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

# Figure 1. Data Figure

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

\* there's obviously a lot of warts in the data - mispair t/n, incomplete, private variants ... no cancer is the best ... altogether this is the best data set but it's far from perfect

* 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))
2. ***Regulation network:*** Gene expression regulation network can be built based on ChIP-seq and EnhancerSeq data ([FigM 1](#qbsccindw6ao) (D))

# Figure 2. Burden Figure (middle scale prioritization)

BMR = background mutation rate

1. ***Reason for BMR correction***: Mutation rate is confounded by various genomic features
   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 (to be disc)
2. ***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)
3. ***Value for ENCODE data for BMR***:High correlation among ENCODE features
   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. \*\*\* we do the pca and the leading the component contains a mix of many different - not just replication
   4. [[[ show hela replication timing [lawrence et al], then show best matched replic timing, then pc1, then pc1+pc2]]
4. ***Value for ENCODE data for annotation***: Extended gene definition helps to SUGGEST more sensible cancer driver candidates [[too strong]]
   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 driver prediction [FigM 2](#tzh2erk9pdci)  (D-E)
5. What result - we find XXX(BCL6) burden gene bcl2 in blood cancer

\* part E smaller rectangles & more cancer

Part C[[[ show hela replication timing [lawrence et al], then show best matched replic timing, then pc1, then pc1+pc2]] & rest in suppl.

# Figure 3. Rewiring Figure (Macro Scale prioritization)

1. *dATA FIGRUE* ***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

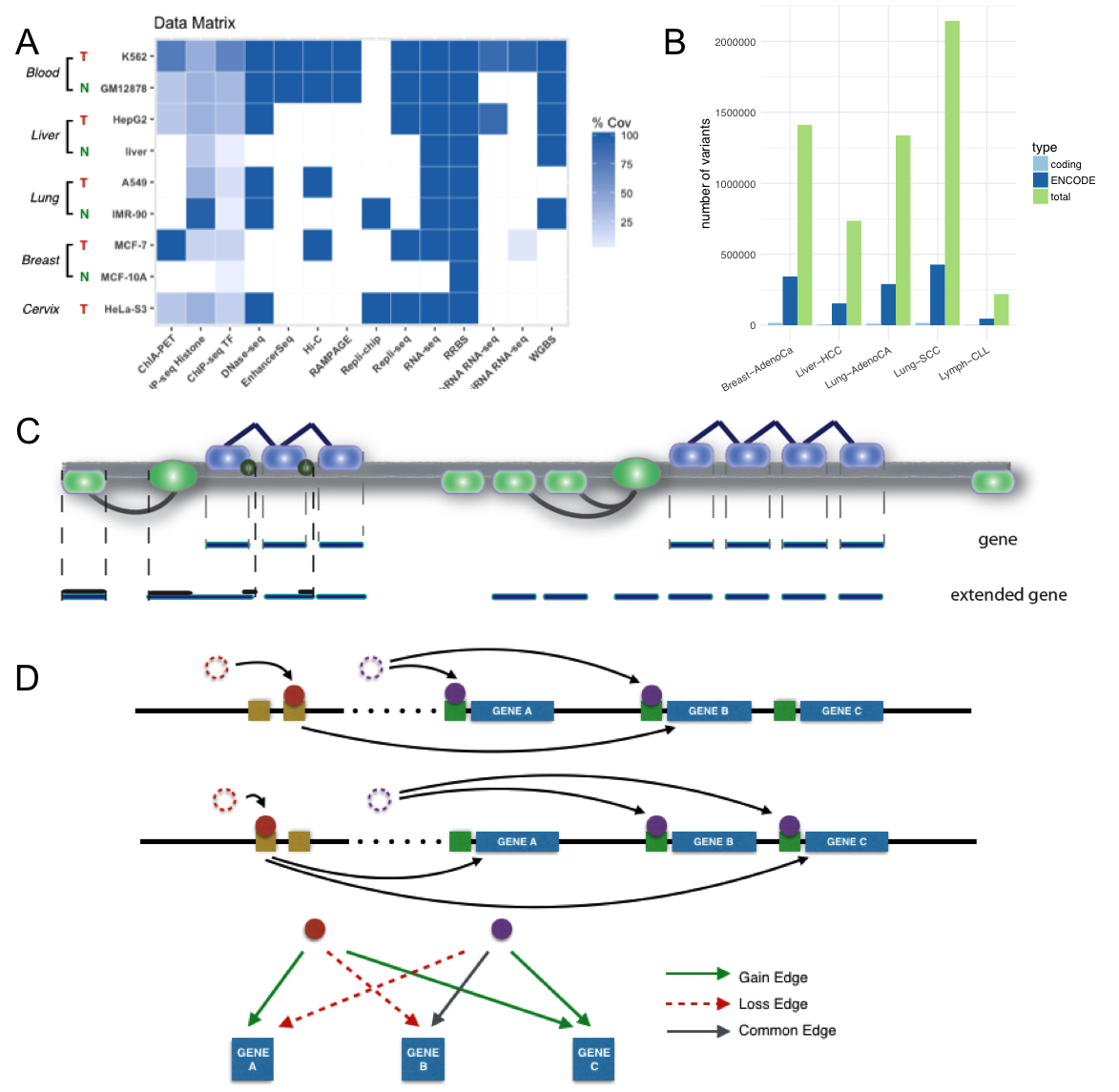
1. Gene target regulation status changes
   1. Classification of regulation status: inactive, suppressed, enhanced
   2. Genes that undergo sharp regulation status changes is usually associated with huge expression change in tumor and normal pairs
   3. These regulation status change is mainly due to chromatin status changes
   4. DIRECT Mutational effect plays only a small part

