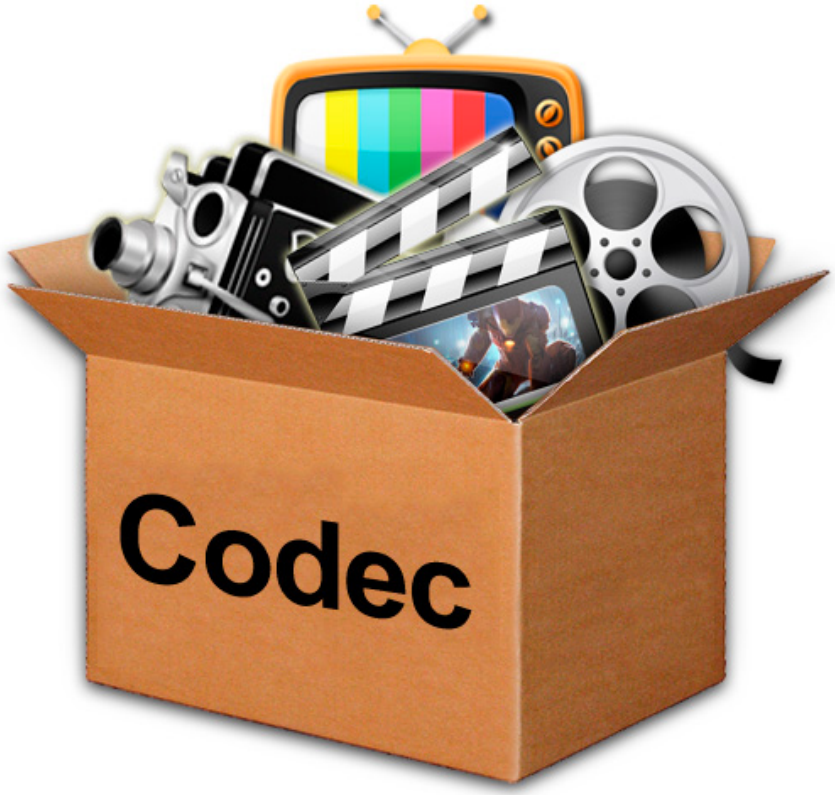
ENCodec: a companion encyclopedia in ENCODE for cancer research



Jing Zhang

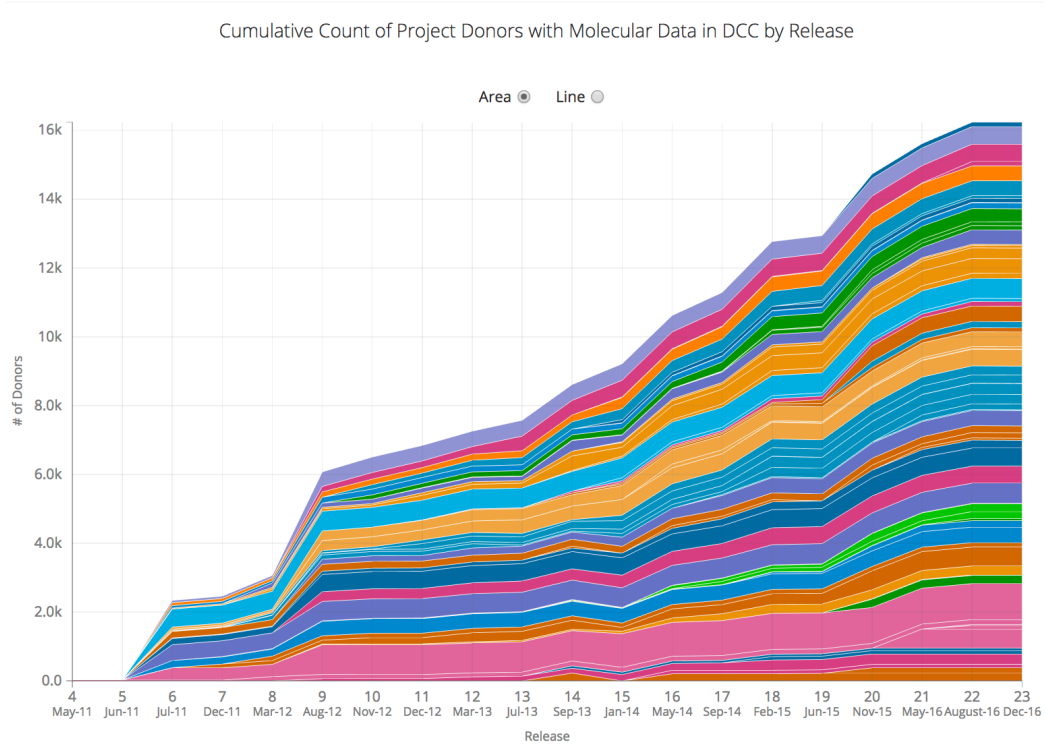
Gerstein Lab

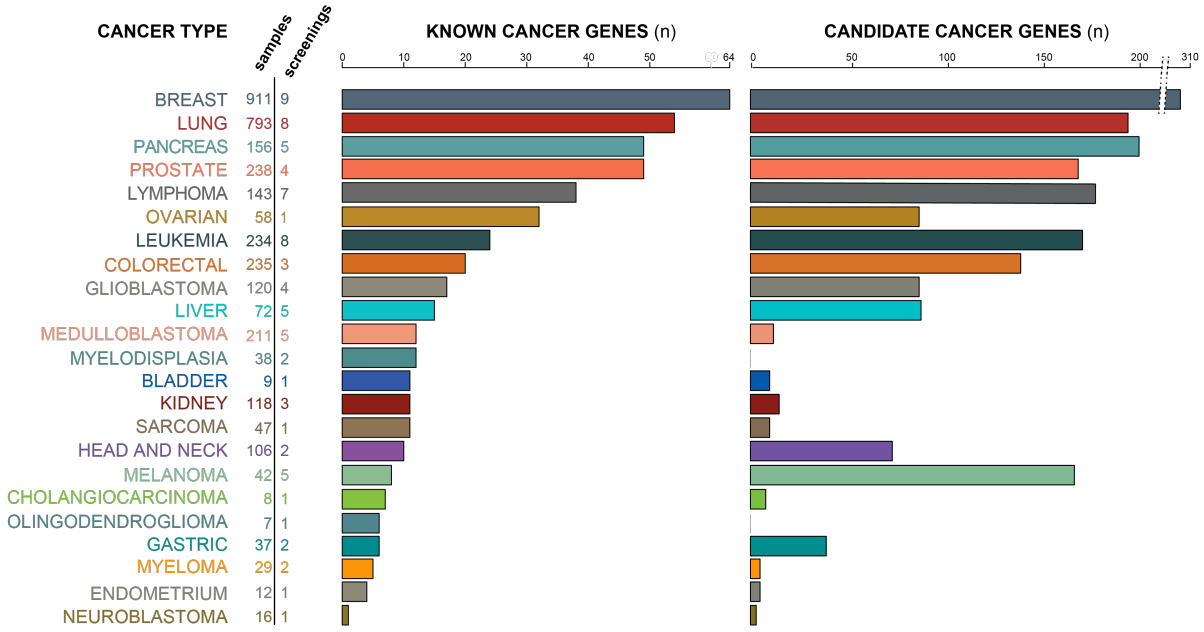
MB&B, Yale

Outline

1. Introduction to EN-codec resource
2. Multi-level integration from ENCODE benefits mutation burden analysis
    - Raw signal level integration
    - Annotation level integration
3. Interpreting transcriptional level regulatory changes through network rewiring analysis
    - Quantification of regulatory changes
    - Effect of highly rewired TFs in cancer
4. Integrating regulatory networks with tumor expression profiles identifies key regulators in cancer
    - Identification of key regulators that driver T/N differential expression
    - Investigating the cooperation pattern between key regulators
5. Variant prioritization scheme and small scale validations

Efforts to interpret cancer genomes





Tens of thousands of patient data released

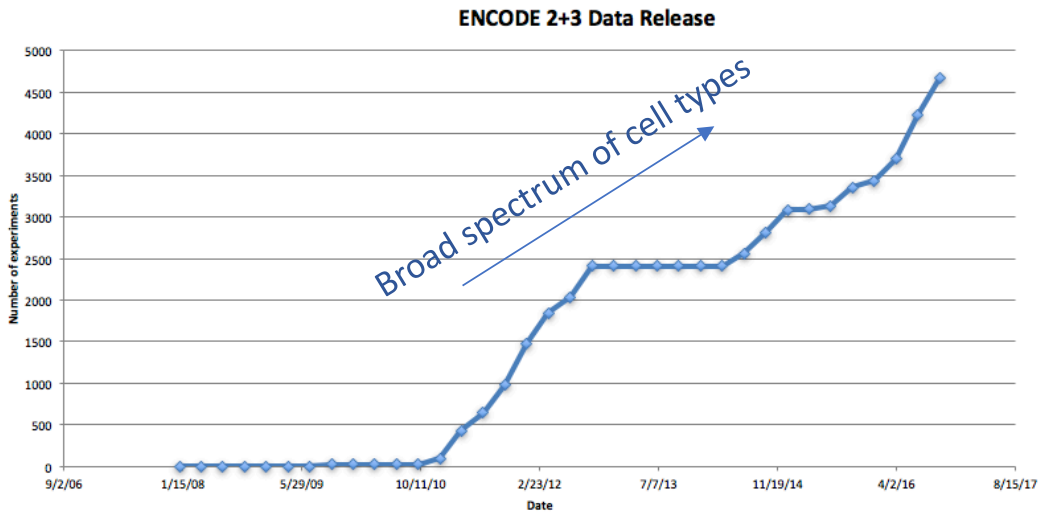
- Thousands of whole genome sequencing
- Millions of mutations
- Tens of thousands of tumor and/or normal expression profiling
- Tens of thousands of patient survival data
- …

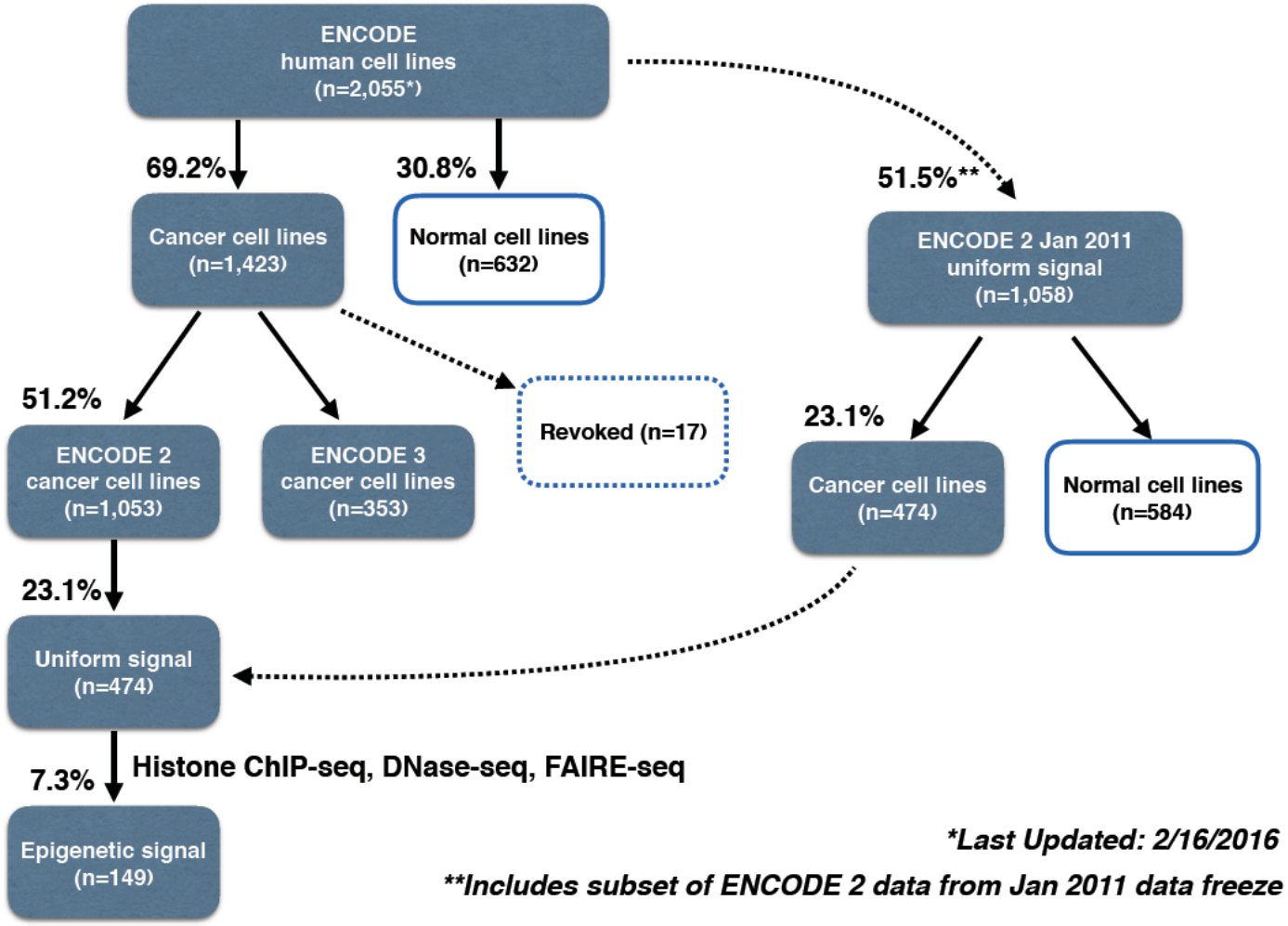
How to interpret the "dark" part of cancer genomes?

Our understanding of the genome

- Focused on hundreds of cancer associated genes
- Annotation of 20k genes, 1-2% of genome
- Limited knowledge of non-coding regulatory regions
- …

The majority of ENCODE cell lines are cancerous



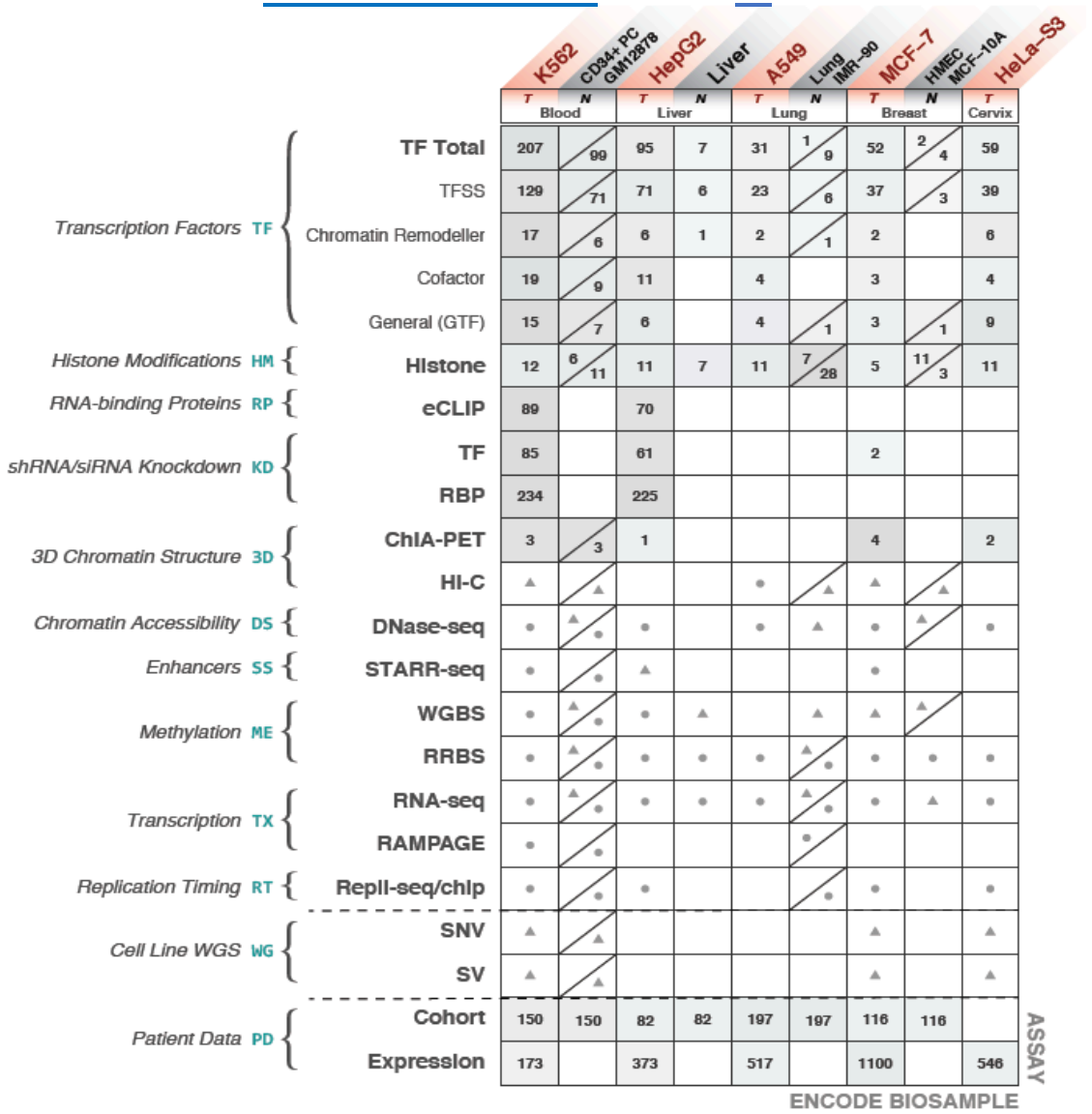


- Dataset almost doubled in ENCODE3 as compared with ENCODE2
- ~69% of cell types in ENCODE are cancerous
- Several top tier cell lines are enriched with various types of assays
- Many new experimental assays in top tier cell lines to help genome functional characterization

