# Supplementary text

## Table of Contents

*Supplementary Figures*

*Supplementary Tables*

An overarching objective of our study is to leverage ENCODE data in order to provide novel insights and resources for cancer research. We aim integrate ENCODE and cancer genomic data to gain a more comprehensive understanding of the non-coding elements involved in oncogenesis, their associated linkages to protein-coding genes and the background mutation rates therein, and the global regulatory nature of TFs in the context of matched tumor-normal cell lines. The recent ENCODE data release provides a rich source of information for investigating questions both in basic biology and human disease. In large part, this wealth of information derives from the multiple genomic annotations provided across multiple cell lines. In addition to providing new opportunities, however, the very richness of this data provides considerable challenges in terms of data integration and organization. In addition to the complexity of this data resource, our analyses relies on an array methodologies, the details for which are difficult to include within the main text of this paper. As such, the purpose of this Supplementary document is to provide a clear and organized reference to support and explain the datasets, pipelines, and analyses associated with this study. In addition to supplementary text, supplementary figures and tables provide additional information not included in the main figures.

Our study is broadly organized into 4 main parts: a description of the assays, the construction of enhancer-target gene linkages, the workflow for variant prioritizing key genomic features associated with cancer, and concluding remarks. This supplement is presented in roughly a parallel fashion to the main text. The supplement is also connected to main text through the major results presented in the form of main text figures – captions associated with main text figures point to relevant sub-sections within the supplement. We have written our study in roughly a hierarchical fashion, and aim to present data and results (including predications) in an organized way. The main

text lies at the top of this hierarchy, and synthesizes everything in a broad fashion. It refers to more detailed descriptions of our methods and datasets, as provided in the supplement.

Part 1 provides in-depth documentation of the ENCODE data we use, along with the subsidiary steps (including ENCODE data processing, enhancer and enhancer-target predictions, and extended gene definitions). Part 2 provides details on our recurrence analyses. Part 3 provides in-depth discussions and data regarding our TF network construction and analyses. Part 4 aims to expand on our expression aggregation analysis. Finally, Part 5 deals with the validation of prioritized SNVs.

# 1 Details about data summary from ENCODE

## 1.1 Summary of the cancer-related encyclopedia companion resource

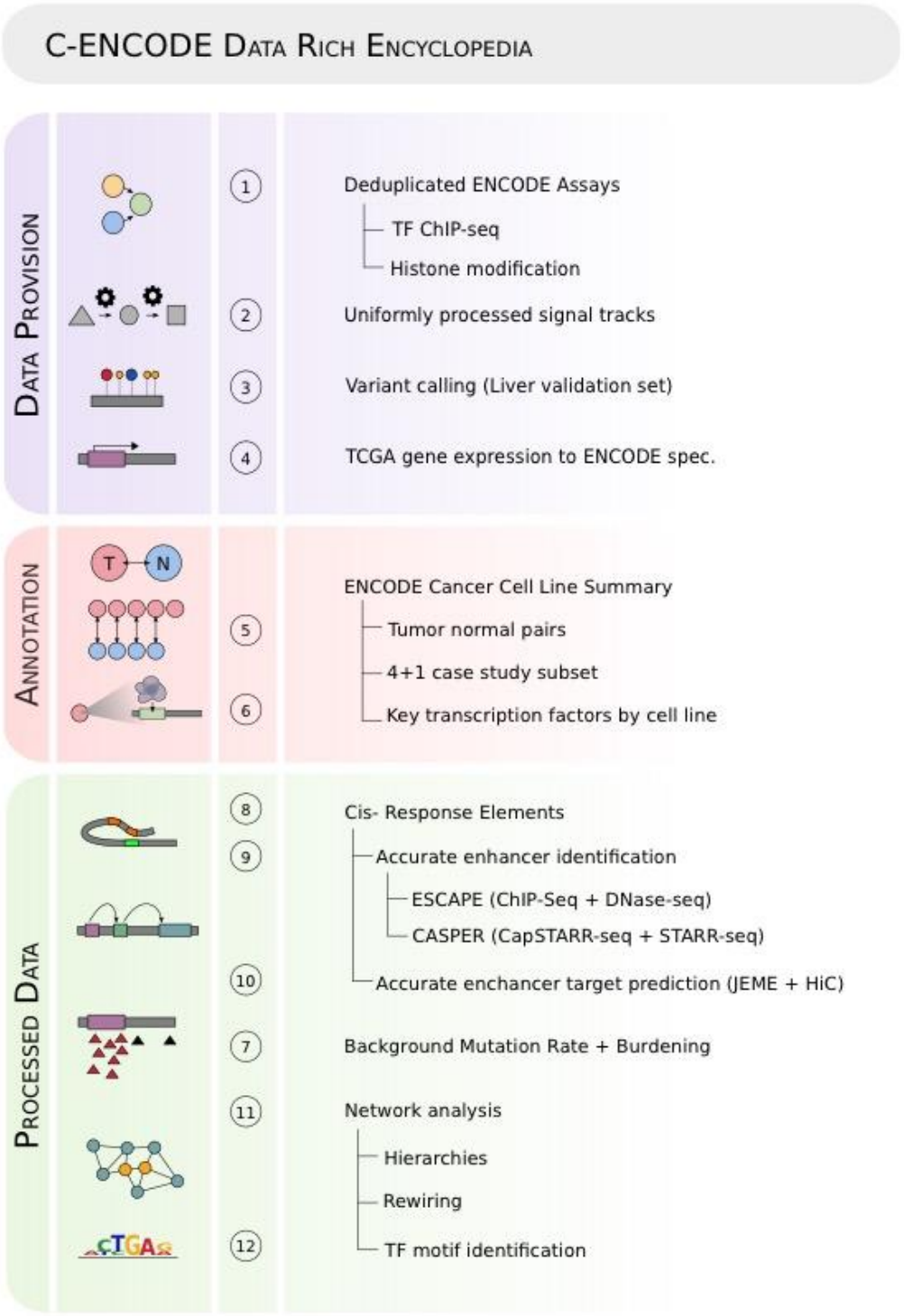
Mutations associated with cancer have been well characterized in a many key oncogenes and tumor suppressors. However, the overwhelming bulk of mutations in cancer genomes – particularly those discovered from the recent large-scale cancer genomics initiatives – lie within non-coding regions. Whether these mutations drive cancer development or progression, or simply emerge as byproducts of genomic instability remains an open question. Newly-released data from the ENCODE Consortium can help address this question by providing comprehensive characterization of non-coding genomic elements, as well as by linking such elements to well-known cancer associated genes.

Here, we endeavor to provide a companion resource to the main ENCODE encyclopedia by building a “cancer related encyclopedia companion” resource. The main encyclopedia is oriented toward breath of the annotations to describe elements over hundreds of cell lines. In contrast, we focus on top cell lines with a wide variety of profiles available. Most of these cell lines are

associated with cancers of the blood, liver, lung, cervix, and breast. We show that these cell lines can be used to provide a better understanding of oncogenesis, and we provide a resource for interpreting the wealth of mutational and transcriptional profiles produced by the cancer community. We summarized our efforts in Figure S 1-1. This encyclopedia companion mainly provides three layer of resource: 1) Data provision: carefully collected and de-duplicated signal tracks from various experimental assays both within and outside ENCODE; 2) pairing cell lines and datasets to cancer types; 3) Detailed Annotations: enhancers and their gene linkages, tissue specific and generalized networks, network hierarchies, rewiring status, gene expression regulating potentials, predicted mutation rates, and motif identifications.

[[JZ2PDM: figure S 1-1 to be updated]]

*Figure S 1-1 Summary of the resources in cancer related encyclopedia companion*



## 1.2 Detailed annotation of TFs

In this study, we collected a total of 344 transcription related factors and abbreviate them all as TFs in the main text for simplicity. For our main analyses, we further classified them into four major classes: 282 sequence-specific TFs, which bind DNA at particular motifs to regulate gene expression; 16 general TFs, which comprise that segment of the cell's transcriptional machinery that complexes with DNA; 19 chromatin-associated TFs, which comprise complexes that bind to and remodel chromatin; and 27 co-factors, which support the function of other TFs, do not directly bind DNA, and do not belong to another class. Detailed classification was given in supplementary Table1.

We further extracted 68 common TFs between K562 and GM12878, annotated in Table S 1-1. We searched the COSMIC Cancer Gene census \cite{15188009} and an authoritative list of cancer genes by Vogelstein *et al.* \cite{23539594} to identify TFs associated with cancer. We further listed whether a TF has been reported to regulate the ABL gene or BCR-ABL transcript, or the BCR-ABL KEGG pathway \cite{18287706}, because of the dominant role this fusion gene plays in CML and K562 \cite{3023859, 12476301}.

*Table S 1-1 Detailed annotations of 68 common TF in K562 and GM12878*

| TF | Class | FAMILY | TF in COSMIC | TF in Vogelstein | Targets ABL | Targets BCR-ABL pathway | Targets Vogelstein gene |
|---|---|---|---|---|---|---|---|
| **ATF3** | TFSS | bZIP | 0 | 0 | 1 | 1 | 1 |
| **BCLAF1** | TFSS | bZIP | 0 | 0 | 0 | 1 | 1 |
| **BHLHE40** | TFSS | HLH | 0 | 0 | 0 | 0 | 0 |
| **CBX5** | chromatin | | 0 | 0 | 0 | 0 | 0 |
| **CEBPB** | TFSS | bZIP | 0 | 0 | 0 | 1 | 1 |
| **CEBPZ** | TFSS | bZIP | 0 | 0 | 0 | 0 | 0 |
| **CHD1** | chromatin | Homeodomain | 0 | 0 | 0 | 0 | 0 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **CHD2** | chromatin | Homeodomain | 0 | 0 | 0 | 0 | 0 |
| **CTCF** | TFSS | ZNF | 1 | 0 | 0 | 1 | 1 |
| **E2F4** | TFSS | wHTH | 0 | 0 | 0 | 1 | 1 |
| **EGR1** | TFSS | ZNF | 0 | 0 | 0 | 1 | 1 |
| **ELF1** | TFSS | ETS | 0 | 0 | 0 | 1 | 1 |
| **ELK1** | TFSS | ETS | 0 | 0 | 0 | 0 | 0 |
| **EP300** | general | | 1 | 1 | 1 | 1 | 1 |
| **ETS1** | TFSS | ETS | 0 | 0 | 0 | 1 | 1 |
| **ETV6** | TFSS | ETS | 1 | 0 | 0 | 0 | 0 |
| **EZH2** | chromatin | | 1 | 1 | 0 | 0 | 0 |
| **FOS** | TFSS | bZIP | 0 | 0 | 0 | 1 | 1 |
| **GABPA** | TFSS | ETS | 0 | 0 | 0 | 0 | 0 |
| **HDGF** | TFSS | | 0 | 0 | 0 | 0 | 0 |
| **IKZF1** | TFSS | ZF-C2H2 | 1 | 0 | 0 | 0 | 0 |
| **JUNB** | TFSS | bZIP | 0 | 0 | 0 | 0 | 0 |
| **JUND** | TFSS | bZIP | 0 | 0 | 0 | 1 | 1 |
| **MAFK** | TFSS | bZIP | 0 | 0 | 1 | 1 | 1 |
| **MAX** | TFSS | HLH | 1 | 0 | 0 | 1 | 1 |
| **MAZ** | TFSS | HLH | 0 | 0 | 0 | 0 | 0 |
| **MEF2A** | TFSS | MADs-box | 0 | 0 | 0 | 0 | 0 |
| **MLLT1** | TFSS | | 1 | 0 | 0 | 0 | 0 |
| **MTA2** | TFSS | ZF-GATA | 0 | 0 | 0 | 0 | 0 |
| **MXI1** | TFSS | HLH | 0 | 0 | 0 | 1 | 0 |
| **MYC** | TFSS | HLH | 1 | 0 | 0 | 1 | 1 |
| **NBN** | TFSS | | 1 | 0 | 0 | 0 | 0 |
| **NFE2** | TFSS | bZIP | 0 | 0 | 0 | 1 | 0 |
| **NFYA** | TFSS | CBF-NFY | 0 | 0 | 0 | 1 | 1 |
| **NFYB** | TFSS | CBF-NFY | 0 | 0 | 0 | 1 | 1 |
| **NR2C2** | TFSS | NR | 0 | 0 | 1 | 1 | 1 |
| **NRF1** | TFSS | bZIP | 0 | 0 | 1 | 1 | 1 |
| **PML** | cofactor | | 1 | 0 | 0 | 0 | 0 |
| **POLR2A** | general | | 0 | 0 | 0 | 0 | 0 |
| **POLR3G** | general | | 0 | 0 | 0 | 0 | 0 |
| **RAD21** | chromatin | | 1 | 0 | 1 | 1 | 1 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **RCOR1** | TFSS | MYB | 0 | 0 | 0 | 0 | 0 |
| **REST** | TFSS | ZNF | 0 | 0 | 0 | 1 | 1 |
| **RFX5** | TFSS | wHTH | 0 | 0 | 0 | 0 | 1 |
| **SIN3A** | general | | 0 | 0 | 0 | 1 | 1 |
| **SIX5** | TFSS | Homeodomain | 0 | 0 | 0 | 1 | 1 |
| **SMAD5** | TFSS | MH1 | 0 | 0 | 0 | 0 | 0 |
| **SMC3** | chromatin | | 0 | 0 | 1 | 1 | 1 |
| **SP1** | TFSS | ZNF | 0 | 0 | 0 | 1 | 1 |
| **SPI1** | TFSS | ETS | 0 | 0 | 0 | 1 | 1 |
| **SRF** | TFSS | MADs-box | 0 | 0 | 0 | 1 | 1 |
| **STAT5A** | TFSS | STAT | 0 | 0 | 0 | 0 | 0 |
| **SUZ12** | chromatin | ZNF | 1 | 0 | 0 | 1 | 1 |
| **TAF1** | general | | 0 | 0 | 0 | 1 | 1 |
| **TARDBP** | TFSS | | 0 | 0 | 0 | 0 | 0 |
| **TBL1XR1** | cofactor | | 1 | 0 | 0 | 0 | 0 |
| **TBP** | general | | 0 | 0 | 0 | 0 | 1 |
| **UBTF** | TFSS | HMG | 0 | 0 | 0 | 0 | 0 |
| **USF1** | TFSS | HLH | 0 | 0 | 0 | 1 | 1 |
| **USF2** | TFSS | HLH | 0 | 0 | 1 | 1 | 1 |
| **YBX1** | TFSS | CSD | 0 | 0 | 0 | 0 | 0 |
| **YY1** | TFSS | ZNF | 0 | 0 | 0 | 1 | 1 |
| **ZBED1** | TFSS | ZNF | 0 | 0 | 0 | 0 | 0 |
| **ZBTB33** | TFSS | ZNF | 0 | 0 | 1 | 1 | 1 |
| **ZBTB40** | TFSS | ZNF | 0 | 0 | 0 | 0 | 0 |
| **ZNF143** | TFSS | ZNF | 0 | 0 | 0 | 0 | 0 |
| **ZNF274** | TFSS | ZNF | 0 | 0 | 1 | 1 | 1 |

## 1.3 Matching of ENCODE cell lines to major cancer types

Despite the comprehensive catalog of functional characterization assays in ENCODE, integrating its associated data into cancer research remains challenging for two main reasons. First, cancer is such a heterogeneous disease that it is necessary to use data from optimally-matched cell lines. ENCODE is imperfect for such analysis. We observe that there are only loosely matched tumor-normal pairs for some cancer types, and most cell lines lack data from certain experimental assays (Fig 1A). Therefore, it is necessary to create biologically relevant tumor-normal pairs, as well as to develop appropriate algorithms to learn from sub-optimally matched data. The second challenge arises as a result of the heterogeneous nature of the raw data from various experimental assays. The data must undergo de-duplication, unified processing, and proper normalization before accurate large-scale integration can be achieved. Here we endeavor to match the ENOCDE data to most relevant cancer types. A detailed matching summary has been summarized in Table S 1-2.

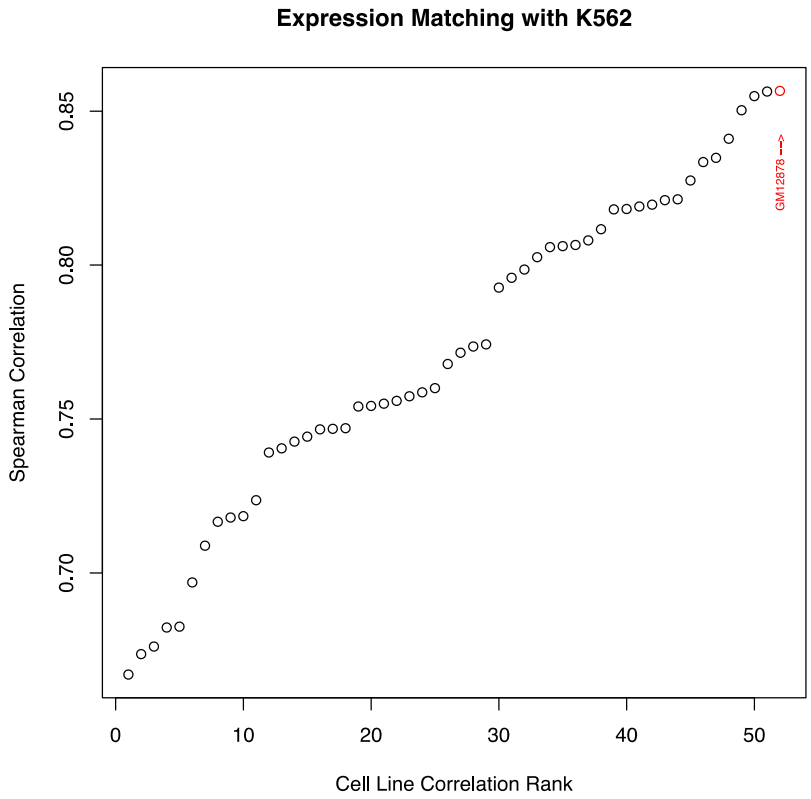
The key feature of the ENCODE annotation is that it relies on a wide variety of diverse assays. Admittedly, some of the matchings in Table S 1-2 are imperfect. These samples are not as accurate as if one directly did these assays on tissue from a patient. However, it's not possible, at least at this moment, to do such a wide variety of assays on actual tissue, so we still believe that this matching provides a valuable opportunity for large scale data integration to interpret cancer genome.

*Table S 1-2. Summary of cell line and cancer type matching*

| Cancer Type | Abbreviation | ENCODE cell line | |
|---|---|---|---|
| Breast | BRCA | Tumor | MCF-7 |
| | | Normal | MCF-10A |
| Liver | LIHC | Tumor | HepG2 |
| | | Normal | Liver |
| Lung | LUAD | Tumor | A549 |
| | | Normal | IMR-90 |
| Blood | CML[ CLL/AML] | Tumor | K562 |
| | | Normal | GM12878 |
| Cervix | CESC | Tumor | HeLa-S3 |

Wherever possible, we have matched each ENCODE cancer cell line with a data-rich ENCODE normal cell line that derives from the same cell-type as the cancer. Exact matching was not possible with K562: the cancer cell-line derives from a myeloid cell, but there is no data-rich noncancerous myeloid cell included in ENCODE. GM12878 is a data-rich ENCODE cell-line that derives from a closely related lineage, the lymphoid lineage. Supporting this choice, we determined that of all non-cancerous cell-lines among uniformly processed Roadmap Epigenome and GTEX cell-lines provided by Roadmap Epigenome, GM12878 has the highest Spearman correlation with K562 from expression data, as shown in Figure S 1-2. Hence, we used GM12878 as a rough pair for K562.

*Figure S 1-2 Expression correlations with K562 from many other cell lines*




MCF-7 is the most studied human breast cancer cell line, which has been reported by nearly 25,000 scientific publications \cite{25828948}. It is a human cell line from a pleural effusion derived from a breast carcinoma \cite{4357757}. MCF7 is one of a very few cell lines that express substantial levels of estrogen receptor (ER) that widely used to mimic ER-positive invasive human breast cancers. It is also a stable cell line for understanding intracellular binding constants, transport mechanism and defining DNA binding sites of ER in target genes \cite{25828948}. Besides, T47D is another ER-positive cell line derived from pleural effusion that has been widely used to study breast cancer \cite{228940}. Unlike MCF-7, it is mutant for the tumor suppressor gene TP53 \cite{8562478}.

MCF10A is a human breast epithelial cell line that most commonly used in vitro model for studying normal breast cell function and transformation \cite{26147507}. It derived from human benign fibrocystic mammary tissue and spontaneously immortalized, which is not tumorigenic and

dose not express ER \cite{1975513; 26147507}. Numerous studies have utilized both MCF7 and MCF10A cell lines to facilitate the development of breast cancer treatment and therapy, via comparing differential response of these two cell lines under multiple experimental settings. One study characterized distinct dynamic behaviors of MCF7 and MCF10A cells in ultrasonic field, and determined a specific frequency of ultrasound for induction of cell ablation with minimum cytotoxicity \cite{26241649}. Similarly, another study illustrated that silver nanoparticles are effective photothermal agents by comparing the differential response of MCF7 and MCF10A \cite{25144821}. Additionally, the MCF10A cell line was used to represent healthy cells to determine the level of safety of the use of one compound, in comparison with MCF7 and MDA-MB-231 cell lines \cite{27668797}.

However, one recent study challenged MCF10A as a representative model for normal mammary cells and demonstrated that this cell line exhibit some phenotypes and expression profiles that have not been observed in mammary gland tissues \cite{26147507}. But the paper also mentioned that whether MCF10A cells represent a suitable model for human mammary epithelial cells warrants further investigation. In any case, given the wealth of ENCODE data on MCF-7 and the breast cancer's status as one of the most frequent cancers, we consider the pairing of MCF-7 and MCF-10A worthwhile so that breast cancer can be included in our analysis though we cannot exclude differences between these lines being due to different ER status or other differences unrelated to malignant transformation. Inclusion of T47D as another breast cancer cell line, adds to this analysis.

