Using the ENCODE regulatory data to interpret non-coding somatic variants in cancer

# 

[**Abstract**](#_ige7asid7u3z)[**1**](#_ige7asid7u3z)

[**Figure 1. Data Figure**](#_rah5h626hyhw)[**2**](#_rah5h626hyhw)

[**Figure 2. Burden Figure (Middle Scale prioritization)**](#_uxevip35mgqy)[**2**](#_uxevip35mgqy)

[**Figure 3. Rewiring Figure (Macro Scale prioritization)**](#_86369v1ak8ic)[**3**](#_86369v1ak8ic)

[**Figure 4. Chromatin Figure**](#_px5wu3472qj4)[**4**](#_px5wu3472qj4)

[**Figure 5. Validation Figure**](#_a28gzpe5twsq)[**4**](#_a28gzpe5twsq)

[**Main Figure Section**](#_jr4d9xruc5d8)[**4**](#_jr4d9xruc5d8)

[**Supplementary Figure section**](#_y4i8z1la0blz)[**10**](#_y4i8z1la0blz)

# 

# Abstract

We understand the impact of somatic mutations well in a very limited number of cancer genes; in contrast, the overwhelming number of mutations in cancer genomes occur in non-coding regions. The new release of the ENCODE data allow us to bridge these two facts. First, the new ENCODE data enables precise tissue-matched genome-wide background mutation rate calibration in a variety of tumors by separating the effect of well-known confounders, such as replication timing and chromatin status. Furthermore, by integrating large scale ChIP-seq, DNase-seq, Enhancer-seq, Hi-C, and ChIA-PET data from ENCODE, we are able to define with high confidence distal and proximal regulatory elements and their linkages to annotated genes. This enables us to create extended gene definitions, and we are able to show these are more sensitive than coding regions in terms of burdening analysis. In particular in leukemia, in addition to well-known drivers such as TP53 and ATM, it allows us to pick up other key genes such as BCL6, which can then be associated with patient prognosis. Second, we integrated the ENCODE data to build up a high confidence TF-gene regulatory network. This enabled us to identify highly rewired (i.e. target changing) TFs, such as NRF1 and MYC by comparing tumor and normal samples. By integrating large-scale chromatin features, we demonstrated that such massive rewiring events between tumor and normal cell lines are mainly attributable to the chromatin structure changes instead of direct mutational effect. Furthermore, we also found that TFs with more mutationally burdened binding sites (e.g., EZH2 and NR2C2) tend to be located at the bottom hierarchy of the TF regulation network. Third, using the ENCODE regulatory network, we developed integrative scoring workflow to prioritize key elements (and mutations in them) according to their role in cancer and then validated these in small-scale studies. In particular, we prioritized ZNF687 as a key TF for breast cancer and SUB1 as a key RNA binding protein for liver and lung cancer and validated them through siRNA knockdown experiments. Finally, we identified key enhancers and mutations in them in breast cancer and then validated their functional effect through luciferase assays.

# Introduction

What's the key background

How to put this in context

Why should you be interested

Recent developments of whole genome sequencing (WGS) and personal genomics have provided unprecedented opportunity to identify deleterious mutations that are importation for tumor genesis, which in turn would allow for targeted therapies in clinical studies. However, although hundreds to thousands of WGS data were provided through collaborative effort from many consortiums, the overwhelming number of mutations occur in non-coding regions, where their functional impacts still remain challenging. Deciphering how these noncoding regions interplay and how they are perturbed in cancerous cells are key to understand cancer.

Since the inception of ENCODE project, deep-sequencing of entire human genome allowed us to identify many functional regulatory regions and link these regions to the better understood coding regions to uncover the underlying biological mechanism. However, it is still challenging to directly incorporate the ENCODE data in an effective way to study the effect of somatic variants in cancer genomes. First, while ENCODE provides one of the most comprehensive functional genomics data, data sets from various experiments are quite heterogeneous and require careful processing and thinking to be merged together. Second, due to the heterogeneous nature of various cancer types, it is important to data from most relevant cell line when evaluating the somatic variant effect in different cancer types. However, tissue matching is still a challenging problem. Lastly, none of the available cancer genomics data is complete in any cancer cell line. Hence, how to maximize the utility of ENCODE data from other tissue is an important topic.

We here presented an integrative framework to tailor the whole ENCODE data for cancer analysis. Specifically, we first integrated highly heterogeneous functional genomic data into signal matrices to dissect the somatic mutational landscape. Then we integrated the most comprehensive non coding annotations and precisely linked them to well-known coding genes to better interpret the non coding variants. We finally set up loosely matched tumor and normal gene regulation network to understand dramatic gene expression regulation changes during the transition from tumor to normal. Experimental validation at different scales demonstrated the effectiveness of this data to pinpoint the key elements/variants in various cancer types.

# Figure 1. Data Figure

1. ***Cancer cell lines data***:
   1. ENCODE includes extensive functional genomics data for cancer cell lines ([FigM 1](#qbsccindw6ao) (A))

* Data types used in this paper: RepliSeq, CHIP-seq (TF & Histone), DNAse-seq, RNA-seq, RRBS/WGBS, CHIA-PET, Enhancer-seq, Hi-C, shRNA RNA-seq, siRNA RNA-seq,
* Three key problems to integrate data
  + Raw data from ENCODE is heterogeneous, how to incorporate these data (solutions: data matrix at different resolution)
  + The most relevant functional genomic data usually provides best performance, but such tissue matching option is usually not obvious. (computationally select the best from all)
  + None of the cell line has a complete list of all genomic data, but there is usually correlation among different cell types. Our framework could allow to pick up information from unmatched tissue for rarer cell lines but still provide decent performance.
  1. We have integrated uniformly processed and quality-controlled datasets from ENCODE and Roadmap Epigenomics Consortium to build one of the most comprehensive representation of how functional regulatory elements interplay in human “cancer” genome.

1. *Extended gene definition:* High-confidence extended gene definition could be built up based on these data ([FigM 1](#qbsccindw6ao) (C))
   1. Difference of our annotation from the official ENCODE list is that we provided the most reliable gene target prediction by integrating various data types. It helps to bridge the noncoding annotation with several cancer genes.
2. *Regulation network:* Gene expression regulation network can be built based on ChIP-seq and EnhancerSeq data ([FigM 1](#qbsccindw6ao) (D))
   1. We build up the network based on the the personal genome for the key cell lines

# Figure 2. Burden Figure (Middle Scale prioritization)

**BMR = background mutation rate**

Population level analysis, which looks for regions mutated more frequently than expected, is one of the most powerful ways to identify deleterious mutations for diseases. Recent developments of whole genome sequencing (WGS) and personal genomics have provided unprecedented statistical power to perform such analyses. However, mutation burden tests for somatic variants in cancer research remain challenging for several reasons.

[[SETUP]] The first step in her cancer analysis is to really understand the mutational recurrence properly and to one of the tricky bits of this of course is that this is compounded by many genomic factors which affect the underlying background mutation rate against which you have to find the of the richest mutations for the right regions see it works enriched with mutations

1 - integrates many features - many matched features are necessary

* Comparison w just a feature & the unmatched

2 - allows for differences w respect to local nt

Numerous genomic features have been reported to largely affect the mutation process, necessitating careful correction in burden analysis. The large cohort of functional genomics data in ENCODE provided us a chance to build up an integrative model to better dissect the somatic mutational landscape. Specifically, we collected the most comprehensive features from ENCODE and processed these heterogeneous data into a covariate matrix to predict the local mutation rate with high precision through regression.

[[SETUP]] Why can't most of cancer analysis to now has focused just with coding regions obviously those mutations and regions linked genes such as their promoters enhancers can potentially also have an effect on the gene by using the Incarnation we are unable to of this is pretty yucky

In addition, we also customizes the most comprehensive non coding annotations from ENCODE and link them to the well characterized protein coding genes to defined the extended gene region with high confidence. Our model allows to pick up weak mutation burdens from various coding and noncoding regions for an extended gene in a collaborative way to better define mutation burden.