En-Codec: a companion encyclopedia in ENCODE for cancer research

Data enriched cell types

Cell_Type	Biosample type	Total assays
K562	immortalized cell line	680
HepG2	immortalized cell line	356
GM12878	immortalized cell line	281
K562 RNAi	immortalized cell line	276
HepG2 RNAi	immortalized cell line	230
MCF-7	immortalized cell line	162
HEK293	immortalized cell line	153
H1-hESC	stem cell	131
HeLa-S3	immortalized cell line	129
SK-N-SH	immortalized cell line	79
A549	immortalized cell line	78
endothelial cell of umbilical vein	primary cell	59
keratinocyte	primary cell	53
HCT116	immortalized cell line	49
IMR-90	primary cell	46
fibroblast of lung	primary cell	42
A549 ethanol	immortalized cell line	40
liver	tissue	37
foreskin fibroblast	primary cell	31
mammary epithelial cell	primary cell	31



En-Codec: Structure of the resource

Data Provision

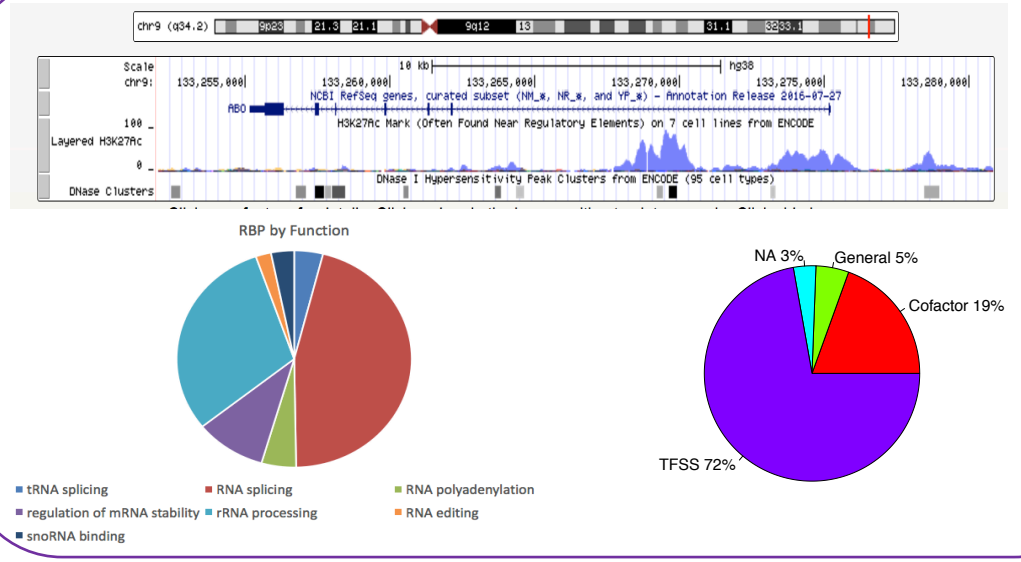
1. Deduplicated ENCODE Assays
   - TF ChIP-seq
   - Histone modification
2. Uniformly processed signal tracks
3. Variant calling (Liver validation set)
4. TCGA gene expression to ENCODE spec.

Annotation

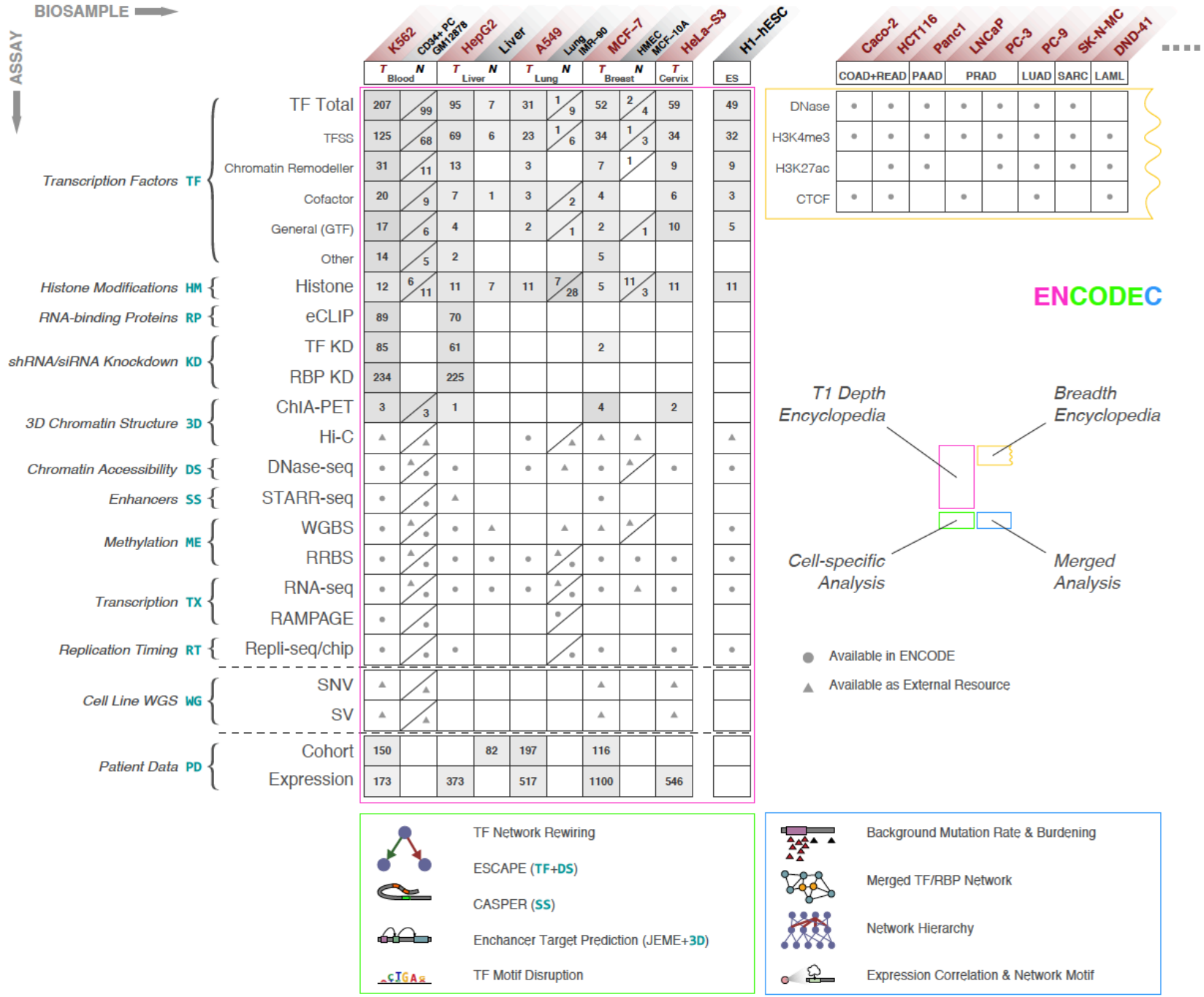
5. ENCODE Cancer Cell Line Summary
   - Tumor normal pairs
   - 4+1 case study subset
6. Key transcription factors by cell line

Processed Data

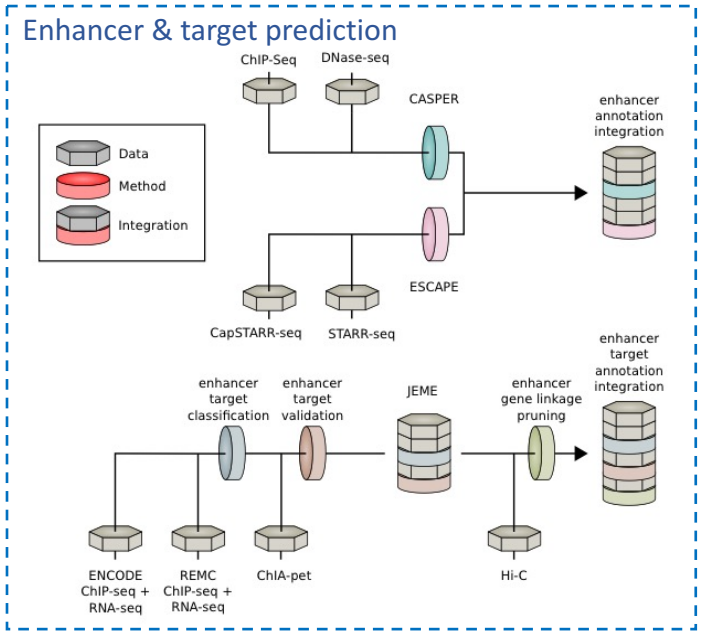
8. Cis- Response Elements
9. Accurate enhancer identification
   - ESCAPE (ChIP-Seq + DNase-seq)
   - CASPER (CapSTARR-seq + STARR-seq)
10. Accurate enchancer target prediction (JEME + HiC)
7. Background Mutation Rate + Burdening
11. Network analysis
    - Hierarchies
    - Rewiring
12. TF motif identification

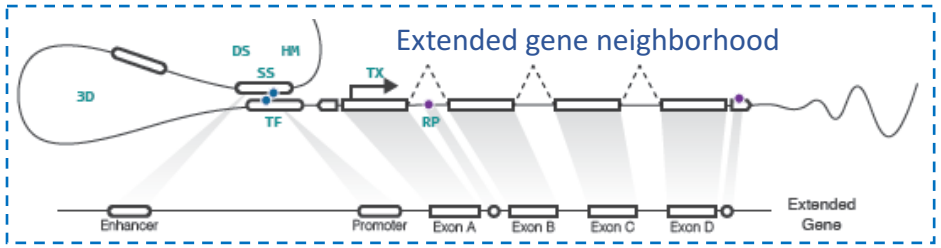


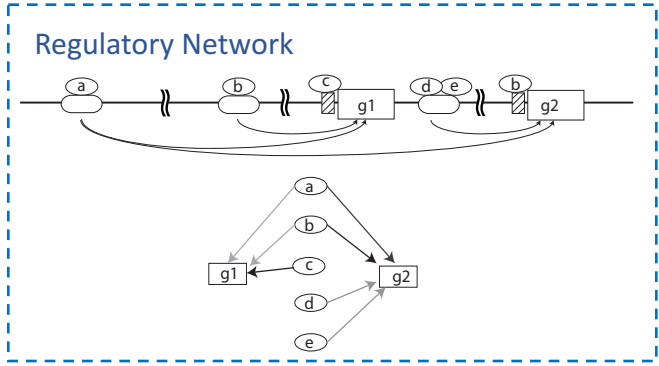
Cancer	Abbr.	Tumor	Normal
Breast	BRCA	MCF-7	MCF-10A
Liver	LIHC	HepG2	Liver
Lung	LUAD	A549	IMR-90
Blood	CML[CLL/AML]	K562	GM12878
Cervix	CESC	HeLa-S3	NA

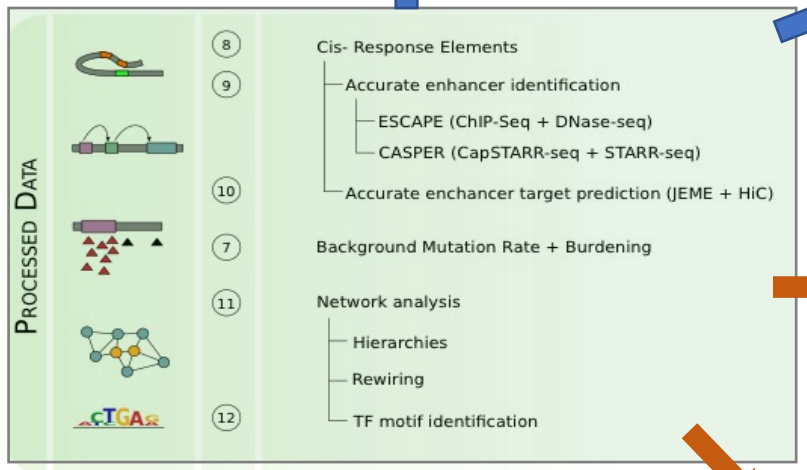

BIOSAMPLE
ASSAY
K562
CD34+ PC GM12878
HepG2
Liver
A549
Lung IMR-90
MCF-7
HMEC MCF-10A
HeLa-S3
H1-hESC
Caco-2
HCT116
Panc1
LNCaP
PC-3
PC-9
SK-N-MC
DND-41
T N Blood
T N Liver
T N Lung
T N Breast
T Cervix
ES
COAD+READ
PAAD
PRAD
LUAD
SARC
LAML
Transcription Factors TF
TF Total
TFSS
Chromatin Remodeller
Cofactor
General (GTF)
Other
Histone Modifications HM
Histone
RNA-binding Proteins RP
eCLIP
shRNA/siRNA Knockdown KD
TF KD
RBP KD
3D Chromatin Structure 3D
ChIA-PET
Hi-C
Chromatin Accessibility DS
DNase-seq
Enhancers SS
STARR-seq
Methylation ME
WGBS
RRBS
Transcription TX
RNA-seq
RAMPAGE
Replication Timing RT
Repli-seq/chip
Cell Line WGS WG
SNV
SV
Patient Data PD
Cohort
Expression
DNase
H3K4me3
H3K27ac
CTCF
ENCODEC
T1 Depth Encyclopedia
Breadth Encyclopedia
Cell-specific Analysis
Merged Analysis
Available in ENCODE
Available as External Resource
TF Network Rewiring
ESCAPE (TF+DS)
CASPER (SS)
Enchancer Target Prediction (JEME+3D)
TF Motif Disruption
Background Mutation Rate & Burdening
Merged TF/RBP Network
Network Hierarchy
Expression Correlation & Network Motif

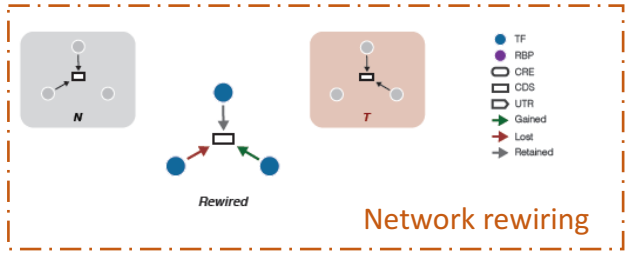
En-Codec: a cancer encyclopedia resource from ENCODE