# Figure 5. Validation Figure

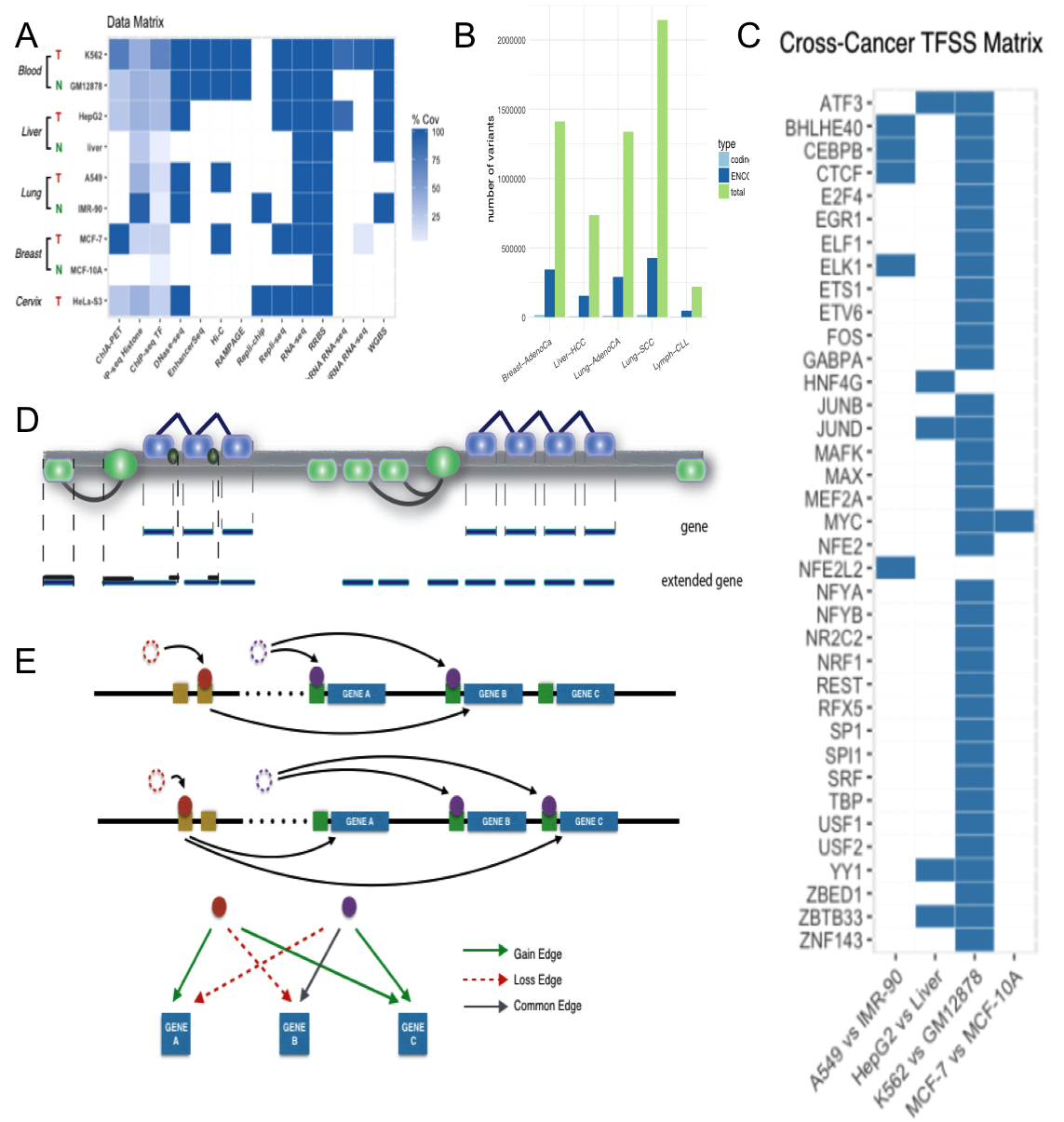
1. ***Flowchart*** of Validation experiment - macro middle & micro
2. Experimental result of Validation
   1. ***Macro Scale validation***:
   2. From our network/expression we can identify key regulators in oncogenesis & we can validate their effects w knockdown
   3. ZNF687 for MCF-7, SUB1 for HepG2 and A549 as key elements that drives tumor/normal validation ([FigM 5.](#l3vm36nolov9), B)
   4. ***Middle Scale validation***: Active cis-regulatory element identification using ENCODE data ([FigM 5.](#l3vm36nolov9), A)
3. We can validate the functional activity of non-coding elements assoc w cancer
   * 1. promoter like regions near APP gene, which is a well-known cancer gene in breast cancer
     2. Enhancers in intron and intergenic regions that are regulating cancer genes
   1. ***Micro Scale validation***: Key SNVs in the elements discovered in Middle Scale validation ([FigM 5.](#l3vm36nolov9), A) -- known prioritized cancer mutations can be shown to have a clear functional effect