Something like figure XXX demonstrates that using the extended gene model is more sensitive than not

1. ***Reason for BMR correction***:

Somatic background mutation rate in cancer genome is mainly confounded by two factors - large scale genomic features and smaller scale local context. Without precise background mutation rate calibration, mutation rate usually changes up to several orders even in the same sample, which would severely affect the mutation burden calculation and generate numerous false positive and negatives.

* 1. Replication timing, DHS, expression and histone modification all highly correlated with background mutation rate. ([FigureS 2.1](#p24c64z0baxt))
  2. Without correction, BMR changes up to several orders across different regions of the genome ([FigureS 2.2](#d9dexd53raik))
  3. Local context effect also significantly affect mutation rate in various cancer types ([FigureS 2.3](#mnxv7i8x9ce7))
  4. Could claim if there is not proper correction, there is many false positive and negatives and add up a schematic figure, but mutSigCV already included one figure like this. Suggest to remove it even for supplementary figure

1. ***Reason for collaborative BMR correction***: Cumulative effect of different features

* 1. Joint mutation rate estimation improves BMR mutation rate estimation ([FigureS 2.4](#dycezwb4nfyl) , [FigM 2](#tzh2erk9pdci)  (B-C))
  2. Matched tissue usually provides better estimation performance (Action: replication timing analysis)

1. ***Value for ENCODE data for BMR***: High correlation among ENCODE features

It has been reported that the most accurate local mutation rate prediction can be achieved by using features from matched tissue. However, biologically meaningful tissue matching remains challenging and usually is not an obvious process for researchers without enough domain knowledge. Specifically, if samples of distinct hidden subtypes were pooled together for a certain disease, tissue matching would be more difficult. Furthermore, even after the optimally matched tissue has been identified, we frequently need to handle missing features in that tissue. We noticed that many genomic features are highly correlated both within and across tissues , which leads to suboptimal but still decent regression performance. This is extremely helpful when processing WGS from diseases without matched features. For example, there are no prostate related features in ENCODE, but features in other tissues still help to estimate the local mutation rates.

* 1. Different features are highly correlated ([FigureS 2.5](#b24t5hfbok8g))
  2. ENCODE data is still value for new cancer types
     1. Set of complete features is often missing for many cancer types (data table 1)
     2. To know the complete sensible matching tissue is difficult (Shirley’s enhancer paper?)
     3. Due to the correlation of features, BMR estimation still provides decent performance even if there is no matched data (Prostate cancer example)
  3. PCA analysis of the covariate matrix showed that the first PC contains a mix of many different features - not just replication
  4. Matched tissue replication timing usually provides best performance, however increasing number of PCs of the remaining covariate matrix continue to contribute to BMR estimation precision significantly

1. ***Value for ENCODE data for annotation***: Extended gene definition helps to identify burdened genes by gathering mutation burdening for both coding and noncoding regulatory regions of a gene.
   1. ENCODE annotation link the non-coding elements to known coding genes ([FigM 2](#tzh2erk9pdci)  (D-E))
   2. Extended gene annotation helps to gather weak signal from multiple regulatory elements of the gene and provide better burdening analysis [FigM 2](#tzh2erk9pdci)  (D-E)
2. We find BCL6 burden gene in blood cancer to be associated with patient survival

# Figure 3. Rewiring Figure (Macro Scale prioritization)

1. *Data Figure* ***Network Setup***: Gene-gene expression regulation network integration by integrating both distal and proximal regulation signals ([FigureS 3](#pd2wm7fbjzzz))
2. ***Rewiring analysis***
   1. Identify Key TFs that sharply rewires in between tumor and normal pairs ([Figure 3](#1gqcm8kn2gfp))
      1. Look across different tumor normal pairs in ENCODE ([FigureS 3.x](#3zeywtmrlzda)), while CTCF has similar between blood and lung, JUND, MYC, and BHLHE40 have different rewiring profile across cell types.
   2. Focus on K562-GM12878 pair, which has the most abundant TF ChIP-seq data, classification of TFs according to their rewiring changes from the network ([Figure 3](#1gqcm8kn2gfp) d)
   3. Table of chromatin and mutational effects ([Figure 4](#6t0uz16x9ric))
3. ***Co-association changes***: Identify key TFs that changes their co-association relationship in K&G
   1. ZNF274 as an example of co-association changes ([Figure 3 v2](#8uwtkiiw3nqk))
   2. Disruption of well-known pairs within the network
4. *TF-TF network hierarchy* analysis
   1. Highly rewiring TFs are usually associated with hierarchy change in the network ([Figure 3](#1gqcm8kn2gfp) c)
   2. The TFs in higher layer are usually more significantly associated with tumor normal expression change
   3. TFs with most frequently burdened TFBS are usually found in the bottom layer of the TF-TF hierarchy

# Figure 4. Chromatin Figure

~~Target analysis, not TF analysis, serve as middle layer prioritization~~

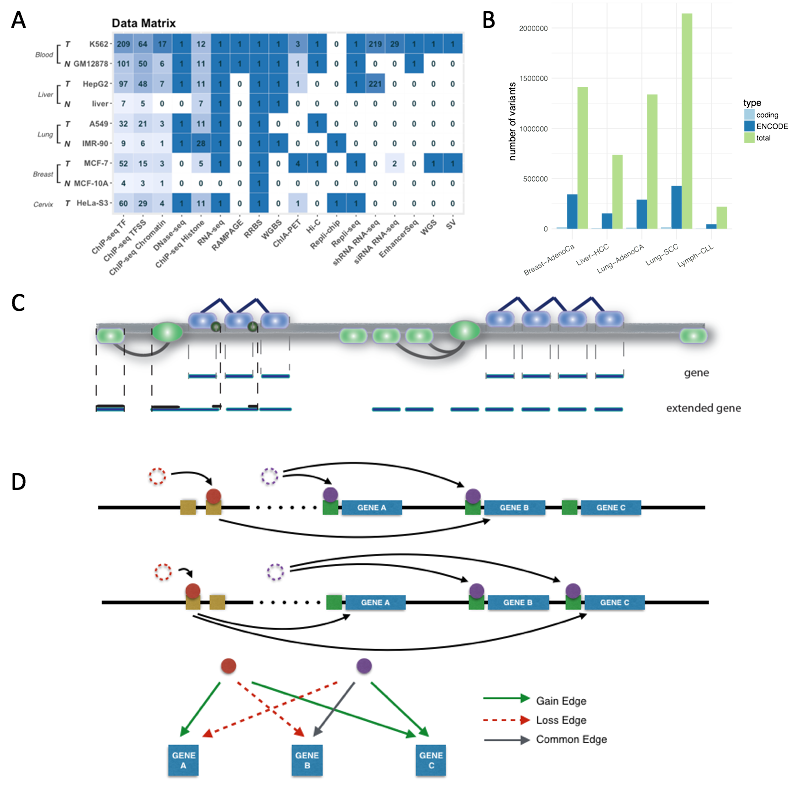
1. ~~Gene target regulation status changes~~
   1. ~~Classification of regulation status: inactive, suppressed, enhanced~~
   2. ~~Genes that undergo sharp regulation status changes is usually associated with huge expression change in tumor and normal pairs~~
   3. Target gene expression changes are accompanied by chromatin change. These regulation status change is mainly due to chromatin status changes.
   4. Mutational burdening analysis shows target genes do not harbor more mutations than others. DIRECT Mutational effect plays only a small part.