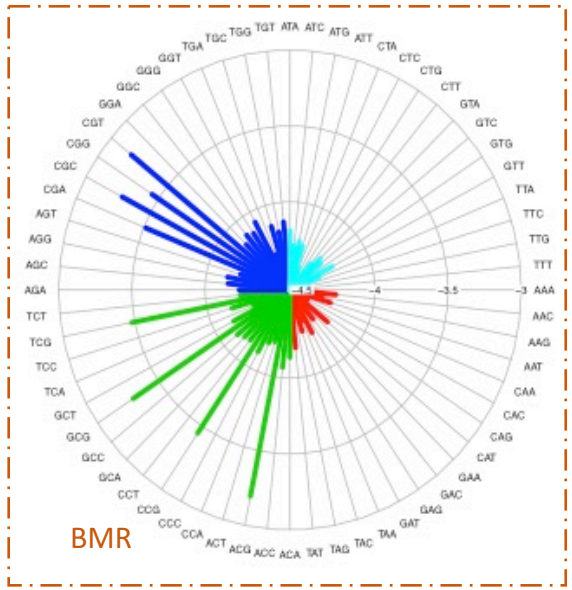


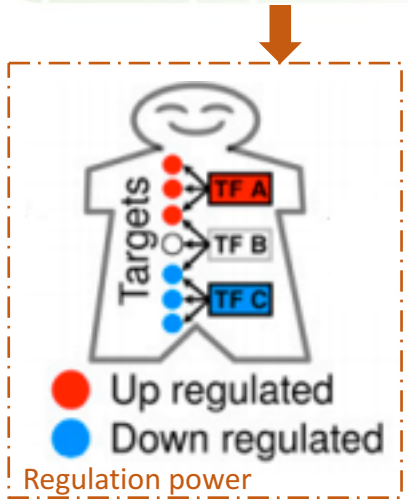


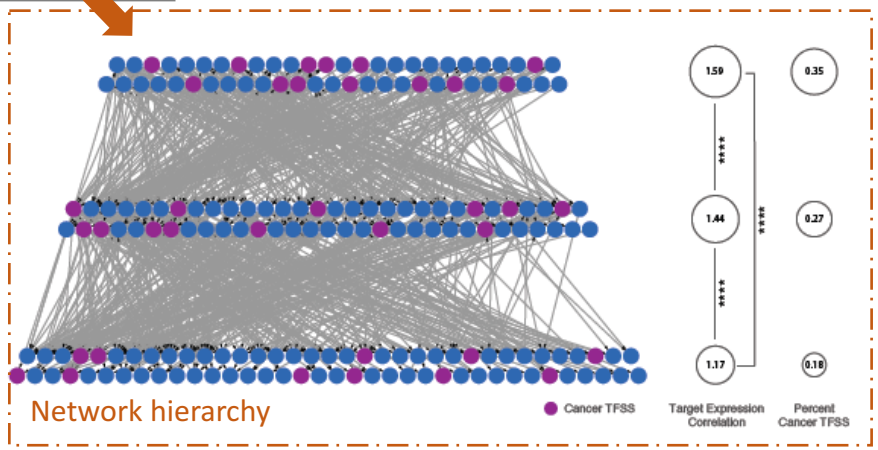




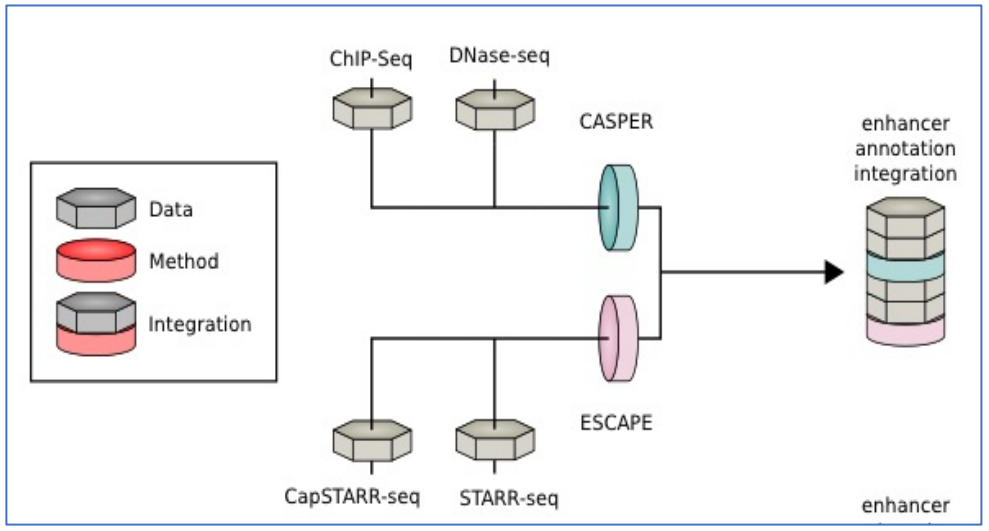


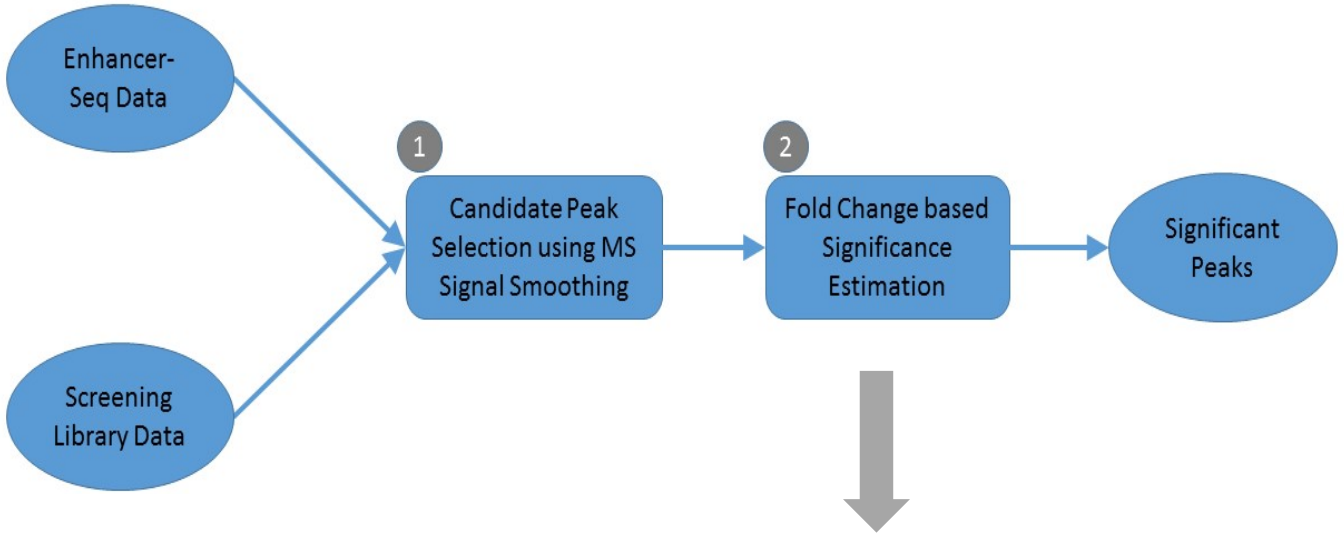


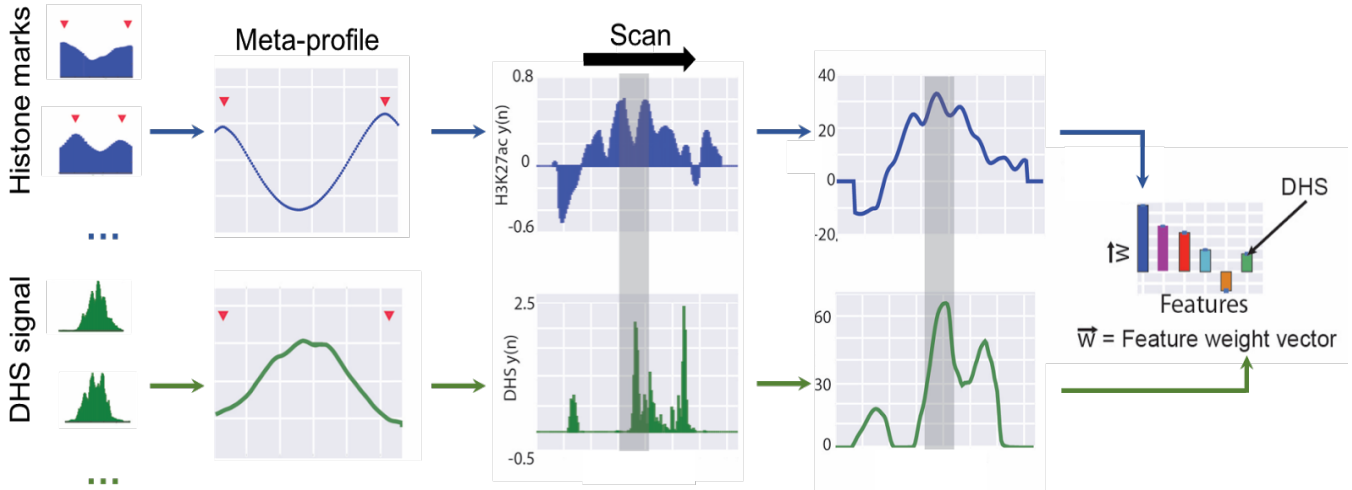




ESCAPE+CASPER: integrative approach for enhancer prediction







intersection

High confidence enhancers in different cell types

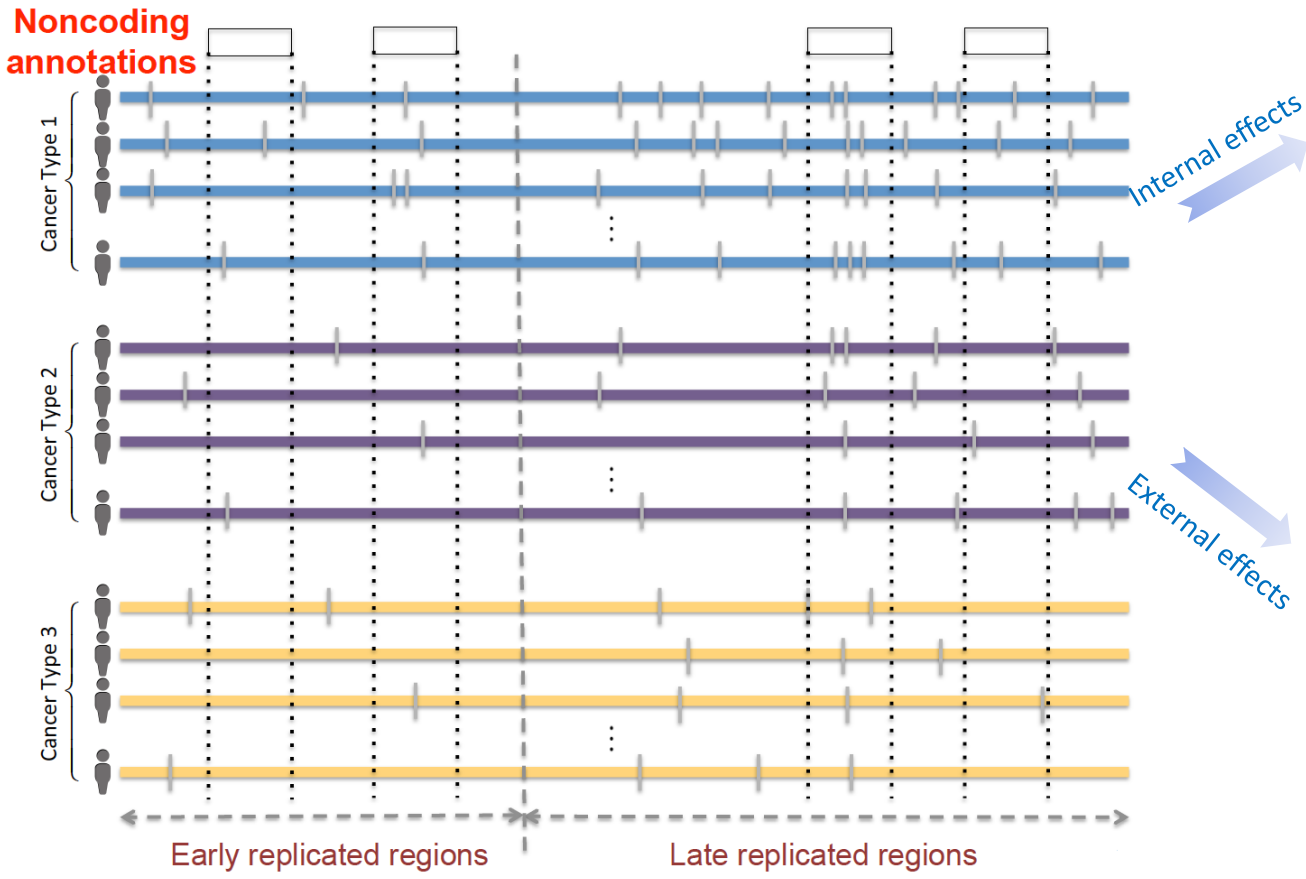
Outline

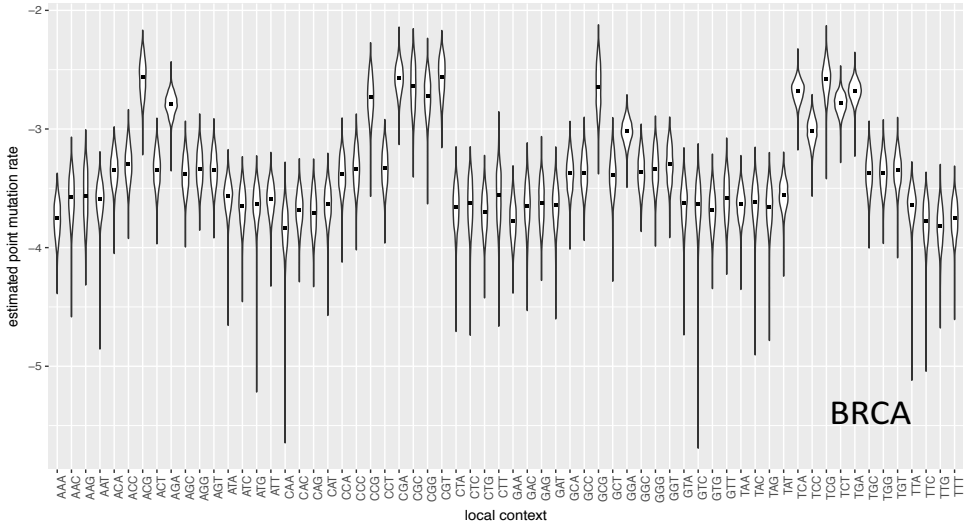
1. Introduction to EN-codec resource
2. Multi-level integration from ENCODE benefits mutation burden analysis
    - Raw signal level integration
    - Annotation level integration
3. Interpreting transcriptional level regulatory changes through network rewiring analysis
    - Quantification of regulatory changes
    - Effect of highly rewired TFs in cancer
4. Integrating regulatory networks with tumor expression profiles identifies key regulators in cancer
    - Identification of key regulators that driver T/N differential expression
    - Investigating the cooperation pattern between key regulators
5. Variant prioritization scheme and small scale validations

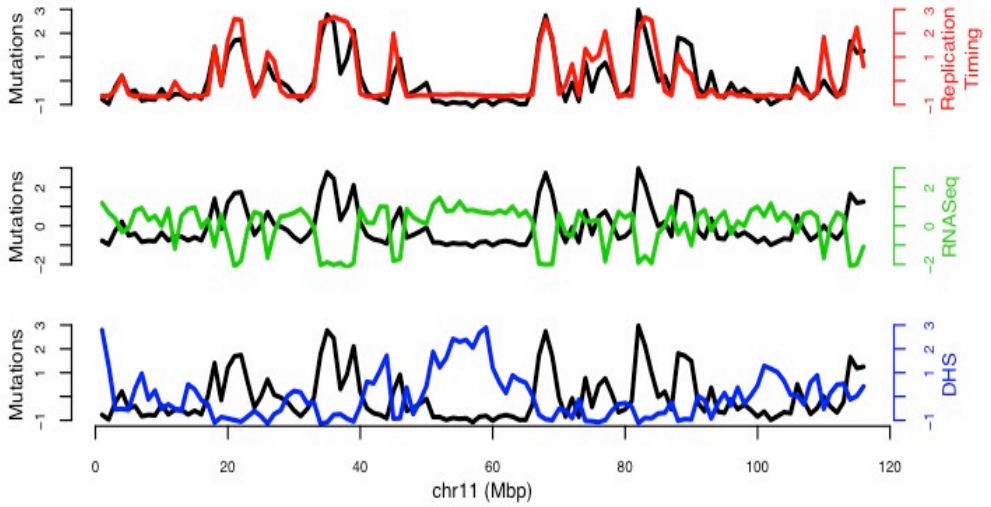
Multi-level data integration better enables recurrent variant analysis

How to evaluate the somatic mutation burden in cancer?