# Main Figure Section

FigM 1 ENCODE data summary related with Cancer



FigM 1 V1 ENCODE data summary related with Cancer



Notes

C in suppl

par t A

Number of SeqSpecifc TF

Histone

Pol

Matched maybe on line

Inclu Variants = WGS - number - #

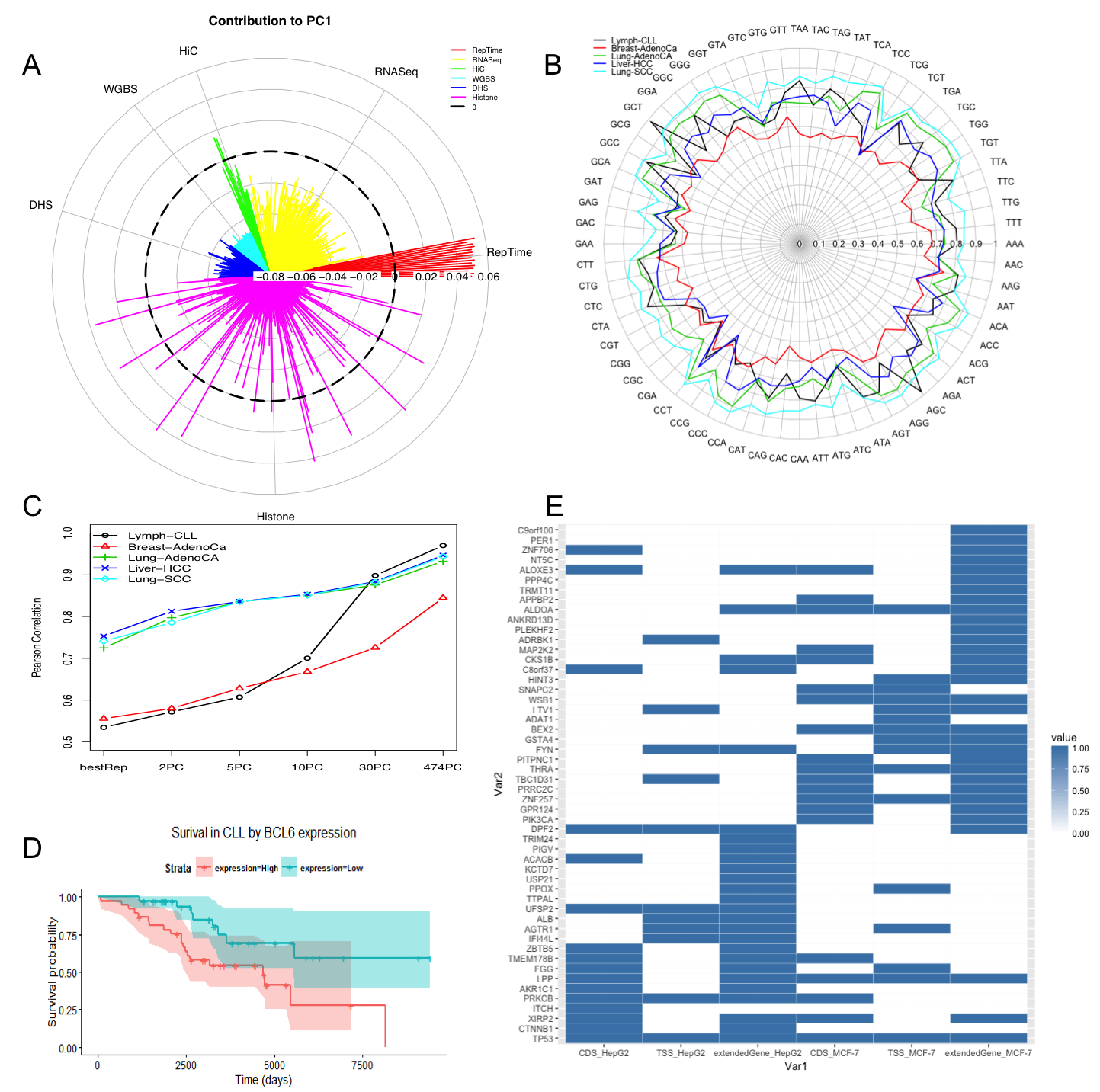
Whole genome seq

SVs

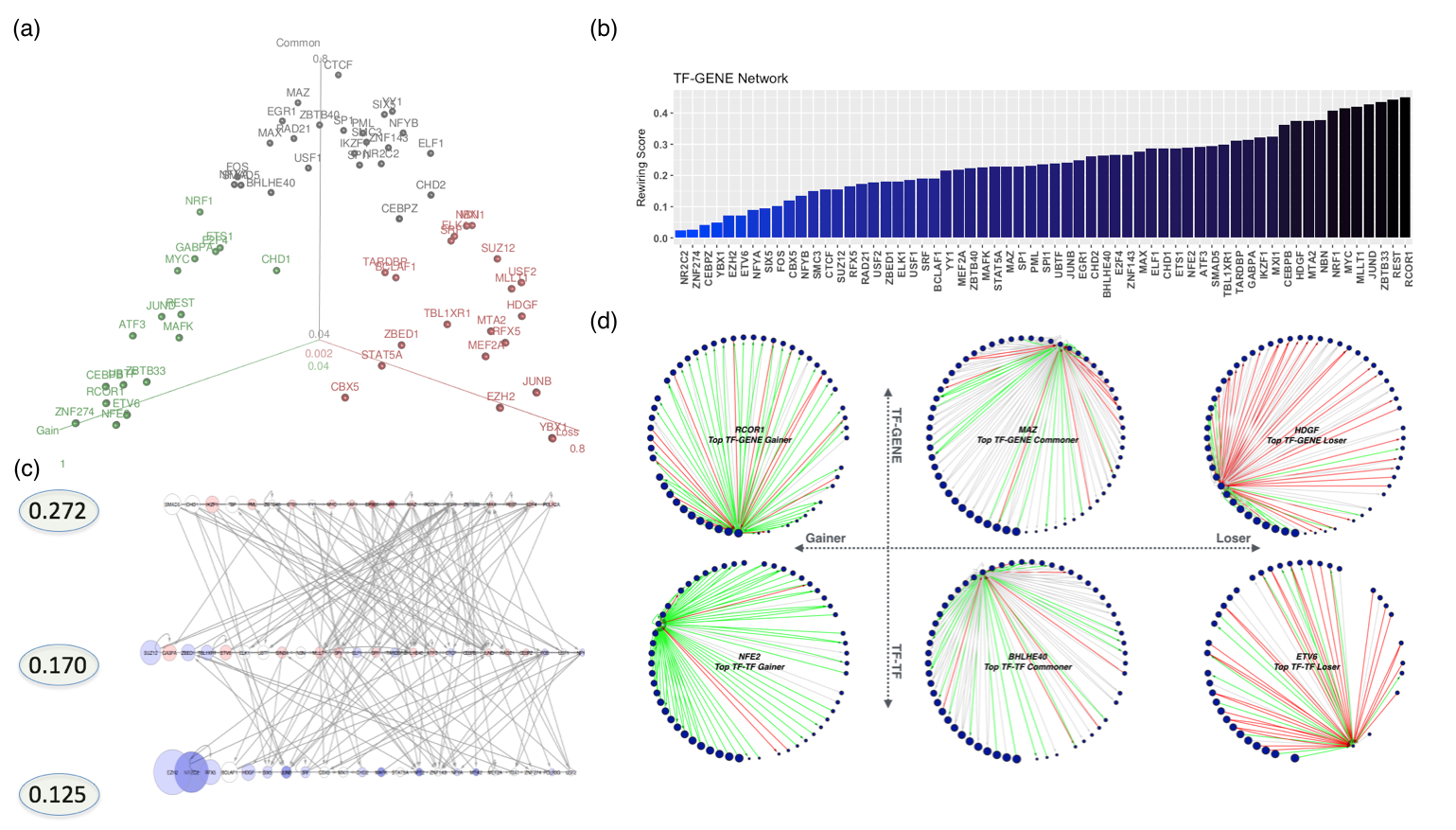
Maybe the real numbers instead of shading of color