# Figure 5. Validation Figure

1. ***Flowchart*** of variant prioritization scheme
   1. Macro-level: selecting functional regulatory elements at gene-level
   2. Middle-level: selecting functional cis-regulatory regions and hotspots
   3. Micro-level: selecting critical variants that disrupt cis-regulatory elements at base-pair resolution
2. Experimental validation
   1. ***Macro Scale validation***:
      1. From our network/expression we can identify key regulators in oncogenesis & we can validate their effects w knockdown
      2. We identified ZNF687 for MCF-7, SUB1 for HepG2 and A549 as key elements that drives tumor/normal validation ([FigM 5.](#l3vm36nolov9), B)
   2. ***Middle Scale validation***: We can validate the functional activity of non-coding elements associated with cancer. Active cis-regulatory element identification using ENCODE data ([FigM 5.](#l3vm36nolov9), A).
      1. Promoter-like regions near APP gene, which is a well-known cancer gene in breast cancer
      2. Predicted enhancers in intron and intergenic regions that are regulating cancer genes
   3. ***Micro Scale validation***: Known prioritized cancer mutations can be shown to have a clear functional effect ([FigM 5.](#l3vm36nolov9), A)
      1. We prioritized and validated 8 non-coding variants that are predicted to disrupt cis-regulatory elements in MCF-7 (two experimentally failed) ([FigureS. 5.X](#761lkr2prlvg)). 7 out of 8 samples with variants showed significant change in luciferase expression compared to wildtype allele.

