 Medline

41. Y. H. Eom, H. S. Kim, A. Lee, B. J. Song, B. J. Chae, BCL2 as a subtype-specific prognostic marker for breast cancer. *J. Breast Cancer* **19**, 252–260 (2016). doi:10.4048/jbc.2016.19.3.252 Medline

42. S. W. Cho, J. Xu, R. Sun, M. R. Mumbach, A. C. Carter, Y. G. Chen, K. E. Yost, J. Kim, J. He, S. A. Nevins, S.-F. Chin, C. Caldas, S. J. Liu, M. A. Horlbeck, D. A. Lim, J. S. Weissman, C. Curtis, H. Y. Chang, Promoter of lncRNA gene *PVT1* is a tumor-suppressor DNA boundary element. *Cell* **173**, 1398–1412.e22 (2018). doi:10.1016/j.cell.2018.03.068 Medline

43. T. C. Silva, S. G. Coetzee, L. Yao, D. J. Hazelett, H. Noushmehr, B. P. Berman, Enhancer linking by methylation/expression relationships with the R package ELMER version 2. bioRxiv 148726 [Preprint]. 11 June 2017. https://doi.org/10.1101/148726.

44. A. G. Robertson, J. Kim, H. Al-Ahmadie, J. Bellmunt, G. Guo, A. D. Cherniack, T. Hinoue, P. W. Laird, K. A. Hoadley, R. Akbani, M. A. A. Castro, E. A. Gibb, R. S. Kanchi, D. A. Gordenin, S. A. Shukla, F. Sanchez-Vega, D. E. Hansel, B. A. Czerniak, V. E. Reuter, X. Su, B. de Sa Carvalho, V. S. Chagas, K. L. Mungall, S. Sadeghi, C. S. Pedamallu, Y. Lu, L. J. Klimczak, J. Zhang, C. Choo, A. I. Ojesina, S. Bullman, K. M. Leraas, T. M. Lichtenberg, C. J. Wu, N. Schultz, G. Getz, M. Meyerson, G. B. Mills, D. J. McConkey, J. N. Weinstein, D. J. Kwiatkowski, S. P. Lerner; TCGA Research Network, Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell* **171**, 540–556.e25 (2017). doi:10.1016/j.cell.2017.09.007 Medline

45. M. A. A. Castro, I. de Santiago, T. M. Campbell, C. Vaughn, T. E. Hickey, E. Ross, W. D. Tilley, F. Markowetz, B. A. J. Ponder, K. B. Meyer, Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nat. Genet.* **48**, 12–21 (2016). doi:10.1038/ng.3458 Medline

46. V. Thorsson, D. L. Gibbs, S. D. Brown, D. Wolf, D. S. Bortone, T.-H. Ou Yang, E. Porta-Pardo, G. F. Gao, C. L. Plaisier, J. A. Eddy, E. Ziv, A. C. Culhane, E. O. Paull, I. K. A. Sivakumar, A. J. Gentles, R. Malhotra, F. Farshidfar, A. Colaprico, J. S. Parker, L. E. Mose, N. S. Vo, J. Liu, Y. Liu, J. Rader, V. Dhankani, S. M. Reynolds, R. Bowlby, A. Califano, A. D. Cherniack, D. Anastassiou, D. Bedognetti, A. Rao, K. Chen, A. Krasnitz, H. Hu, T. M. Malta, H. Noushmehr, C. S. Pedamallu, S. Bullman, A. I. Ojesina, A. Lamb, W. Zhou, H. Shen, T. K. Choueiri, J. N. Weinstein, J. Guinney, J. Saltz, R. A. Holt, C. E. Rabkin, A. J. Lazar, J. S. Serody, E. G. Demicco, M. L. Disis, B. G. Vincent, L. Shmulevich; The Cancer Genome Atlas Research Network, The immune landscape of cancer. *Immunity* **48**, 812–830.e14 (2018). doi:10.1016/j.immuni.2018.03.023 Medline

47. K. Yoshihara, M. Shahmoradgoli, E. Martínez, R. Vegesna, H. Kim, W. Torres-Garcia, V. Treviño, H. Shen, P. W. Laird, D. A. Levine, S. L. Carter, G. Getz, K. Stemke-Hale, G. B. Mills, R. G. W. Verhaak, Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013). doi:10.1038/ncomms3612 Medline

48. S. L. Carter, K. Cibulskis, E. Helman, A. McKenna, H. Shen, T. Zack, P. W. Laird, R. C. Onofrio, W. Winckler, B. A. Weir, R. Beroukhim, D. Pellman, D. A. Levine, E. S.