- Background mutation rate (BMR) modeling
- mutational heterogeneity control

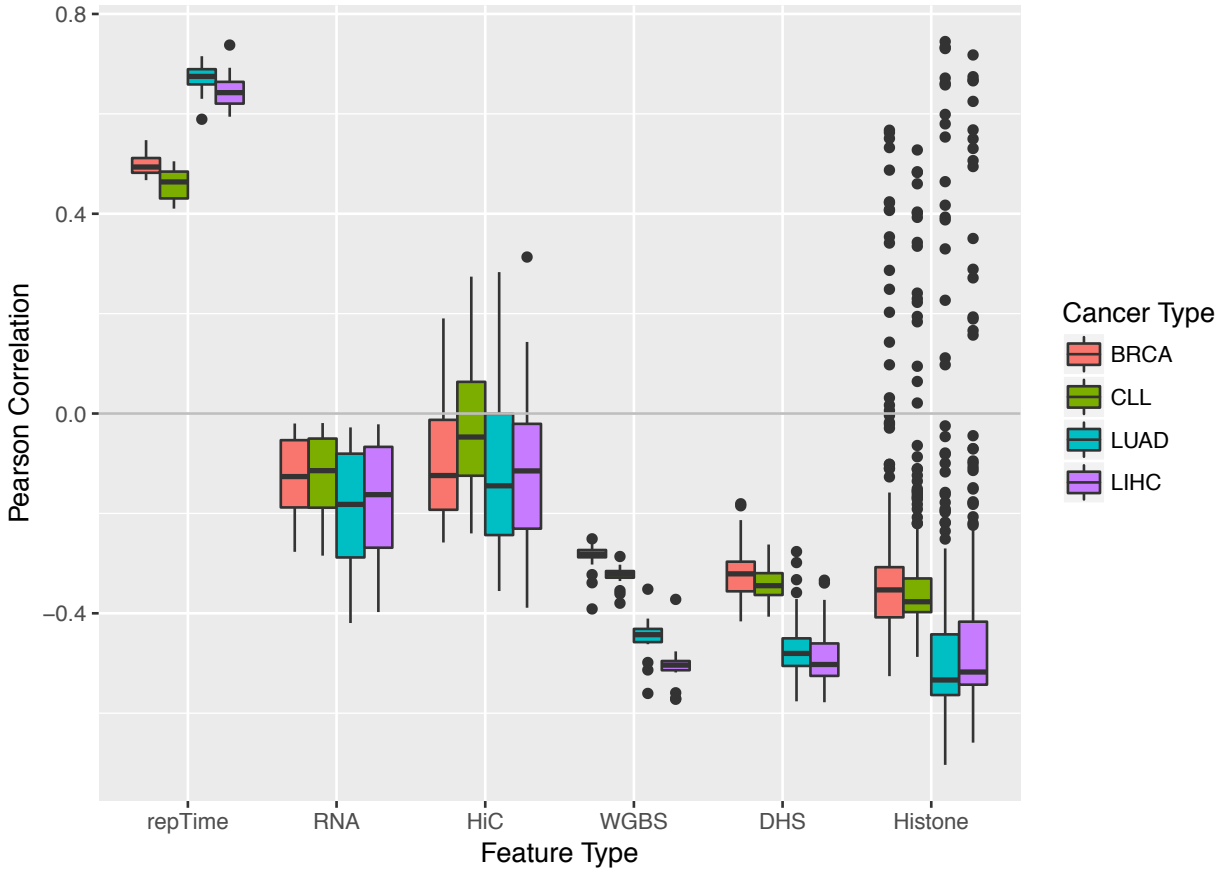


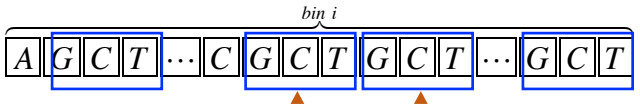




Raw signal level integration enables accurate BMR modeling

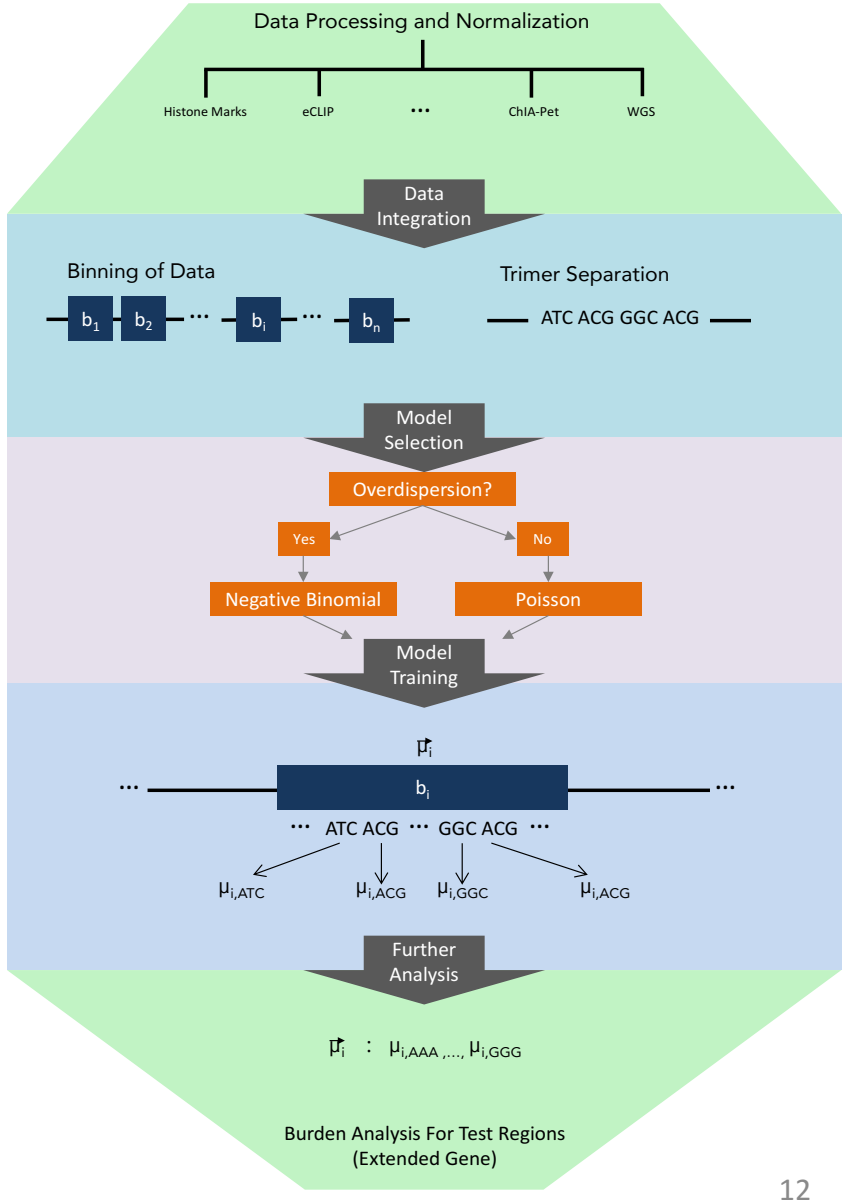
Many features are correlated with BMR

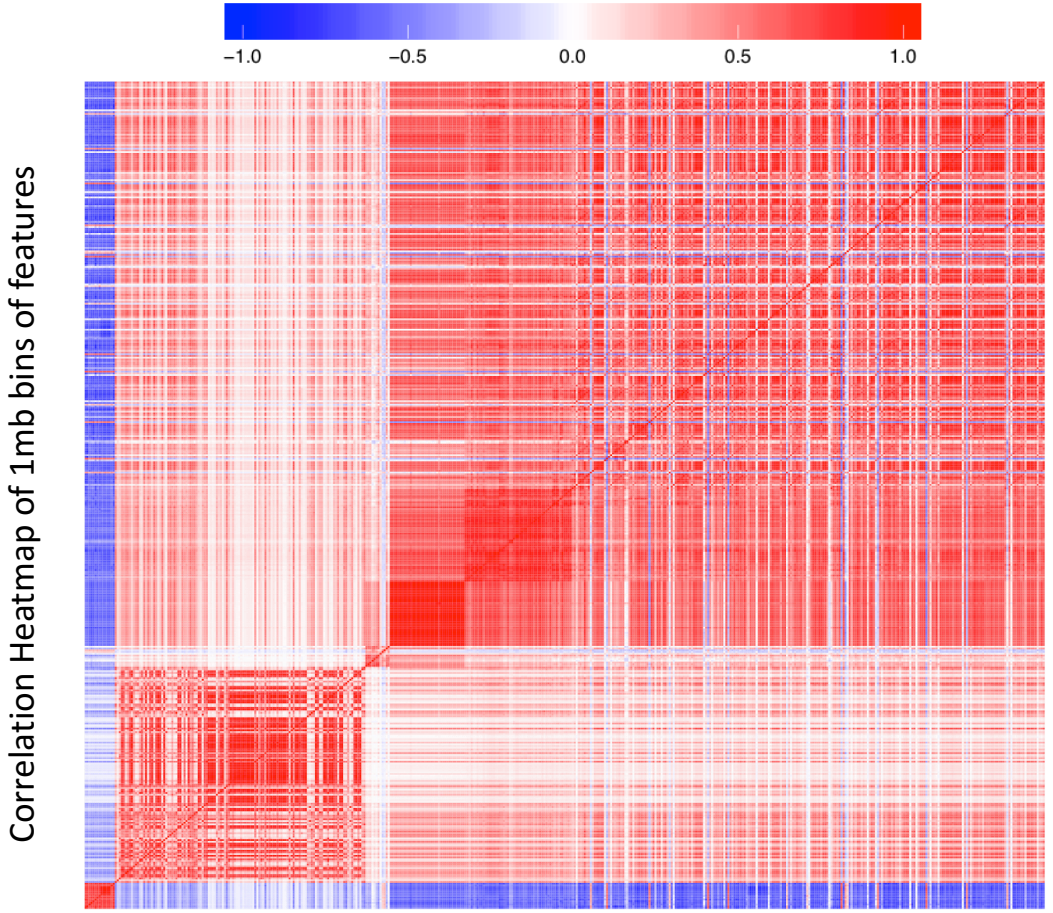


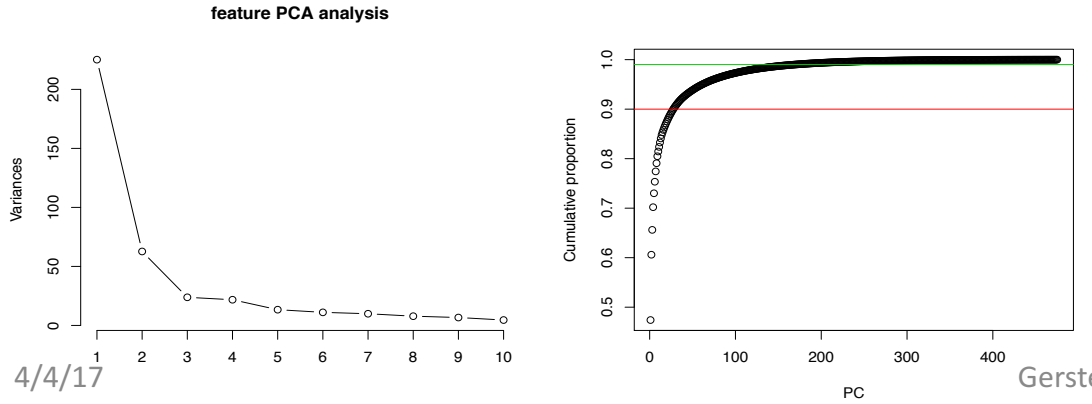


	Bin	All	Mutated
GCT	1Mb	3000	15

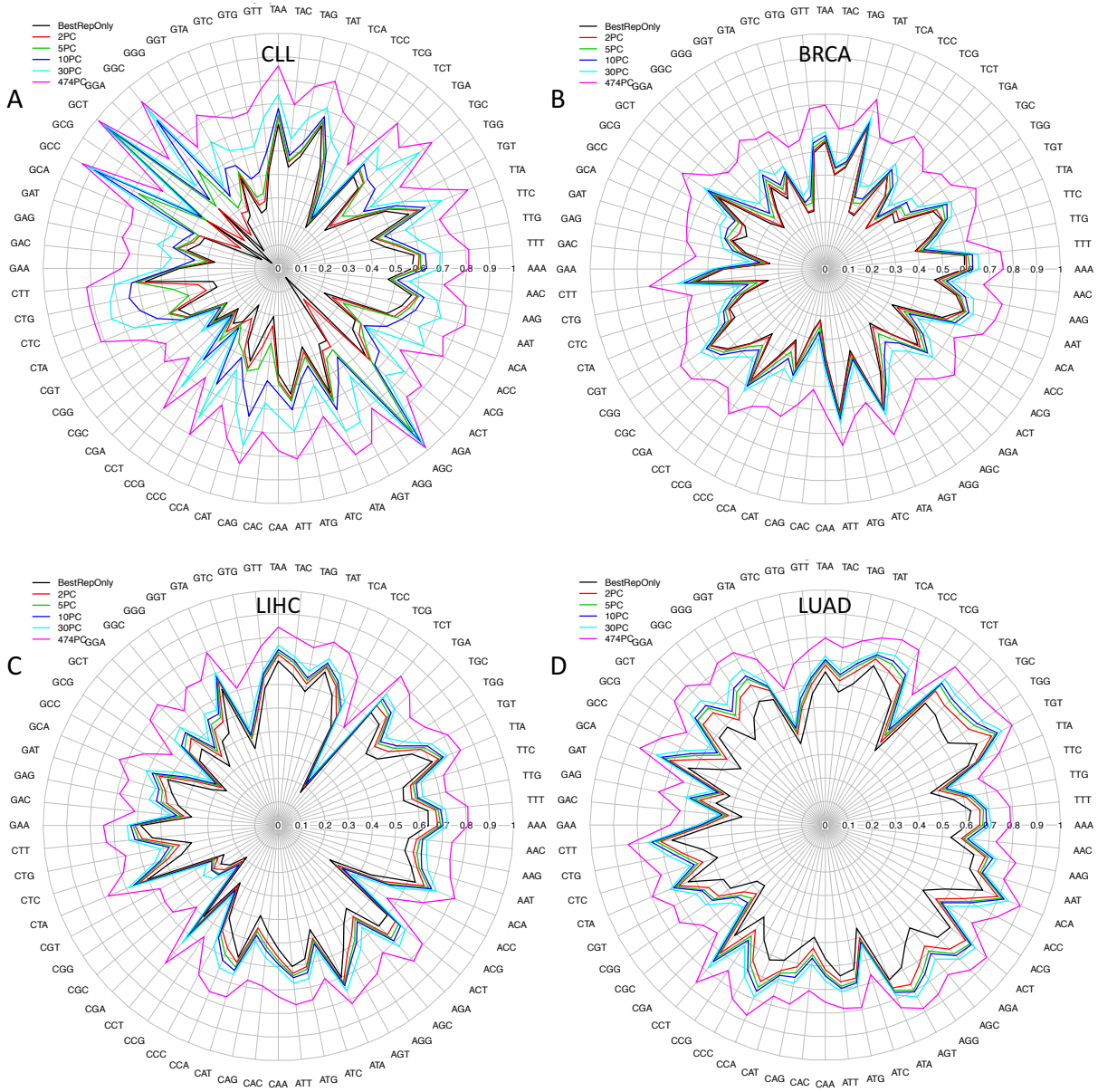
$$\log\left(\mu^{i}_{GCT}\right) = \beta_{0,GCT} + \beta_{1,GCT}x_{i,1} + \cdots + \beta_{j,GCT}x_{i,j} + \cdots + \beta_{J,GCT}x_{i,J}$$