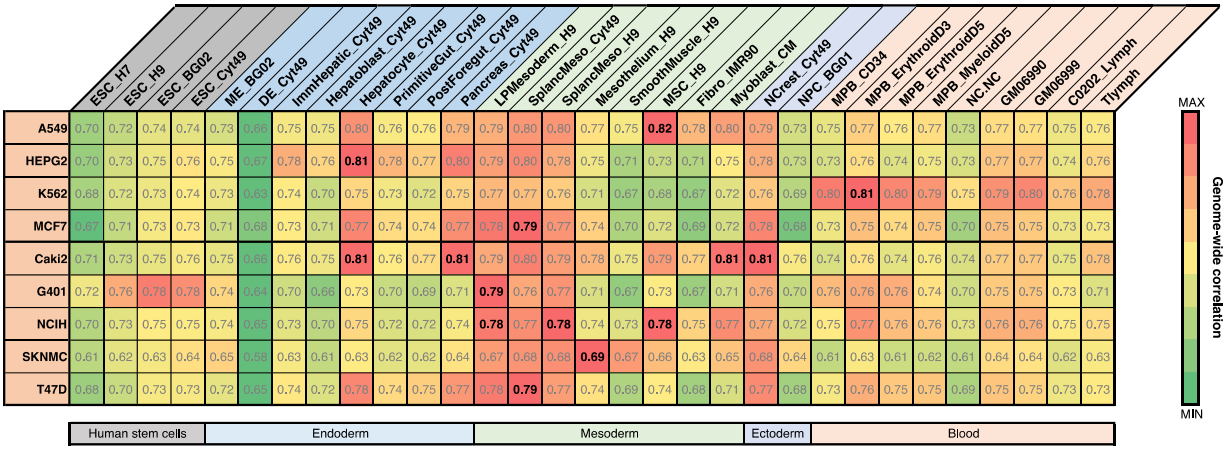
A549 is a carcinomic lung epithelial cell line \cite{9743595} and IMR90 is a normal lung fibroblast cell line \cite{841339}. Lung fibroblasts and lung epithelial cells are closely related cell types, and conversion between these cell types is common and meaningful in tumor cells and

normal cells \cite{26560033,12189386}. Lung fibroblasts like IMR90, are mesenchymal cells that arise in embryologic development subsequent to epithelial to mesenchymal transition (EMT). The dedifferentiation of mesenchymal cells into secondary epithelial tissue following mesenchymal to epithelial transition (MET) is also observed and is best characterized in kidney development \cite{10508232}. It has been postulated that the dedifferentiation and metastasis of epithelial lung cancer cells, may occur through EMT and/or MET \cite{20533280,18376396, 19759262,19010860}. Such a process has been observed in other cancers \cite{12189386}. Indeed, exposure of A549 epithelial cells to chemotherapeutic agents or TGF-B, causes differentiation to a mesenchymal phenotype, and EMT is thought to play a role in chemotherapeutic resistance of lung adenocarcinoma \cite{18599154,16123809}. These cellular relationships support the utility of a tumor normal comparison between A549 cancer cells and IMR90 normal cells.

## 1.4 Normal to Tumor cell line matching using replication timing data.

It is well known that replication timing significantly affect the mutational landscape in both germline and normal cells \cite{24598232}. We also made a genome-wide correlation of replication timing data (excluding ChrX and ChrY to avoid gender differences) between the cancer cell lines and several candidate normal cell types. Results are listed in Figure S 1-3. As expected, the best matching normal data for K562 and HepG2 are Hepatocytes and Erythroid progenitors. However, we also noticed that replication timing data in A549 and MCF-7 shows the highest correlation with those in Mesenchymal Stem cells and Splanchnic mesorderm. However, our proposed matching normal cell lines, such as like IMR-90 for A549, still showed decent correlation in terms of their replication timing profiles.

*Figure S 1-3. Comparison of several Tumor cell lines with normal ones with replication timing data*




## 1.5 Summary of data from each experimental assay from ENCODE

We have integrated uniformly processed and quality-controlled datasets from ENCODE and Roadmap Epigenomics Mapping Consortium (REMC) to build one of the most comprehensive representation of how functional regulatory elements interplay in human genome. All dataset used in the analysis were mapped to a standardized version of the GRCh37 (hg19) reference human genome. We used ENCODE dataset that were submitted and released up to October 31st, 2016 (Oct 2016 freeze).

### 1.5.1 Collection of RNA-seq data

### 1.5.2 Preprocessing of Repli-seq data

The raw signal of 90 Repli-seq data sets for 15 different tissue or cell lines were downloaded from the ENCODE data portal (link here). For each tissue/cell line, in cell cycle phases G1, S1, S2, S3, S4, and G2, newly replicated DNA positions were analyzed by massively parallel sequencing were sequenced \cite{21957152}. Simiar to \cite{21957152}, we added up the signal

strength in 1mb bins by comparing the (G1 + S1) with the (S4 + G2) datasets by measuring the inverse tangent (arctangent) for each data point \cite{20359321}.

### 1.5.3 Deduplication of CHIP-seq data

We collected 1,040 TF ChIP-seq experiments released for ENCODE. There are 888 released TF ChIP-seq experiments for ENCODE 2. We used a subset of 801 experiments that either had no treatment or ethanol treatment only. There were 570 TF ChIP-seq experiments released for ENCODE 3, which had no treatment.

For a common TF target in top-tier cell lines, ENCODE has multiple of the same experiments from different labs. We carefully de-duplicated dataset by selecting one TF ChIP-seq experiment per each sample by the following prioritization scheme. When ENCODE 3 experiment was available, it was prioritized over ENCODE 2 experiment. When there was the same type of experiments were done by different labs, we prioritized using the following order determined by the total number of ChIP-seq experiments deposited on ENCODE: stanford, haib, broad, usc, uw, uta, uchicago, hms, yale. We removed epitope-tagged experiment if endogenous antibody was available. After deduplication, there are 860 unique TF ChIP-seq experiments.

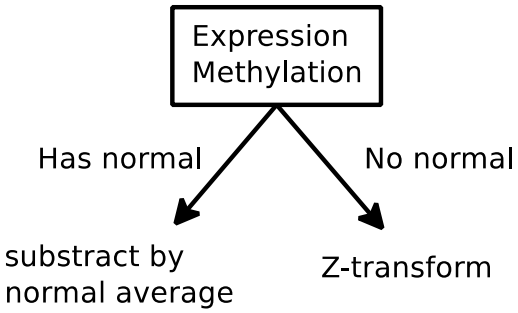
## 1.6 External data

We deeply integrated our ENCODE functional characterization data with data from external cohorts to interpret cancer genome. Specifically, we downloaded both expression and WGS data from external cohorts.

### 1.6.1 Expression data from external Cohort

All TCGA expression, methylation and mutation data were downloaded from GDAC firehose (http://gdac.broadinstitute.org) with data version of 2016_01_28. For cancer types with normal control samples profiled, the expression values of each gene are subtracted with the average value of all normal controls. For cancer types without any normal samples profiled, the expression profile of each gene is transformed to zero mean and unit deviation (see Figure S 1-4). The DNA methylation values are also normalized in the same way as RNA-Seq data, according to the availability of normal control samples in each cancer type. For copy number alteration (CNA), GDAC firehose doesn't provide standardardized data and we downloaded the data matrix from cBioportal with data version of 2016_10_20 (http://www.cbioportal.org).

*Figure S 1-4 Schemetics of RNA-seq data processing*




### 1.6.2 WGS data

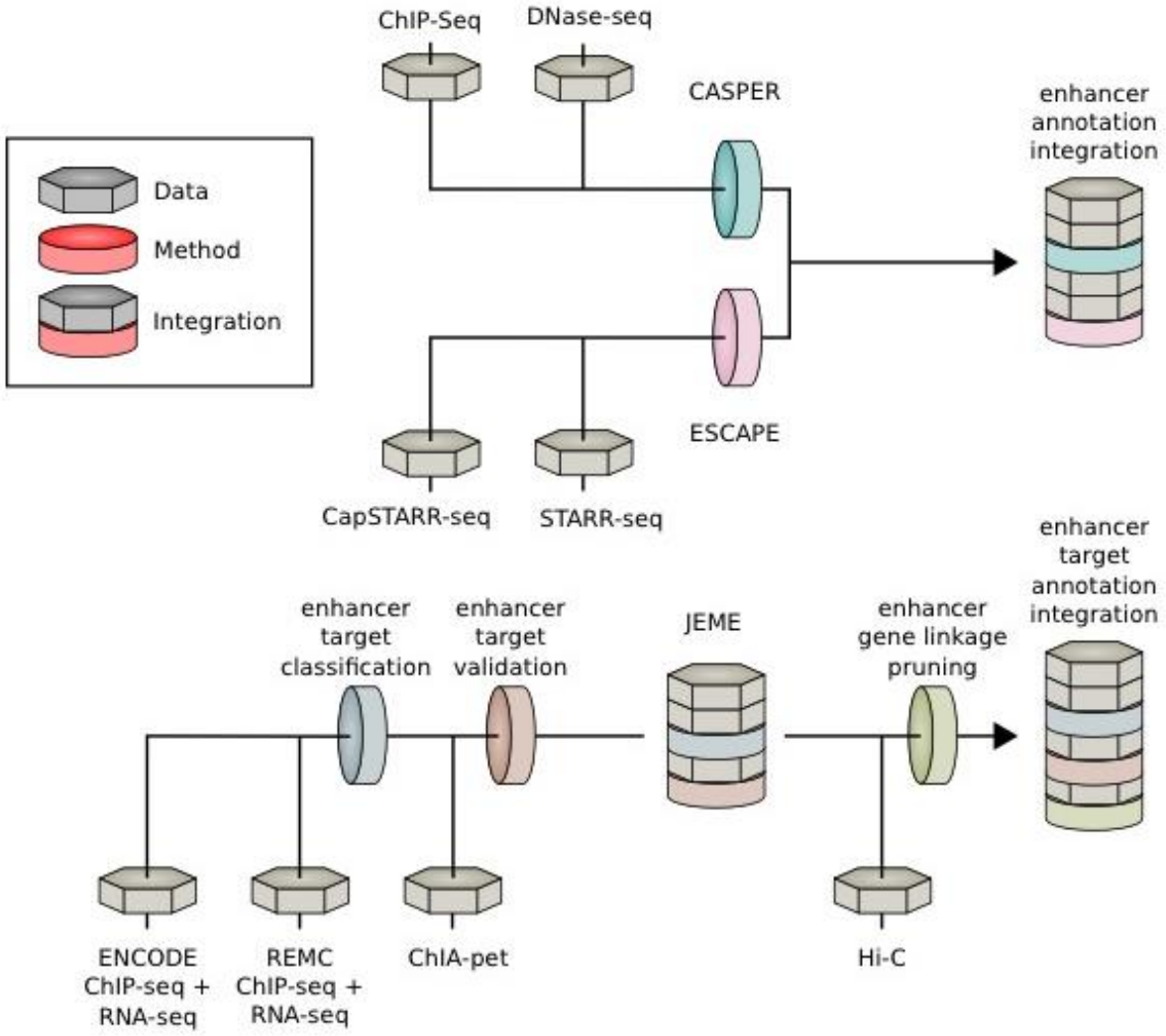
2709 WGS samples were collected for 5 cancer types (BRCA, LAML, LUAD, LIHC, UCEC).

## 1.7 An Ensemble to predict enhancers and their gene linkages

In contrast to previous approaches to enhancer annotations (many of which use only histone modification and chromatin accessibility data \cite{22373907}), we proposed an ensemble method to accurately pinpoint active enhancers and link them to protein coding genes. It composes three computational pipelines (CASPER, ESCAPE, and JEME) to integrate tens of datasets from six different experimental assays, including ChIP-Seq, DNase-Seq, STARR-Seq (CapSTARR-Seq), RNA-seq, ChIA-pet, and Hi-C for higher accuracy. The overall schematic has been summarized into Figure S 1-5.

For the enhancer prediction part, our scheme combines large-scale STARR-seq experimental data with computational predictions based on pattern recognitions of histone marks (Figure S 1-5). Here we developed two pipelines CASPER and ESCAPE for each of them. Eventually we assemble results from these two pipelines for accurate enhancer identification. Enhancer targets were then predicted using JEME and further pruned by the Hi-C results.

*Figure S 1-5 Overall schematic of enhancer and gene linkage prediction by large scale data integration*




### 1.7.1 Enhancer prediction Pipeline based on ChromAtin Shape PattErn Recognizer (CASPER)

We first developed a framework to impute enhancer regions across the genome through aggregated signals of epigenetic features. The unprecedented large number of massively parallel

reporter assays (MPRA) has demonstrated that regulatory regions are generally depleted of histone proteins while regions around it tends to contain histone proteins with certain post-translational modifications \cite{26072433}. This characteristic is revealed in many ChIP-Seq experiments as enriched peak-trough-peak (double peak) signal at the distal regulatory regions for many activating histone marks. A supervised machine-learning model is well suited to identify this pattern.

For each histone modification, we aggregated the ChIP-Seq signals around STARR-seq identified peak regions. The two maxima in each region is aligned, interpolated and smoothened before averaged to generate meta profile. An additional flipping step was applied to maintain the asymmetry of the two maxima since it might be associated with the directionality of transcription. The meta profile is then used to scan the whole genome to find matched patterns through a shape-matching filter. A 10-fold cross validation is performed to assess the accuracy of prediction through this method. In predicting active STARR-Seq peaks, H3K27ac is the most accurate feature for predicting active regulatory regions (AUROC=0.92). Other features including H3K4me1, H3K4me2 also achieved high performance.

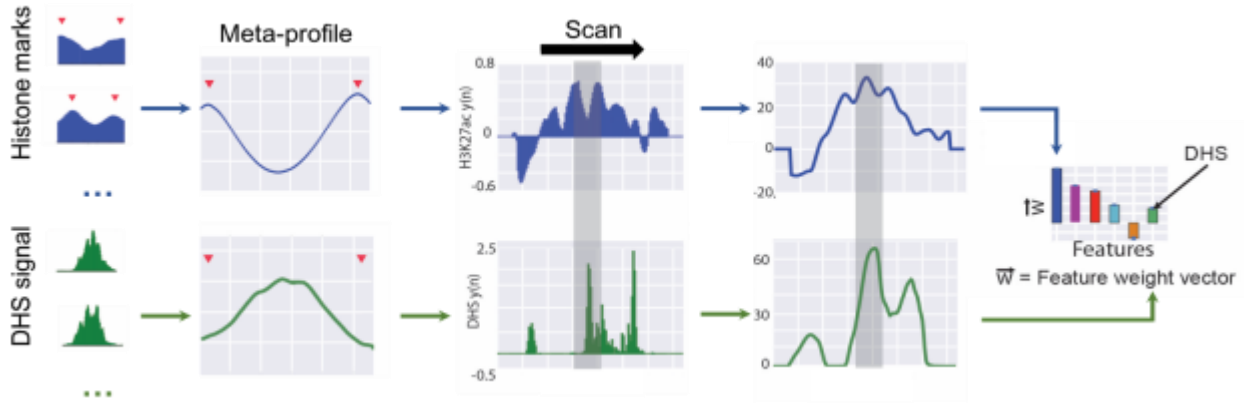
To achieve higher accuracy, we further developed an ensemble method to combine the normalized pattern-matching result from several different epigenetic marks with linear SVM (Figure S 1-5). This include ChIP-Seq signals for H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K9ac and DHS signals associated with active regulatory regions. The ChIP-Seq data is available through ENCODE Consortia (https://www.encodeproject.org) and Roadmap Epigenomics (http://www.roadmapepigenomics.org). The integrated model performs better than each of the individual histone marks, and different integration methods perform similarly. We use linear SVM to assemble the signals to form a discriminant function, where the sign of the result value is used

to predict whether a specific region is an enhancer. The resultant enhancers have been summarized in Table S 1-3.

*Table S 1-3 Number of enhancers predicted by histone-shape based method*

| | GM2878 | HepG2 | K562 | MCF7 |
|---|---|---|---|---|
| Number of Enhancers | 45202 | 61005 | 45801 | 59827 |

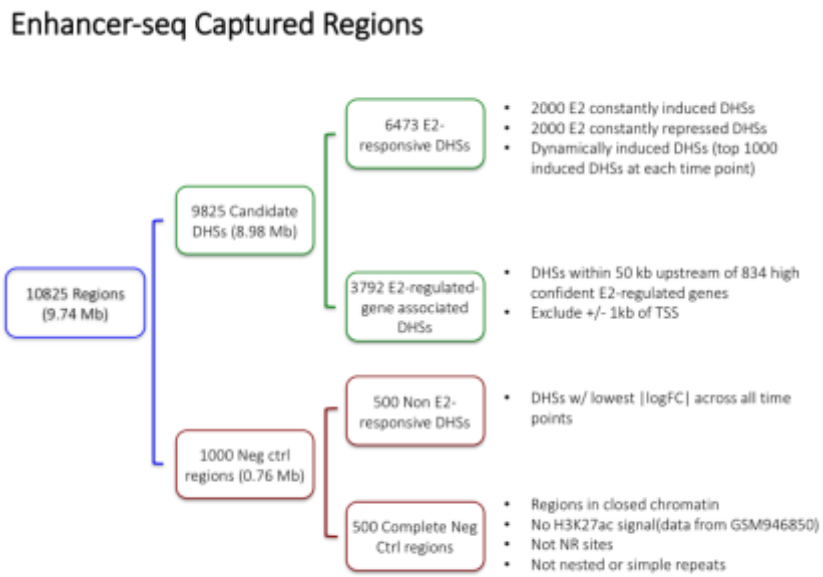
*Figure S 1-6 schematic of shape based enhancer prediction method*




## 1.7.2 Enhancer prediction by EnhancerS Peak CAlling PipEline from STARR-seq (ESCAPE)

The whole-genome STARR-seq was performed using a protocol conceptually similar to the previously published STARR-seq technique that was done in the *Drosophila* melanogaster genome \cite{23328393}. The CapSTARR-seq is a variant of STARR-seq technique which combines STARR-seq with genome capturing technology \cite{25872643}. In brief, the genomic DNA from each cell line was fragmented into ~500 bp by sonication and built into plasmid library, which was named as screening library. The screening library was subjected for Next Generation Sequencing. We verified the sequence complexity and genome coverage of screening libraries, which were then

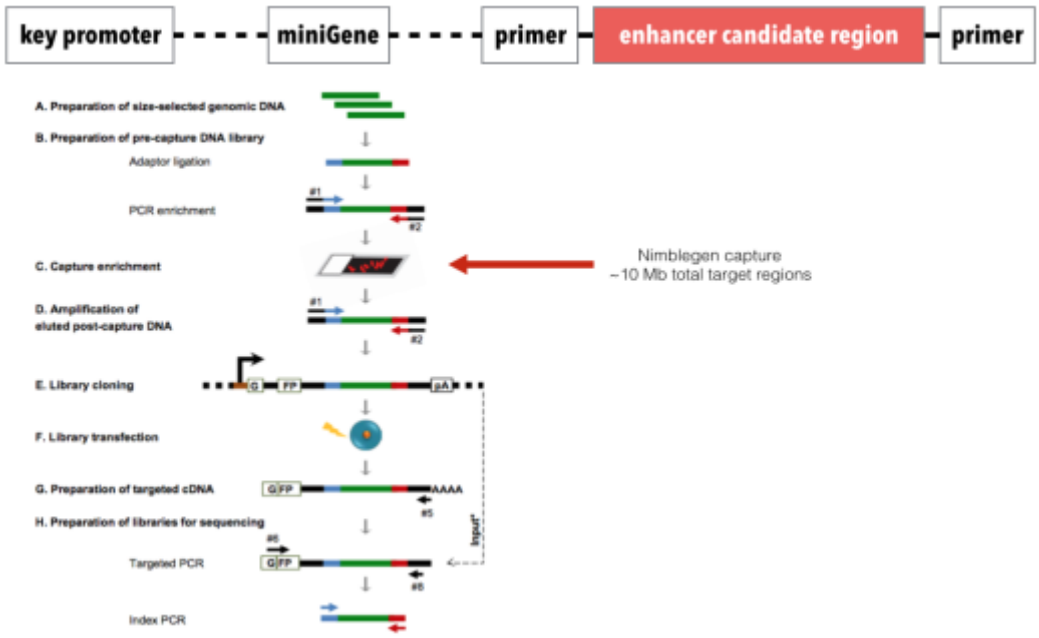
*Figure S 1-7 flowchart of capture EnhancerSeq target region selection procedure*




transfected into GM12878, K562 or MCF7 cells by electroporation. After 24 hours of transfection, the plasmid-specific mRNA was purified, reverse transcribed and PCR amplified. The PCR products, which are the so-called STARR-seq libraries, were subjected to sequencing. Both screening library and STARR-seq libraries were sequenced as 100 bp paired-end on the Illumina HiSeq 2500/4000 platforms. The general workflow of the MCF7 CapSTARR-seq is similar to the whole-genome STARR-seq, however, we captured ~10,000 DNase I hypersensitivity sites (a total length of 9.7 Mb) from fragmented genomic DNA to build the screening library. Compared to the published STARR-seq work, we'd like to note the following innovation and improvement: (1) We significantly increased the complexity of the screening libraries to ensure comprehensive coverage to the human genome; (2) We significantly increased the electroporation scale and efficiency to maximize the size of screening library that got into the cells; (3) We introduced an extra multiplexing step to minimize the bias introduced by PCR duplicates. For the capture based assay for MCF-7 cell line, total of 10,825 target regions consisting of 9,825 candidate enhancer regions

and 1,000 negative control regions were selected tested for regulatory potential. Candidate enhancer regions were selected based on DHS peaks excluding both 1kb upstream and downstream of TSS. Negative control regions were selected from 500 randomly selected regions and 500 non-E2-responsive DHS regions. Details of the selection procedure can be found in Figure S 1-6. (L. Ma et al for GM12878 and K562 whole-genome STARR-seq; S. Yu et al for MCF7 CapSTARR-seq, in preparation). Candidate enhancer regions were primed and inserted into 3' UTR. Schematics of the experimental procedure can be found in Figure S 1-6.