Tumor and normal pair add a common row

FigM 2 Somatic Recurrence Analysis



FigM 3 v1

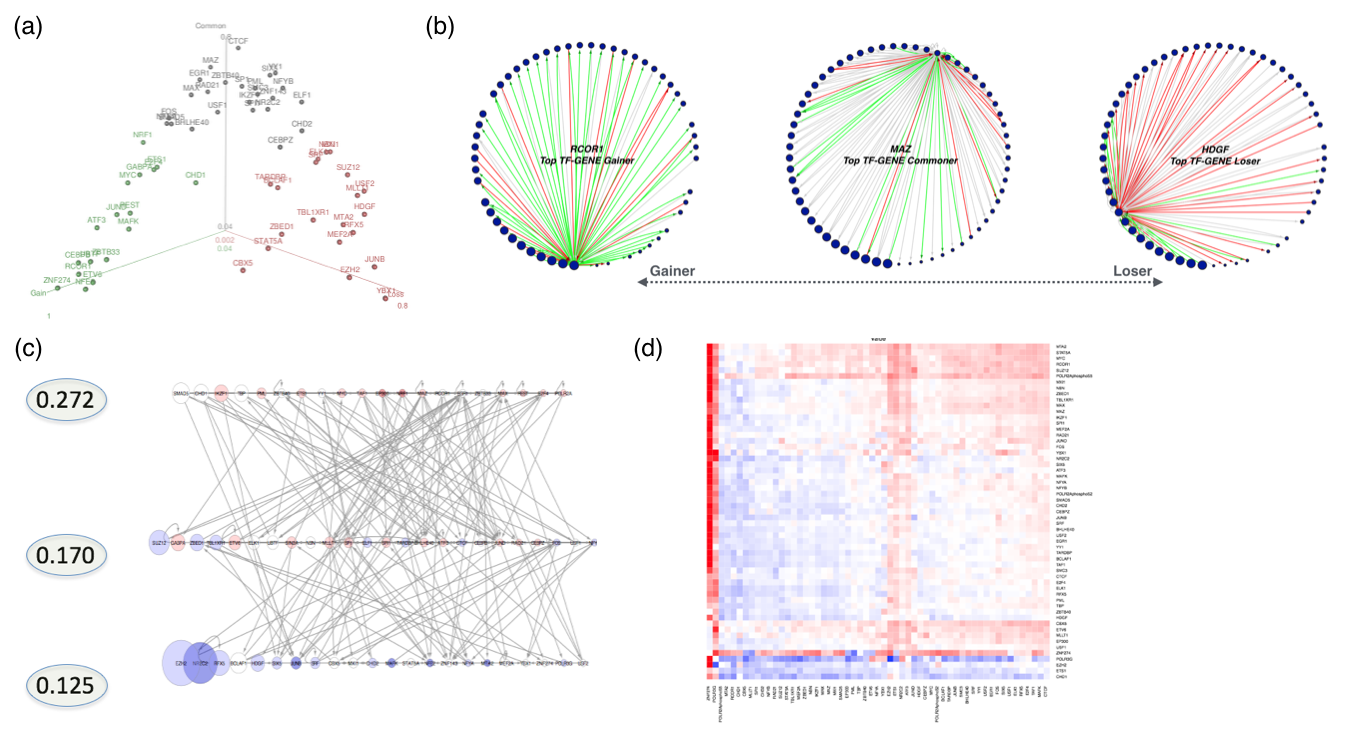


Part A - unsure - discuss & think

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



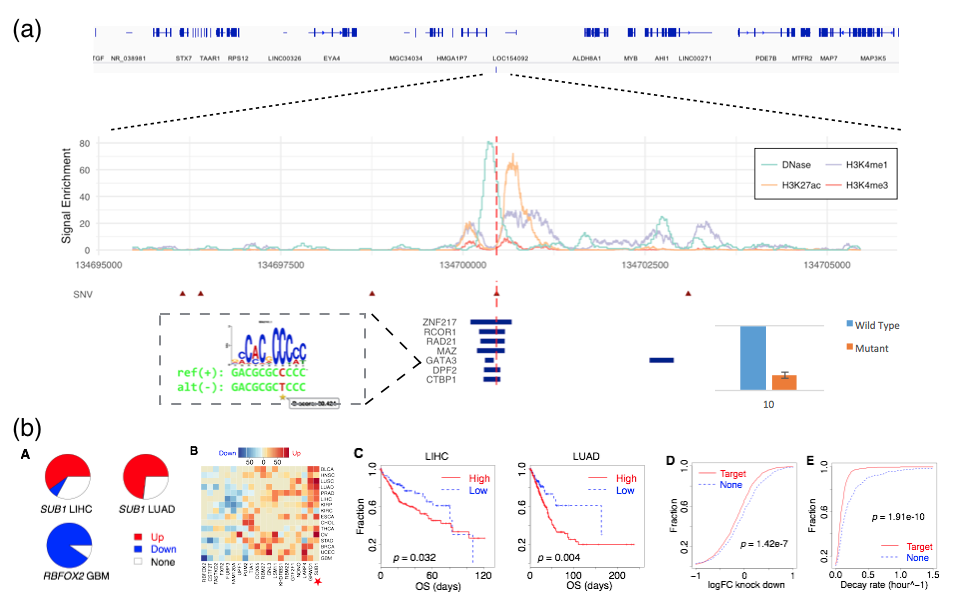
FigM 4 v1



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

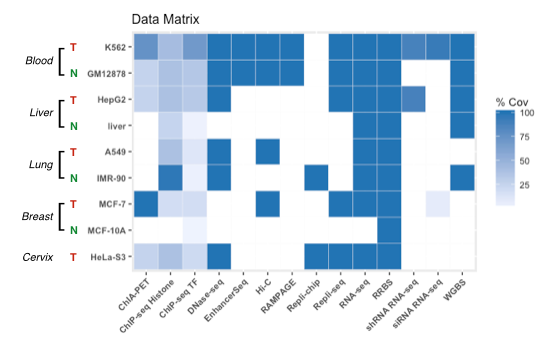
FigM 5. Cis Regulatory Element Validation (Hi-C result and enhancer target linkage to be included)