# Main Figure Section

FigM 1 ENCODE data summary related with Cancer



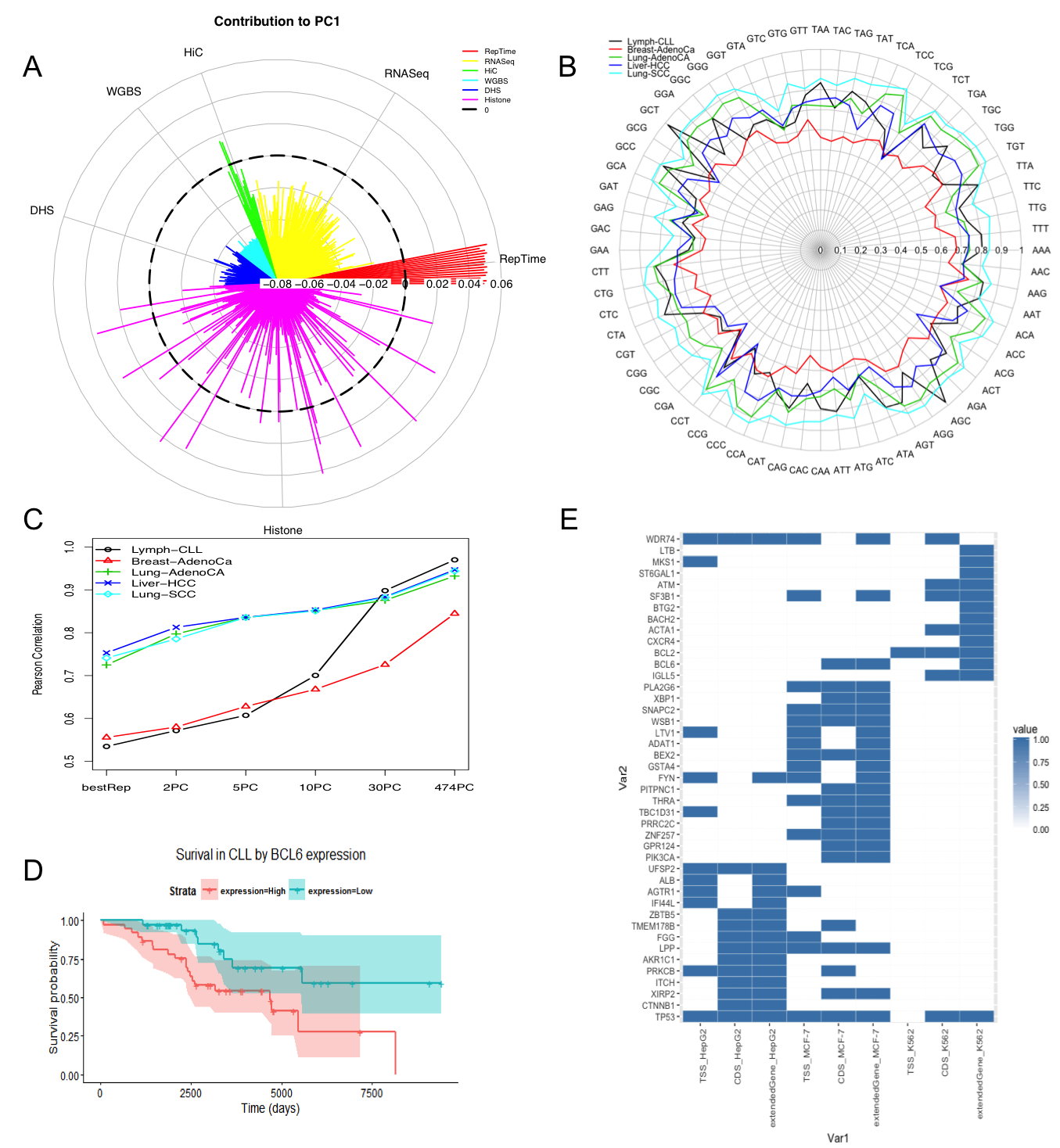
Notes

Tumor and normal pair add a common row

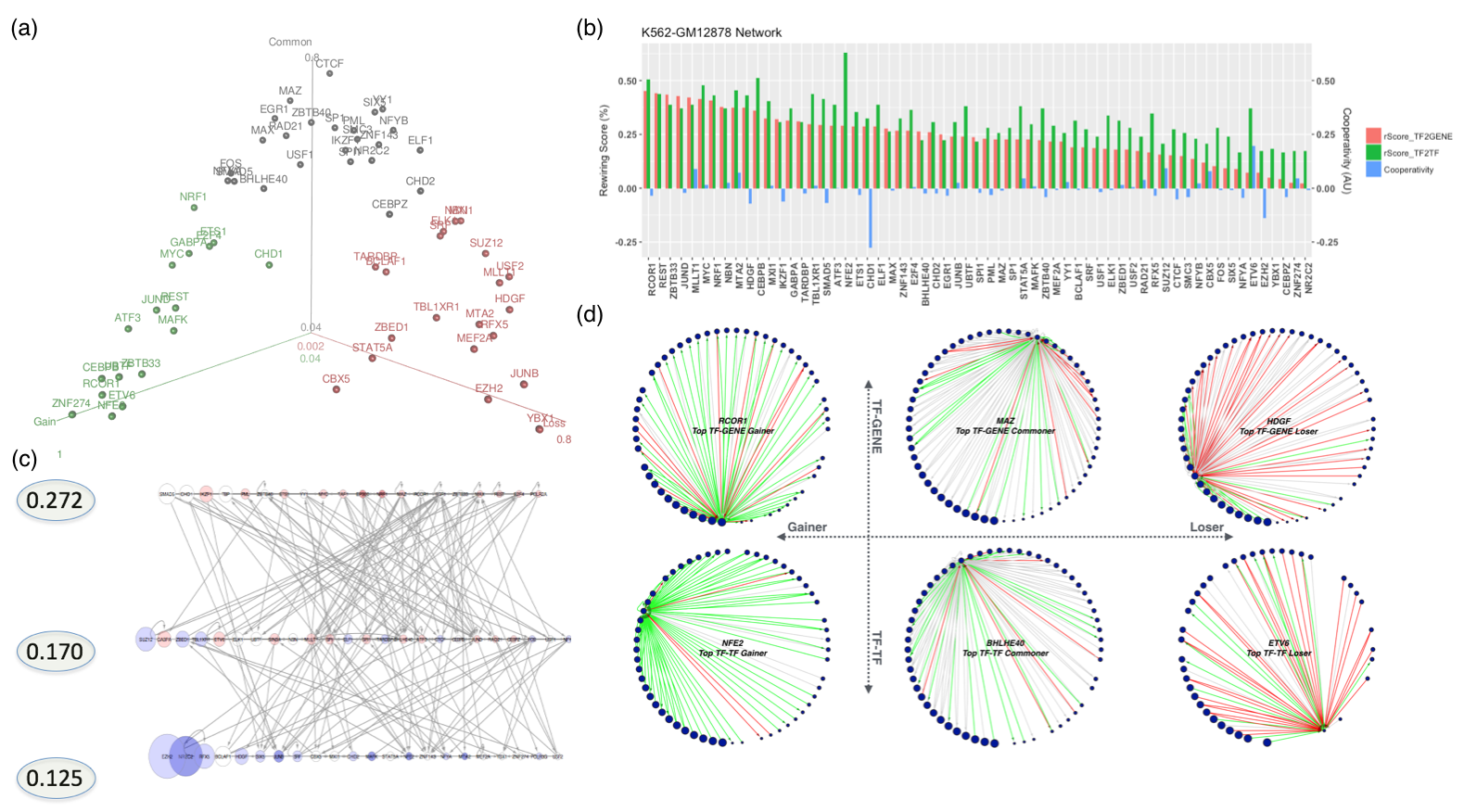
Cohort size

WGS SNV/SV for HeLa and GM

FigM 2 Somatic Recurrence Analysis



FigM 3 v1

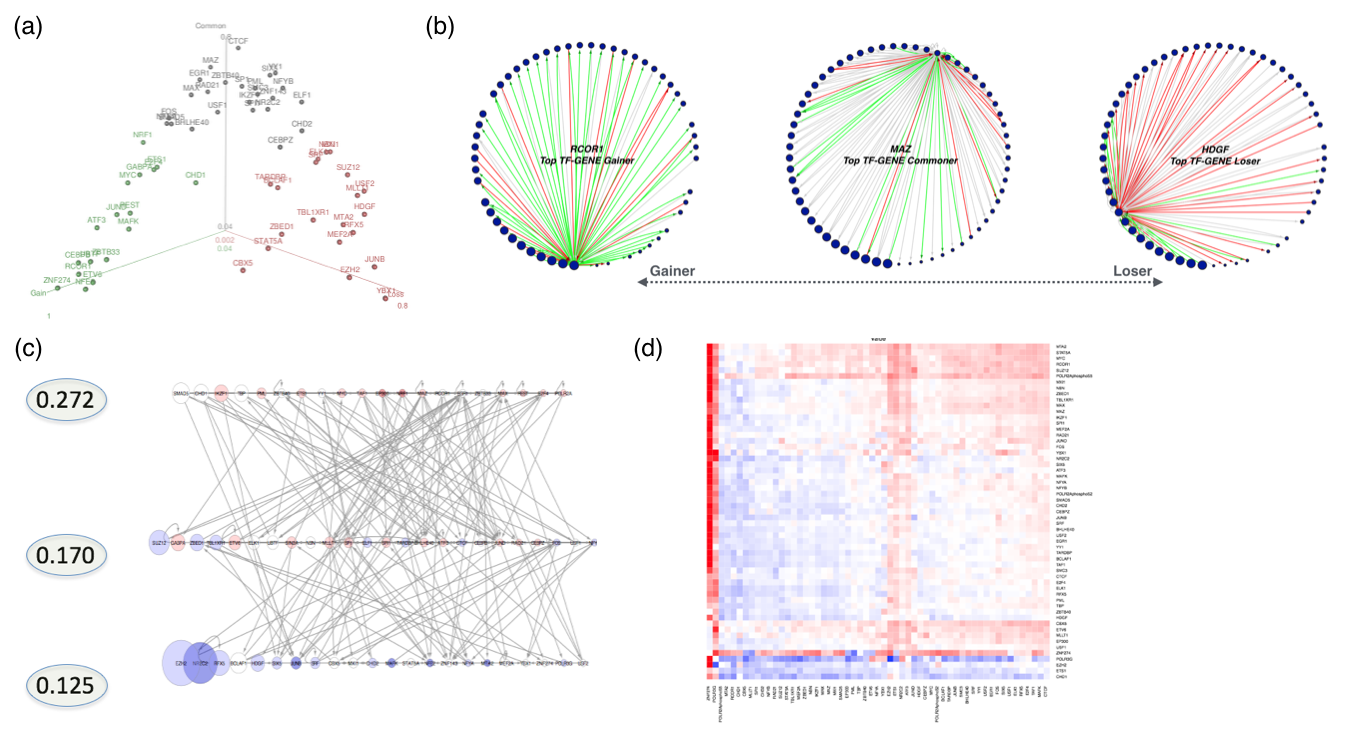


Part A - unsure - discuss & think

~~Part B - put more in here... rank tFs by how much they in/out rewire but also by how much they in co-assoc.~~

Part C - gene expre driven at bottom , size of the pts relates to mut & color relates to gene expr change

FigM 3 v2

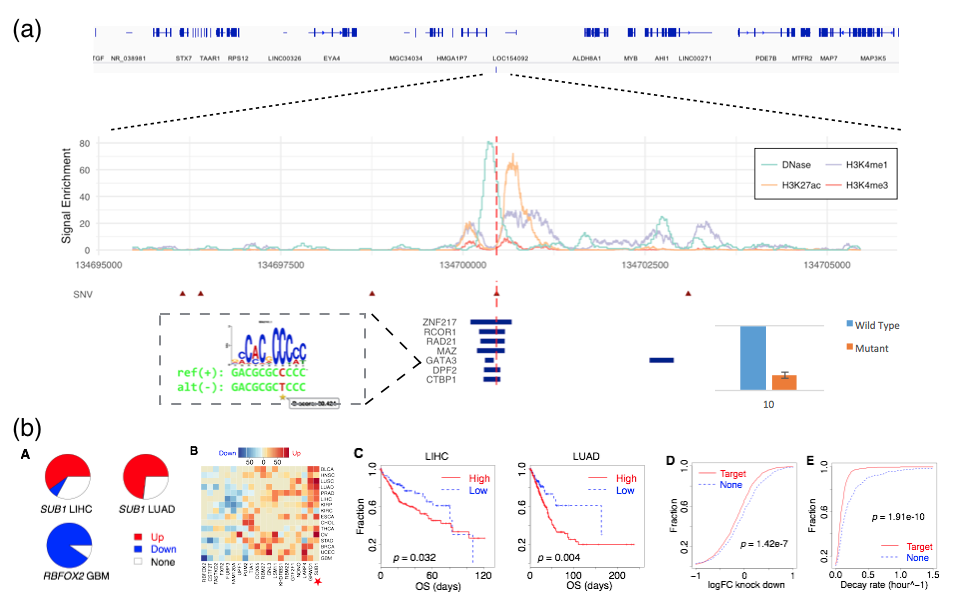


FigM 4 v1



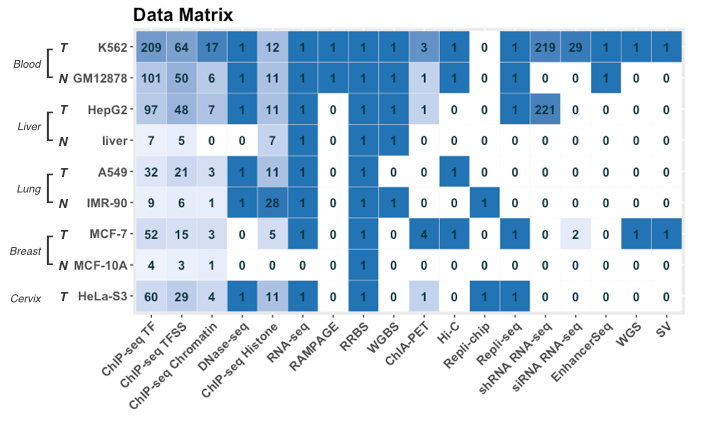
\*\* we're going to ened to work on part A ... leave for now... part b how is diff from fig 3... part c fine