*Figure S 1-8 Capture STARR-seq experiment design*

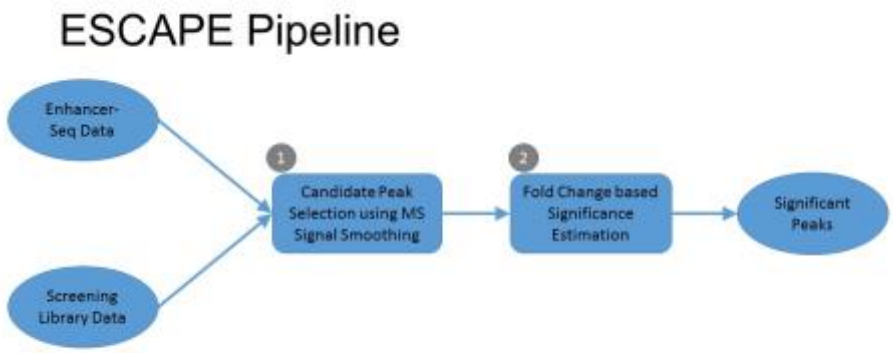



*1.7.2.1 Whole genome Enhancer-seq data processing*

To uniformly process the data from whole genome and capture-based EnhancerSeq assays, we developed a new analysis pipeline named ESCAPE (Figure S 1-9). The pipeline is tailored for optimally processing the output from EnhancerSeq experiments. The output from an EnhancerSeq experiment is two datasets for each cell line. First is screen library that contains the sequencing of plasmids from which the enrichment is performed. This screen library serves as a control in the

EnhancerSeq analysis. Second is the actual enhancer-seq enriched sequencing data that contains the actual enhancer signal. We have removed low quality reads and mapped them using BWA version 0.7.12 \cite{19451168}. We have used the reference genome from 1000 Genomes Project's decoy genome\cite{20981092}. ESCAPE then removes the reads with mapping quality lower than 20 and removes PCR duplicates and estimates fragment length distribution using cross-correlation between the strands (Figure S 1-10). Then the enhancer-seq signal tracks are generated and library and performed peak calling. The Enhancer-Seq signal shows lower fold change characteristics compared to ENCODE ChIP-Seq datasets (Figure S 1-10).

*Figure S 1-9 Schematic of ESCAPE pipeline*




For peak calling, ESCAPE uses the following strategy: First the peak candidates are identified. For the whole genome assay, ESCAPE uses a multiscale decomposition based peak calling strategy \cite{25292436}. For this, we have decomposed the signal using smoothing filters with lengths varying between 100 and 2000 base pairs. The filtering can be summarized with following formula:

$$x_i^s = \text{median}\left(\{\tilde{x}_a\}_{a\in\left[i-\frac{l_s}{2},i+\frac{l_s}{2}\right]}\right), l_s \in (l_{start}, \lfloor l_{start}\times\sigma\rfloor, \cdots, l_{end})$$

where $x_i^s$ is the $i^{th}$ signal level at scale decomposition $s$. The smoothing window length is $l_s$. Then we identified the local minima in the smoothed signal profiles and used these as possible enriched regions. For this, ESCAPE first estimates the derivative at each point:

$$dx_i^s = (x_i^s - x_{i-1}^s)$$

where $dx_i^s$ is the derivative of the smoothed signal $x_i^s$. The local extrema are found as the points where the derivative flips its sign:

$$I_{min} = \{i \mid dx_i^s < 0, dx_{i-1}^s > 0\}$$

$$I_{max} = \{i \mid dx_i^s > 0, dx_{i-1}^s < 0\}$$

where $I_{min}$ and $I_{max}$ are the sets of positions of minima and maxima of $x_i^s$, respectively. The scale specific candidate enriched regions of $x_i^s$ are identified as the regions between the consecutive minima. The multiscale decomposition approach identifies enriched regions at different length scales that correspond to punctate features like enhancers. Then, ESCAPE computes the fold change on each peak candidate as the ratio of total signal in the enhancer-seq signal and screening library signal. We refer to this as $FC$:

$$FC = \frac{\sum_{i=s}^{e} x_i^s}{\sum_{i=s}^{e} y_i^s}$$

where $y_i^s$ represents the value of screening library signal profile at position $i$. For capture based assay, ESCAPE uses a more focused analysis to identify candidate peak regions. For each capture region, ESCAPE selects a bins size that balances the peak calling sensitivity and specificity. To set a threshold for the fold change to select candidate peaks, we exchanged screening library and enhancer-seq and we computed the fold change on the candidate peaks, which we refer to as $FC_{random}$:

$$FC_{random} = \frac{\sum_{i=s}^{e} y_i^s}{\sum_{i=s}^{e} x_i^s}$$

These fold change scores serve as a random distribution of fold change scores. We use this distribution for selecting a fold change threshold. For a $FC$ threshold ***fc***, we estimated the false discovery rate as the ratio of number of peaks that for which $FC_{random}$ > ***fc*** and the number of peaks for which $FC$ > ***fc***. We set the FDR threshold at 0.1% and filtered the peaks that do not satisfy the $FC$ threshold selected using this FDR threshold. For capture based assay, ESCAPE uses the candidate enriched regions with top 10% $FC$ values.

*Figure S 1-10. Whole genome Enhancer-Seq signal enrichment properties*

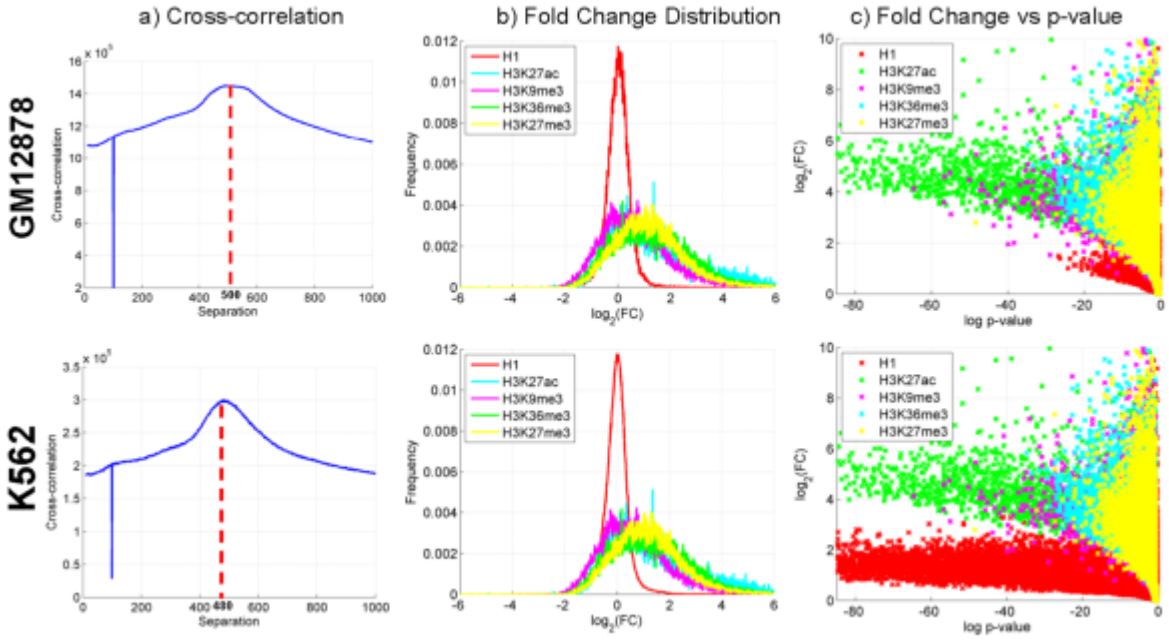
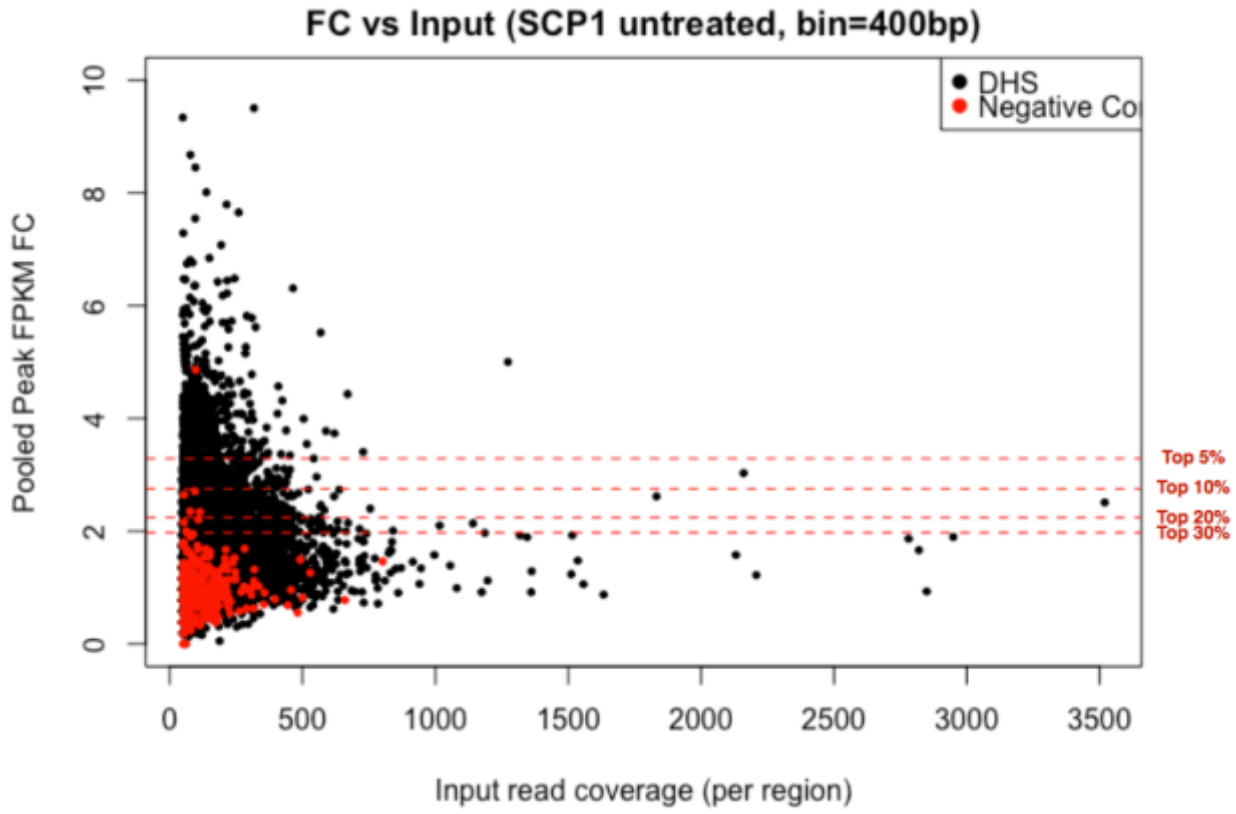


*Figure S 1-11 Capture STARR-Seq experiment properties*




## 1.7.3 Enhancer Target prediction

### *1.7.3.1 Enhancer Gene linkage prediction using JEME*

Enhancer targets were predicted using JEME (Joint Effect of Multiple Enhancers, Cao *et al.*, under review), which involves two main steps (Figure S 1-12). In the first step, the transcript levels around each transcription start site (TSS) in 49 ENCODE and Roadmap Epigenomics cell lines (Table S 1-4) were modeled based on histone modification data at nearby enhancers without requiring any known enhancer-target pairs as examples. Specifically, for each enhancer feature $i$ the expression level $y$ of a TSS is modeled as $y = a_{i0} + \sum_j a_{ij}\, x_{ij}$, where the summation is over all enhancers $j$ within 1Mbp from the TSS, and $x_{ij}$ is the value of feature $i$ of enhancer$j$. The

coefficients $a_{ij}$ of the enhancers are learned by LASSO, which minimizes the regression error over all samples while selecting a small number of enhancers to have non-zero coefficients. The features considered include H3K4me1, H3K27ac and H3K27me3 (A separate model involving only the latter two features was built when constructing the enhancer-target network in MCF7 since H3K4me1 data were unavailable).

In the second step, single-enhancer error terms were first computed. Specifically, an error term is computed to check how much the expression $y_k$ of the TSS in sample $k$ can be explained by considering each feature $i$ of each enhancer $j$, i.e.,$e_{ijk} = |y_k - (a_{i0} + a_{ij}x_{ijk})|$, where $x_{ijk}$ is the value of feature $i$ of enhancer $j$ in sample $k$ and $a_{i0}$ and $a_{ij}$ are the coefficients learnt in the first step. These error terms were then combined with genomic distance and cell-line-specific data (i.e. the levels of histone modifications across the enhancer, the TSS and the window between them in sample $k$) to predict the enhancers that regulate a TSS in a particular cell line using a Random Forest model. The parameter values of these second-level models were learned from published ChIA-PET data from K562 and MCF7 cell lines. A 5-fold cross-validation procedure was used to evaluate the accuracy of the predicted enhancer-target pairs. The model was then applied to those samples without ChIA-PET data.

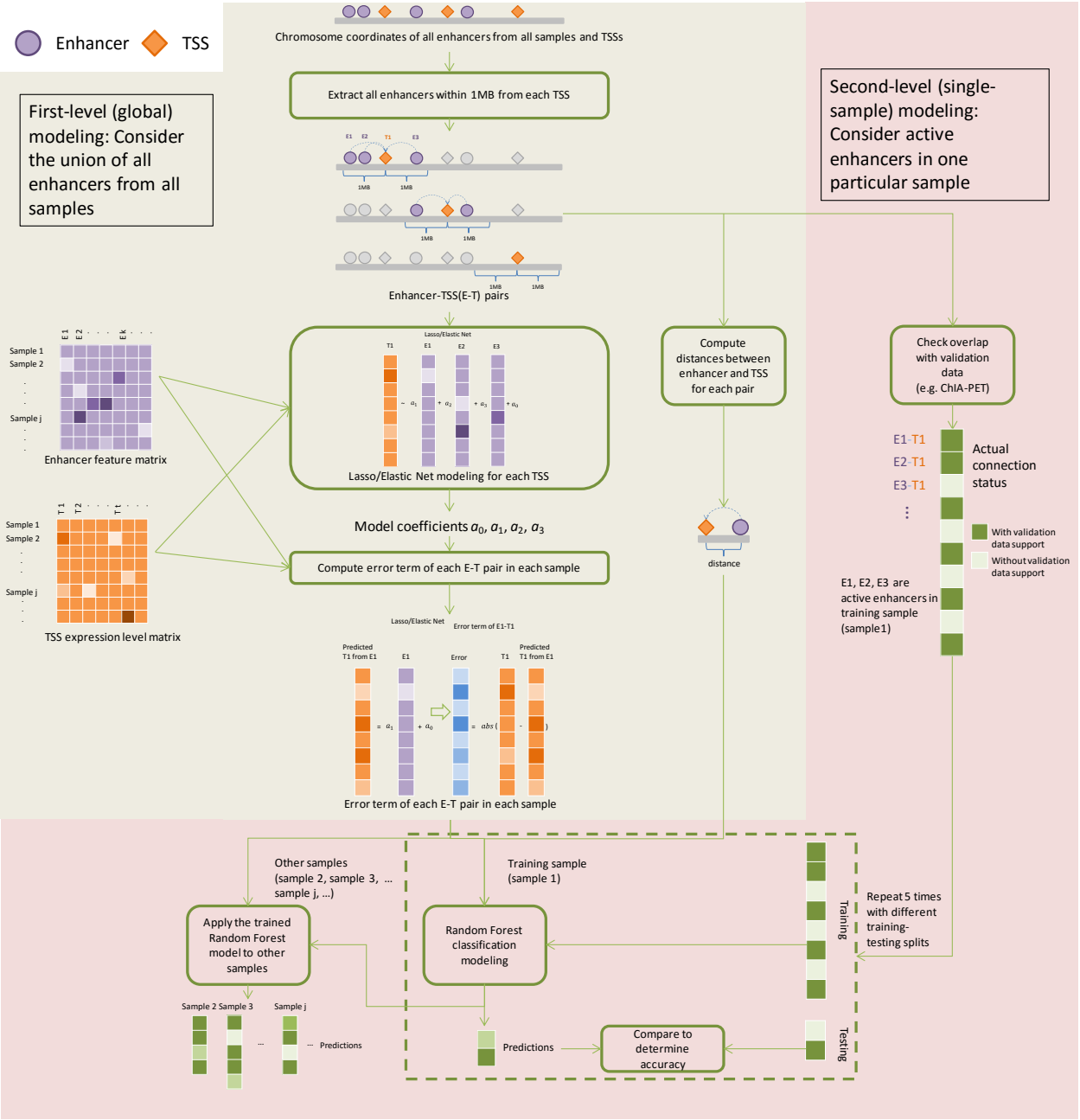
*Table S 1-4 The 49 ENCODE and Roadmap Epigenomics cell lines used to construct enhancer-target networks by JEME*

| Data source | Cell lines |
|---|---|
| ENCODE | GM12878, HepG2, K562, MCF7 |
| Roadmap Epigenomics | E003,E004,E005,E006,E007,E011,E012,E013,E016,E037,<br>E038,E047,E050,E055,E056,E058,E059,E061,E062,E065,<br>E066,E071,E079,E084,E085,E087,E094,E095,E096,E097, |

| | |
|---|---|
| | E098,E100,E104,E105,E106,E109,E112,E113,E114,E117,<br>E119,E120,E122,E127,E128 |

### *1.7.3.2 Enhancer gene linkage pruning using Hi-C data*

*Figure S 1-12 Schematic of JEME*



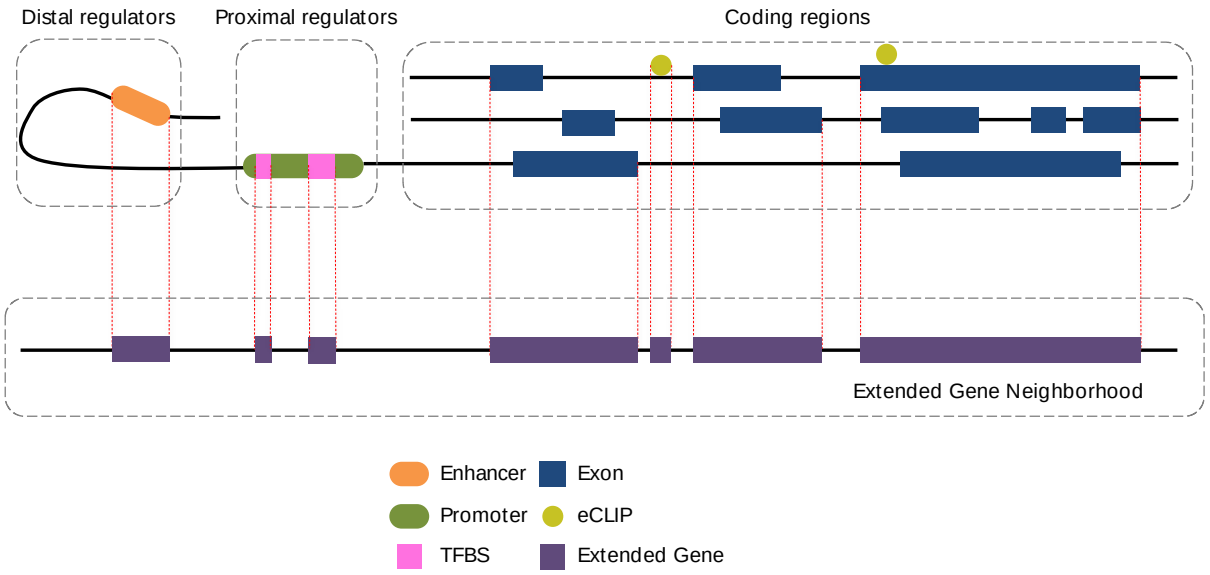

Enhancer target predictions are further filtered by using Hi-C data. Contact maps of individual chromosomes (in 5kb bins) for both K562 and GM12878 cell lines were obtained from (Rao et al. Cell 2014). MCF7 contact maps (40kb) were obtained from (Barutcu et al. Genome Biol. 2015).

Element (i,j) in a contact map represents the frequency of interactions between genomic loci i and j. For all possible (i,j), we used the tool Fit-Hi-C to estimate the statistical significance the contact frequency based on the coverage of the loci as well as their genomic distance (Ay Ges. Res. 2014) and kept the interactions with q-value<0.1. We then used the list of significant loci to filter the enhancer-target predictions. Only enhancer-gene pairs in which enhancer and gene are respectively belong to a pair of significantly interacting loci are kept for further analysis.

## 1.8 Extended gene neighborhood generation

Here we generated the extended gene neighborhoods by combing the coding region with the key non-coding proximal and distal regulatory elements together for a joint mutation burdening quantification. Details of the schematic is given in Figure S 1-13.

*Figure S 1-13 Schematic of extended gene definition*



## 1.9 TF/RBP networks

### 1.9.1 TF network

[JZ2DL: to be added here]

### 1.9.2 RBP network

[JZ2Peng: Please add something here how the RBP network was built]

# 2 Details about recurrence analysis

## 2.1 Variant calling

### 2.1.1 Germline

We called germline single nucleotide variants (SNVs) for a set of 88 liver cancer samples (Table 1) that were whole genome DNA sequenced at the Beijing Genomics Institute (BGI) Shenzhen for a mutation analysis published in [ref: PMID 23788652]. The authors made the raw sequence data available in FASTQ format from the European Nucleotide Archive (ENA) under accession ERP001196. We downloaded these files and conducted a germline variant calling procedure in accordance with the Broad Institute's Best Practices for read-to-variant workflows (https://software.broadinstitute.org/gatk/best-practices/index.php). Read alignments were generated using the Burrows-Wheeler Aligner (BWA v0.7.15; http://bio-bwa.sourceforge.net/), using the BWA-MEM algorithm. After that, we proceeded with preprocessing for variant calling, including cleaning out duplicate reads using Picard tools (MarkDuplicates tools v2.6.0), and base recalibration with the Genome Analysis Tookit (GATK; v3.6.0). Variant calls for individual samples were derived with the GATK HaplotypeCaller, followed by joint genotyping with the GenotypeGVCFs tool. The final variant set was subjected to standard quality filtration in accordance with the standard configuration of the GATK VariantFiltration tool. Each step was performed on the Mt Sinai Minerva scientific compute cluster, and utilized hundreds of CPU cores per compute step. Table S 2-1 summarizes the distribution of germline variant calls per sample.