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

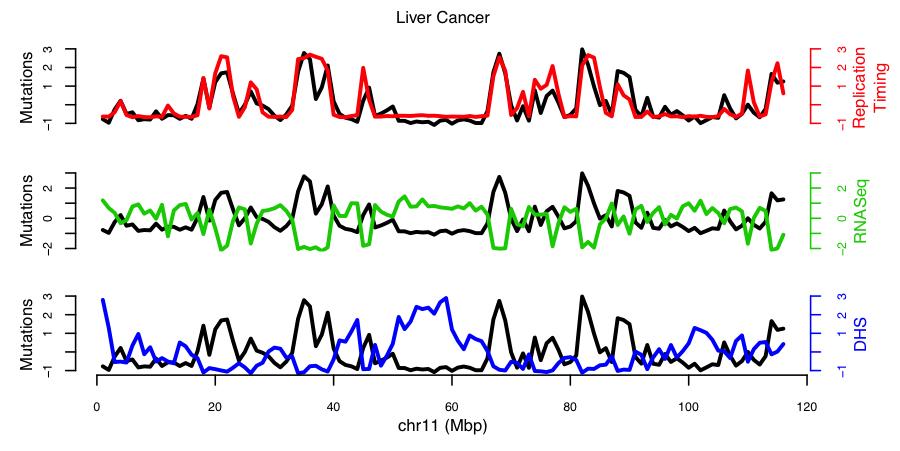


# Supplementary Figure section

FigureM 1.1 encode data matrix



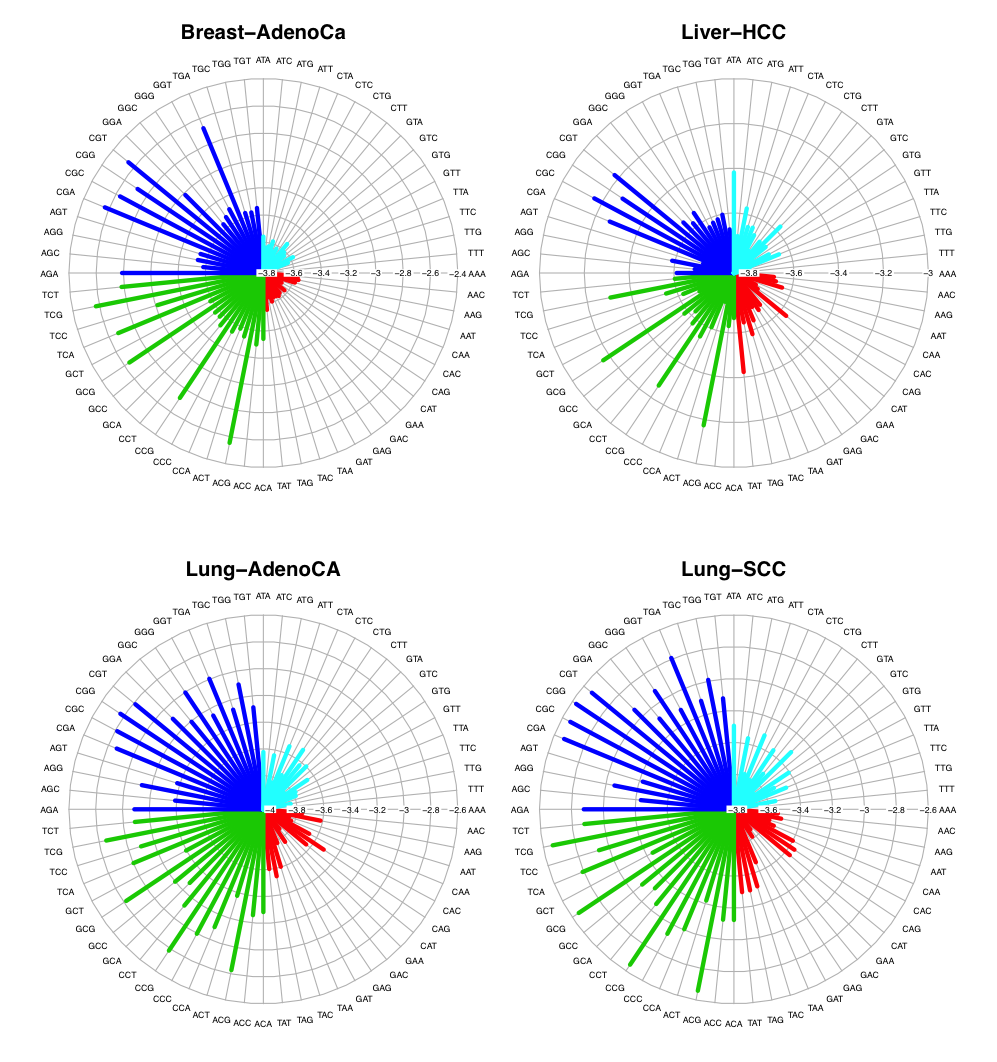
FigureS 2.1 correlation of mutation rates with some genomic features