FigM 5. Cis Regulatory Element Validation (Hi-C result and enhancer target linkage to be included)

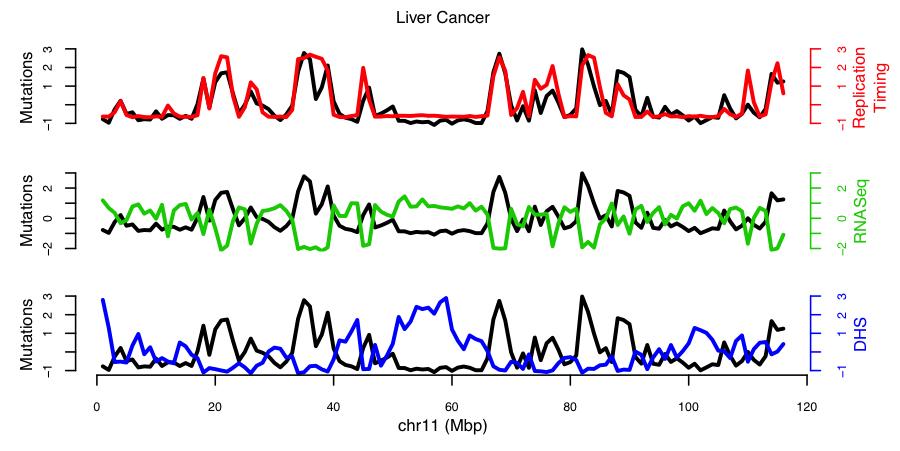


# Supplementary Figure Section

FigureM 1.1 Encode data matrix



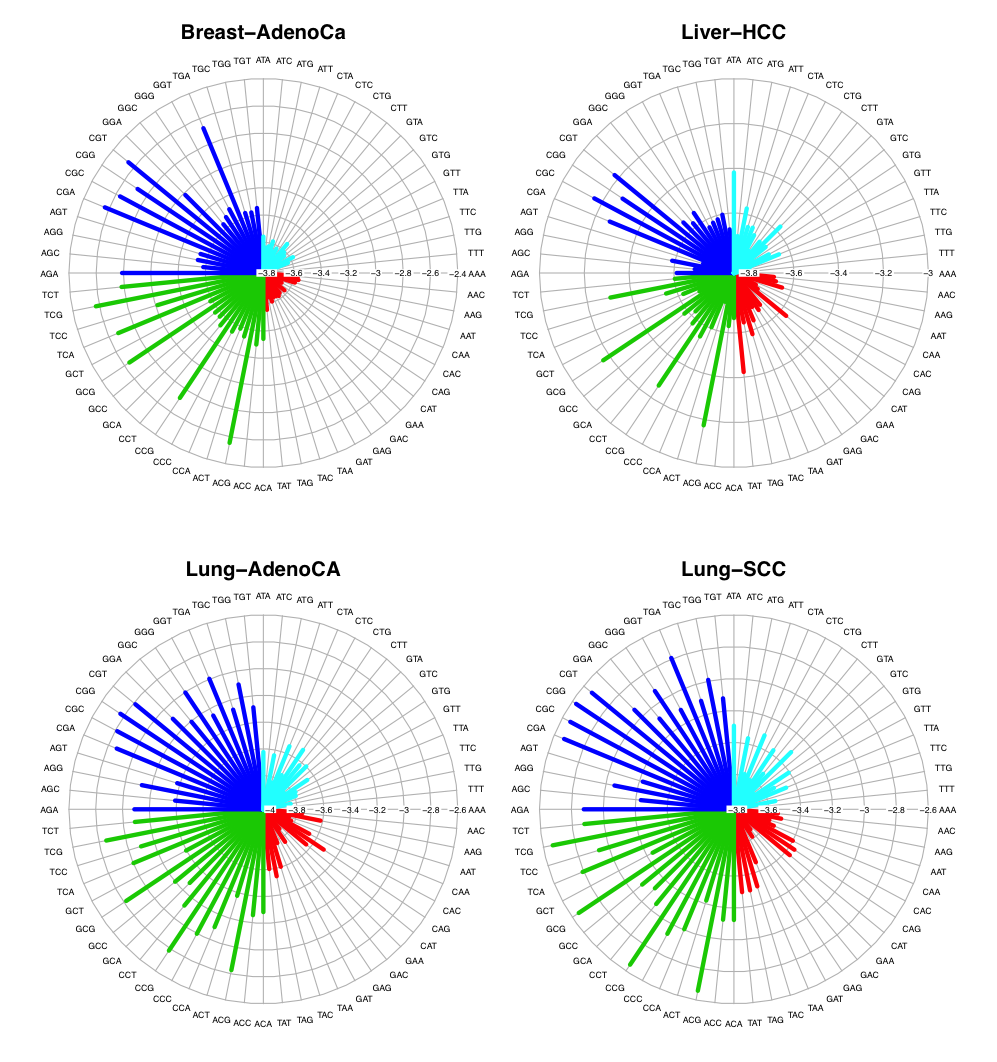
FigureS 2.1 correlation of mutation rates with some genomic features



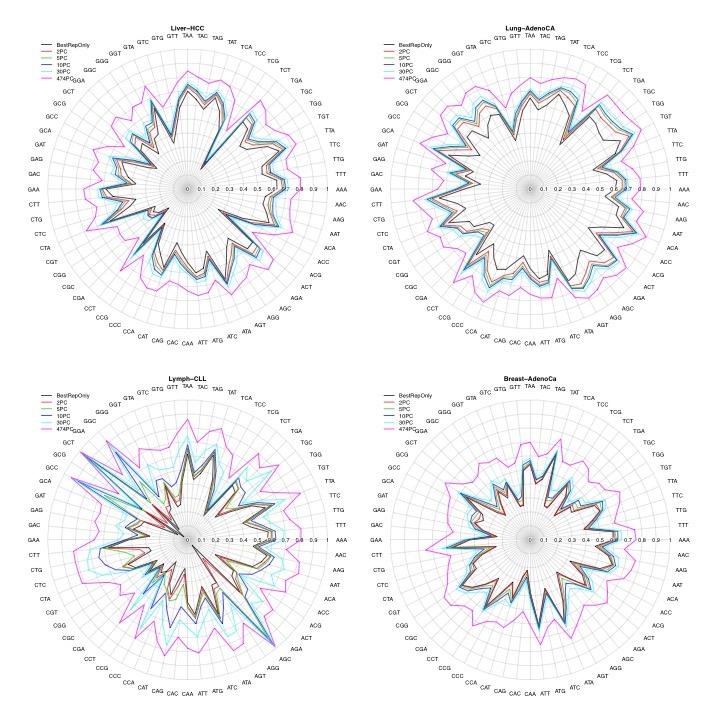
FigureS 2.2 Mutational Heterogeneity across the genome and across different local context in various cancer types