Lander, M. Meyerson, G. Getz, Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012). doi:10.1038/nbt.2203 Medline

49. M. S. Rooney, S. A. Shukla, C. J. Wu, G. Getz, N. Hacohen, Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160**, 48–61 (2015). doi:10.1016/j.cell.2014.12.033 Medline

50. M. M. Makowski, E. Willems, J. Fang, J. Choi, T. Zhang, P. W. T. C. Jansen, K. M. Brown, M. Vermeulen, An interaction proteomics survey of transcription factor binding at recurrent TERT promoter mutations. *Proteomics* **16**, 417–426 (2016). doi:10.1002/pmic.201500327 Medline

51. R. Katainen, K. Dave, E. Pitkänen, K. Palin, T. Kivioja, N. Välimäki, A. E. Gylfe, H. Ristolainen, U. A. Hänninen, T. Cajuso, J. Kondelin, T. Tanskanen, J.-P. Mecklin, H. Järvinen, L. Renkonen-Sinisalo, A. Lepistö, E. Kaasinen, O. Kilpivaara, S. Tuupanen, M. Enge, J. Taipale, L. A. Aaltonen, CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.* **47**, 818–821 (2015). doi:10.1038/ng.3335 Medline

52. D. Hnisz, A. S. Weintraub, D. S. Day, A.-L. Valton, R. O. Bak, C. H. Li, J. Goldmann, B. R. Lajoie, Z. P. Fan, A. A. Sigova, J. Reddy, D. Borges-Rivera, T. I. Lee, R. Jaenisch, M. H. Porteus, J. Dekker, R. A. Young, Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458 (2016). doi:10.1126/science.aad9024 Medline

53. R. D. Hawkins, G. C. Hon, L. K. Lee, Q. Ngo, R. Lister, M. Pelizzola, L. E. Edsall, S. Kuan, Y. Luu, S. Klugman, J. Antosiewicz-Bourget, Z. Ye, C. Espinoza, S. Agarwahl, L. Shen, V. Ruotti, W. Wang, R. Stewart, J. A. Thomson, J. R. Ecker, B. Ren, Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6**, 479–491 (2010). doi:10.1016/j.stem.2010.03.018 Medline

54. T. K. Kelly, Y. Liu, F. D. Lay, G. Liang, B. P. Berman, P. A. Jones, Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.* **22**, 2497–2506 (2012). doi:10.1101/gr.143008.112 Medline

55. A. Mo, E. A. Mukamel, F. P. Davis, C. Luo, G. L. Henry, S. Picard, M. A. Urich, J. R. Nery, T. J. Sejnowski, R. Lister, S. R. Eddy, J. R. Ecker, J. Nathans, Epigenomic signatures of neuronal diversity in the mammalian brain. *Neuron* **86**, 1369–1384 (2015). doi:10.1016/j.neuron.2015.05.018 Medline

56. S. Picelli, Å. K. Björklund, B. Reinius, S. Sagasser, G. Winberg, R. Sandberg, Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014). doi:10.1101/gr.177881.114 Medline

57. J. D. Buenrostro, B. Wu, H. Y. Chang, W. J. Greenleaf, ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 1–9 (2015). Medline

58. B. Adamson, T. M. Norman, M. Jost, M. Y. Cho, J. K. Nuñez, Y. Chen, J. E. Villalta, L. A. Gilbert, M. A. Horlbeck, M. Y. Hein, R. A. Pak, A. N. Gray, C. A. Gross, A. Dixit, O. Parnas, A. Regev, J. S. Weissman, A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–1882.e21 (2016). doi:10.1016/j.cell.2016.11.048 Medline

59. A. R. Perez, Y. Pritykin, J. A. Vidigal, S. Chhangawala, L. Zamparo, C. S. Leslie, A.

Ventura, GuideScan software for improved single and paired CRISPR guide RNA design. *Nat. Biotechnol.* **35**, 347–349 (2017). doi:10.1038/nbt.3804 Medline

60. H. Jiang, R. Lei, S. W. Ding, S. Zhu, Skewer: A fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* **15**, 182 (2014). doi:10.1186/1471-2105-15-182 Medline

61. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012). doi:10.1038/nmeth.1923 Medline

62. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). doi:10.1093/bioinformatics/btp352 Medline

63. J. M. Korn, F. G. Kuruvilla, S. A. McCarroll, A. Wysoker, J. Nemesh, S. Cawley, E. Hubbell, J. Veitch, P. J. Collins, K. Darvishi, C. Lee, M. M. Nizzari, S. B. Gabriel, S. Purcell, M. J. Daly, D. Altshuler, Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40**, 1253–1260 (2008). doi:10.1038/ng.237 Medline

64. Y. Newton, A. M. Novak, T. Swatloski, D. C. McColl, S. Chopra, K. Graim, A. S. Weinstein, R. Baertsch, S. R. Salama, K. Ellrott, M. Chopra, T. C. Goldstein, D. Haussler, O. Morozova, J. M. Stuart, TumorMap: Exploring the molecular similarities of cancer samples in an interactive portal. *Cancer Res.* **77**, e111–e114 (2017). doi:10.1158/0008-5472.CAN-17-0580 Medline