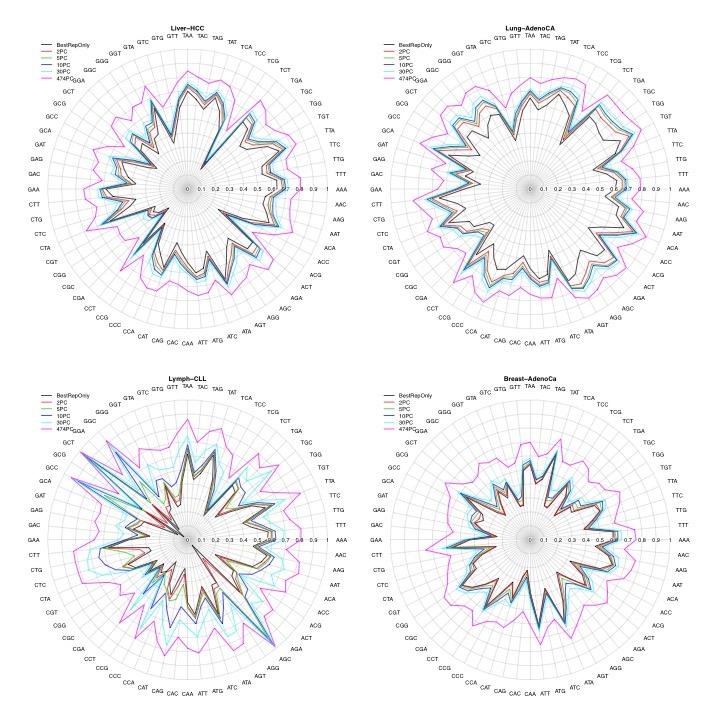
FigureS 2.2 Mutational Heterogeneity across the genome and across different local context in various cancer types



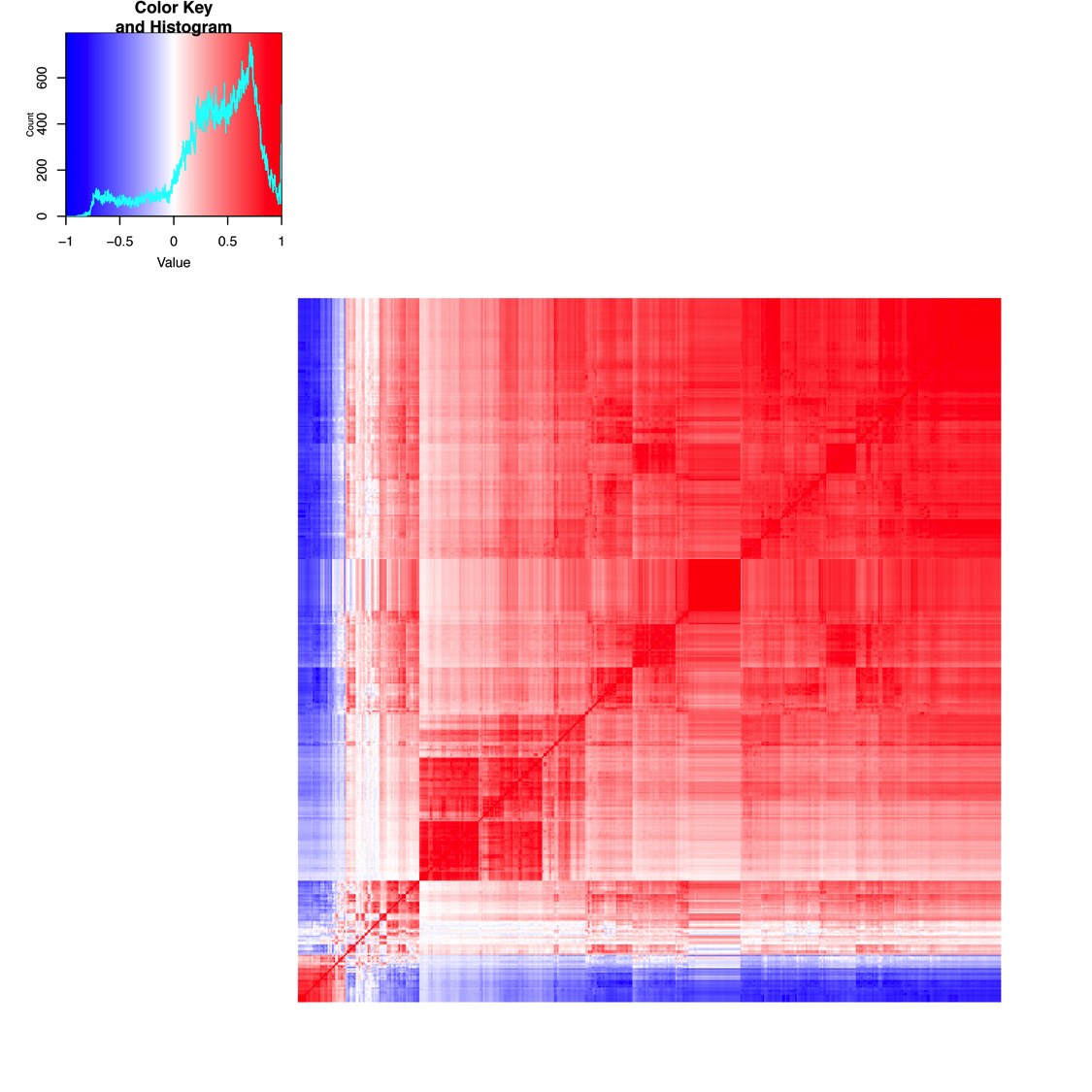
FigureS 2.3 Local context effect on mutation rate in various cancer types