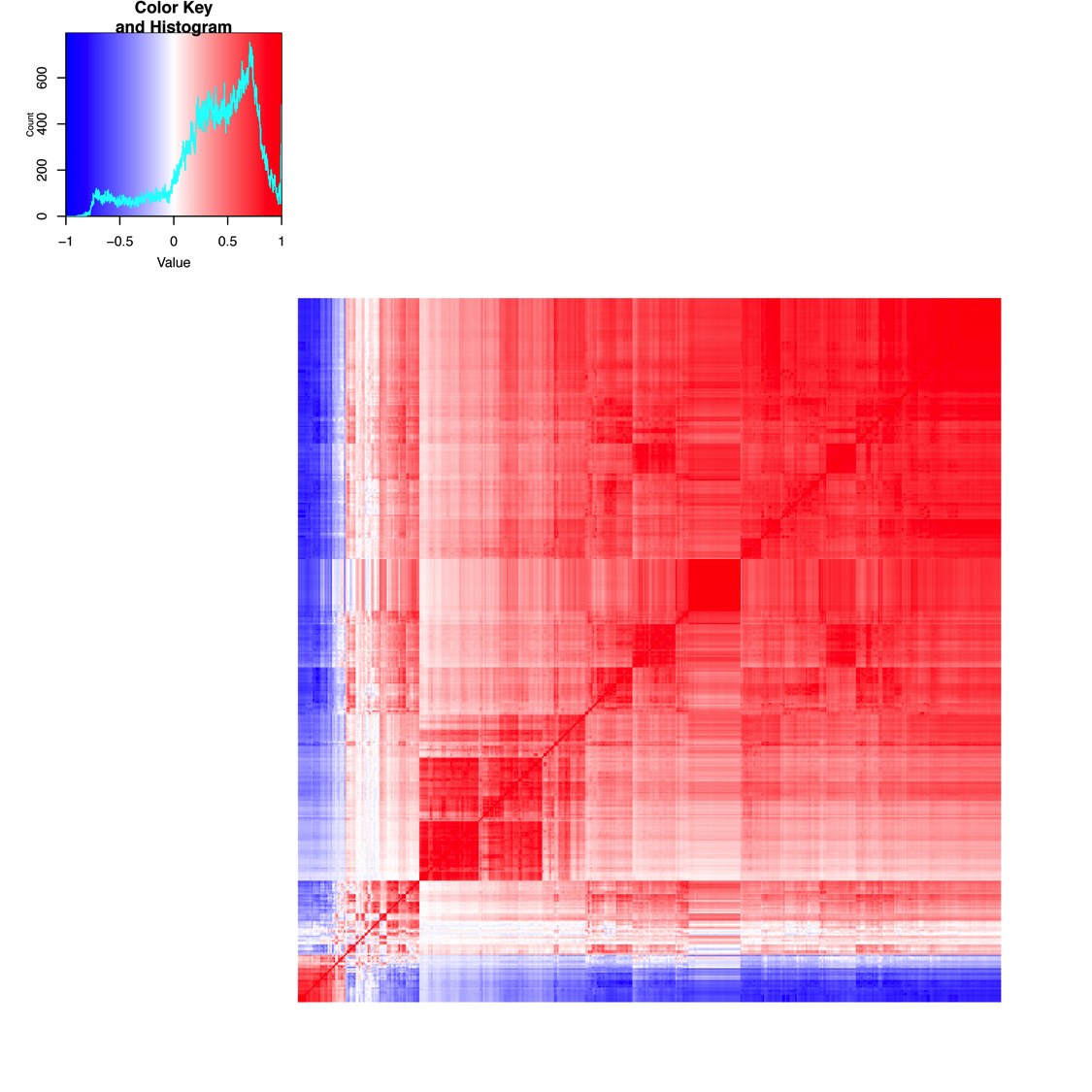
FigureS 2.3 Local context effect on mutation rate in various cancer types



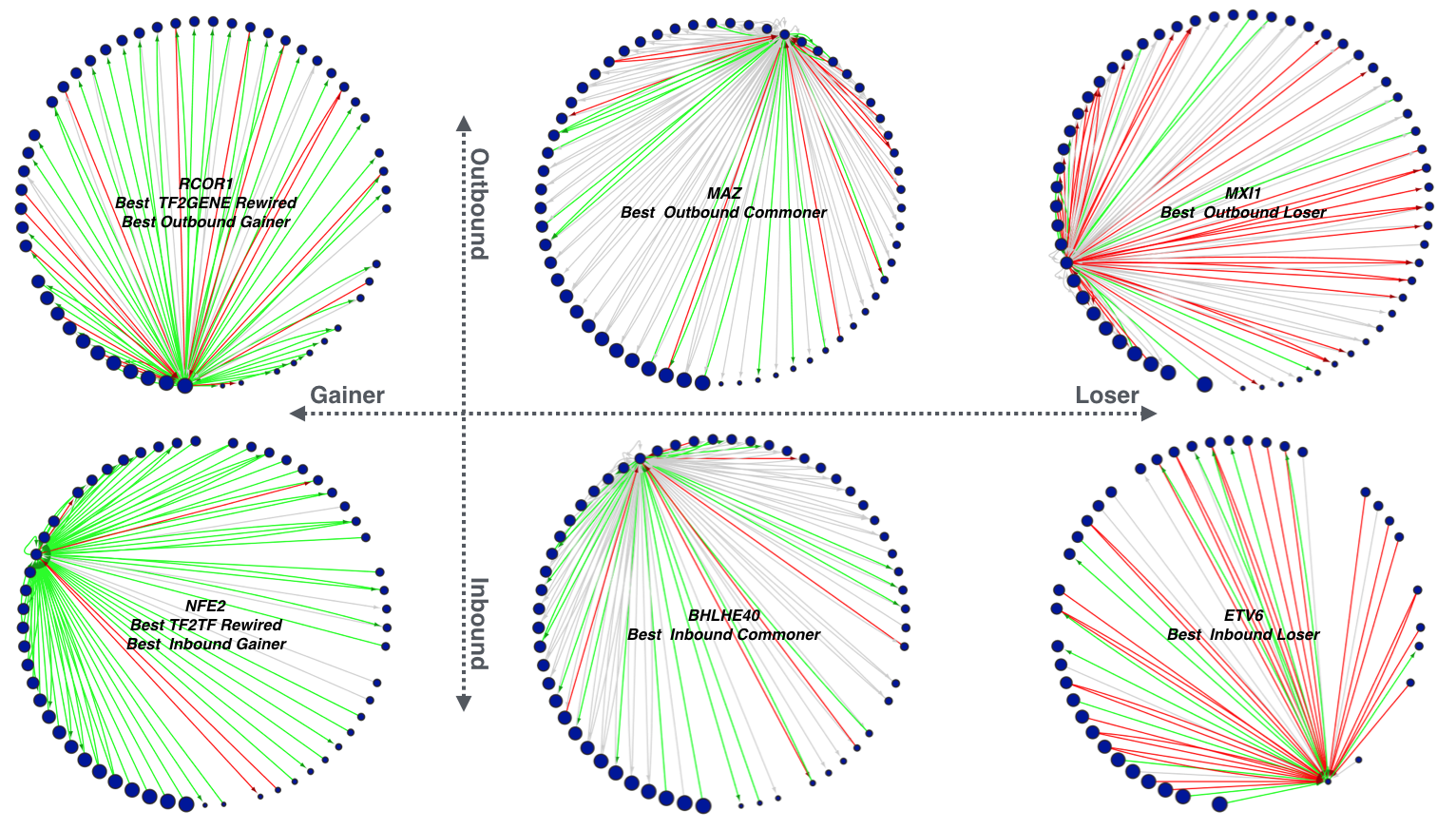
FigureS 2.4 improvement of BMR estimation using multiple features



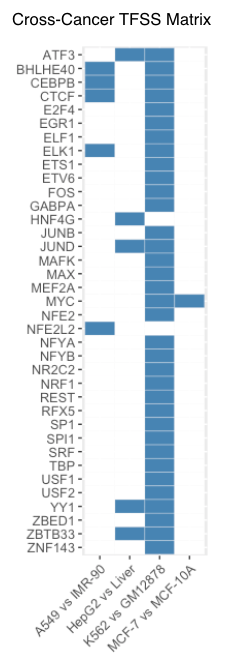
FigureS 2.5 correlation heatmap of ENCODE genomic features