*Table S 2-1 List of cancer whole genome DNA sequence data obtained for variant calling*

| Cancer type | Number of samples | Median number of variants per sample | Source |
|---|---|---|---|
| Liver - germline | 88 | XXX | BGI Shenzhen (Kan *et al.* 2013) |
| Liver - somatic | 88 | XXX | BGI Shenzhen (Kan *et al.* 2013) |
| Breast | 116 | 8485 | TCGA |
| Lung | 197 | 83,402 | TCGA |
| Chronic lymphocytic leukemia (CLL) | 150 | XXX | ICGC |

### 2.1.2 Somatic

In addition to the aforementioned liver cancer samples, we obtained the BAM files for 116 Breast Invasive Carcinoma whole genomes, and 197 lung cancer whole genomes (147 Lung Adenocarcinoma, 50 Lung Squamous Cell Carcinoma). Furthermore, BAM files corresponding to 150 chronic lymphocytic leukemia (CLL) whole genomes were obtained from the International Cancer Genome Consortium (ICGC) via the European Genome-Phenome Archive (EGA). Somatic variant calls were derived from the Broad Institute's Mutect (v1.1.4) and Strelka (v1.0.15). This variant calling compute was performed on the Mt Sinai Minerva scientific compute cluster, and utilized hundreds of CPU cores per compute step. Table S 2-2 summarizes the distribution of somatic variant calls per sample.

*Table S 2-2 Summary of distribution of variant calls per cancer sample*

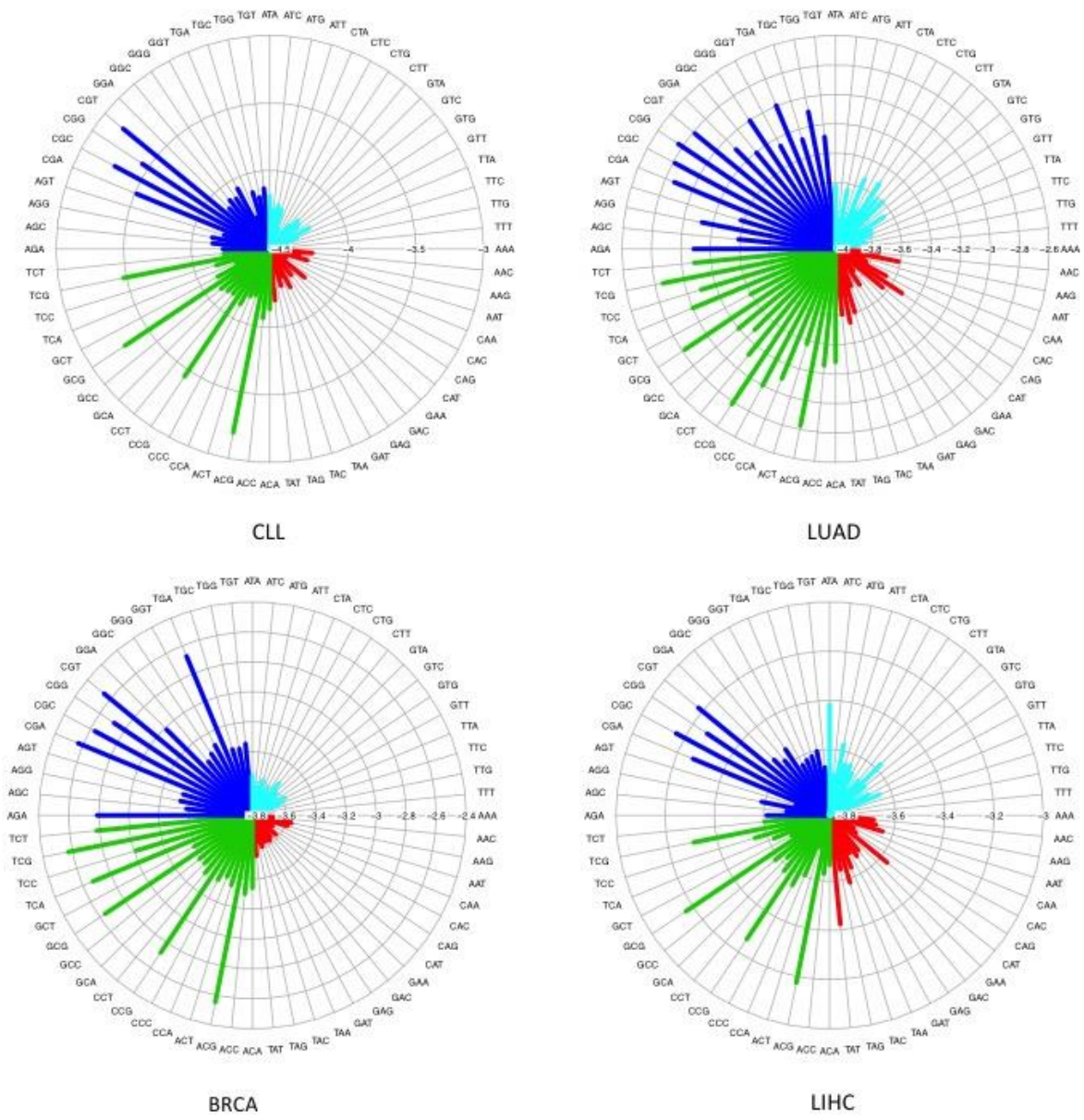
| Cancer Type | Summary statistics of variants per sample | | | | | |
|---|---|---|---|---|---|---|
| | Min | 1st Qu | Median | Mean | 3rd Qu | Max |

| | | | | | | |
|---|---|---|---|---|---|---|
| Liver - germline | XXX | XXX | XXX | XXX | XXX | XXX |
| Liver - somatic | XXX | XXX | XXX | XXX | XXX | XXX |
| Breast | 1898 | 5779 | 8485 | 13,290 | 14,370 | 294,100 |
| Lung | 159 | 40,490 | 83,400 | 229,200 | 295,000 | 2,127,000 |
| Chronic lymphocytic leukemia (CLL) | XXX | XXX | XXX | XXX | XXX | XXX |

## 2.2 Local context effect significantly affect local mutation rate (JZ)

We observed that BMR is significantly associated with local context effect in all cancer types up to several orders, which largely contributes to the mutation rate heterogeneity. Details are given in Figure S 2-1. For example, the average pooled mutation rate ranges from 2.92e-03 to 1.58e-04 (1.8 fold). The observed mutation has been plotted in the following radial plot for each cancer type. In general, G/C positions are more prone to mutations as compared to A/T positions, but the local context effect within G/C positions still has strong effect (2.40e-04 and 2.40e-04 vs. 1.21e-03 and 1.20e-03). In addition, we also observed that the local context effect varies significantly across multiple cancer types. Hence, it is important to separate cancer types during the BMR estimation process.

*Figure S 2-1 Local context severely confounds BMR in multiple cancer types*

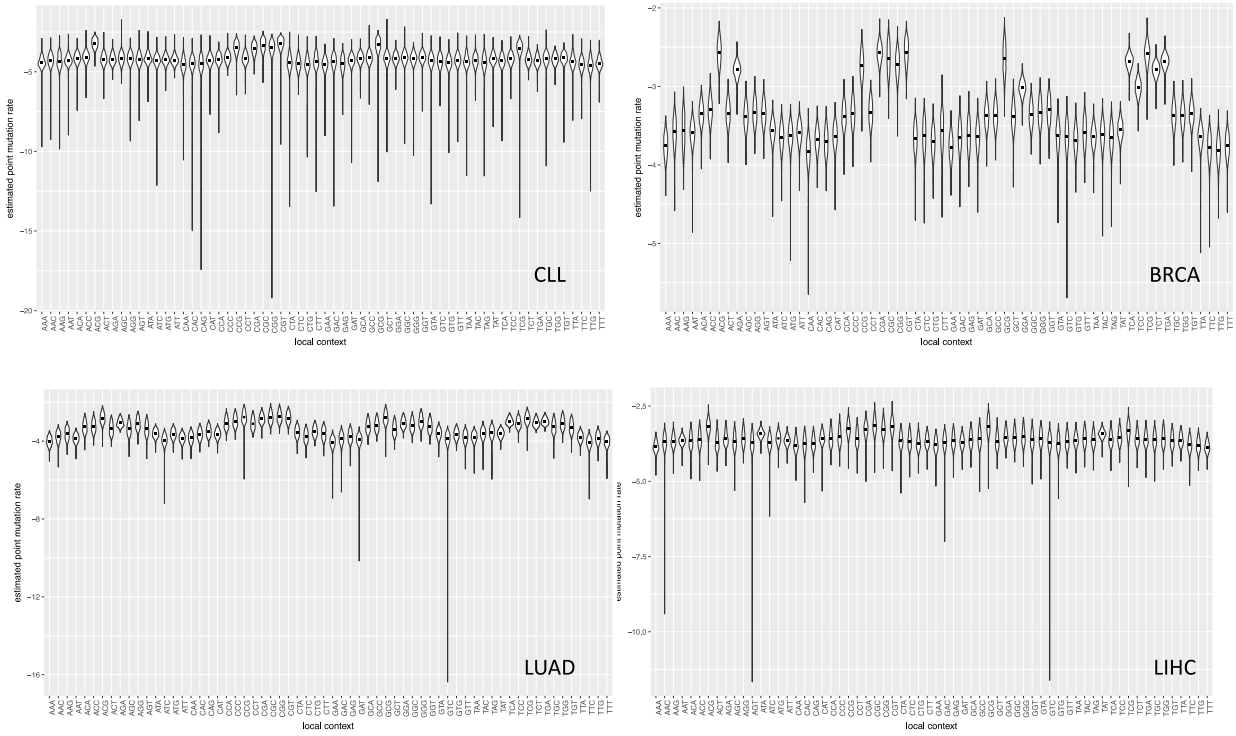



## 2.3 Local mutation rates are highly correlated with many genomic features

Consistent with previous literatures, we observed huge mutation heterogeneity over the genome for all 3mers in all cancer types \cite{23770567}. As seen in Figure S 2-2, the mutation

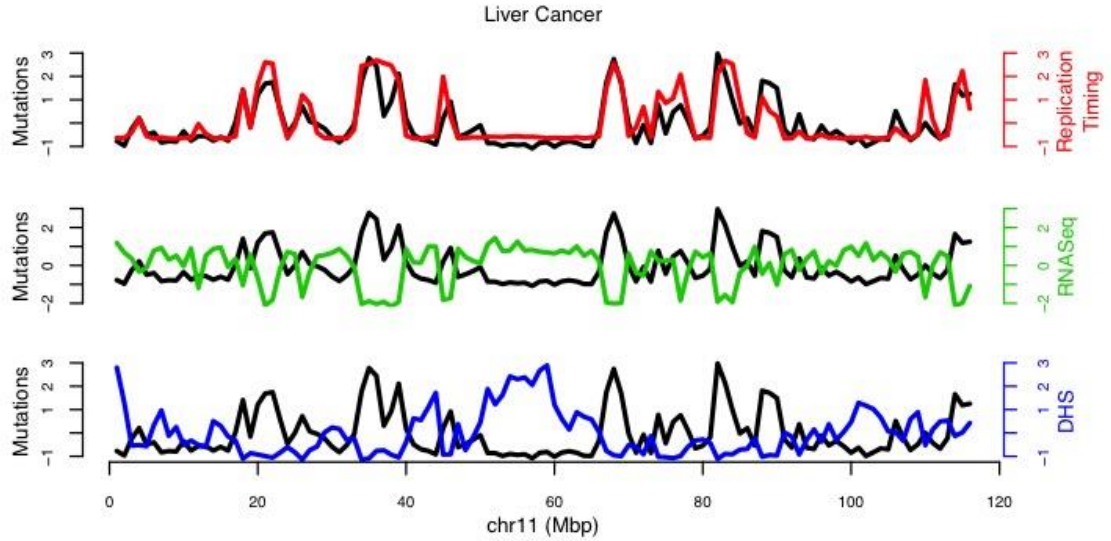
rate changes significantly over different region of the genome (large region of each violin bar) and over different local contexts.

*Figure S 2-2 violin plot of estimated mutation rate over local context and genomic locations in all four cancer types*




It is well-known that the somatic mutational process is affected by various external effects, such as replication timing and chromatin status. We also observed this phenomenon in many cancer types. For example, the normalized pooled mutation rates in the 1mb bins are given in Figure S 2-3 chromosome 11. It correlated replication timing data quite well. On the contrary, it has negative correlation with both RNA-Seq and DHS signal in liver cancer. Hence, it is important to correct BMR against the confounding effects from these external genomic features.

*Figure S 2-3 example of external effects on Local mutation rate*



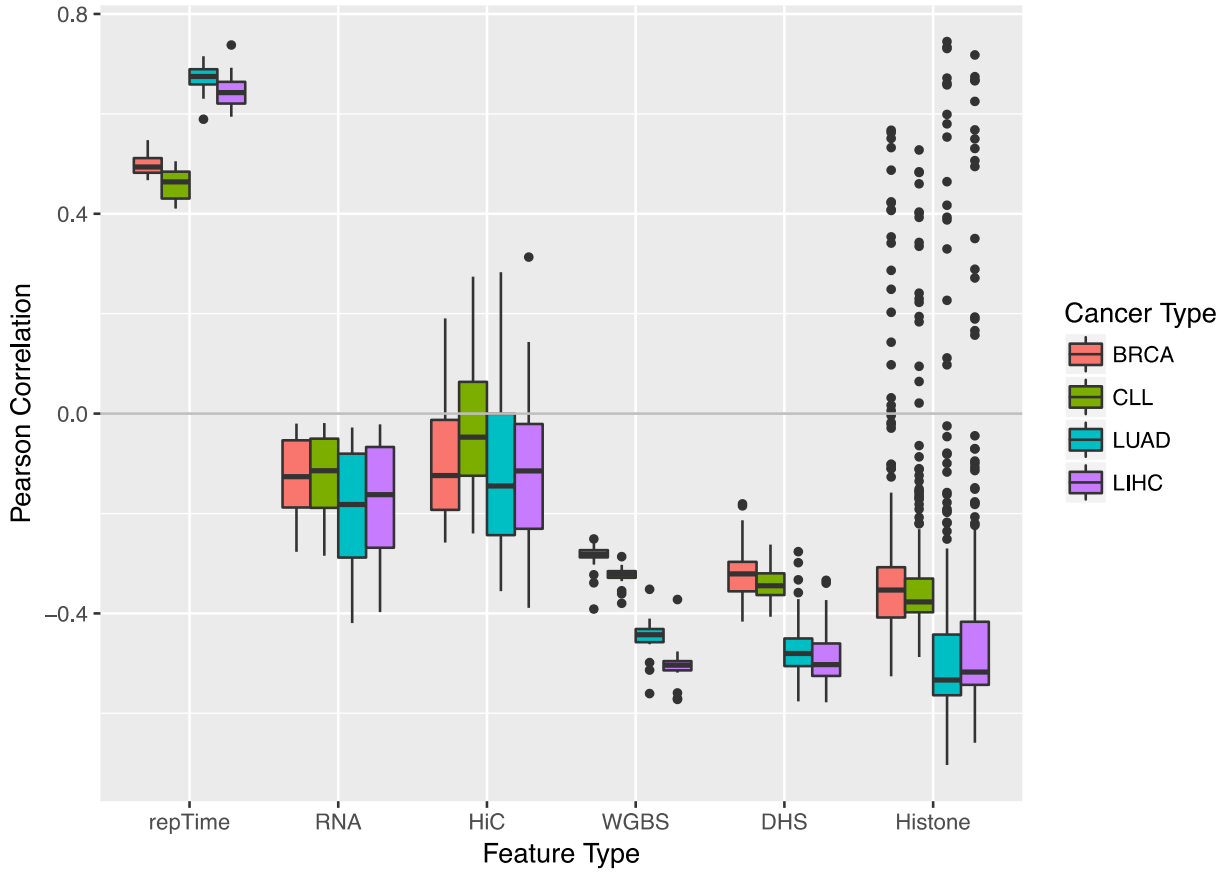

We systematically explored the effect of multiple genomic features including replication timing, DHS, WGBS, RNA-seq and Hi-C and their correlation with the overall mutation rate in multiple cancer types is given in Figure S 2-4. For example, in breast cancer, the correlation of replication timing and mutation rate ranges from 0.4673 to 0.5474, while correlation from DHS signals ranges from -0.4162 to -0.1806. However, we observed an increased correlation of mutation rates to these features in liver cancer (0.5943 to 0.7378 for replication timing and -0.5781 to -0.3337 for DHS, details in Table S 2-3). Hence it is important to correct the effect of external features in a cancer specific way to achieve better burden analysis.

*Table S 2-3 summary of correlation of mutation rate at 1mb bins with different external features in multiple cancer types*

| | | BRCA | CLL | LUAD | LIHC |
|---|---|---|---|---|---|
| repTime | Min | 0.467300058 | 0.410292073 | 0.589242196 | 0.594308714 |
| repTime | Median | 0.493598234 | 0.4637348 | 0.67495952 | 0.642464943 |

| | | | | | |
|---|---|---|---|---|---|
| repTime | Max | 0.547374543 | 0.504968332 | 0.715431489 | 0.737831334 |
| RNA | Min | -0.276898241 | -0.284573043 | -0.419458551 | -0.397504844 |
| RNA | Median | -0.126228292 | -0.114417754 | -0.181953795 | -0.162453668 |
| RNA | Max | -0.019998871 | -0.019088747 | -0.027760515 | -0.021616959 |
| HiC | Min | -0.258214534 | -0.240053286 | -0.355389238 | -0.388864195 |
| HiC | Median | -0.124167655 | -0.047070425 | -0.144698057 | -0.114903485 |
| HiC | Max | 0.190402416 | 0.274050935 | 0.283105375 | 0.31333984 |
| WGBS | Min | -0.391201031 | -0.379756186 | -0.560373539 | -0.572058872 |
| WGBS | Median | -0.28162047 | -0.321745409 | -0.442804565 | -0.503776857 |
| WGBS | Max | -0.250846402 | -0.28621958 | -0.35169477 | -0.37209731 |
| DHS | Min | -0.416168532 | -0.406787562 | -0.576209738 | -0.578067199 |
| DHS | Median | -0.321014839 | -0.344702489 | -0.480339053 | -0.502492181 |
| DHS | Max | -0.180644113 | -0.262259514 | -0.276360875 | -0.333674803 |
| Histone | Min | -0.525809902 | -0.487246457 | -0.703446634 | -0.658923263 |
| Histone | Median | -0.35312576 | -0.376874169 | -0.533489718 | -0.517573323 |
| Histone | Max | 0.567295147 | 0.527731502 | 0.74483745 | 0.717959133 |

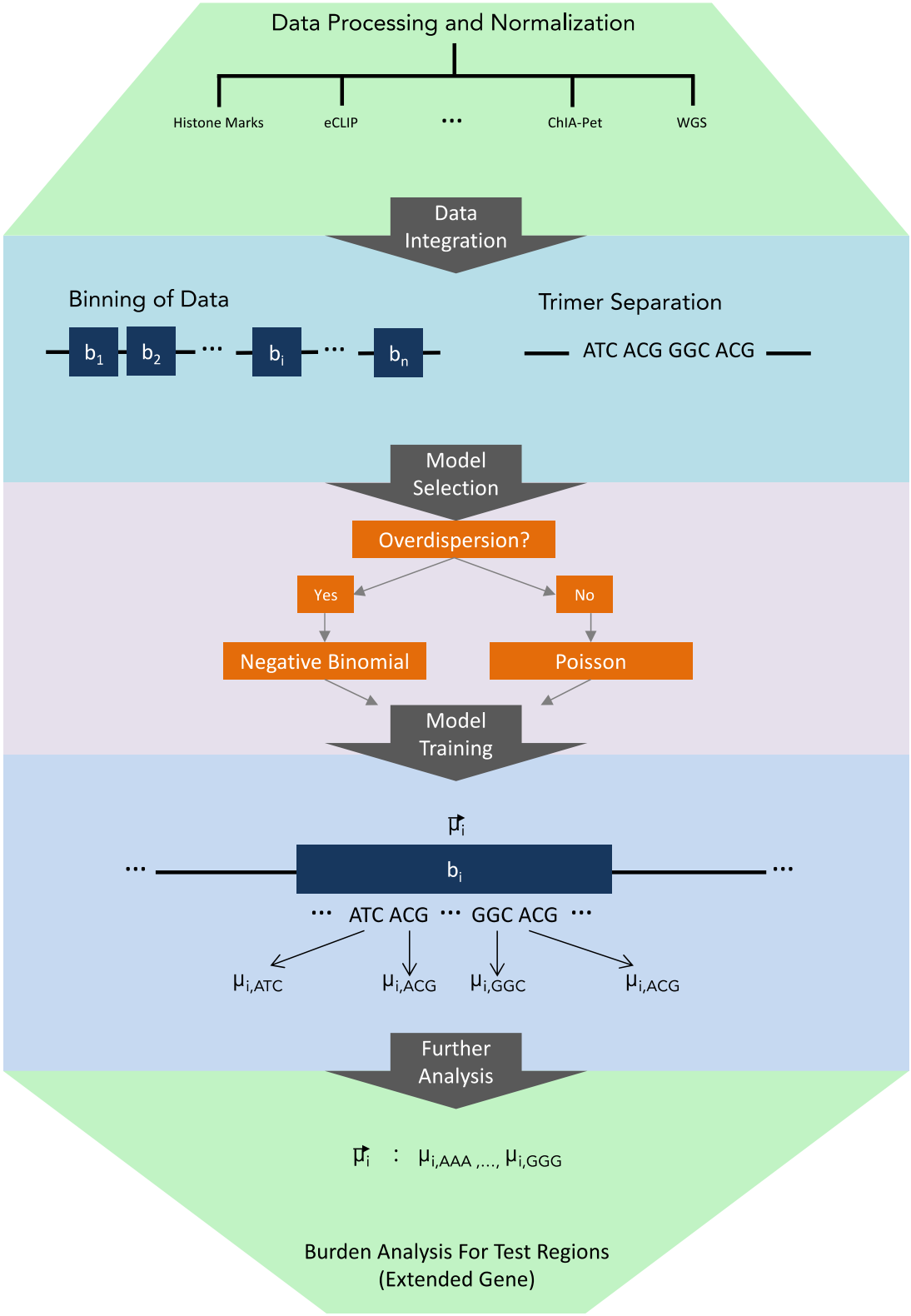
*Figure S 2-4 correlation of mutation rate and external features across multiple cancer types*




## 2.4 Background mutation rate estimation and P value calculation

Here we proposed a regression based somatic mutation recurrence analysis in cancer. The schematic of this method is shown in Figure S 2-5.