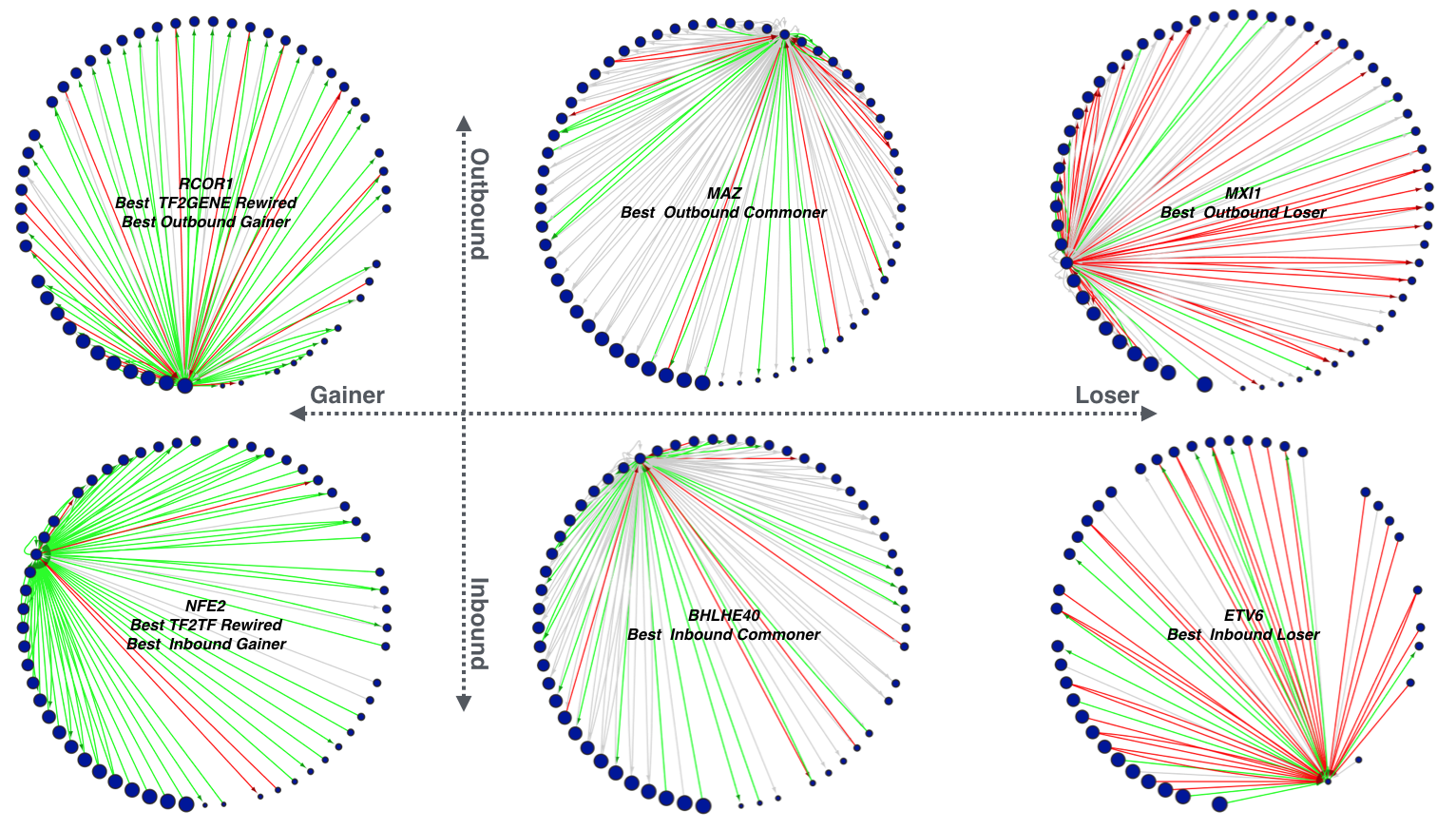
FigureS 2.4 improvement of BMR estimation using multiple features



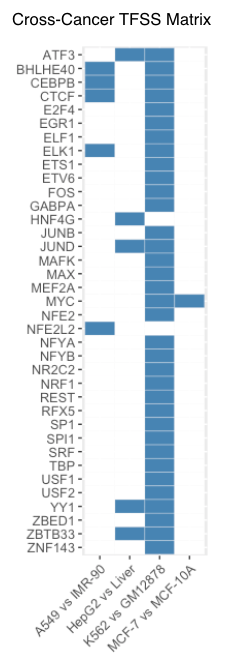
FigureS 2.5 correlation heatmap of ENCODE genomic features



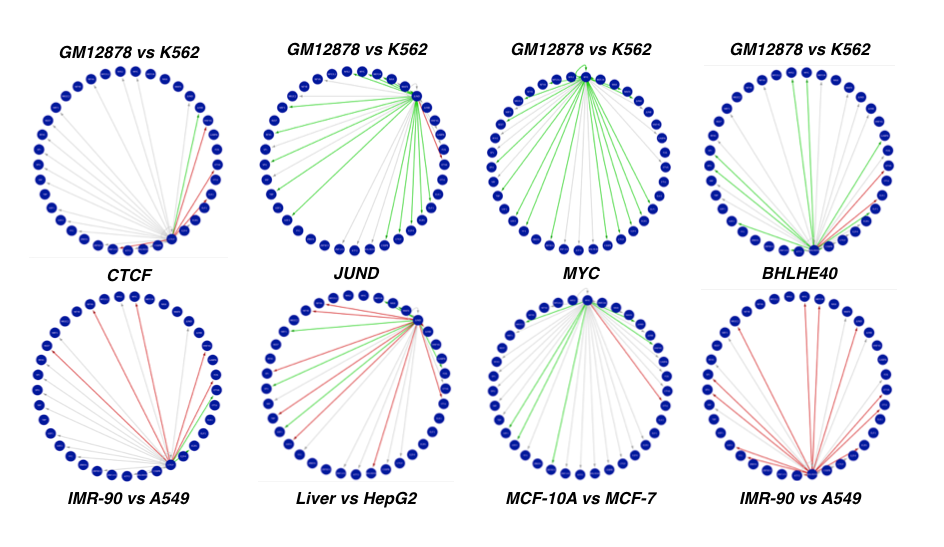
FigureS 3.x K562-GM12878 rewiring in TF-TF network



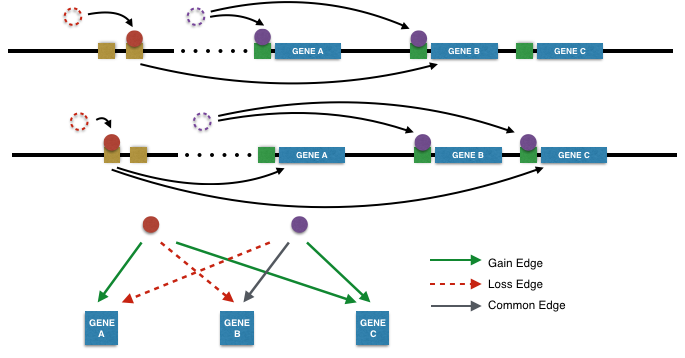
FigureS 3.x Cross cancer-normal TFSS data matrix