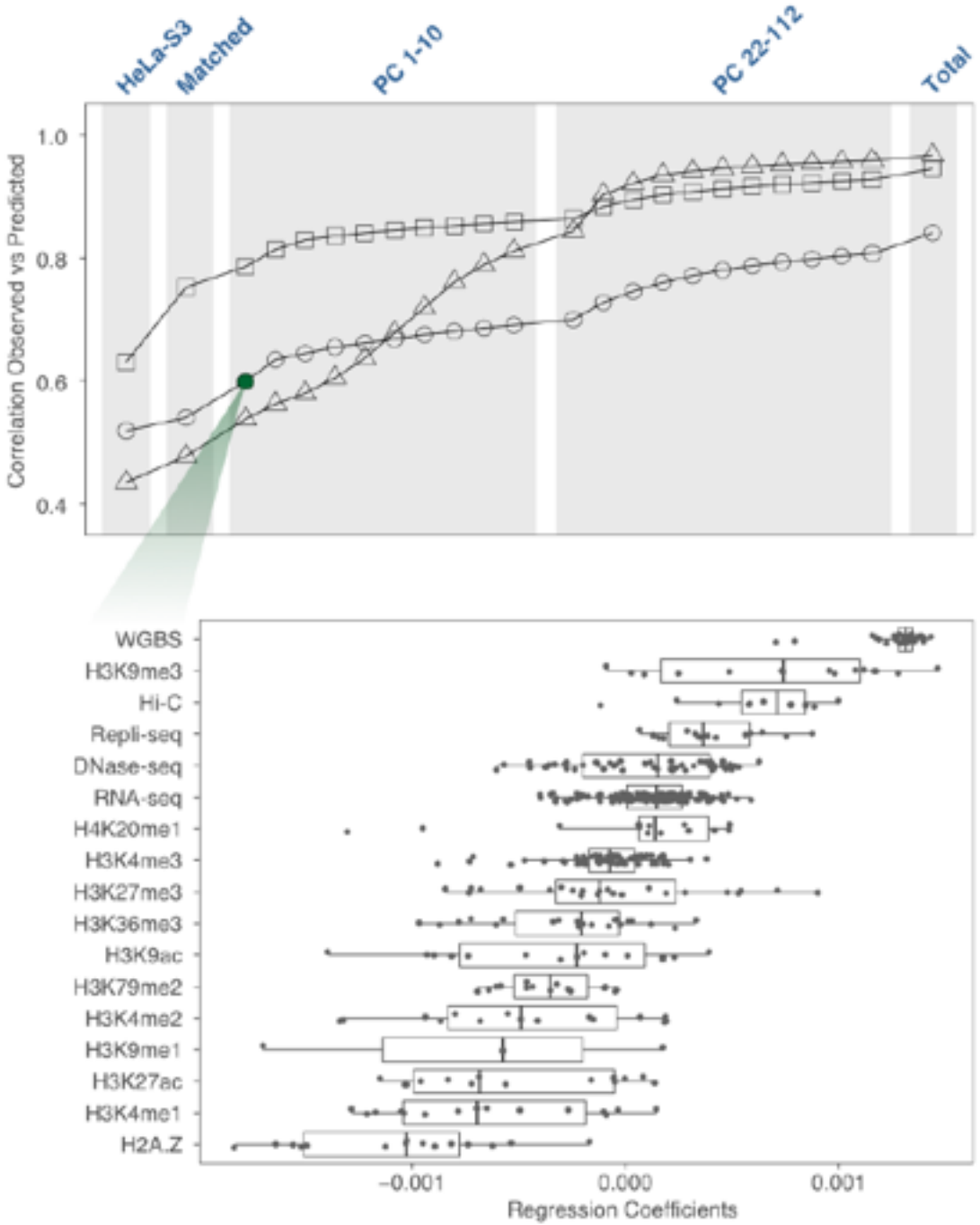




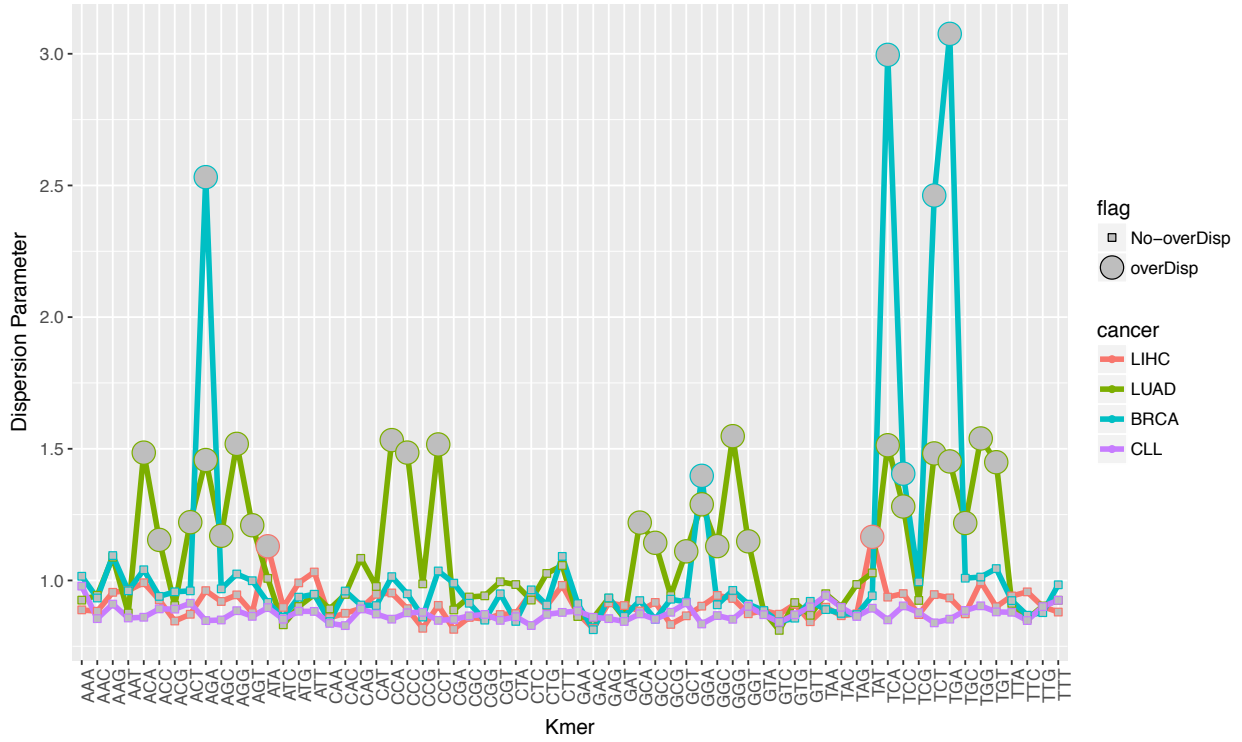


Performance in four different cancer types

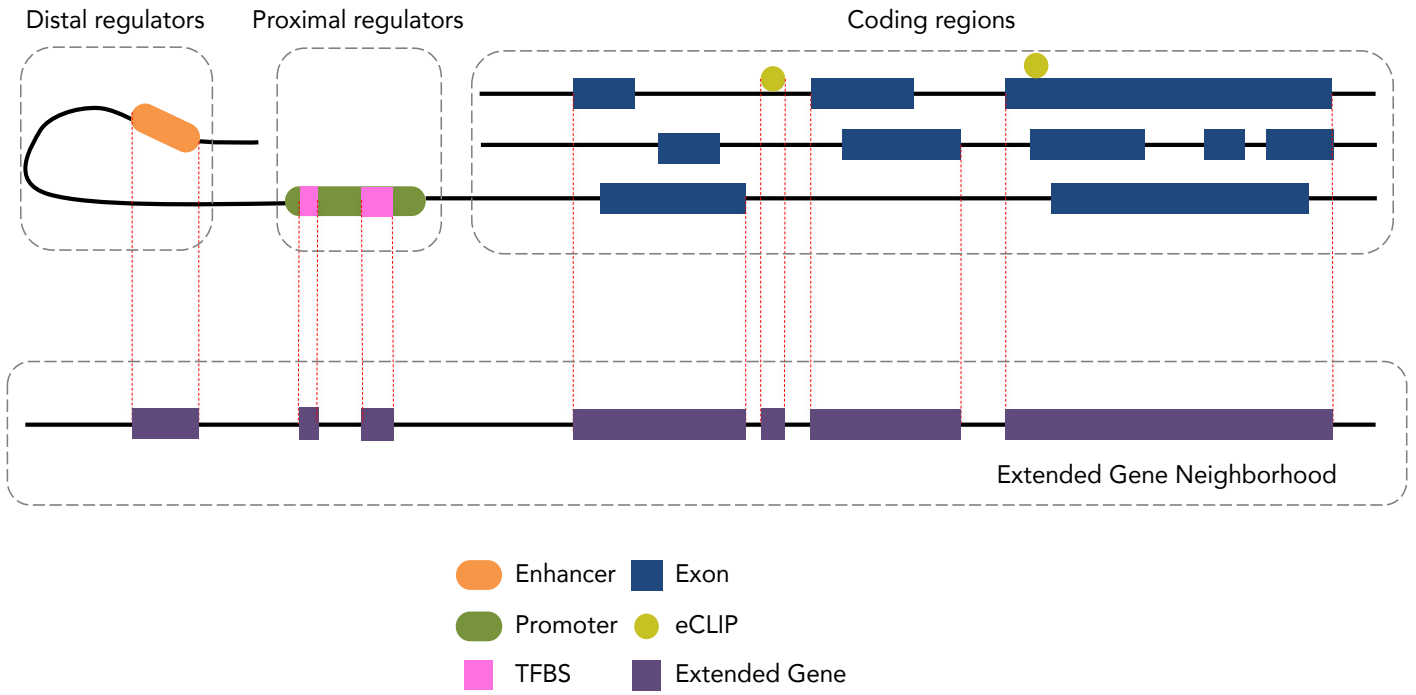


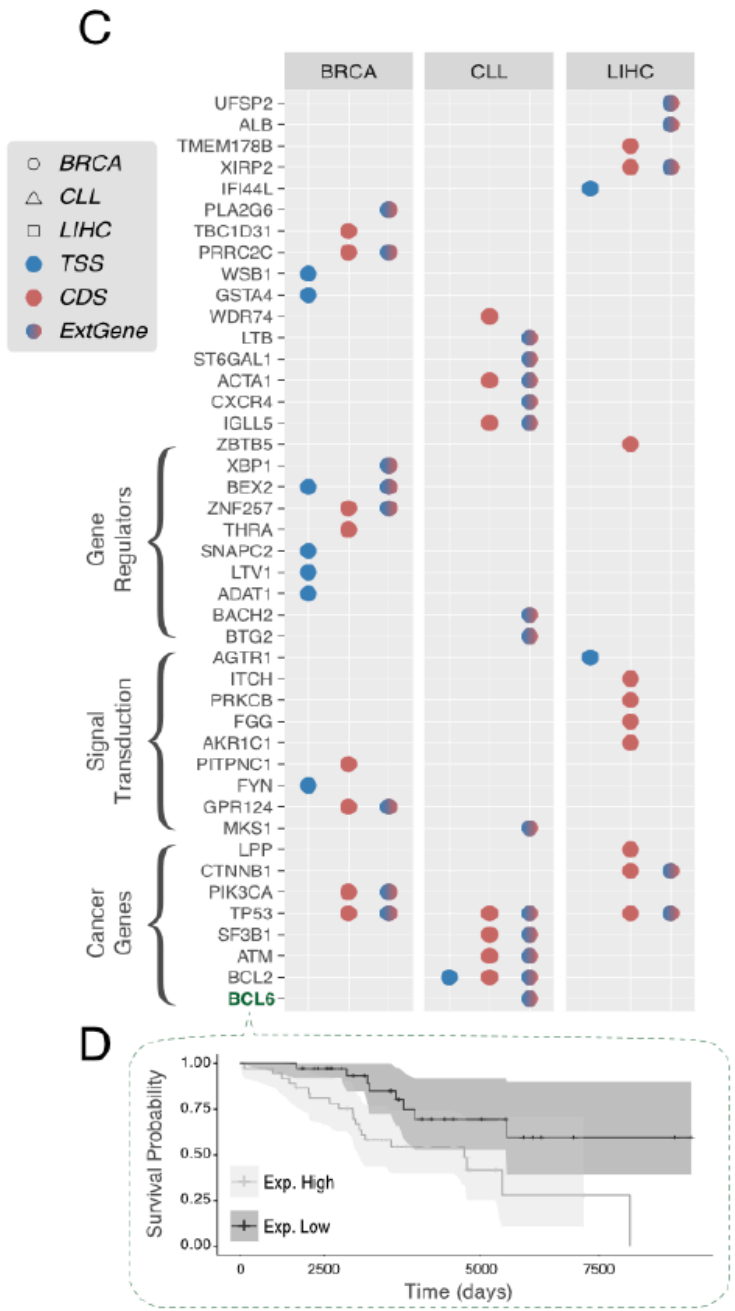


- How to improve BMR estimation accuracy
  - ✓ Use matched tissue data
  - ✓ Combine more features
  - ✓ Detect overdispersion and correct if necessary



Annotation level integration in mutation burden analysis





Extended gene neighborhood

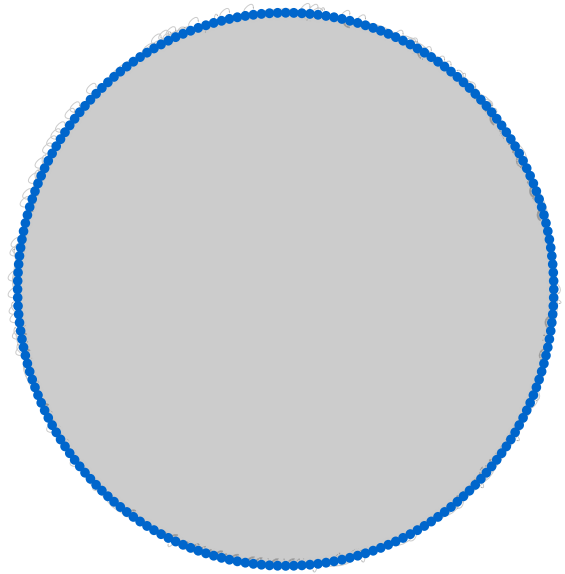
- Picks up distributed mutation signals from multiple regions
- Benefits variant impact interpretation
- Discovers genes (BCL6) with prognostic value but missed by CDS analysis