*Figure S 2-5 Schematic of the recurrence analysis*



### 2.4.1 Covariate data collection

We collected uniformly processed and non-redundant set of confounding genomic features across different cell types from both ENCODE to build the master covariate matrix that were used to correct for the background mutation rate (BMR). To ensure that the covariate matrix is not affected by processing bias and artifacts, we manually curated the dataset to have processed in the latest uniform processing pipeline and de-duplicated signal tracks from either untreated or ethanol-treated experiments. To build a covariate matrix, we then averaged the signal over specified 1mb bin size.

### 2.4.2 Covariate table creation

We aim to provide effective training of our model that is convenient for users. Different from the calibrated training data selection mentioned in \{cite 23770567}, we divided the whole genome into bins with fixed length, such as 1mb, 100kb, 50kb, etc. Only autosomal chromosomes and chromosome X were included in our analysis to remove the gender imbalance in mutation data or covariates.

Repetitive regions on human genome are known to generate artifacts in high throughput sequencing analysis mainly due to their low mappability. We downloaded the mappability consensus excludable table used in the ENCODE project from http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDacMapabilityConsensusExcludable.bed.gz. Any fixed length bins that overlap with this table would be removed from the training process. We also downloaded the gap regions of hg19 from the UCSC genome browser, which include gaps from telomere, short_arm, heterochromatin,

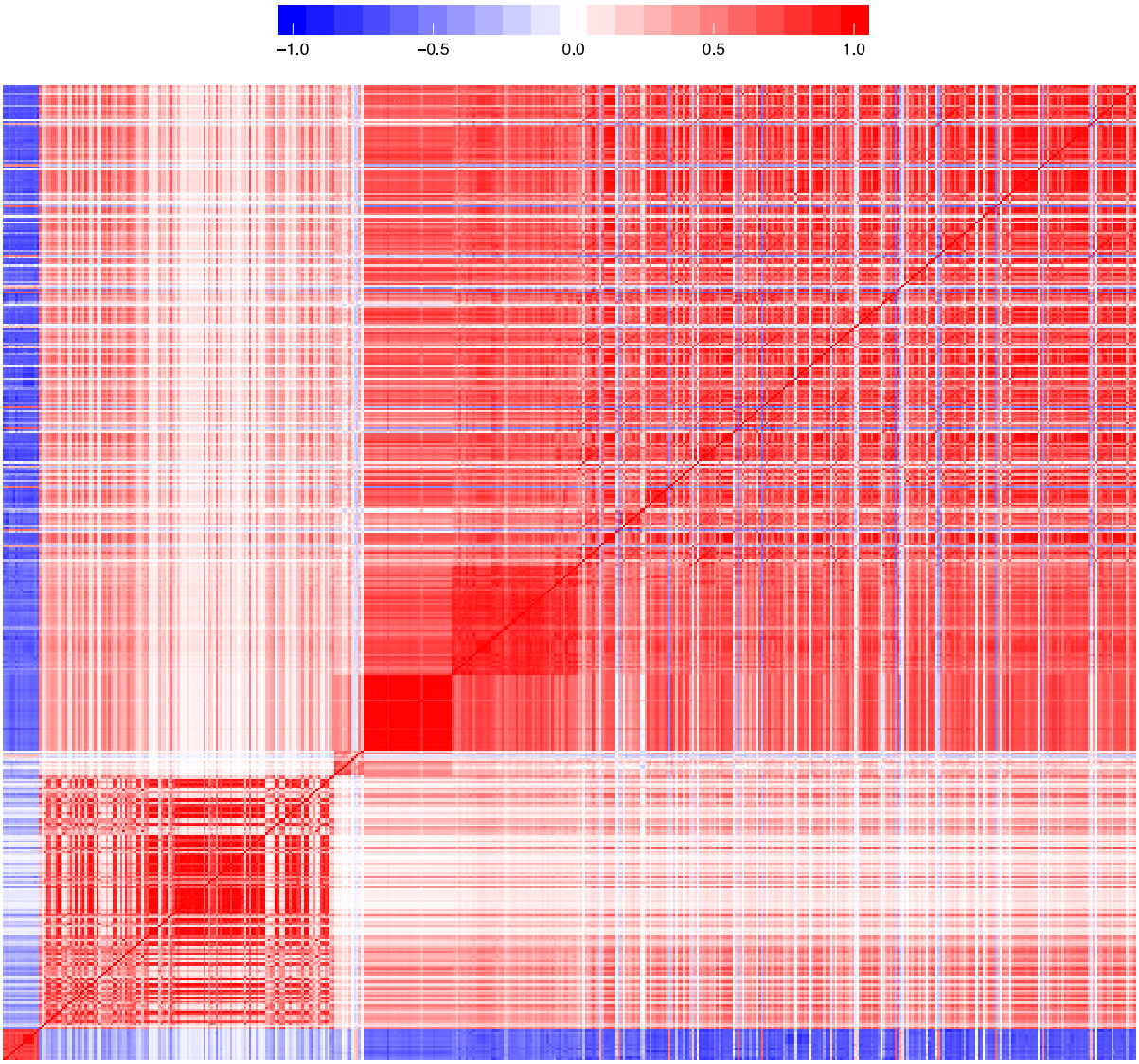
contig, and scaffold. The fixed length bins that intersect with these gap regions were also removed in our analysis.

All the bigWig files generated in step one were used to calculate the average signal using the bigWigAverageOverBed tool for each fixed length bin we generated above. In the end, we summarized all the covariates values in each bin into a covariate table, with 475 columns indicating different features and rows representing different training bins.

## 2.5 PCA analysis of the covariate matrix

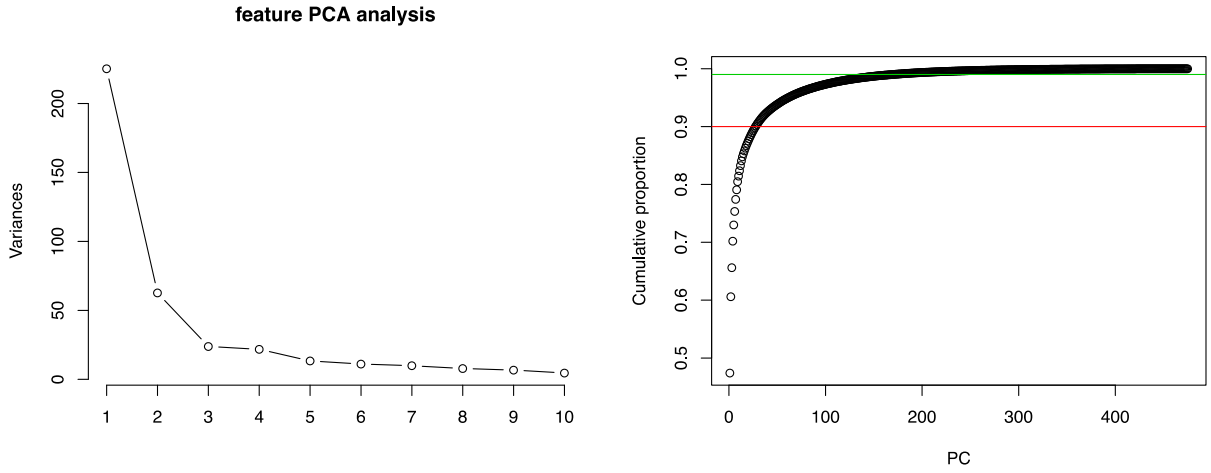
It has been reported that many genomic signal tracks demonstrate noticeable correlations across features and tissues. The heatmap of the pearson correlation of the 475 features have been given in Figure S 2-6. We observed strong correlations among the used features. For example, Pearson correlation of colors ranges from -0.874 to 0.998 at the 1mb bins.

*Figure S 2-6 Heatmap of feature correlations*



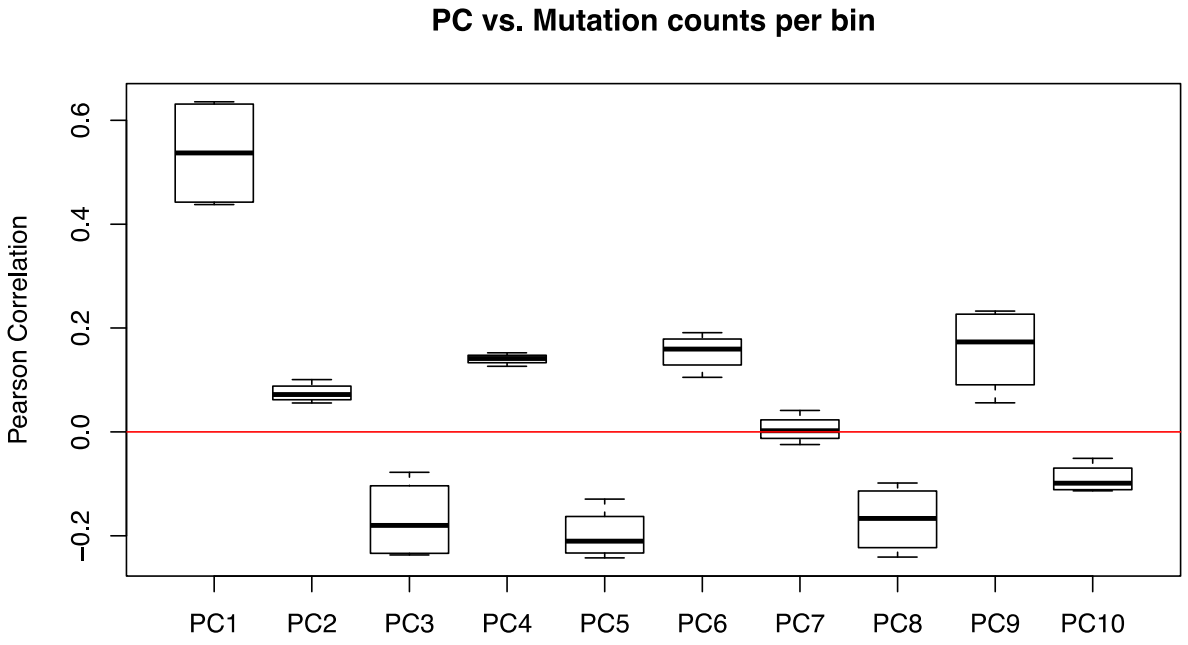

Hence we first centered and scaled the covariate matrix $X$ and then performed PCA on it to obtain $\acute{X}$. Then the cumulative proportion of variance explained by the PCs was given in Figure S 2-7. As expected, there is lots of redundancy in the covariate table. The first PC may explain as much as 47.41% of variance, while the 2nd PC explains an additional 13.19%. And it takes up to 28 and 169 PCs to capture 90% and 99% of variance (details in Figure S 2-7).

*Figure S 2-7 Summary of feature PCA analysis*




We also calculated the Pearson correlation of PC $j$ with mutation counts in cancer type $d$ as $\rho_j^d$. Then the absolute correlation value $\left|\rho_j^d\right|$ were averaged over different cancer types as $\widehat{\rho_j}$ to rank the PCs. The top 20 PCs with highest $\widehat{\rho_j}$ were selected and boxplot for each of the PCs was given in **Error! Reference source not found.**.

*Figure S 2-8. Boxplot of Pearson correlations of top PCs to mutation counts data in different cancer types*




## 2.6 Training model details

First we divide the whole genome into bins with fixed length $l$. In this stage, $l$ is usually large, such as 1 Mb. Any bins overlapping either of the two blacklist regions are removed. Then, 381 features are extracted from both REMC and ENCODE, and the average signal in the bins is calculated. We let $x_{i,j}$ denote the average signal strength for the $i^{th}$ bin and $j^{th}$ covariate, where $i = 1, \cdots, n$ and $j = 1, \cdots, m$.

Suppose there are $d = 1, \cdots, D$ different diseases (or disease types) in the collected WGS data, and $s = 1, \cdots, s_d$ unique samples, for example different patients, for each disease (or disease type such as liver cancer or lung cancer) $d$. Let $y_i^{d,s}$ and $\lambda_i^{d,s}$ denote the observed mutation count and rate for the $i^{th}$ bin defined above for sample $s$ in disease $d$. In previous efforts, scientists assume that mutation rate $\lambda_i^{d,s}$ is constant across different regions of the human genome, samples, and

diseases, so they have that $\lambda_i^{d,s} \triangleq \lambda$ for $\forall\, i, d, s$. Hence $y_i^{d,s}$ follows a Poisson distribution with the probability mass function (PMF) given in equation (1).

$$p\left\{Y_i^{d,s} = y_i^{d,s}\right\} = \frac{e^{-\lambda_i^{d,s}}\left(\lambda_i^{d,s}\right)^{y_i^{d,s}}}{y_i^{d,s}!} \triangleq \frac{e^{-\lambda}\lambda^{y_i^{d,s}}}{y_i^{d,s}!} \tag{1}$$

However, somatic genomes are highly heterogeneous because mutation rates vary considerably among various diseases, samples, and regions of the same genome, severely violating the assumption in equation (1). As a result, fitting of $y_i^{d,s}$ is usually very poor because overdispersion is often observed \cite{26304545}. Simply assuming a constant mutation rate will generate numerous false positives. Instead, in our model we assume that different $\lambda_i^{d,s}$ are random variables that follow a Gamma distribution with probability density function (PDF)

$$P\left\{I_i^{d,s} = x\right\} = \frac{1}{\Gamma\left(c_i^d\right)\left(v_i^d\right)^{c_i^d}} x^{\left(c_i^d - 1\right)} e^{-\frac{x}{v_i^d}} \tag{2},$$

where $c_i^d > 0$ and $v_i^d > 0$. In equation (2), $c_i^d$ and $v_i^d$ are the shape and scale parameters respectively. Assume that $\lambda_i^d = \sum_{s=1}^{s_d} \lambda_i^{d,s}$ is the overall mutation rate from all samples in bin $i$ of disease $d$. Its distribution can be readily obtained through convolution as

$$P\left\{I_i^d = x\right\} = \frac{1}{\Gamma\left(s_d c_i^d\right)\left(v_i^d\right)^{s_d c_i^d}} x^{\left(s_d c_i^d - 1\right)} \exp\left(-\frac{x}{v_i^d}\right) \tag{3}.$$

If we let $y_i^d = \sum_{s=1}^{s_d} y_i^{d,s}$ represent the total mutation counts in region $i$ from all disease samples, $d$, then the conditional distribution of $y_i^d$ given $\lambda_i^d$ can be written as

$$P\left(y_i^d \middle| I_i^d\right) = \frac{\left(I_i^d\right)^{y_i^d} \exp\left(-I_i^d\right)}{\left(y_i^d\right)!} \tag{4}.$$

By integrating (3) into (4), the marginal distribution of $y_i^d$ can be denoted as a negative binomial distribution ([15], page 50 in [16]).

$$P\left(y_i^d \middle| c_i^d, u_i^d\right) = \left(\frac{1}{1+u_i^d}\right)^{s_d c_i^d} \frac{\Gamma\left(s_d c_i^d + y_i^d\right)}{\Gamma\left(s_d c_i^d\right)\left(y_i^d\right)!}\left(\frac{u_i^d}{1+u_i^d}\right)^{y_i^d} \quad (5a).$$

Equation (5a) is the PDF of a negative binomial distribution with $E\left(y_i^d\right) = s_d c_i^d v_i^d$ and $Var\left(y_i^d\right) = s_d c_i^d v_i^d\left(1 + v_i^d\right)$. To better interpret (5a), we define $v_i^d = \mu_i^d \sigma_i^d$ and $s_d c_i^d = 1/\sigma_i^d$. Then equation (5a) can be rewritten as (5b).

$$p_{Y_i^d}\left(y_i^d \middle| m_i^d, s_i^d\right) = \left(\frac{1}{1+s_i^d m_i^d}\right)^{1/s_i^d} \frac{\Gamma\left(y_i^d + 1/s_i^d\right)}{\Gamma\left(1/s_i^d\right)\Gamma\left(y_i^d + 1\right)}\left(\frac{s_i^d m_i^d}{1+s_i^d m_i^d}\right)^{y_i^d} \quad (5b)$$

The mean and variance of $y_i^d$ from (5b) can be described as $\mu_i^d$ and $\mu_i^d\left(1 + \mu_i^d \sigma_i^d\right)$ respectively. Our model in equation (5b) is convenient due to its explicit interpretability. First, it assumes that the individual mutation rates are heterogeneous by modeling $\lambda_i^{d,s}$ as i.i.d. Gamma distributed random variables. Unlike the constant mutation rate assumption where $Var\left(y_i^d\right) = E\left(y_i^d\right)$, our model captures the extra variance of $y_i^d$ due to population heterogeneity. Our model in (5b) also clearly separates the two main parameters $\mu_i^d$ and $\sigma_i^d$ with physically interpretable meanings: the mean and overdispersion, respectively. Here a larger $\sigma_i^d$ indicates a more severe degree of overdispersion, which is usually due to larger differences in mutation rates.

After modeling $y_i^d$ with a negative binomial distribution, we then estimate the local mutation rate by correcting the covariate matrix $\boldsymbol{X}$ described above. Again $x_{i,j}$ denotes the average signal strength in the $i^{th}$ bin and $j^{th}$ covariate, where $i = 1, \cdots, n$ and $j = 1, \cdots, m$. Because the genomic features in the covariate matrix are highly correlated and may introduce multicollinearity if directly used in regression, we first apply principal component analysis (PCA) to matrix $\boldsymbol{X}$. We define $\boldsymbol{X}'$ to be the covariate matrix after PCA and $x'_{i,j}$ as each element in $\boldsymbol{X}'$.

A generalized regression scheme is used here. Suppose $g_1$ and $g_2$ are two link functions. We then use linear combinations of covariate matrix $\boldsymbol{X}'$ to predict the transformed mean parameter, $\mu_i^d$, and overdispersion parameter, $\sigma_i^d$, as

$$\begin{aligned} g_1\left(\mu_i^d\right) &= \log\left(\mu_i^d\right) = \beta_0^d + \beta_1^d x'_{i,1} + \cdots + \beta_j^d x'_{i,j} + \cdots + \beta_j^d x'_{i,m} \\ g_2\left(\sigma_i^d\right) &= \log\left(\sigma_i^d\right) = \alpha_0^d + \alpha_1^d x'_{i,1} + \cdots + \alpha_j^d x'_{i,j} + \cdots + \alpha_m^d x'_{i,m} \end{aligned} \quad (6).$$

Here we use a log link function for both $g_1$ and $g_2$, so the regression model in (6) is a negative binomial regression. Note that $\boldsymbol{X}$ contains 381 genomic features in all available tissues. In the following analysis, we use all features to run the regression in (6) to achieve better performance. The GAMLSS package in R is used to estimate the parameters in (6) as $\hat{\alpha}_0^d, \cdots, \hat{\alpha}_m^d, \hat{\beta}_0^d, \cdots, \hat{\beta}_m^d$. Generally, there are biological reasons to explain how $\mu_i^d$ changes with covariates. For example, single-stranded DNA in the later replicated regions usually suffers from accumulative damage resulting in larger $\mu_i^d$. It is more difficult to interpret such a relationship with $\sigma_i^d$. Hence, we simplify equation (6) by assuming $\sigma_i^d$ is constant in our real data analysis. In order to separate the local context effect, we separate the 64 local 3 mers to train 64×2 parameter during the training process.

## 2.7 Testing details

Suppose there are *K* regions to be tested. We use the local mutation rate to evaluate the mutation burden. For the $k^{th}$ target region ($k = 1, \cdots, K$), one way of calculating the covariates is to extend it into length $l$ (illustrative figure given in Fig. S2). Then we calculate the average signal for feature $j$ as $x_{k,j}, j = 1, \cdots m$ for this extended bin, and after PCA projection let $x'_{k,j}$ represent the value for the $j^{th}$ PC. The local mutation parameters $\hat{\mu}_k^d$ and $\hat{\sigma}_k^d$ in the extended bin for the $k^{th}$ target region can be calculated as

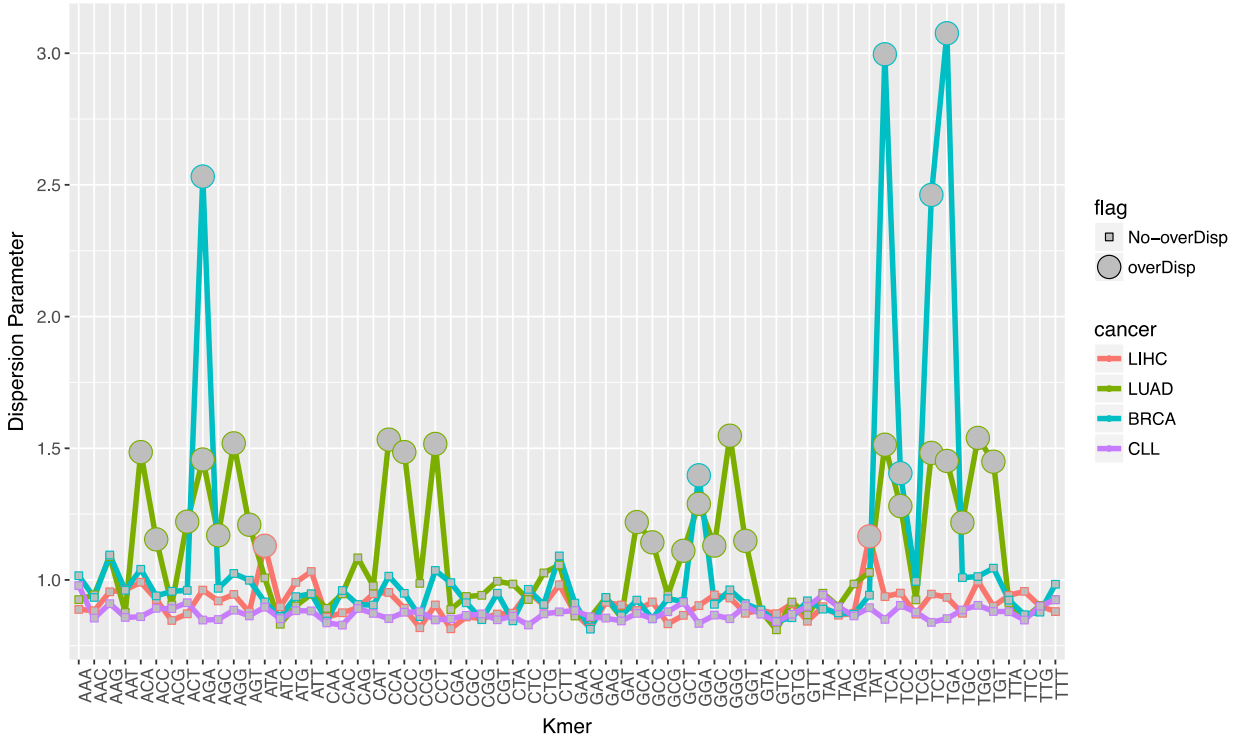
$$\hat{\mu}_k^d = \exp\left(\hat{\beta}_0^d + \hat{\beta}_1^d x'_{k,1} + \cdots + \hat{\beta}_j^d x'_{k,j} + \cdots + \hat{\beta}_m^d x'_{k,m}\right)$$
$$\hat{\sigma}_k^d = \exp\left(\hat{\alpha}_0^d + \hat{\alpha}_1^d x'_{k,1} + \cdots + \hat{\alpha}_j^d x'_{k,j} + \cdots + \hat{\alpha}_m^d x'_{k,m}\right) \quad (7).$$

In real data analysis, the length of the $k^{th}$ test region $l_k$ is much shorter than the length of the training bins (up to 1Mb). Hence $\hat{\mu}_k^d$ needs to be adjusted by a factor of $l_k/l$. Then $\hat{\sigma}_k^d$ and the adjusted $\hat{\mu}_k^d$ can be used to calculate the disease specific P value, $p_k^d$. This above scheme is usually computationally expensive because there are usually millions of target regions to be tested. Therefore, we also propose an approximation method alternatively to replace the optimal $\hat{\mu}_k^d$ and $\hat{\sigma}_k^d$ in our analysis. In order to separate the local context effect, we separate the 64 3mers and run individual regression models for each 3mer.