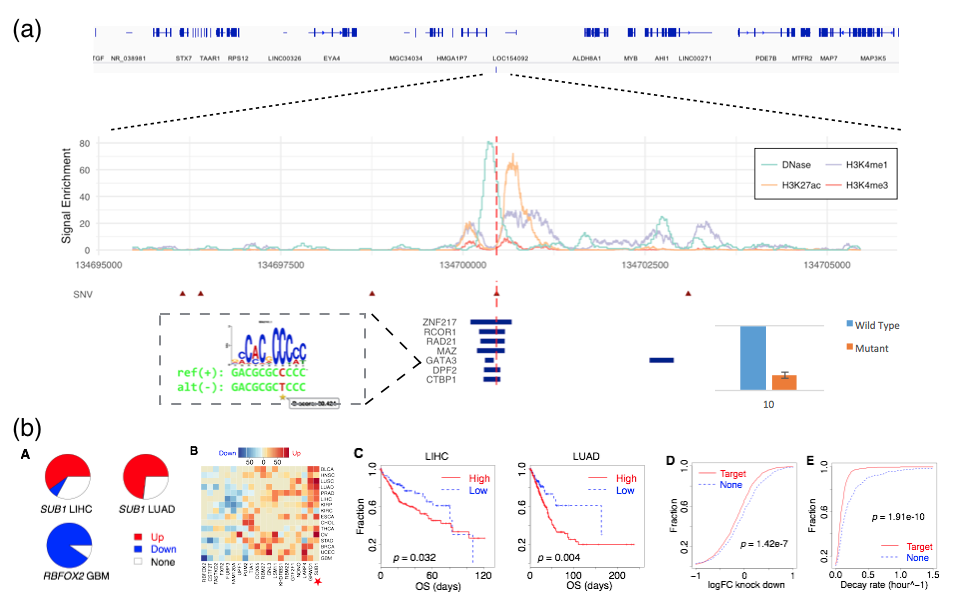
FigureS 3.x Cross cancer-normal comparison of TFSS network (Peng’s score to be included)



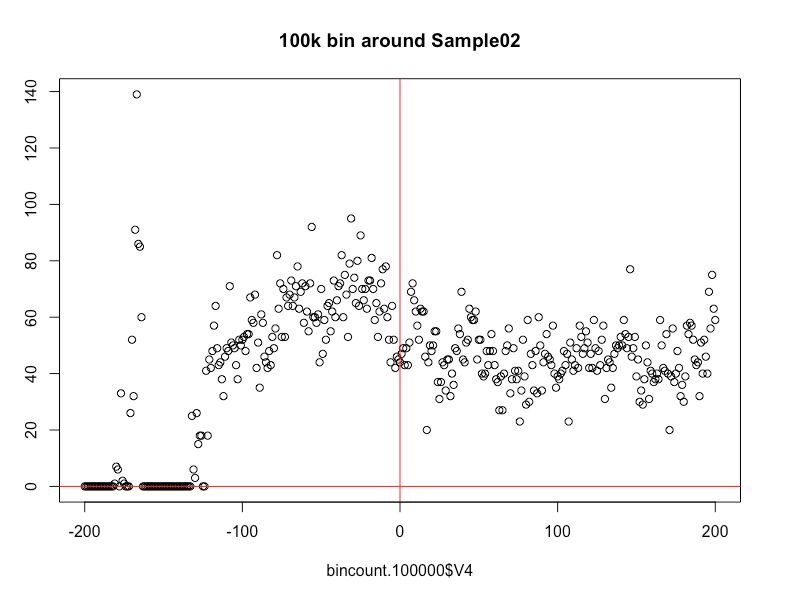
FigureS 3. Network schematic



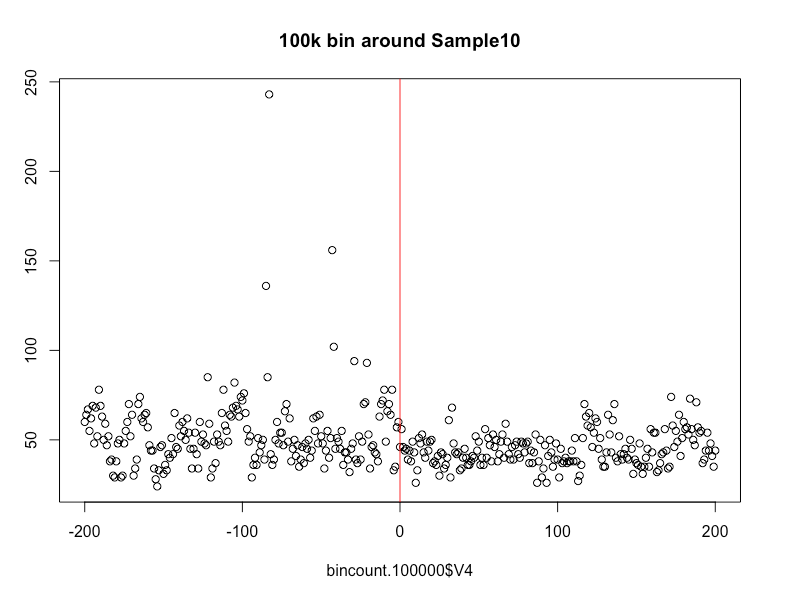
FigureM 5. Cis Regulatory Element Validation (Hi-C result and enhancer target linkage to be included)



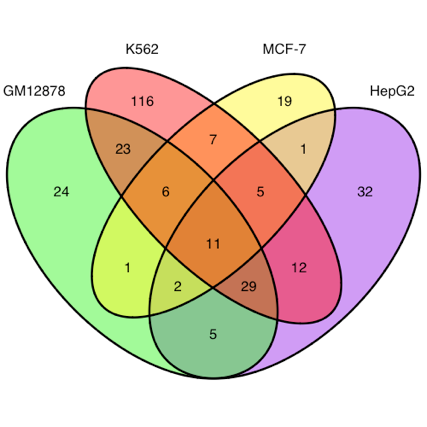
FigureS X. Mutation around validated SNV (Sample02)



FigureS X. Mutation around validated SNV (Sample10)



FigureS X. TF ChIP-seq experiments by cell lines



TableS X. Transcription factor classification