Outline

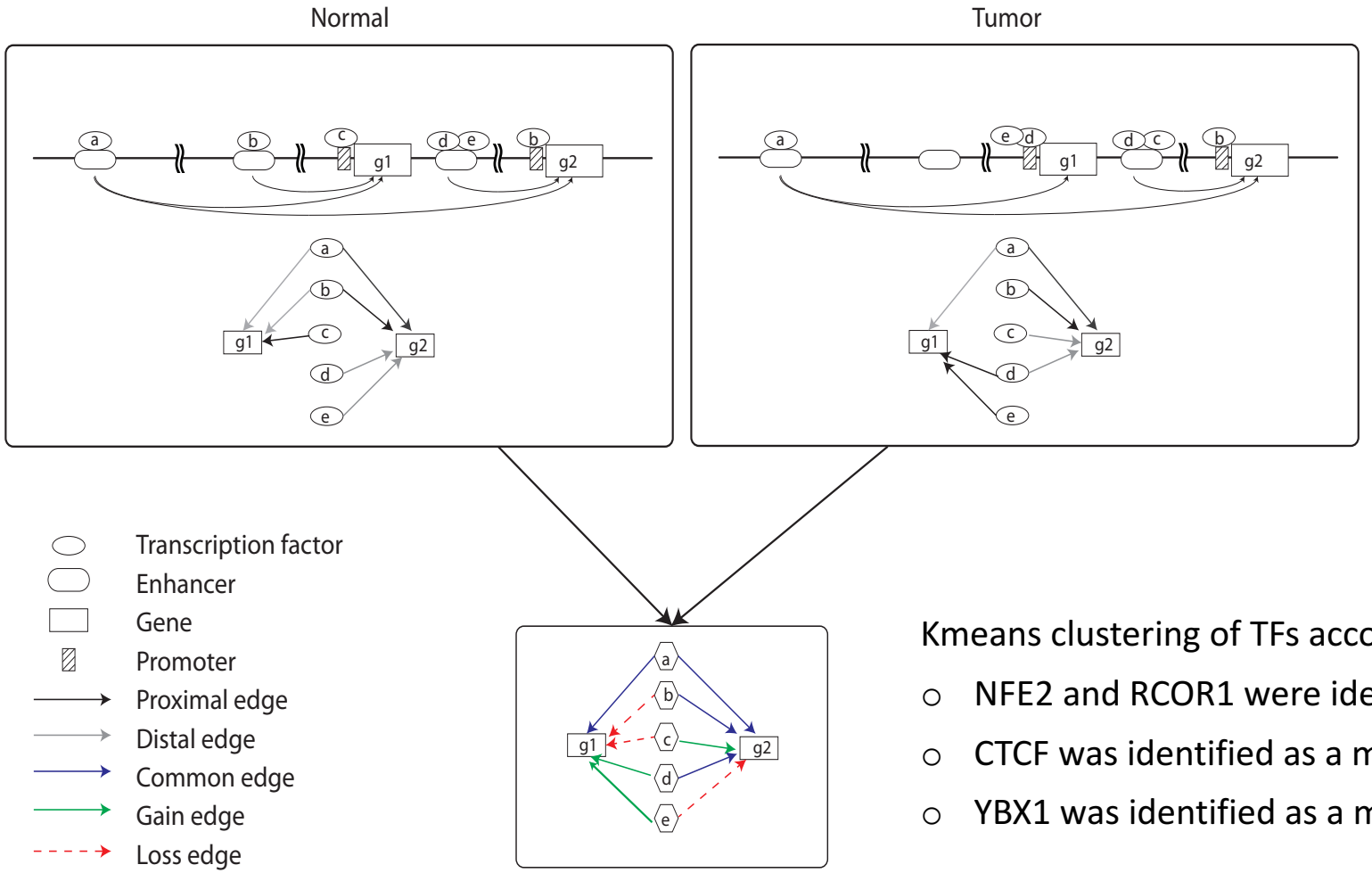
1. Introduction to EN-codec resource
2. Multi-level integration from ENCODE benefits mutation burden analysis
    - Raw signal level integration
    - Annotation level integration
3. Interpreting transcriptional level regulatory changes through network rewiring analysis
    - Quantification of regulatory changes
    - Effect of highly rewired TFs in cancer
4. Integrating regulatory networks with tumor expression profiles identifies key regulators in cancer
    - Identification of key regulators that driver T/N differential expression
    - Investigating the cooperation pattern between key regulators
5. Variant prioritization scheme and small scale validations

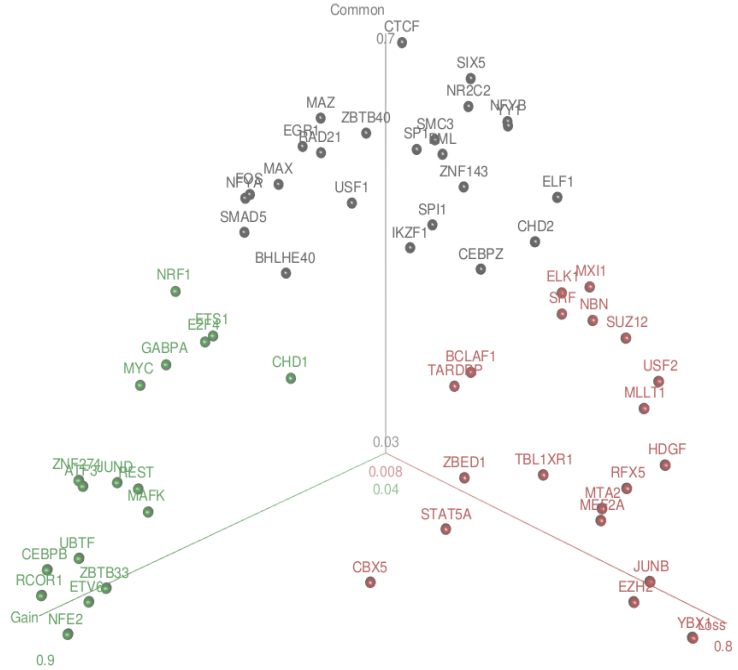
Cell Types specific network analysis helps to pinpoint regulation changes

Biosample	ChIP TF count
A549	32
GM12878	101
H1-hESC	50
HeLa-S3	60
HepG2	97
IMR-90	9
K562	209
liver	7
lung	1
mammary-epithelial-cell	2
MCF-10A	4
MCF-7	51



TF network rewiring schematic



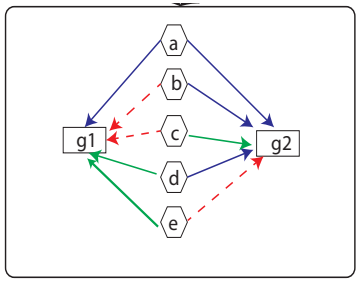


Kmeans clustering of TFs according to percentage of gain/loss edges

- NFE2 and RCOR1 were identified as one of the strongest member of gained group
- CTCF was identified as a member of common group
- YBX1 was identified as a member of loss group

Direct Rewiring Index:

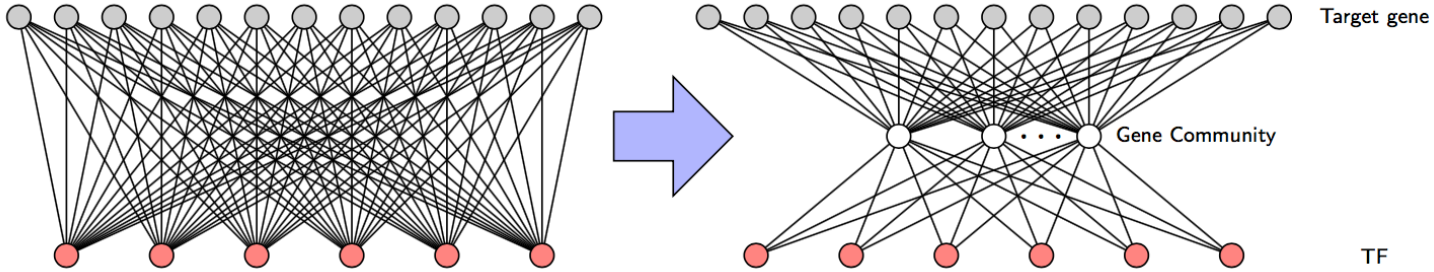
potential of loss and gain edges when transiting from tumor to normal cells

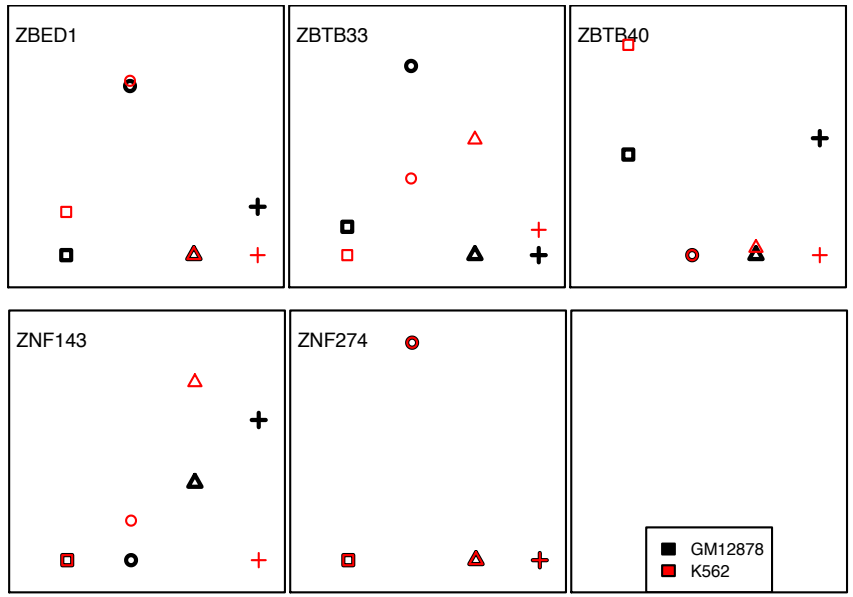


$$n_{\text{fully-connected}} = n_{\text{TF}} * n_{\text{gene}} - 1$$

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

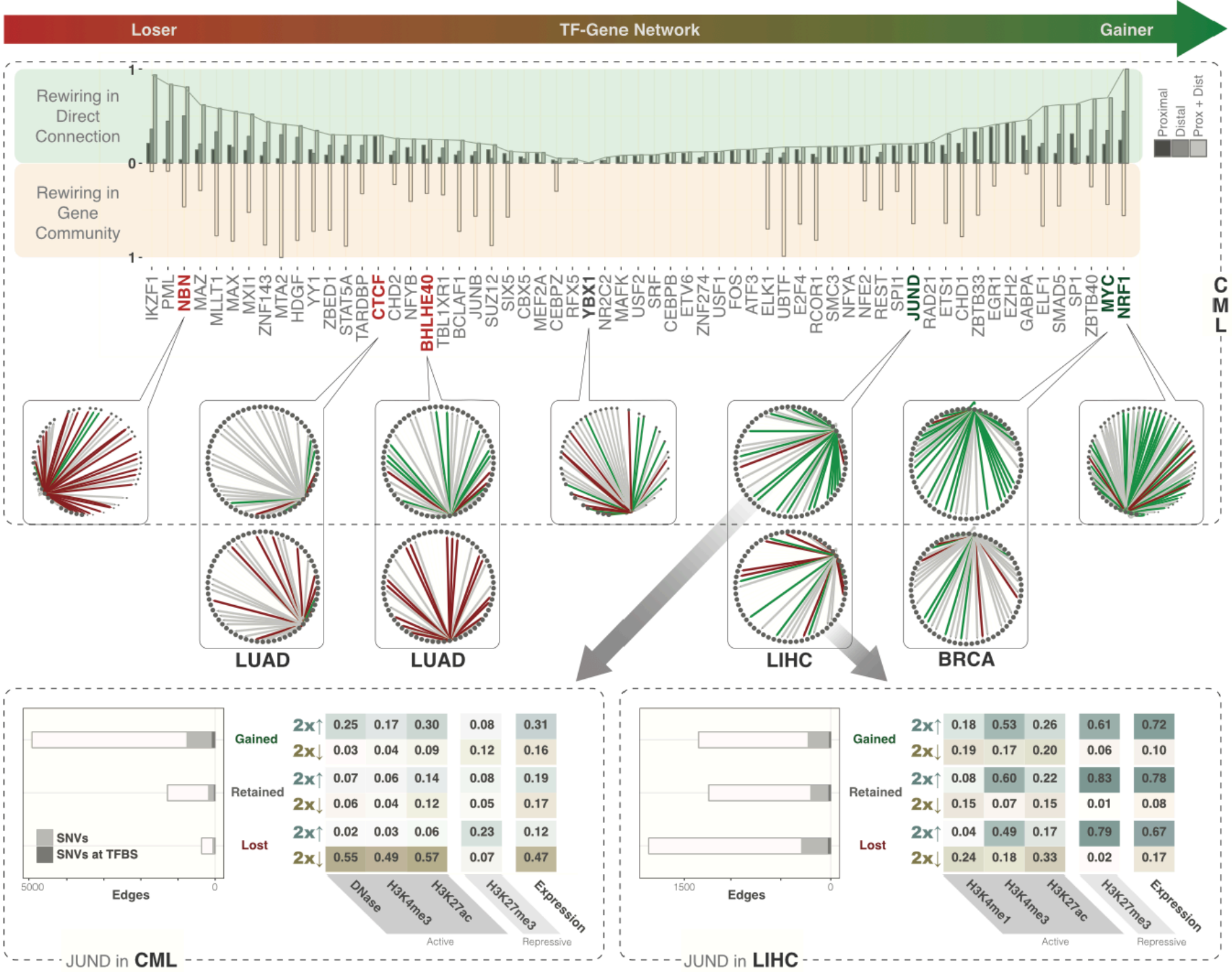
$$rScore_{\text{normalized}} = \frac{rScore_{TF}}{\max_{all}(rScore_{\text{TF}})}$$





Gene community based Rewiring score:

Difference of linkages to different gene communities in T/N cells



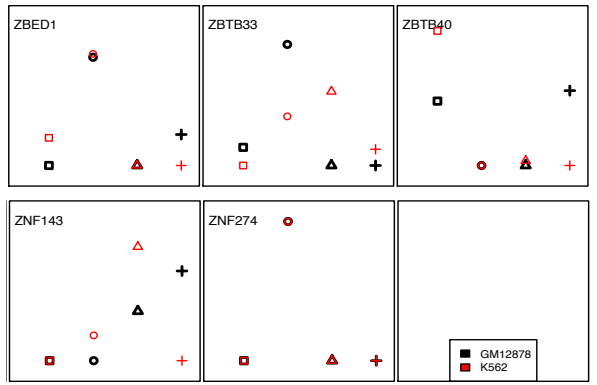
Direct Rewiring index

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

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

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

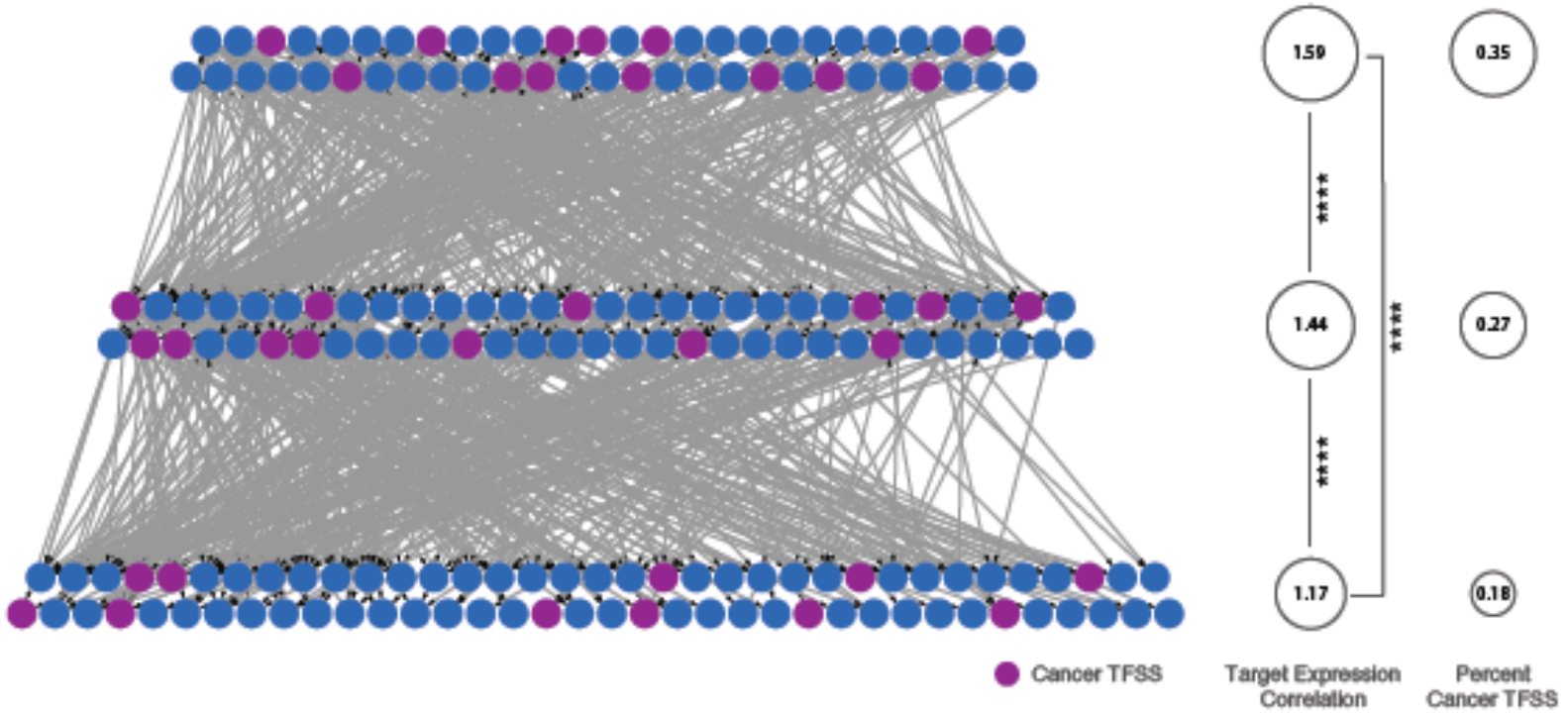
Gene community based rewiring index

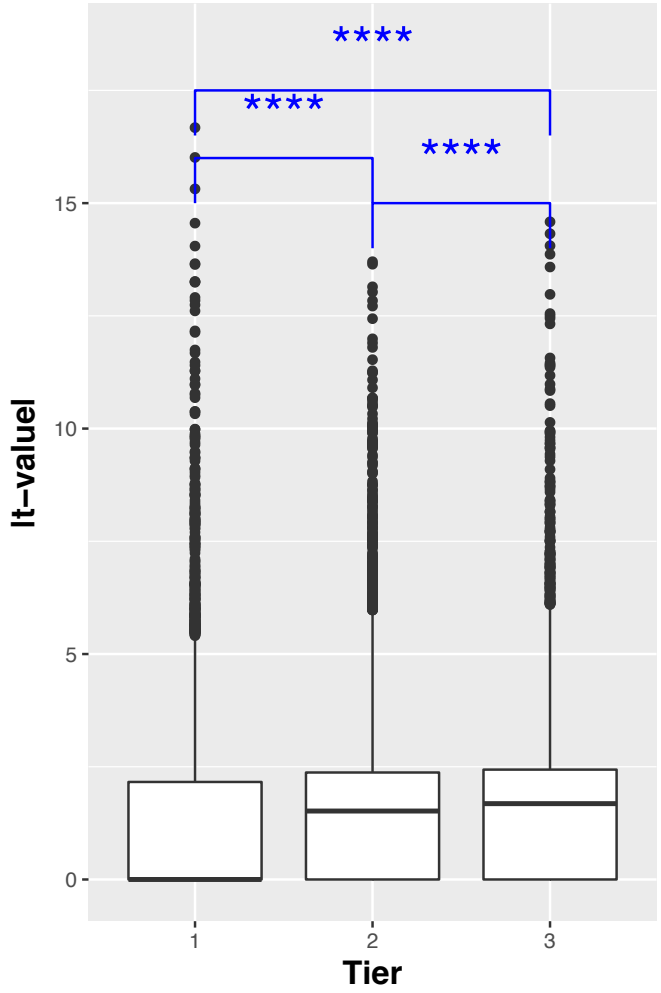


Outline

1. Introduction to EN-codec resource
2. Multi-level integration from ENCODE benefits mutation burden analysis
    - Raw signal level integration
    - Annotation level integration
3. Interpreting transcriptional level regulatory changes through network rewiring analysis
    - Quantification of regulatory changes
    - Effect of highly rewired TFs in cancer
4. Integrating regulatory networks with tumor expression profiles identifies key regulators in cancer
    - Identification of key regulators that driver T/N differential expression
    - Investigating the cooperation pattern between key regulators
5. Variant prioritization scheme and small scale validations

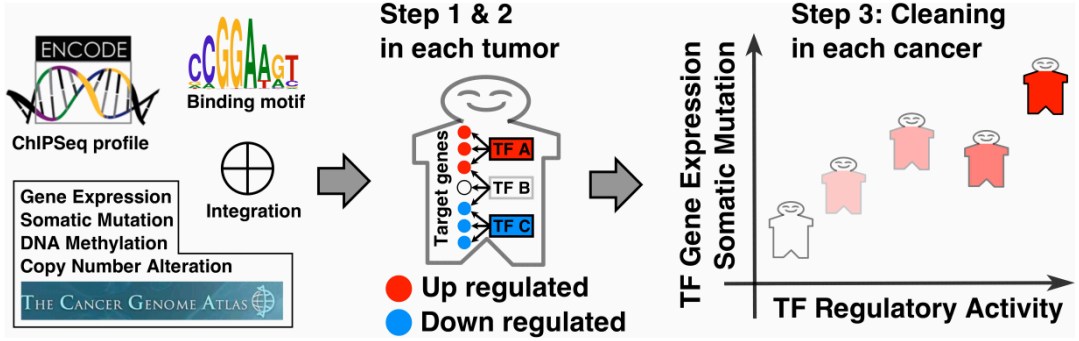
Pinpoint Key Regulators through generalized network analysis





- Top tier TFs are slightly enriched with more cancer associated TFs
- Top tier TFs demonstrates stronger regulation power in cancer genomes

Pinpoint Key Regulators through generalized network analysis



TF regulation Score

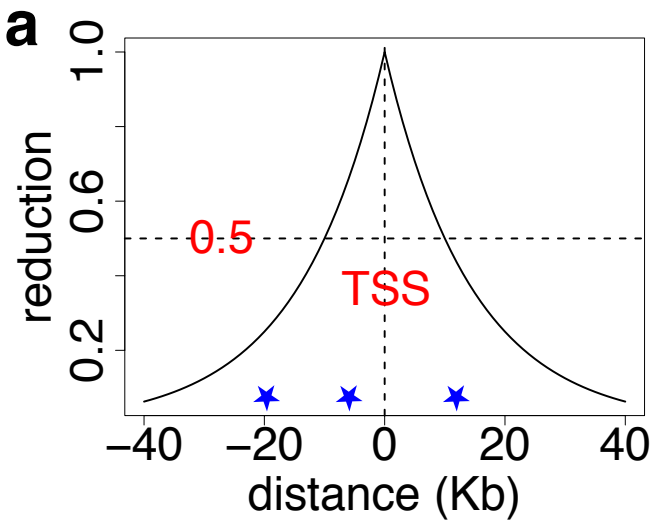
Gene Expression

$$\begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_i \\ \vdots \\ g_n \end{bmatrix} = \beta_0 + \beta_1 \begin{bmatrix} r_{1,1} \\ r_{1,2} \\ \vdots \\ r_{1,i} \\ \vdots \\ r_{1,n} \end{bmatrix} + \cdots + \beta_k \begin{bmatrix} r_{k,1} \\ r_{k,2} \\ \vdots \\ r_{k,i} \\ \vdots \\ r_{k,n} \end{bmatrix}$$

$r_{1,1}$ regulation score, not the expression level of TF1 to gene 1

- Run a regression on each patient
- Do model selection for TFs using forward selection
- Use t value of the final model to measure significance
- Summarize the percentage of patients with significant correlation

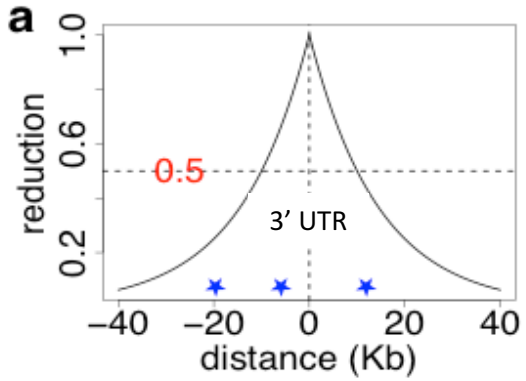
Modeling regulation potential of regulator-gene pair