The negative binomial model mentioned in equation (5) can effectively control the false positives when there is huge overdispersion. However, the negative side on (5) is that when there is little heterogeneity among patients and heterogeneity over different regions of the genome can be completely removed by regressing against the external features, estimation in (7) might fail. In other words, it cannot handle the non over-dispersed data well. In order to solve this problem, we first use Poisson regression which assumes equal mean and variance. Then we run a test using the method mentioned in \cite{Regression-based tests for overdispersion in the Poisson model} for the following hypothesis: provided the regression function is correctly specified and ordinary least squares parameter estimates are consistent, whether variance is equal to the mean. Specifically, we assume $H_0: var\left(y_i^d\right) = \mu_i^d$, and the alternative hypothesis is $H_0: var\left(y_i^d\right) = \mu_i^d + \alpha g\left(\mu_i^d\right)$. In particular, we tested whether $\alpha = 0$. When this test for Poisson regression fails, we swich to negative binomial regression for better fitting. During the implementation stage, we used the AER package in R (the dispersiontest function) to run this test. We provided the summary of estimated overdispersion parameter in multiple cancer types in Figure S 2-9. It clearly shows that different

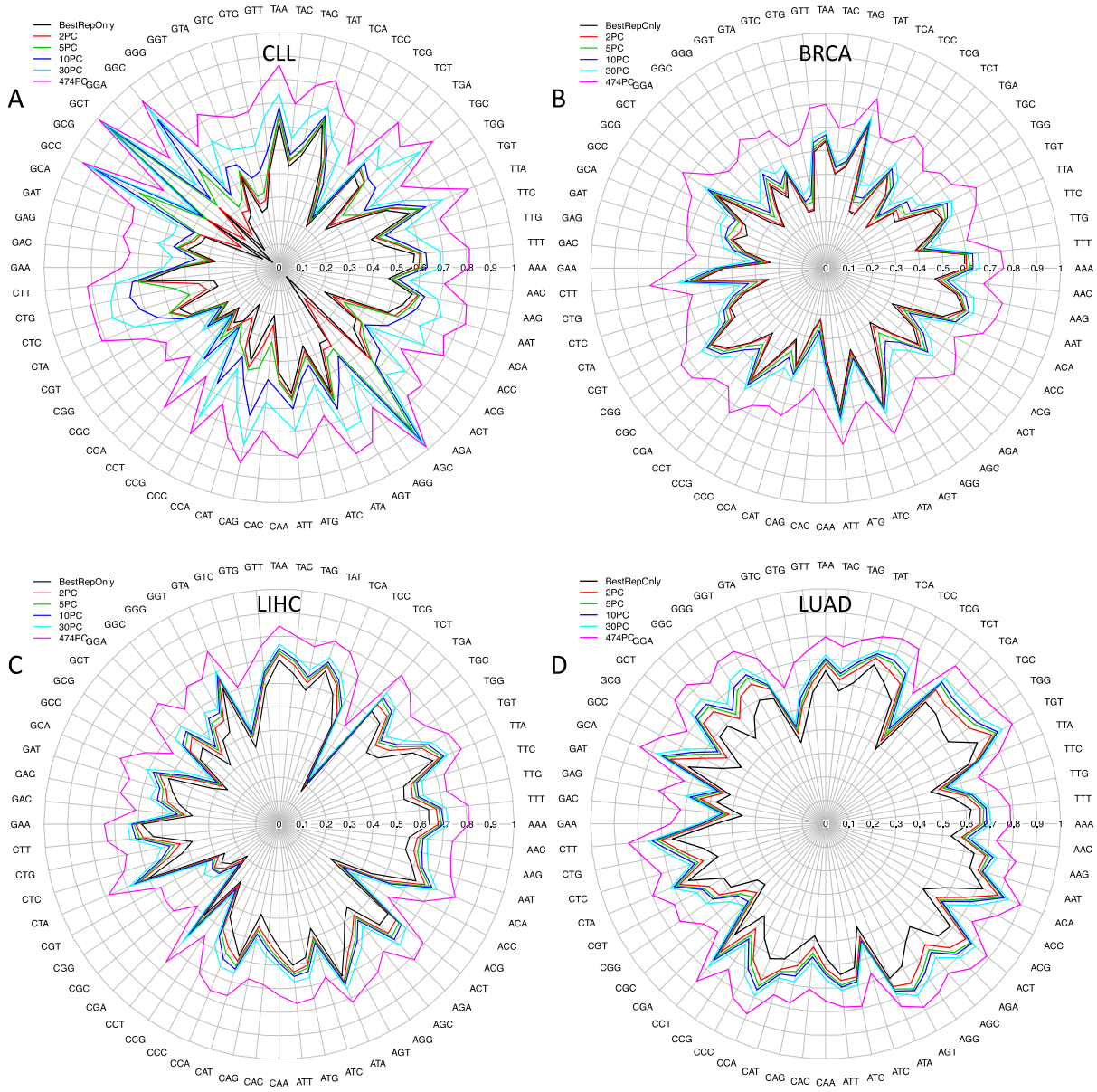
cancer types and local 3mers has distinct overdispersion status. In CLL, the overdispersion parameters ranges from 0.8285 to 0.9784, indicating Poisson regression models for all 3mers are enough during the training process. However, in breast cancer, the overdispersion parameter ranges from 0.8130 to 3.0760. 8 out of the 64 3mers need to use the negative binomial models to handle the extra variance.

*Figure S 2-9 summary of estimated overdispersion parameter in multiple cancer types*




The performance of the model is given in Figure S 2-10. We observed that in all cancer types, using more features significantly improves the BMR estimation precision. However, in order to avoid overfitting, we first run regression using projected PCs on the feature matrix, and then all PCs with adjusted P value greater than 0.05 will be removed during the training process.

*Figure S 2-10. performance of BMR model training using different number of parameters.*




Sometimes it is necessary to analyze several related diseases (or disease types) to provide a combined P value. One typical example is in pan-cancer analysis. In the above section, we

calculated the P value for disease/disease type $d$ as $p_k^d$ for test region $k$. Fisher's method can be used to combine these P values. Specifically, the test statistic is

$$T_k = -2\sum_{d=1}^{D} \ln\left(p_k^d\right) \sim \chi^2(2D) \quad (8).$$

Here $T_k$ follows a centered chi-square distribution with $2D$ degrees of freedom, where $D$ is the total number of diseases/disease types. The final P value, $p_k$, can be calculated from $T_k$.

## 2.8 P value summaries

To check the distribution of P values vs. the theoretical ones, the Q-Q plots were given in Figure S 2-11 to Figure S 2-13.

*Figure S 2-11. Q-Q plots of P values for CLL.*

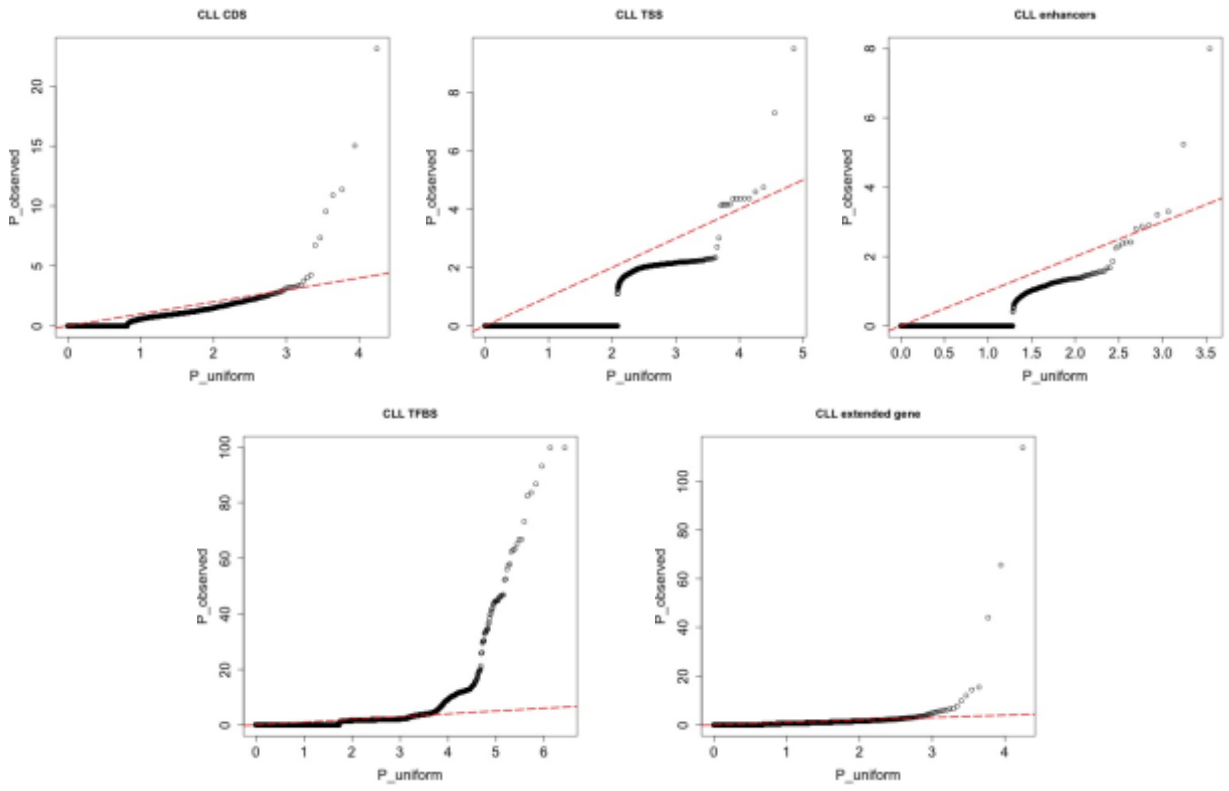


*Figure S 2-12 Q-Q plots of P values for BRCA*

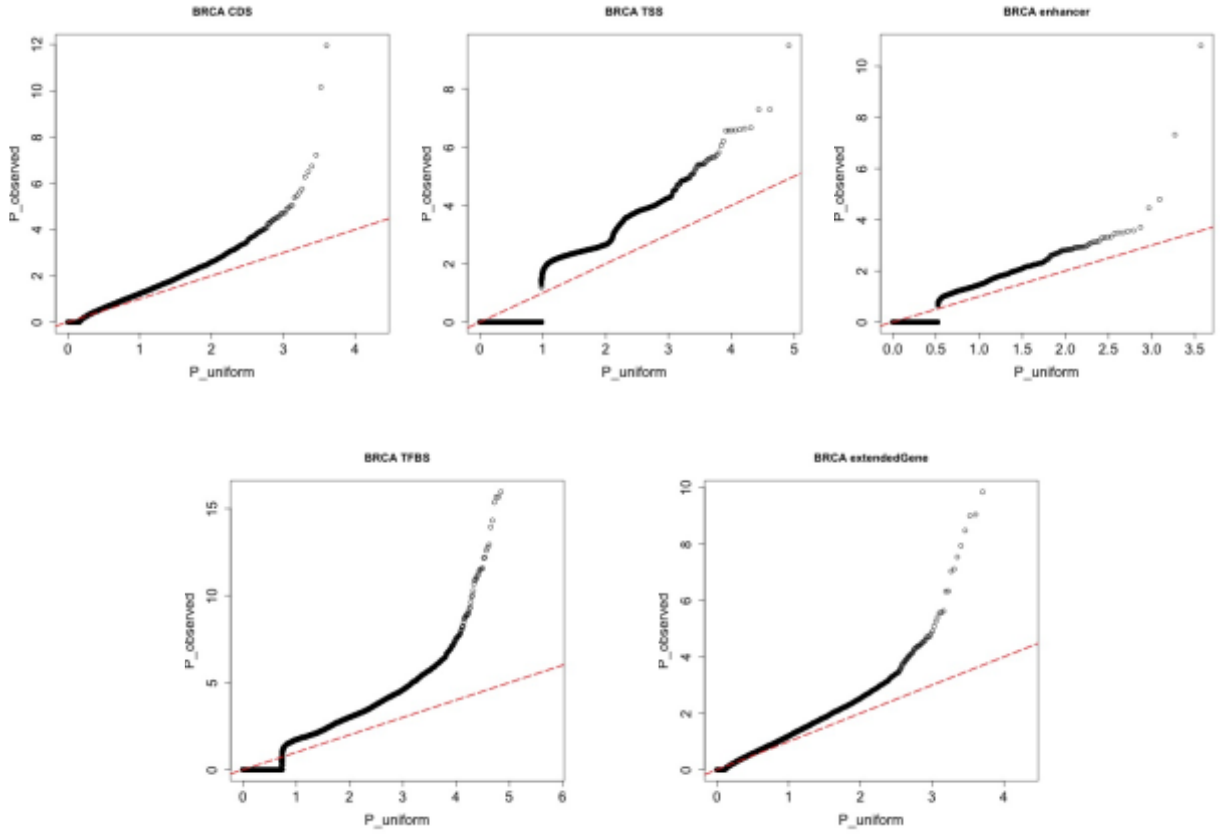
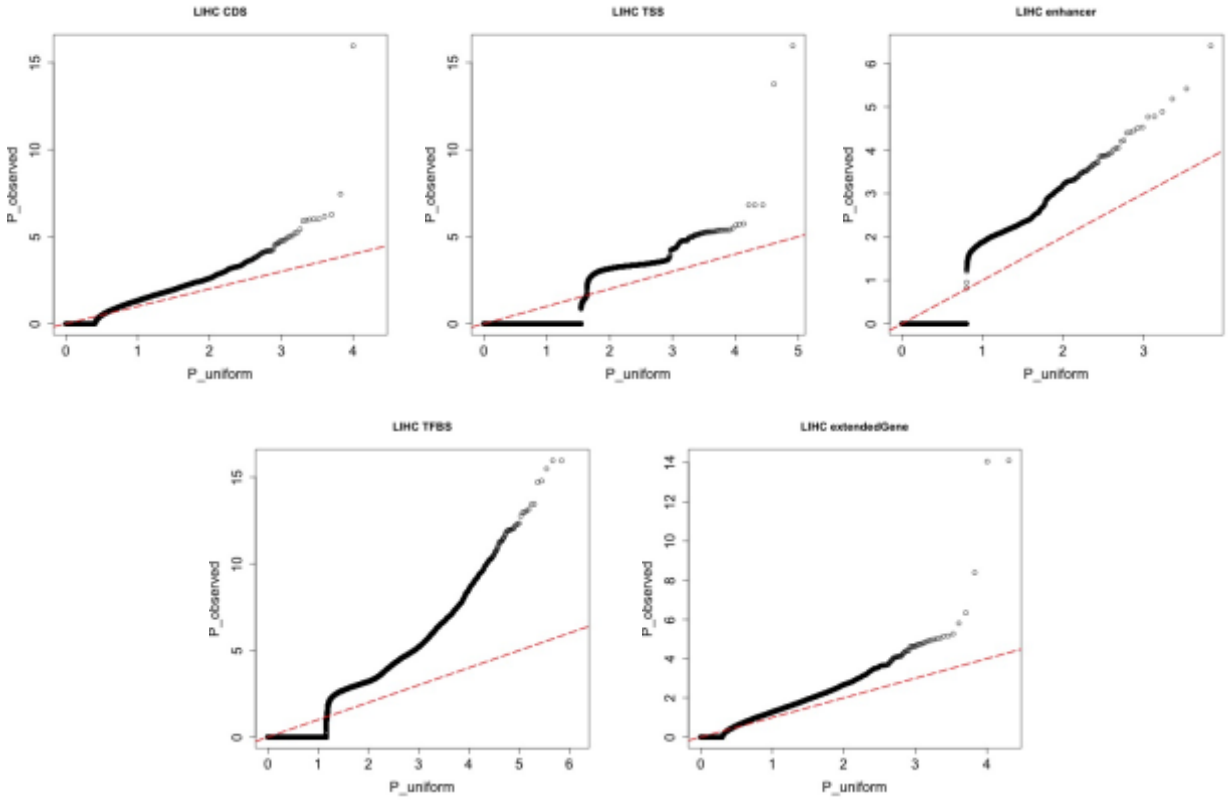


*Figure S 2-13 Q-Q plots of P values for LIHC*



# 3 Details about TF network rewiring analysis

## 3.1 Rewiring analysis based on direct counts

### 3.1.1 TF-gene linkage

We evaluated the rewiring of TF to gene linkages between normal and cancerous cells. To define TF rewiring between cell types, we first defined TF-gene regulatory network in each cell type using simple count based target gene linkage. We used two different methods that examine TF to gene linkages based on their proximities to the TSS. For TSS-based method, we simply used 2,500bp upstream and downstream of transcription start site (TSS) based on Gencode v19 annotation as a boundary for proximal regulatory region. On average, 33.5% of TF ChIP-seq peaks fell into promoter region (See Suppl. Result table S2). We defined a target gene linkage if TF ChIP-seq peak was found within the boundary. However, we discovered, in Gencode annotation, there were numbers of genes that have more than 50 alternative TSS, which gave these genes unfair advantages of having more target gene linkages than others since their proximal regulatory regions can span up to 250kbp. Therefore, we selected one canonical TSS for each gene based on the total number of aggregated ENCODE TF ChIP-seq peaks. While this method is far from perfect, we believe this is the best method to capture the high-level TF network rewiring and quantify epigenetics changes around TSS while minimizing artifacts when counting all TSSs from all possible alternative transcripts.

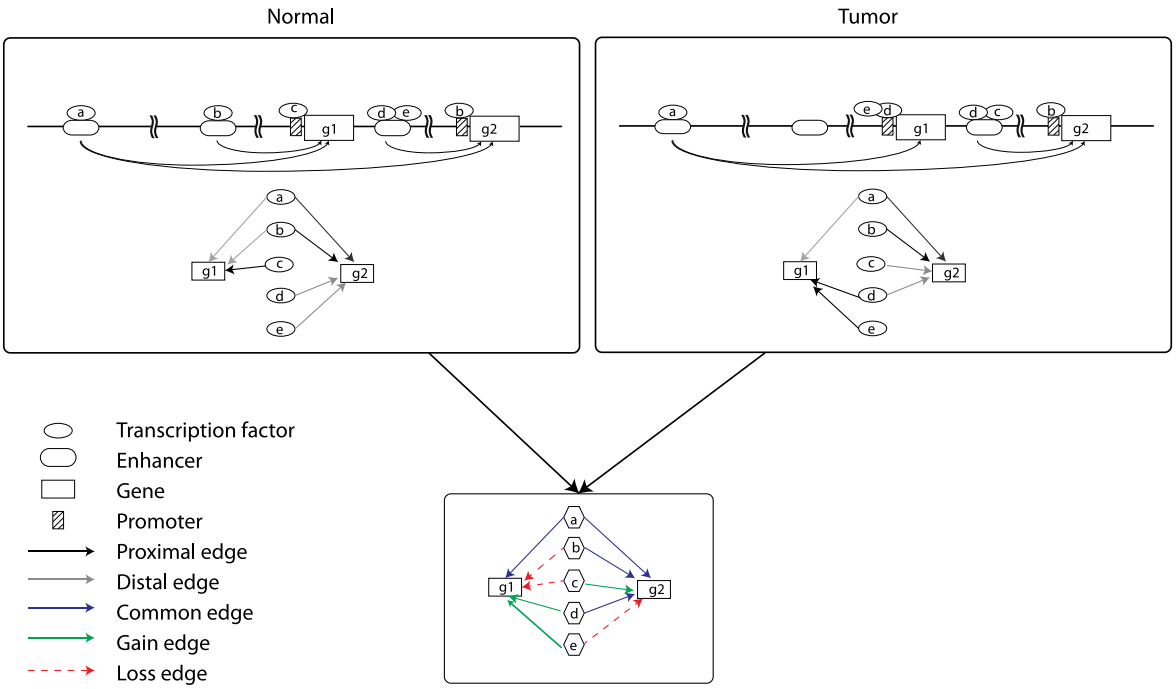
In addition to TSS-based TF-gene linkages, we used target identification from profiles (TIP) method that quantitatively measures the regulatory relationships between TFs and target genes to define a subset of the full TF-gene network. For each TF, TIP model builds a characteristic, averaged profile of binding around the TSS and then uses this to weight the sites associated with

a given gene, providing a continuous-valued 'regulatory' score relating each TF and potential target \cite{ PMID: 22039215 }. We used false discovery rate of 0.1 for cutoff. Since TIP uses narrower promoter definition than TSS-based method, we defiend TIP-based network as a subnetwork of TSS-based network.