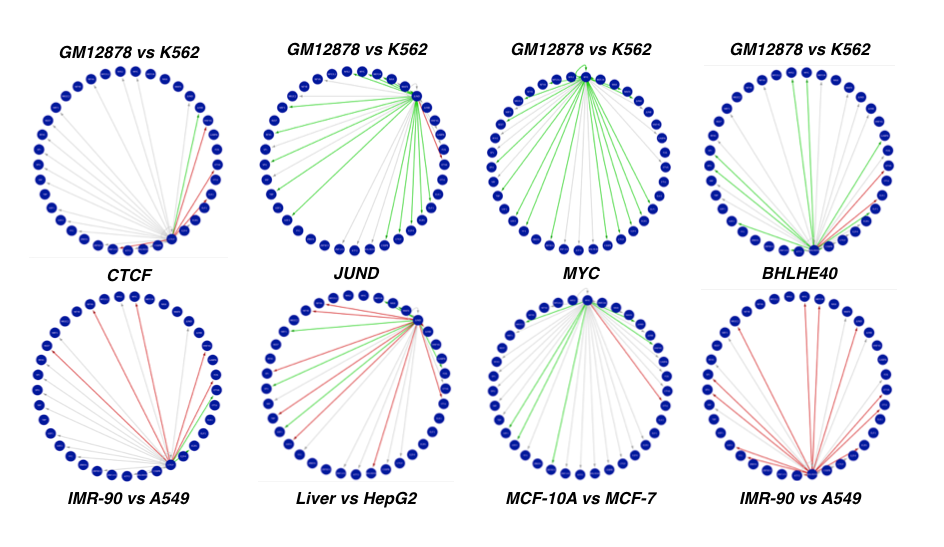
FigureS 3.x K562-GM12878 rewiring in TF-TF network



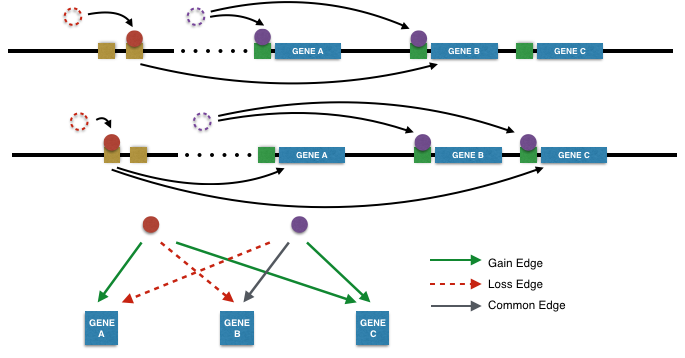
FigureS 3.x Cross cancer-normal TFSS data matrix



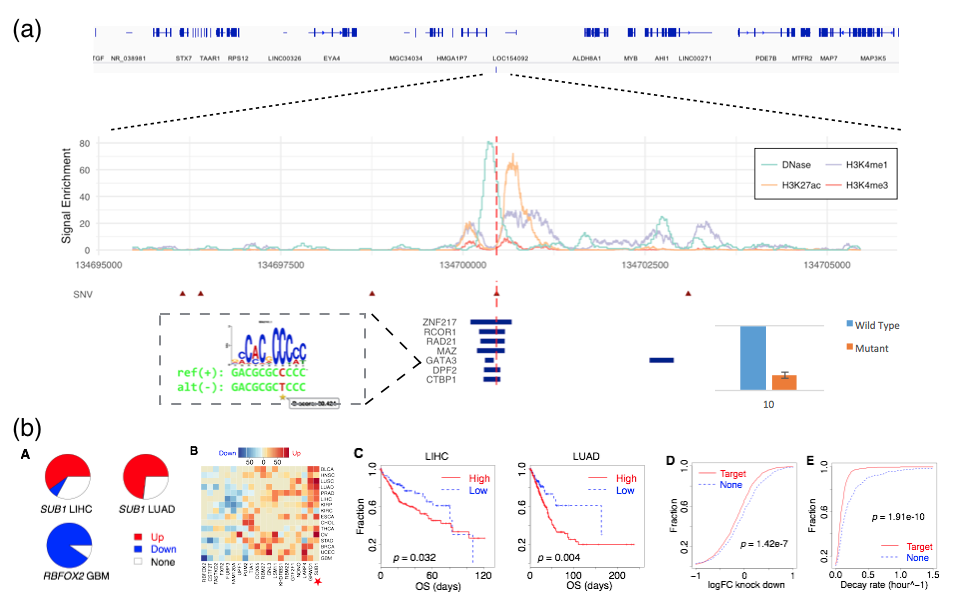
FigureS 3.x Cross cancer-normal comparison of TFSS network (Peng’s score to be included)



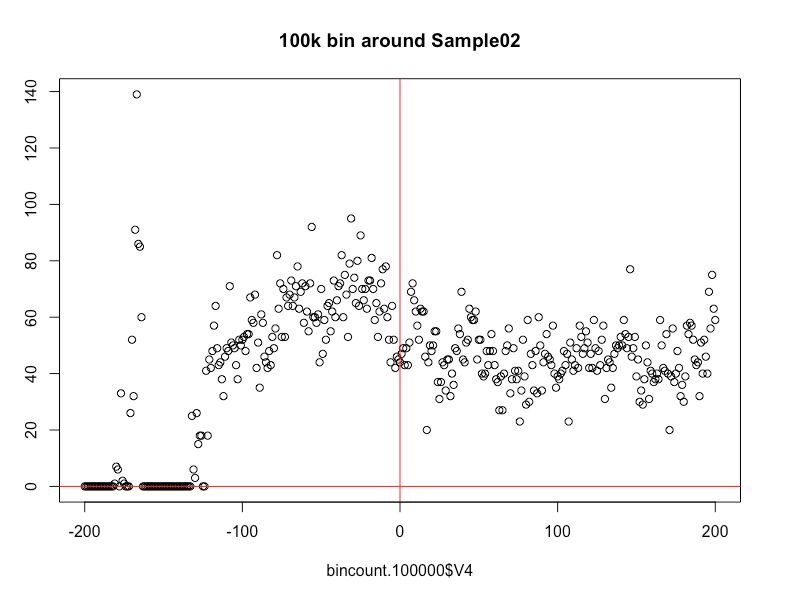
FigureS 3. Network schematic



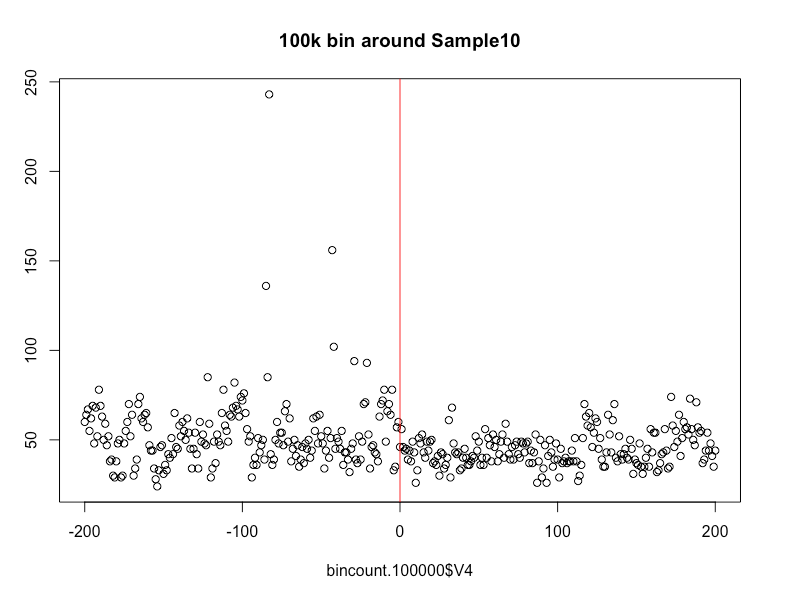
FigureM 5. Cis Regulatory Element Validation (Hi-C result and enhancer target linkage to be included)