65. M. T. Weirauch, A. Yang, M. Albu, A. G. Cote, A. Montenegro-Montero, P. Drewe, H. S. Najafabadi, S. A. Lambert, I. Mann, K. Cook, H. Zheng, A. Goity, H. van Bakel, J.-C. Lozano, M. Galli, M. G. Lewsey, E. Huang, T. Mukherjee, X. Chen, J. S. Reece-Hoyes, S. Govindarajan, G. Shaulsky, A. J. M. Walhout, F.-Y. Bouget, G. Ratsch, L. F. Larrondo, J. R. Ecker, T. R. Hughes, Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014). doi:10.1016/j.cell.2014.08.009 Medline

66. P. Du, X. Zhang, C.-C. Huang, N. Jafari, W. A. Kibbe, L. Hou, S. M. Lin, Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587 (2010). doi:10.1186/1471-2105-11-587 Medline

67. D. M. A. Gendoo, N. Ratanasirigulchai, M. S. Schröder, L. Paré, J. S. Parker, A. Prat, B. Haibe-Kains, Genefu: An R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics* **32**, 1097–1099 (2016). doi:10.1093/bioinformatics/btv693 Medline

68. H. R. Ali, O. M. Rueda, S.-F. Chin, C. Curtis, M. J. Dunning, S. A. J. R. Aparicio, C. Caldas, Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome Biol.* **15**, 431 (2014). doi:10.1186/s13059-014-0431-1 Medline

69. P. Wirapati, C. Sotiriou, S. Kunkel, P. Farmer, S. Pradervand, B. Haibe-Kains, C. Desmedt, M. Ignatiadis, T. Sengstag, F. Schütz, D. R. Goldstein, M. Piccart, M. Delorenzi, Meta-analysis of gene expression profiles in breast cancer: Toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res.* **10**, R65 (2008).

doi:10.1186/bcr2124 Medline

70. J. S. Parker, M. Mullins, M. C. U. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, J. F. Quackenbush, I. J. Stijleman, J. Palazzo, J. S. Marron, A. B. Nobel, E. Mardis, T. O. Nielsen, M. J. Ellis, C. M. Perou, P. S. Bernard, Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009). doi:10.1200/JCO.2008.18.1370 Medline

71. G. Ciriello, M. L. Gatza, A. H. Beck, M. D. Wilkerson, S. K. Rhie, A. Pastore, H. Zhang, M. McLellan, C. Yau, C. Kandoth, R. Bowlby, H. Shen, S. Hayat, R. Fieldhouse, S. C. Lester, G. M. K. Tse, R. E. Factor, L. C. Collins, K. H. Allison, Y.-Y. Chen, K. Jensen, N. B. Johnson, S. Oesterreich, G. B. Mills, A. D. Cherniack, G. Robertson, C. Benz, C. Sander, P. W. Laird, K. A. Hoadley, T. A. King, C. M. Perou; TCGA Research Network, Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* **163**, 506–519 (2015). doi:10.1016/j.cell.2015.09.033 Medline

72. C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Gräf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, A. Langerød, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A.-L. Børresen-Dale, J. D. Brenton, S. Tavaré, C. Caldas, S. Aparicio; METABRIC Group, The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012). doi:10.1038/nature10983 Medline

73. H. H. He, C. A. Meyer, S. S. Hu, M.-W. Chen, C. Zang, Y. Liu, P. K. Rao, T. Fei, H. Xu, H. Long, X. S. Liu, M. Brown, Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat. Methods* **11**, 73–78 (2014). doi:10.1038/nmeth.2762 Medline

74. M. N. C. Fletcher, M. A. A. Castro, X. Wang, I. de Santiago, M. O'Reilly, S.-F. Chin, O. M. Rueda, C. Caldas, B. A. J. Ponder, F. Markowetz, K. B. Meyer, Master regulators of FGFR2 signalling and breast cancer risk. *Nat. Commun.* **4**, 2464 (2013). doi:10.1038/ncomms3464 Medline

75. J. Liu, T. Lichtenberg, K. A. Hoadley, L. M. Poisson, A. J. Lazar, A. D. Cherniack, A. J. Kovatich, C. C. Benz, D. A. Levine, A. V. Lee, L. Omberg, D. M. Wolf, C. D. Shriver, V. Thorsson, H. Hu; Cancer Genome Atlas Research Network, An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400–416.e11 (2018). doi:10.1016/j.cell.2018.02.052 Medline

76. W. Zhou, T. J. Triche Jr., P. W. Laird, H. Shen, SeSAMe: Reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Res.* 10.1093/nar/gky691 (2018).

77. M. R. Mumbach, A. T. Satpathy, E. A. Boyle, C. Dai, B. G. Gowen, S. W. Cho, M. L. Nguyen, A. J. Rubin, J. M. Granja, K. R. Kazane, Y. Wei, T. Nguyen, P. G. Greenside, M. R. Corces, J. Tycko, D. R. Simeonov, N. Suliman, R. Li, J. Xu, R. A. Flynn, A. Kundaje, P. A. Khavari, A. Marson, J. E. Corn, T. Quertermous, W. J. Greenleaf, H. Y. Chang, Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat. Genet.* **49**, 1602–1612 (2017). doi:10.1038/ng.3963 Medline

78. M. J. Machiela, S. J. Chanock, LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557 (2015). doi:10.1093/bioinformatics/btv402 Medline