### 3.1.2 Full regulatory network, merged network, and network rewiring

To build a complete TF-gene network, both promoter-based linkages and enhancer target based linkages were merged into one. For more information about enhancer target based linkages, please refer to section 1.7.3. Two versions of full regulatory networks were constructed; one larger network by concatenating TSS-based network and enhancer-based network and another subnetwork by concatenating TIP-based network and enhancer-based network. In addition, we built a merged network by combining all available ENCODE tissue types.

*Figure S 3-1. Network rewiring schematics*



Rewiring of edges between TF and target genes were compared in normal and tumor cells as shown in Figure S 3-1. If a target gene linkage was found in normal but lost in tumor, the edge was marked as loss edge. Similarly, if a target gene linkage was found only in tumor, it was labeled gain edge, and for edges found in both, they were labeled common or retained edges

### 3.1.3 Rewiring score

To quantify rewiring events, we first calculated rewiring score for each regulators (TFs). The fraction of the number of gain, loss, and common edges to the number of fully connected network edges, where all available TF nodes are fully connected with all available gene targets in the whole network was used to calculate the raw rewiring score.

$$n_{\text{fully-connected}} = n_{\text{TF}} * n_{\text{gene}} - 1$$

$$rScore_{\text{TF}} = \frac{\frac{G_{\text{in}} + G_{\text{out}}}{L_{\text{in}} + L_{\text{out}}}}{\left|\frac{G_{\text{in}} + G_{\text{out}}}{L_{\text{in}} + L_{\text{out}}}\right|} \cdot \frac{(G_{\text{in}} + G_{\text{out}} + L_{\text{in}} + L_{\text{out}})}{n_{\text{fully-connected}}}$$

$$rScore_{\text{normalized}} = \frac{rScore_{TF}}{\max_{all}(rScore_{\text{TF}})}$$

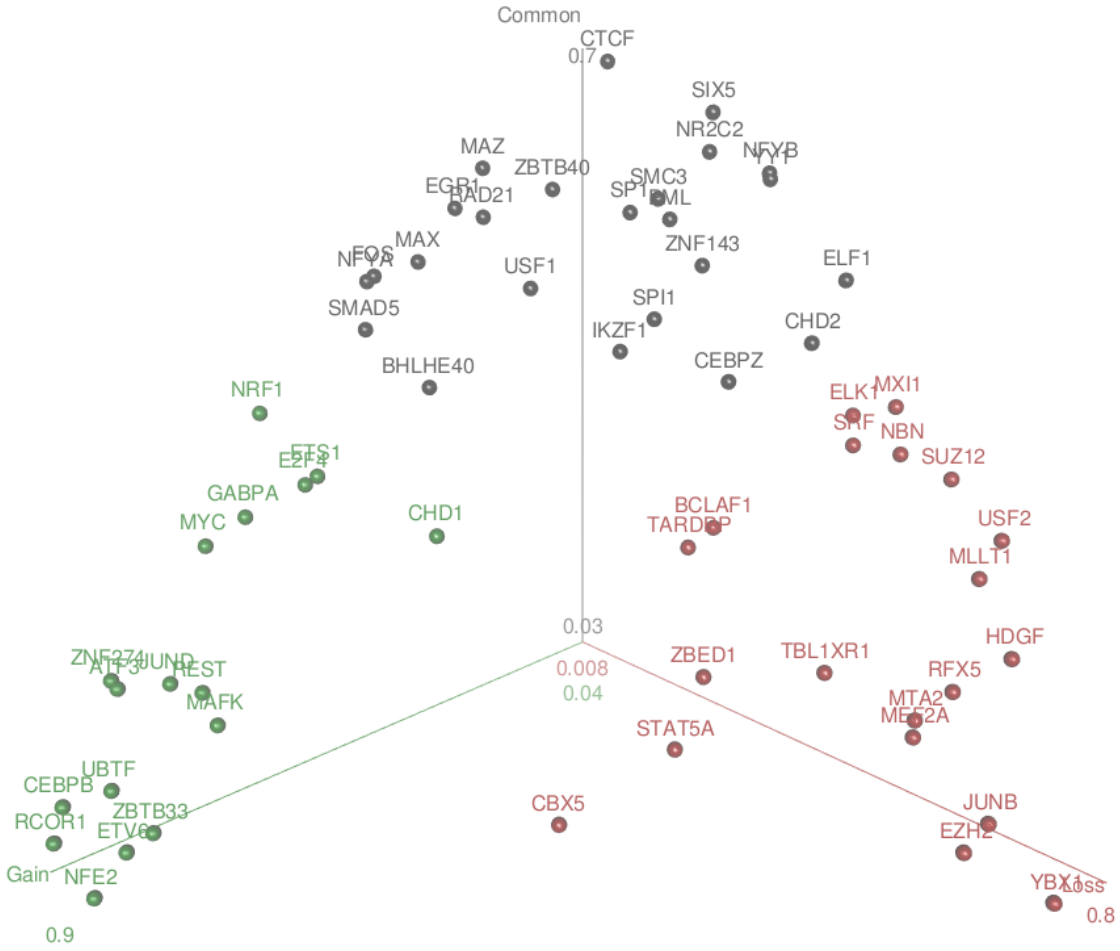
The rewiring score, rScore, after taking normalization over the maximum rScore, was used to rank the TF from the gainer to loser.

### 3.1.4 Clustering of rewired TFs

Based on the fraction of gained, lost, and retained edges with respect to the total number of edges for each TF, rewired TFs were clustered into three groups using kmeans clustering. Hartigan-Wong algorithm with 10 iterations were used \cite{ Hartigan, J. A. and Wong, M. A. (1979). A K-means

clustering algorithm. Applied Statistics 28, 100–108. }. Figure S 3-2 shows the clustering result for rewired TFs between K562 and GM12878. NFE2 and RCOR1 were identified as one of the strongest member of gained group, CTCF was identified as a member of common group, and YBX1 was identified as a member of loss group.

*Figure S 3-2 Kmeans clustering of rewired TFs in K562 and GM12878*

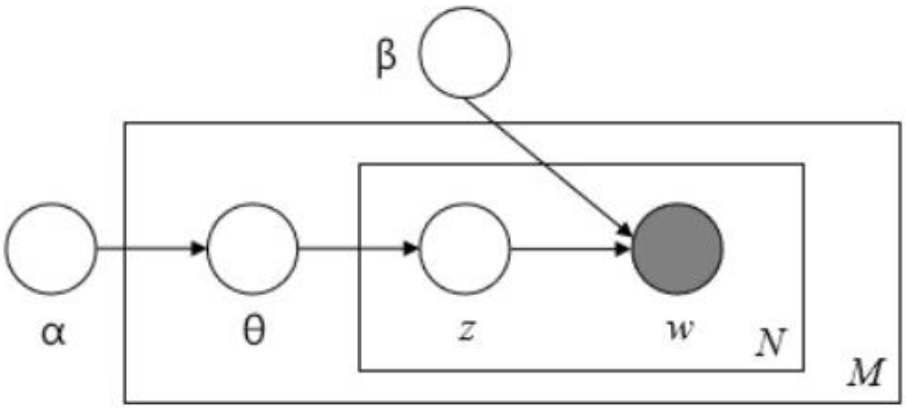


## 3.2 Rewiring analysis based on mixed membership algorithm

We use mixed membership algorithm to investigate the rewiring changes between GM12878 and K562 cell lines (Figure S 3-3). TF-target matrix (M x V) is converted from enhancer and TSS regulatory network, where M is the number of TF and V is the number of unique target genes for all TFs. Each row represents a target gene of a TF i=1,2,…, M. The regulatory pattern of each TF is comprised of K latent communities. Each community includes contributions from N target gene and N varies for different TF.

The target gene of TF $w_i, \vec{w} = \langle w_{i,1}^v, w_{i,2}^v, \dots, w_{i,n}^v \rangle$ and $= 1$ means target, $0$ means non-target. The observation denote TF $i$, target gene $j$ with status $v$. Similarly $Z_{i,j,v}$ is the community distribution of each target gene $j$ for TF $i$ with status $v$. $\beta(M \times V)$ denotes the probability of target gene $j$ belong community $k$, which is parameter of multinomial distribution. $\theta_i$ each and denote the distribution of communities for TF $i$. $\theta \sim \text{Dirichlet}(\alpha)$, where $\alpha$ is the super-parameter of $\theta$.

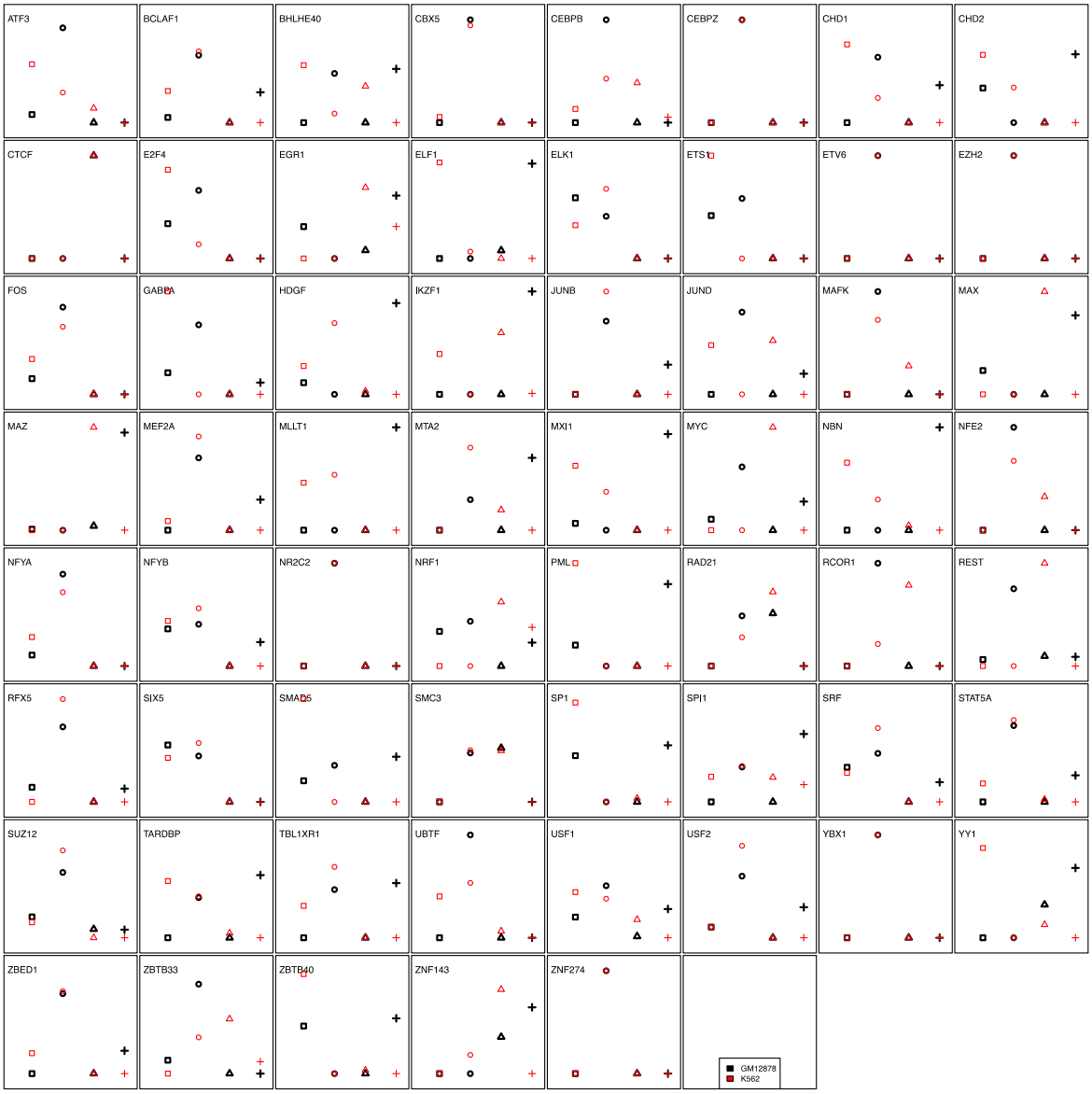
*Figure S 3-3 Schematic of gene community based rewiring analysis*



When inferring the latent gene community model, we are most interested in the communities' parameter $\beta$, the Dirichlet parameter $\alpha$ and the latent community distribution $\theta$ of TF. So the key is to find the posterior distribution of latent variables.

Variational EM algorithm (implemented using mixedMem R package) is used to infer the $\alpha$ and $\theta$ as described in \cite{Blei et al., 2003, Erosheva et al., 2004}. However, computational benefits of EM lead to optimization uncertain and make it easily converge to local maxima. We have no priori knowledge for the $\theta$ and $\alpha$, which is impossible to use near plausible value to find reasonable optimum. To hack this, we repeat multiple times (100) and use median of rewiring changes from all the non-early stops simulation to represent the most optimal regulatory changes of TF. One example of the $\theta$ distribution was given in Figure S 3-4.

*Figure S 3-4. Example of $\theta$ distribution difference in tumor and normal cell lines*




The rewiring of TF regulation is defined by the changes of distribution in K gene communities using $Distance_i = \sqrt[3]{\sum_{i,j}\left[\sqrt[3]{q_{K562,i,j}} - \sqrt[3]{q_{GM12878,i,j}}\right]^{**}3}$, where $q_i$ is the distribution of communities for TF $i$.

## 3.3 Patient survival analysis based on TF activities

In this analysis, we systematically calculated TF activity in 6 different AML datasets using the ENCODE ChIP-seq data. 292 ChIP-seq experiments from K562 (231 TFs) and 120 ChIP-seq experiments from GM12878 (101 TFs) were used to generate TF binding weight profiles from the TIP output. Specifically, the binding score of a TF to each gene (outputted by the TIP algorithm) was z-transformed and a one-sided z-test was carried out to generate p-values corresponding to each TF-gene binding interaction. P-values were -log10-transformed and trimmed at -10 or 10. Weight profiles were re-scaled by subtracting each value in a TF weight profile by the minimum and dividing by the range so that all values fell between 0 and 1. These weight profiles were used as input into the BASE algorithm to calculate TF activity scores for AML patient samples derived from the following gene expression datasets:

| | | | | | |
|---|---|---|---|---|---|
| GEO | -- | GSE37642 | (GPL_96) | (Herold, | n=422) |
| NCI | caArray | -- | willm-0019 | (Wilson | n=170) |
| GEO | -- | GSE14468 | | (Wouters, | n=526) |

Survival analysis was performed for each TF to identify those that were significantly associated with AML patient mortality. Namely, the TF's iRASs (activity scores) across patient samples were used as the independent variable in a Cox proportional hazards model. A hazard ratio <1 indicates that a TF's activity is associated with favorable prognosis and a hazard ratio of >1 indicates that a TF's activity is associated with unfavorable prognosis in AML patient samples. Since a separate model was fit to each TF's iRASs, p-values corresponding to the hazard ratios were adjusted for multiple hypothesis testing by using the Benjamini-Hochberg correction procedure.

In the results, we report the HR, P-value, and Adjusted P-value for each TF and their association with patient survival in each of the 3 AML gene expression datasets. The column labeled "number_datasets_significant_P005" indicates the number of datasets in which the TF's activity was observed to be significantly associated with AML patient prognosis at P<0.05. In particular, the EZH2, STAT1, and NR2C2 TFs were found to be significantly associated with prognosis in all 3 datasets. 15 other TFs were found to be significant in 2 datasets.

### 3.4 Target gene analysis

To evaluate the effect and extent of TF-gene network rewiring, target gene's expression and epigenetic changes were evaluated for genes that have gained and lost edges between normal and tumor samples. For expression, we used RESM quantification of ENCODE DCC uniformly processed long polyA RNA-seq and averaged TPM values over all available replicates. For DNase-seq, histone ChIP-seq, and methylation features, we further processed from fold enrichment signal tracks as follows. We averaged the fold enrichment signal across 200bp upstream and downstream of the unique TSS, the same canonical TSS site used define proximal TF-gene linkage. For all expression, DNase-seq, histone ChIP-seq, and methylation feature was expressed as log2 ratio between tumor to normal samples. To avoid division by zero error, pseudocount of 0.0001 was added to each feature.

### 3.5 Co-binding analysis

[JZ2GG&Jason&DC]

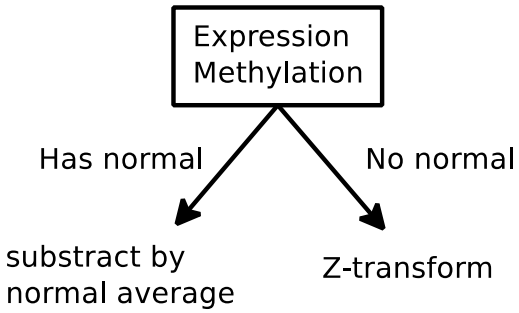
## 4 Details about expression aggregation analysis

[JZ2Peng: Please re-organize this part]

## 4.1 TCGA data collection

All TCGA expression, methylation and mutation data were downloaded from GDAC firehose (http://gdac.broadinstitute.org) with data version of 2016_01_28. For cancer types with normal control samples profiled, the expression values of each gene are substracted with the average value of all normal controls. For cancer types without any normal samples profiled, the expression profile of each gene is transformed to zero mean and unit deviation. The DNA methylation values are also normalized in the same way as RNASeq data, according to the availability of normal control samples in each cancer type. For copy number alteration (CNA), GDAC firehose doesn't provide standardardized data and we downloaded the data matrix from cBioportal with data version of 2016_10_20 (http://www.cbioportal.org).

*Figure S 4-1 Schematic of RNA-seq normalization*



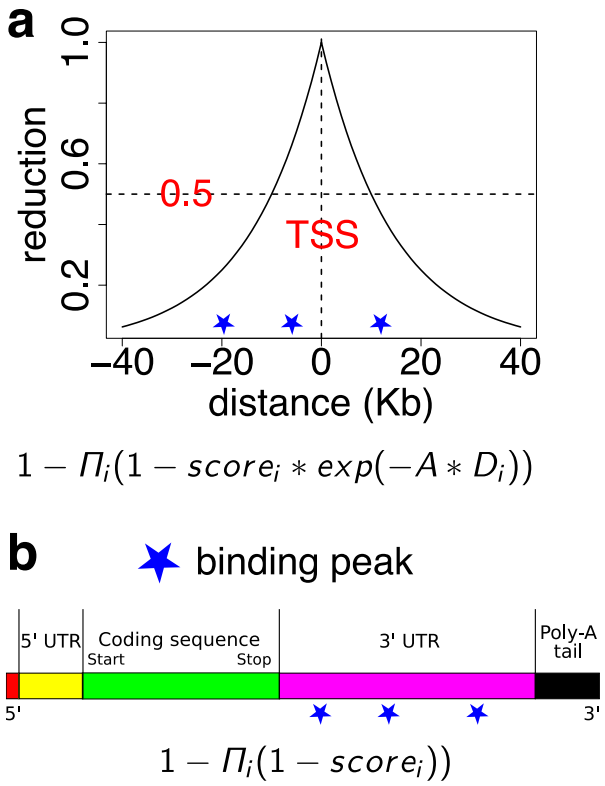

## 4.2 Regulatory network construction from ChIPSeq and eCLIP data

For regulatory analysis, we only considered transcription factors (TF), chromatin regulators (CR), and RNA binding proteins (RBP). In total, there are 978 TF/CR ChIPSeq profiles and 159

RBP eCLIP profiles downloaded from ENCODE DCC until Janurary 4th, 2017 (https://www.encodeproject.org).

All ChIP-seq and eCLIP peak scores are linearly scaled into range (0,1). The regulatory score between TF peaks and gene promoters were built with "connect_host" commands from RABIT package following an exponential decay model (Figure S 4-2 a). The regulatory score between RBPs and genes were built through counting eCLIP peaks within gene 3'UTR regions (Figure S 4-2 b). The following steps were made to construct the network. a) For ChIP-seq data, A regulatory potential score is calculated between each pair of ChIP-seq peak and gene TSS by multiplying the ChIP-seq intensity score with an exponential decay score exp(-A*Distance) of their distance between. The coefficient A is set as log(2)/10K, so that a binding peak 10K bps away from gene TSS will decay by 50%. For each gene TSS, if there are several peaks of a TF nearby, we merged their regulatory potential scores by noisy-or: . (b) For eCLIP data, only binding peaks over gene 3'UTR regions were considered for possible regulatory role of transcript stability. For each gene 3'UTR region, if there are several peaks of a RBP, we merged their regulatory potential scores by noisy-or operation. All regulatory potential scores stay within range (0,1).

*Figure S 4-2. Regulatory network construction*




ChIPSeq and eCLIP profiles were excluded from further analysis if the total sum of regulatory scores across all human genes are less than 100. All general TFs including Pol2 and Pol3 were excluded from further analysis. For certain TF, there exists many ChIP-seq profiles profiled in different conditions. We run a hierarchical clustering among all of its ChIP-seq profiles and cut the hierarcical tree at correlation distance of 0.2. Only profiles in the largest cluster are used for further analysis. The final size of regulatory networks constructed are shown in Table S 4-1. For each data type, column "Profile" represents the number of experimental profiles (ChIP-seq or eCLIP) that passed our quality controls. Column "Regulator" represents the number of regulators (TF, CR or RBP) analyzed. Column "Condition" represents the number of experimental conditions

included in profiles. Column "Target" represents the total number of human genes profiled as targets of analyzed regulators.