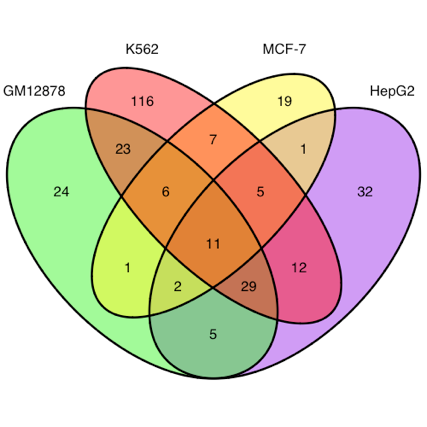
FigureS X. Mutation around validated SNV (Sample02)



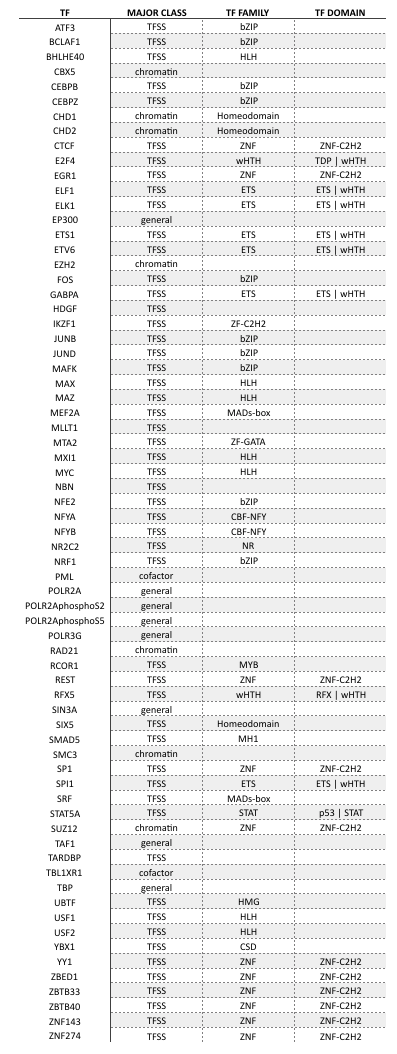
FigureS X. Mutation around validated SNV (Sample10)



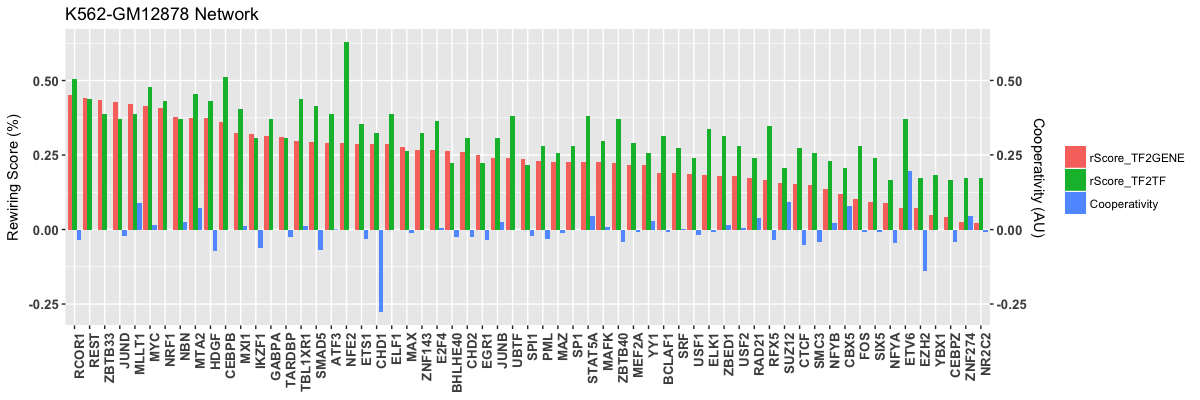
FigureS X. TF ChIP-seq experiments by cell lines



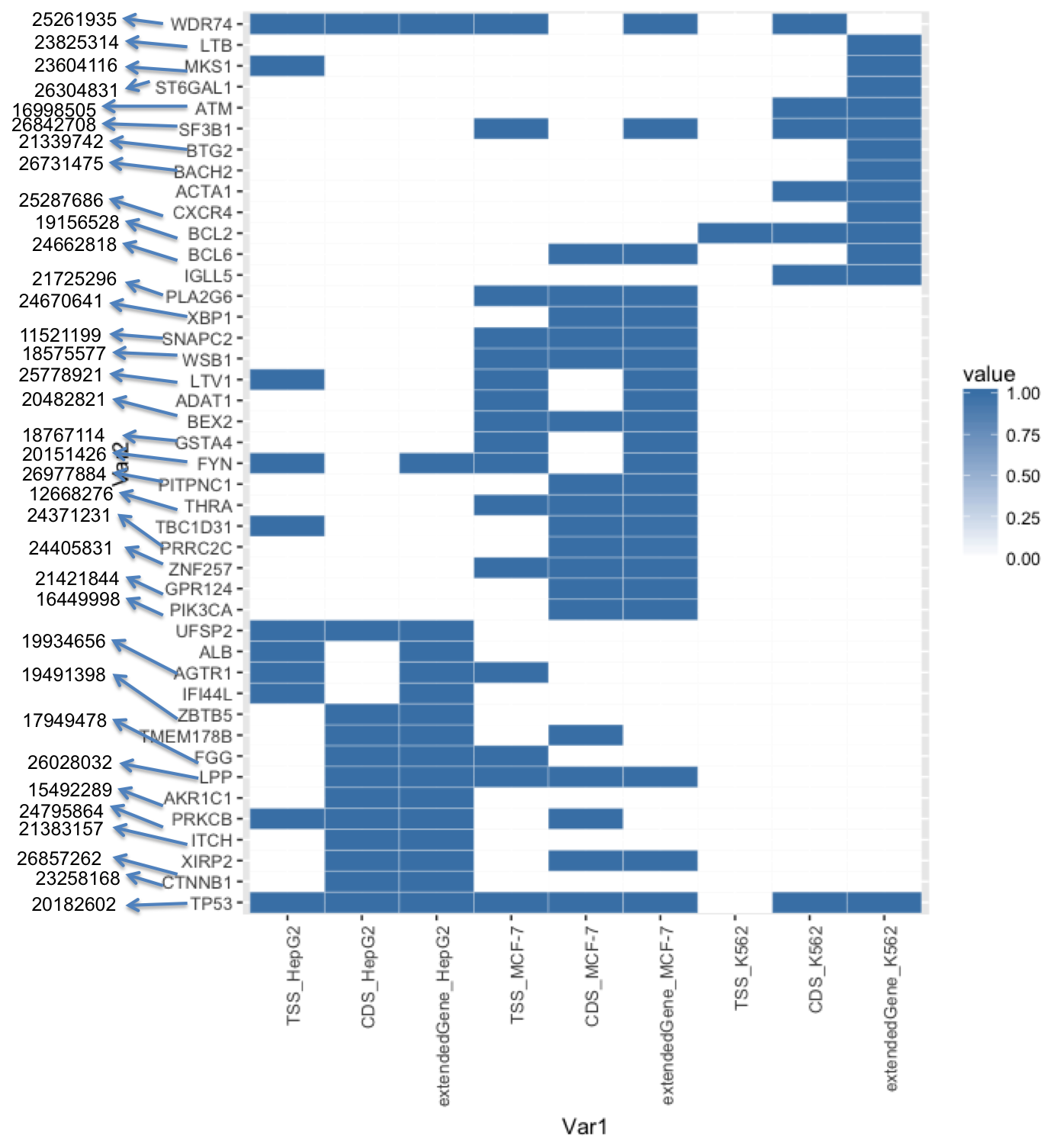
TableS X. Transcription factor classification



FigureS X. Rewiring Score and Cooperativity



FigS 2.0 Burden HeatMap



FigureS. 5.X Luciferase assay validation