*Table S 4-1. Statistics of regulatory networks.*

| | Profile | Regulator | Condition | Target |
|---|---|---|---|---|
| ChIPSeq | 762 | 496 | 44 | 21348 |
| eCLIP | 159 | 112 | 2 | 14593 |

In order to systematically search for transcription factors (TF) that drive tumor specific gene expression patterns, we used a previously developed integration framework RABIT (Regression Analysis with Background InTegration, http://rabit.dfci.harvard.edu). In the RABIT framework, for a given TF ChIP-seq binding profile, candidate target genes are identified by weighting the number of binding sites by their distance to the transcription start site (TSS) of each gene. For a given eCLIP RBP binding profile, candidate genes are identified through searching the binding sites within the gene 3'UTR regions. RABIT uses three steps to identify TFs (or RBPs) that drive tumor specific gene expression patterns at both the individual tumor level and the whole cancer type level. In Step one, RABIT screens for TFs that significantly affect the gene expression patterns in each tumor, and select the most relevant ChIP-seq (or eCLIP) profile if multiple profiles exists for the same regulator. In Step two, RABIT further selected a subset of TFs among those screened in Step one to achieve an optimized model error. In Step three, RABIT investigates how well the public ChIP-seq profiles can capture the active TF targets in each cancer type, and clean up insignificant TFs. The final output of RABIT framework is a set of TFs or RBPs that shape the tumor-specific expression patterns at individual tumor level in each cancer type.

Based on ENCODE ChIPSeq data and TCGA profiles, we applied RABIT framework to identify transcription factors (TF) whose target genes are differentially regulated in cancer. The fractions of patients with TF targets differentially regulated are shown. Only TFs with targets differentially regulated in over 40% patients in at least two cancer types are included and results were summarized into Figure S 4-3. We further extracted those TF with stronger signals to shown in Figure 4. Except the well-known MYC targets showing consistent up-regulation pattern across multiple cancer types, we also found novel TFs such as ZNF687 to be strongly up-regulated in breast and prostate cancer (star in Figure S 4-3). In addition, the breast tumors were further classified into sub types according to PAM50 classification and ER status to show the scores predicted by RABIT for each sub type by boxplots. We further checked in each TCGA cancer type, the fractions of patients detected with different types of ZNF687 alterations (Figure S 4-4 b).

*Figure S 4-3 Heatmap of TF activities in multiple cancer types*

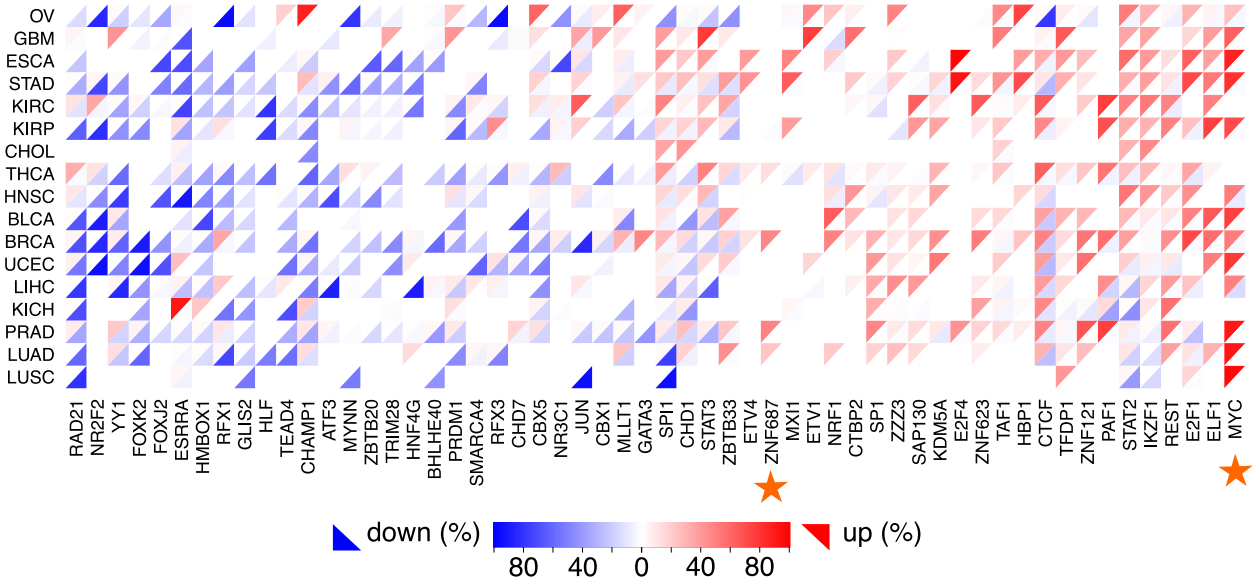
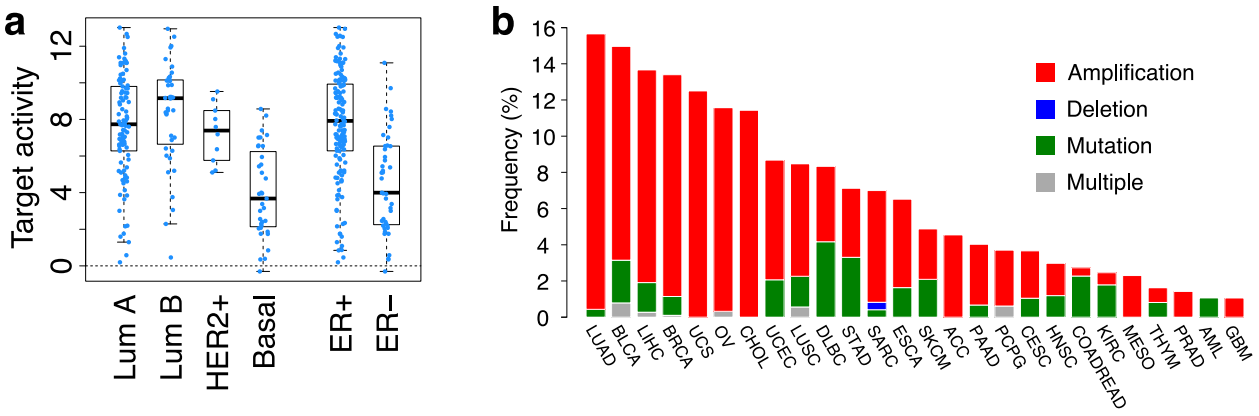


*Figure S 4-4 The potential role of ZNF687 in cancer*




SUB1 was also predicted to be significantly associated with expression changes in multiple tumor types in Figure 4. Here we have listed the full predictions in all cancer types for SUB1. In each cancer type, the association between SUB1 expression and SUB1 regulatory activity predicted by RABIT was tested through t-test in linear regression. Only significant associations above FDR threshold 0.05 are shown in Table S 4-2.

*Table S 4-2 Correlation between SUB1 expression and target activity*

| Cancer | Coef | Stderr | *t*-value | *p*-value |
|---|---|---|---|---|
| **THCA** | 4.79 | 0.46 | 10.46 | 9.03E-23 |
| **OV** | 4.47 | 0.61 | 7.37 | 1.53E-12 |
| **LUAD** | 2.87 | 0.46 | 6.22 | 3.25E-09 |
| **PRAD** | 2.9 | 0.48 | 6.02 | 4.56E-09 |
| **HNSC** | 2.61 | 0.46 | 5.72 | 2.73E-08 |
| **KIRP** | 3.6 | 0.63 | 5.73 | 5.66E-08 |
| **GBM** | 2.93 | 0.54 | 5.47 | 2.60E-07 |
| **LIHC** | 3.22 | 0.64 | 5.02 | 1.18E-06 |
| **BLCA** | 3.21 | 0.66 | 4.83 | 3.91E-06 |
| **LUSC** | 2.8 | 0.66 | 4.25 | 6.39E-05 |
| **KIRC** | 1.91 | 0.5 | 3.83 | 1.62E-04 |
| **STAD** | 2.16 | 0.57 | 3.76 | 2.14E-04 |
| **ESCA** | 1.67 | 0.54 | 3.09 | 2.34E-03 |

| | | | | |
|---|---|---|---|---|
| **UCEC** | 1.47 | 0.52 | 2.84 | 5.28E-03 |
| **KICH** | 1.64 | 0.58 | 2.8 | 6.71E-03 |

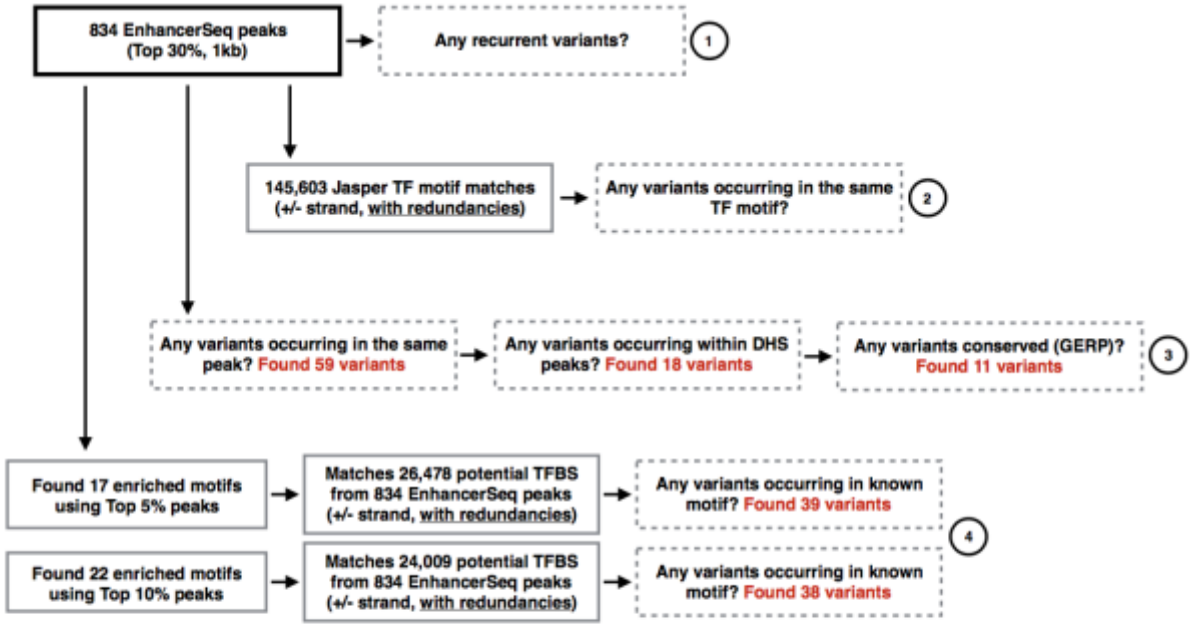
# 5 Variant prioritization

The description of the regulatory network and mutation recurrence analysis provide a way to prioritize key genomic features associated with cancer. The we proposed a step-wise scheme to prioritize the SNVs for small scale validations. First, we start by searching for key regulators that frequently rewired, locate in network hubs or on top of the network hierarchy, or significantly drive expression changes in cancer. We then prioritize functional elements that are associated with top regulators, undergo large regulatory and chromatin changes, or (most importantly) are highly mutated in tumors. Finally, on a nucleotide level, we can pinpoint impactful SNVs for small-scale functional characterization by their ability to disrupt or create specific binding sites, or which occur in positions of particularly high conservation or chromatin changes.

### 5.1.1 Motif analysis using MotifTools (D-score)

To prioritize the variant within high-confidence enhancer sets, we first searched for recurrent non-coding variants or multiple non-coding variant occurring in a known TF motif. However, we could not find any somatic variants that are either recurrent or recurrent within a TF motif (Figure S 5-1).

*Figure S 5-1 Variant prioritization scheme based on Enhancer-seq*




1. Recurrent BRCA non-coding variants within 834 EnhancerSeq peaks => **None**
2. Multiple BRCA non-coding variants occurring in a known TF motif => **None**
3. Multiple BRCA non-coding variants occurring in a EnhancerSeq peak (834) => **59 non-coding variants**
4. BRCA variants in known TF motif with motif breaking power. Same type of analysis was done for E2 induced MCF-7 as well. Combining results from "untreated" and "E2 induced", **46 variants**

Alternatively, we prioritized somatic variants based on its motif breaking power, or D-score, where D stands for disruptive-ness or deleterious-ness. Motif disruption score was calculated based on the difference between sequence specificities of reference to alternative sequence.

$$\text{motif-score}_{\text{ref}} = -10 \cdot \log_{10}(\text{p-value}_{\text{ref}})$$

$$\text{motif-score}_{\text{alt}} = -10 \cdot \log_{10}(\text{p-value}_{\text{alt}})$$

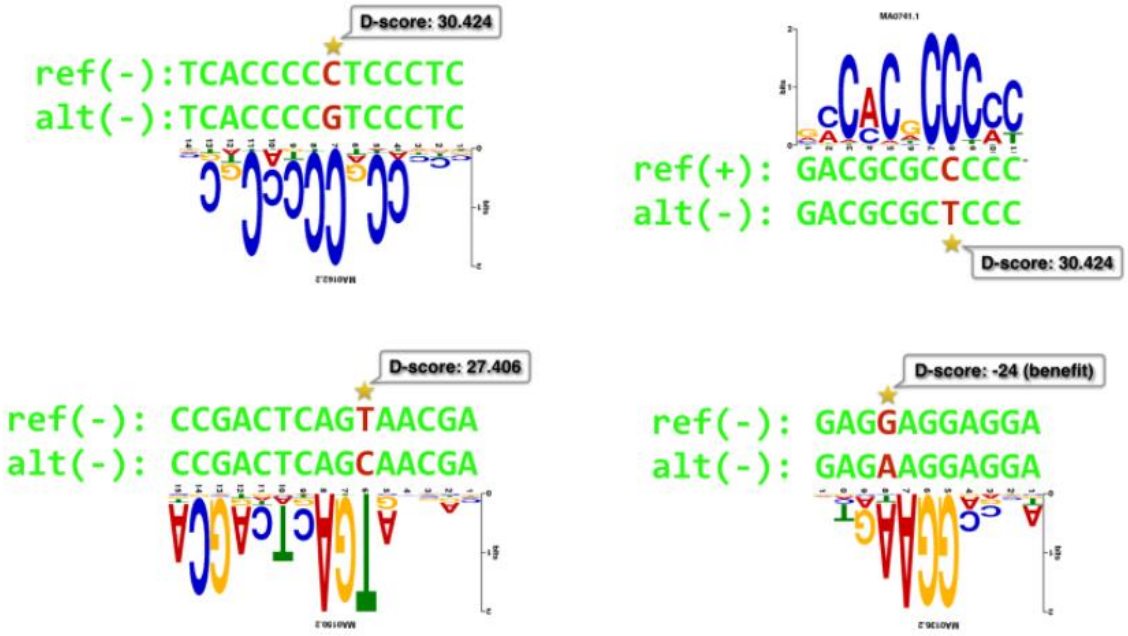
$$\text{D-score (}\mathbf{D}\text{isruptive-ness or }\mathbf{D}\text{eleterious-ness)}$$

$$= \text{motif-score}_{\text{ref}} - \text{motif-score}_{\text{alt}}$$

$$= -10 \cdot \log_{10}\left(\frac{p\text{-}value_{\text{ref}}}{p\text{-}value_{\text{alt}}}\right)$$

Positive D-score denotes a variant is decreasing the likelihood of TF to bind the motif (motif-break), and negative D-score denotes a variant is increasing the likelihood of TF to bind the motif (motif-gain). For assessing D-score, uniform nucleotide background were assumed (A:C:G:T=1:1:1:1), and the p-value threshold of 1e-3 was used. For position weight matrix (PWM), JASPAR TF profiles (2016 core non-redundant vertebrates, http://jaspar.genereg.net/html/DOWNLOAD/JASPAR_CORE/pfm/nonredundant/pfm_vertebrates.txt) were used, and variants that affect multiple TF binding profiles were averaged over all D-scores. More details about the tool and code can be found in https://github.com/hoondy/MotifTools.

*Figure S 5-2 Schematic of Motiftool output*




Somatic variants were further prioritized using conservation score (high positive GERP score).

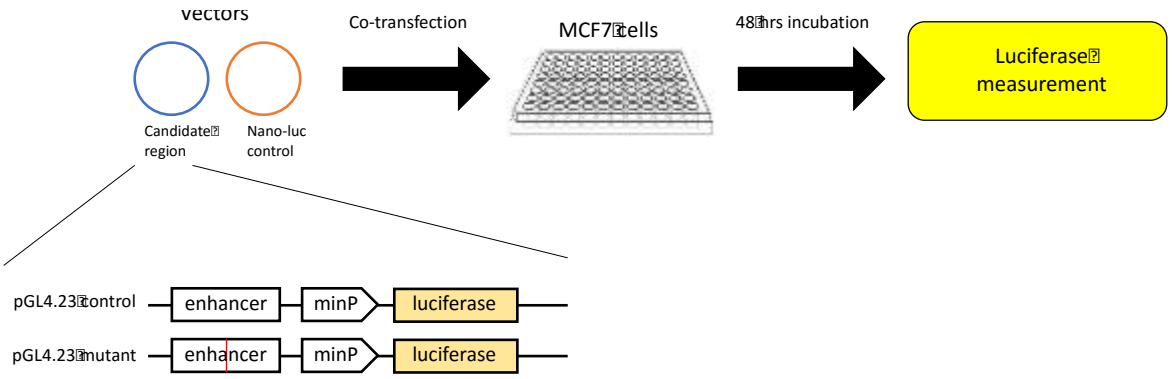
*Table S 5-1 validated mutations in MCF-7 and luciferase assay tested region*

| SAMPLE | CHR | POS | REF | ALT | TEST_START | TEST_END | NOTE |
|---|---|---|---|---|---|---|---|
| Sample01 | chr16 | 85604242 | C | G | 85603992 | 85604491 | issue with plasmid isolation |
| Sample02 | chr21 | 27541982 | G | A | 27541732 | 27542231 | |
| Sample03 | chr8 | 21541726 | A | G | 21541476 | 21541975 | issue with plasmid isolation |
| Sample04 | chr17 | 38474408 | C | G | 38474158 | 38474657 | |
| Sample05 | chr20 | 43971343 | G | C | 43971093 | 43971592 | |
| Sample06 | chr7 | 1598567 | C | T | 1598317 | 1598816 | |
| Sample07 | chr20 | 58563412 | C | T | 58563162 | 58563661 | |
| Sample08 | chr7 | 150759483 | C | G | 150759233 | 150759732 | |
| Sample09 | chr7 | 5596005 | T | G | 5595755 | 5596254 | |
| Sample10 | chr6 | 134700462 | G | T | 134700212 | 134700711 | |

## 5.2 Experiment Details SNV validation

Each regulatory region (both wild and mutant types) was separately synthesized. Enhancer regions were designed in such a fashion where based on the candidate SNV site, 250bp upstream and 250bp downstream was included for each enhancer region. These regions were then cloned into the pGL4.23[luc2/minP] vector (Promega, Cat# E841A). Each candidate region was placed upstream of the minP promoter to determine the effect of each putative enhancer region on luciferase expression. 100ng of each candidate construct and 100ng of Nano-luc control was co-transfected into MCF7 cells (5,000 cells per well in DMEM media containing 10% FBS and 1% Penicillin-Streptomycin antibiotic) using the Lipofectamine 3000 reagent (Thermo Fisher, Cat# L3000001) according to manufacturer's instructions. Cells were incubated for 48 hrs before reading the luciferase signal using Promega Nano-Glo luciferase kit (Promega, Cat# N1521) according to manufacturer's instructions.

*Figure S 5-3. Schematic of SNV validation*




The raw data of the experiment have been listed at Table S 5-2 and Table S 5-3.

*Table S 5-2. Details of SNV replication technical replicate 1*

| | Normal Rep 1 | Normal Rep 2 | Normal Rep 3 | Mutant Rep 1 | Mutant Rep 2 | Mutant Rep 3 |
|---|---|---|---|---|---|---|
| **Background** | 831 | 388 | 416 | 2623 | 1296 | 1065 |
| **2** | 7698 | 5193 | 6893 | 161889 | 132344 | 179837 |
| **4** | 587863 | 778963 | 603304 | 465322 | 408546 | 460135 |
| **5** | 10281 | 16083 | 17192 | 40103 | 63770 | 48912 |
| **6** | 39090 | 20019 | 23419 | 7614 | 6760 | 4959 |
| **7** | 15039 | 18873 | 13468 | 57945 | 47666 | 59931 |
| **8** | 117702 | 115358 | 150245 | 189131 | 295907 | 247173 |
| **9** | 26775 | 30804 | 34042 | 58424 | 104433 | 27587 |
| **10** | 21705 | 22249 | 17162 | 107077 | 31005 | 76174 |
| **Empty** | 61423 | 87225 | 46835 | 774 | 789 | 1111 |
| **Background** | 562 | 1461 | 748 | 4582 | 967 | 473 |
| **Background** | 238 | 500 | 395 | 857 | 635 | 921 |

*Table S 5-3 Details of SNV replication technical replicate 2*

| | Mutant Rep 1 | Mutant Rep 2 | Mutant Rep 3 | Normal Rep 1 | Normal Rep 2 | Normal Rep 3 |
|---|---|---|---|---|---|---|
| **Background** | 11852 | 13823 | 14402 | 15111 | 13245 | 9858 |
| **2** | 1922952 | 1854116 | 1882977 | 2326518 | 1637299 | 1927383 |
| **4** | 1969924 | 1947206 | 2088052 | 1606057 | 1133593 | 1246025 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **5** | 1396532 | 1408962 | 1879464 | 2110566 | 1890350 | 1594218 |
| **6** | 1756884 | 1798060 | 1859447 | 1825321 | 1597249 | 1658538 |
| **7** | 1884514 | 2197614 | 2393865 | 2124074 | 1385636 | 1888050 |
| **8** | 1695866 | 1711603 | 1488882 | 2405882 | 1487463 | 1516048 |
| **9** | 1715909 | 1943040 | 1916404 | 2058790 | 1385673 | 1241105 |
| **10** | 1771446 | 1498757 | 2030086 | 1736458 | 985080 | 1237019 |
| **Empty** | 2575562 | 2699389 | 2494020 | 22537 | 10758 | 6625 |
| **Background** | 12437 | 14855 | 12235 | 7338 | 4629 | 2613 |
| **Background** | 3835 | 4041 | 4182 | 2990 | 1698 | 1009 |