$$1 - \Pi_i(1 - score_i * exp(-A * D_i))$$

Enriched with down-regulated targets

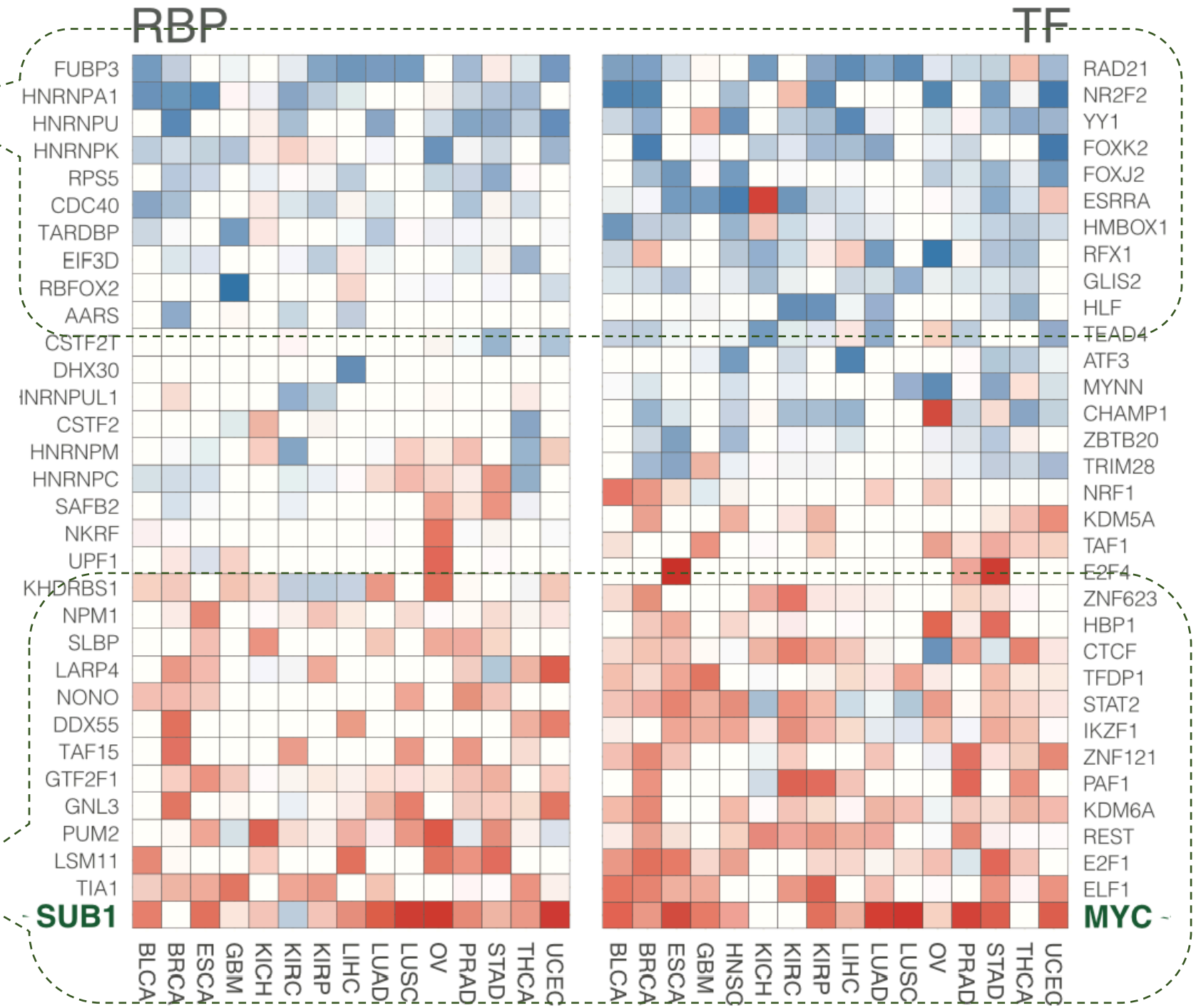
RBP regulation score



$$1 - \Pi_i(1 - score_i * exp(-A * D_i))$$

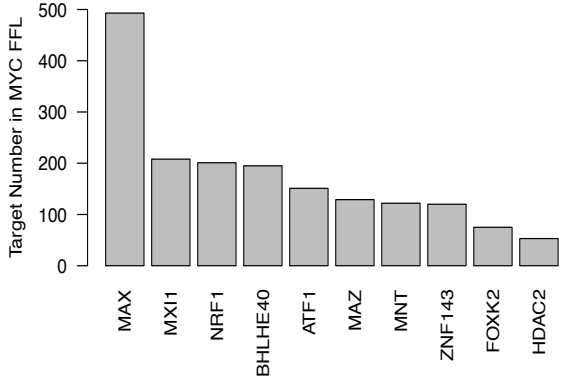
$$1 - \Pi_i(1 - score_i))$$

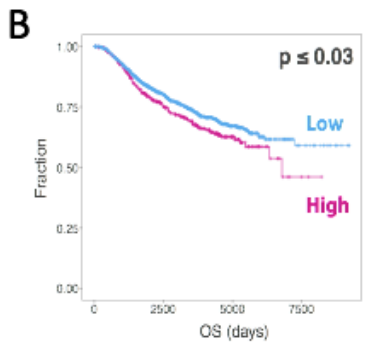
Enriched with up-regulated targets

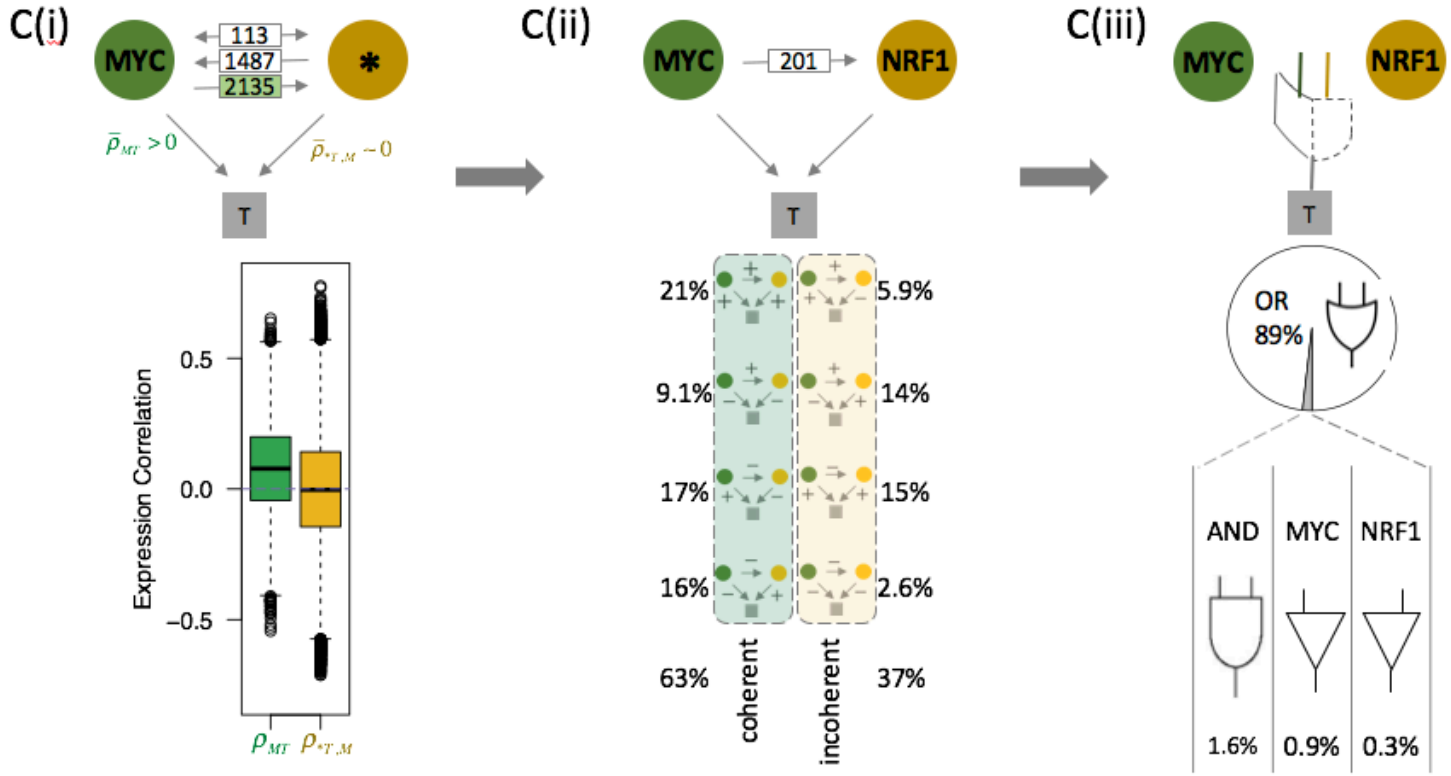


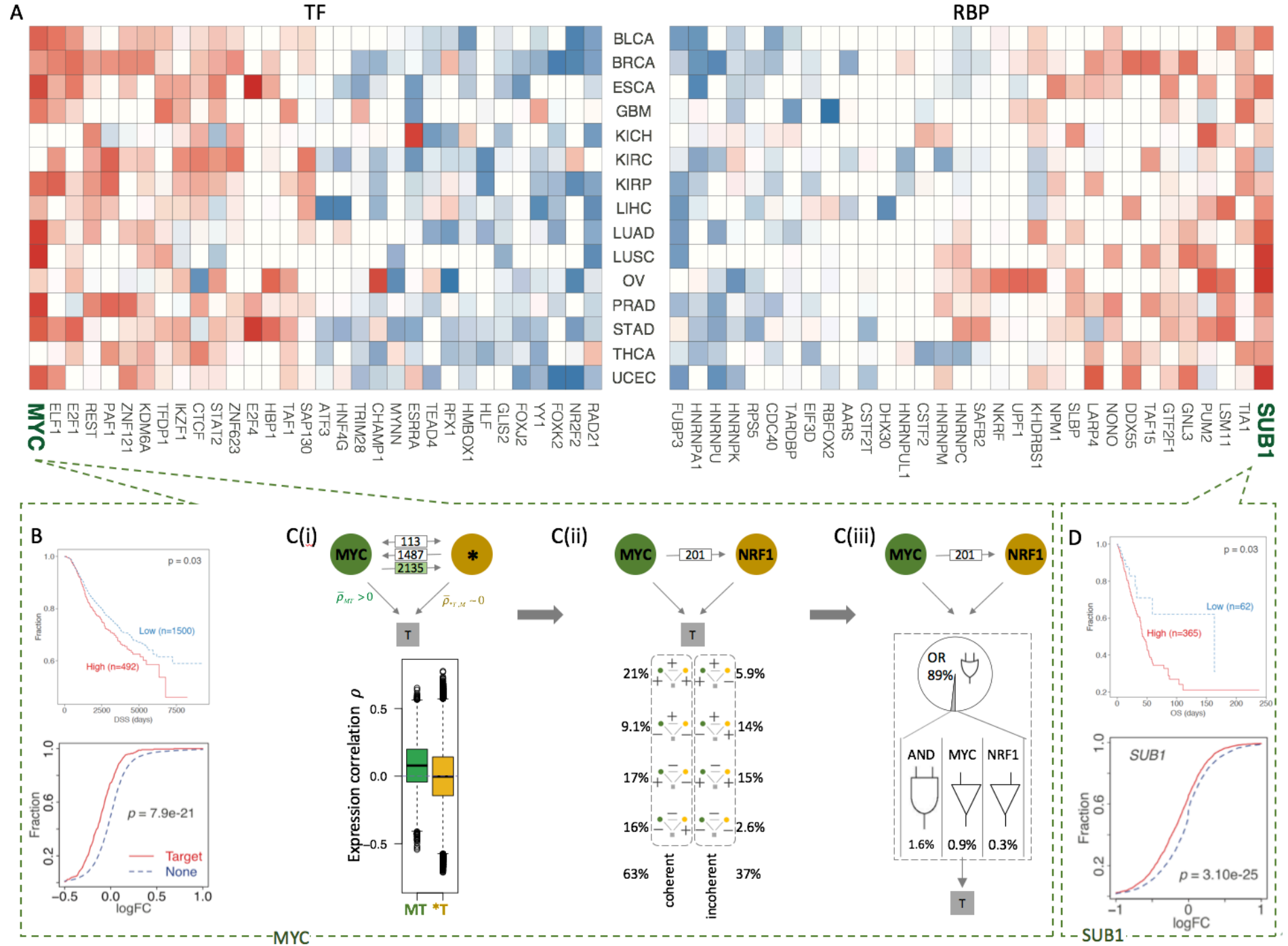
Investigating How MYC work with others regulators

Other TF	212	2	14	38
Triplet	10574	113	1487	2135

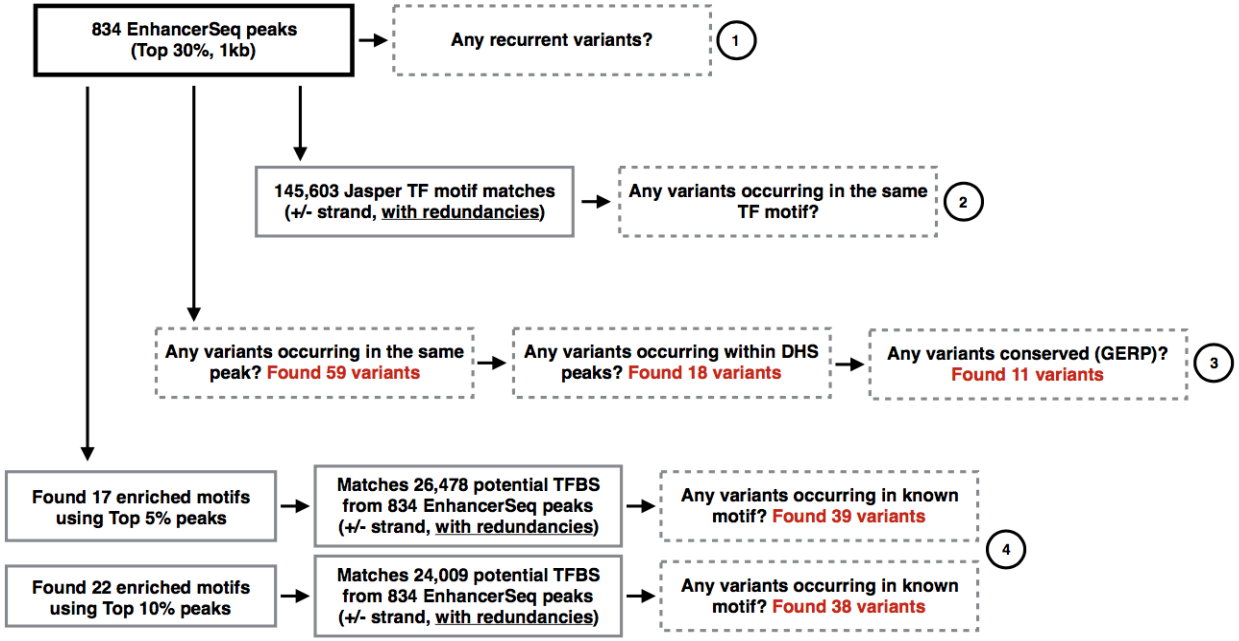






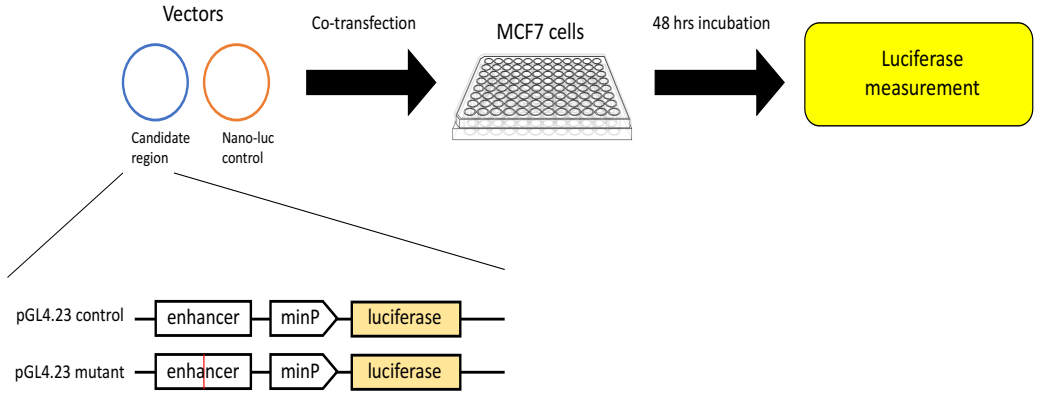

A
TF
RBP
BLCA
BRCA
ESCA
GBM
KICH
KIRC
KIRP
LIHC
LUAD
LUSC
OV
PRAD
STAD
THCA
UCEC
MYC
ELF1 E2F1 REST PAF1 ZNF121 KDM6A TFDP1 IKZF1 CTCF STAT2 ZNF623 E2F4 HBP1 TAF1 SAP130 ATF3 HNF4G TRIM28 CHAMP1 MYNN ESRRA TEAD4 RFX1 HMBOX1 HLF GLIS2 FOXJ2 YY1 FOXK2 NR2F2 RAD21
FUBP3 HNRNPA1 HNRNPU HNRNPK RPS5 CDC40 TARDBP EIF3D RBFOX2 AARS DHX30 HNRNPUL1 CSTF2 CSTF2T HNRNPM HNRNPC SAFB2 NKRF UPF1 KHDRBS1 NPM1 SLBP LARP4 NONO DDX55 TAF15 GTF2F1 GNL3 PUM2 LSM11 TIA1
SUB1
B
p = 0.03
Low (n=1500)
High (n=492)
Fraction
DSS (days)
p = 7.9e-21
Target
None
logFC
C(i)
MYC
113
1487
2135
*
T
Expression correlation ρ
MT *T
C(ii)
201
NRF1
21%
5.9%
9.1%
14%
17%
15%
16%
2.6%
63%
coherent
incoherent
37%
C(iii)
OR
89%
AND
MYC
NRF1
1.6%
0.9%
0.3%
D
p = 0.03
Low (n=62)
High (n=365)
OS (days)
SUB1
p = 3.10e-25
logFC

Variant prioritization

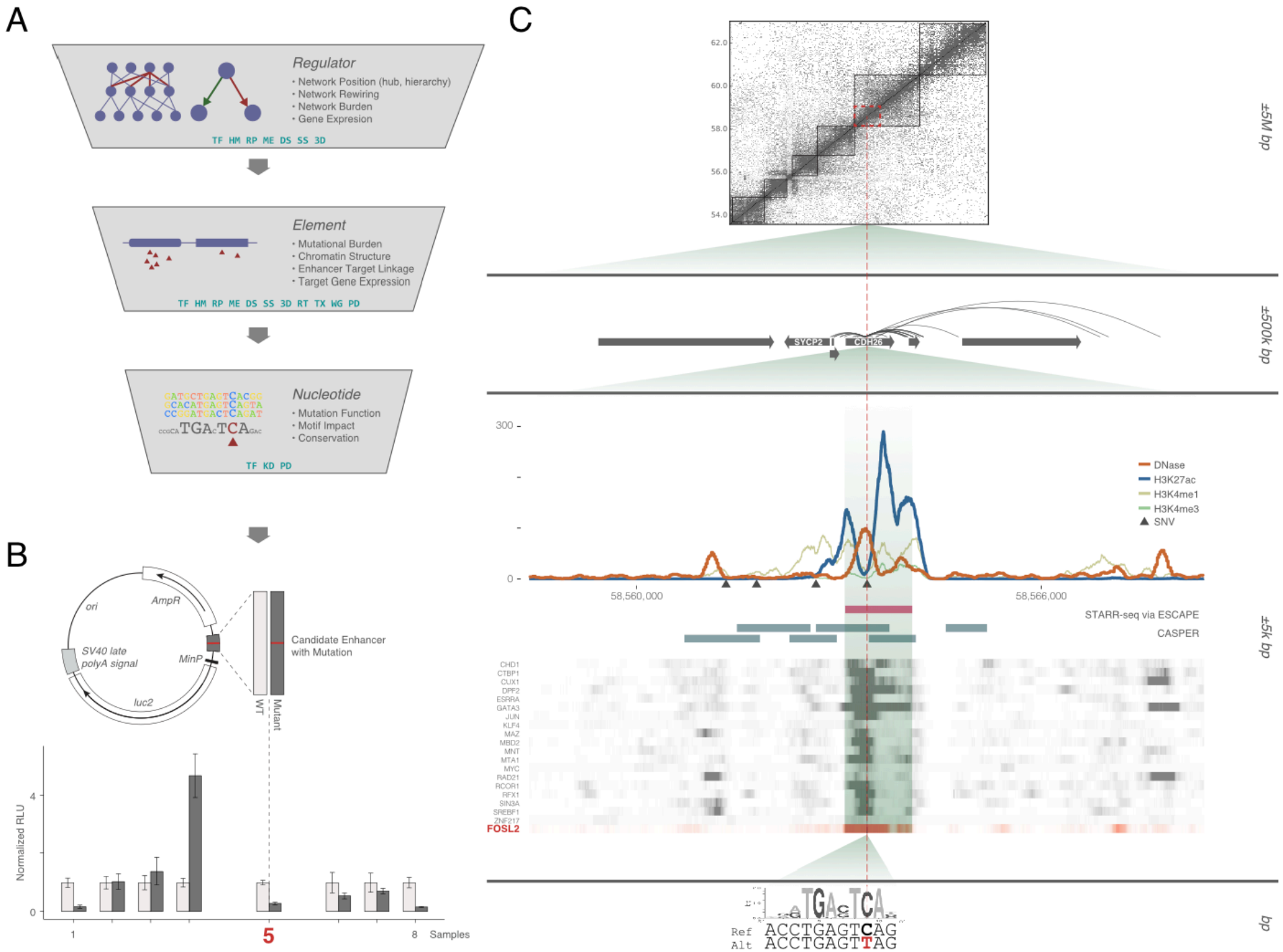


1. Recurrent BRCA non-coding variants within 834 EnhancerSeq peaks => **None**
2. Multiple BRCA non-coding variants occurring in a known TF motif => **None**
3. Multiple BRCA non-coding variants occurring in a EnhancerSeq peak (834) => **59 non-coding variants**
4. BRCA variants in known TF motif with 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**

Variant Validation



SAMPLE	CHR	POS	REF	ALT	TEST_START	TEST_END
Sample01	chr16	85604242	C	G	85603992	85604491
Sample02	chr21	27541982	G	A	27541732	27542231
Sample03	chr8	21541726	A	G	21541476	21541975
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


A
Regulator
• Network Position (hub, hierarchy)
• Network Rewiring
• Network Burden
• Gene Expresion
TF HM RP ME DS SS 3D
Element
• Mutational Burden
• Chromatin Structure
• Enhancer Target Linkage
• Target Gene Expression
TF HM RP ME DS SS 3D RT TX WG PD
Nucleotide
• Mutation Function
• Motif Impact
• Conservation
TF KD PD
B
ori
AmpR
SV40 late polyA signal
MinP
luc2
Candidate Enhancer with Mutation
WT
Mutant
Normalized RLU
Samples
C
±5M bp
±500k bp
SYCP2
CDH26
DNase
H3K27ac
H3K4me1
H3K4me3
SNV
58,560,000
58,566,000
STARR-seq via ESCAPE
CASPER
±5k bp
FOSL2
Ref ACCTGAGTCAG
Alt ACCTGAGTTAG
bp

Summary of resource in ENCodec for cancer research

Raw Data	Signal tracks from de-duplicated ChIP-seq, Dnase-seq, Repli-seq, RNA-seq, WGBS data
	Germline/somatic variant calls from liver cancer patients
	SNV/SV calls for several top tier cell lines
	Annotation of TFs and RBPs
	Tumor-normal pairs of cell types of many cancer types
Annotation	Cell type specific promoter usage
	Enhancer predictions and gene linkages
	TF & RBP network – cell type specific and generalized
	Extended gene neighborhood
Software	Motiftool: motif gain/loss events events
	NIMBus: mutation burden analysis tool
	ESCAPE + Casper: enhancer prediction based on ChIP-seq, Dnase-seq and STARR-seq
	JEME: enhancer target prediction
	Vinas: variant impact on cancer genome analysis tool
Analysis-release	Network hierarchies
	Estimated BMR in various cancer types
	TF rewiring status
	TF/RBP regulation power in T/N cell types
	regulatory status changes per gene
	burdened elements

Acknowledgement

- Mark Gerstein
- Donghoon Lee
- Lucas Lochovsky, Jason Liu, Yanlin Feng
- Joel Rozowsky, Shaoke Lou – ChIP-seq
- Arif Harmanci, Mengting Gu, Koon-Kiu Yan, Anurag Sethi – Enhancer
- William Meyerson, Patrick Mcgillivray, Xiaotong Li – Survival analysis and T/N pairing
- Gamze Gursoy, Daifeng Wang – loregic
- Variation group!

X. Shirley Liu Lab
Biostatistics & Computational Biology

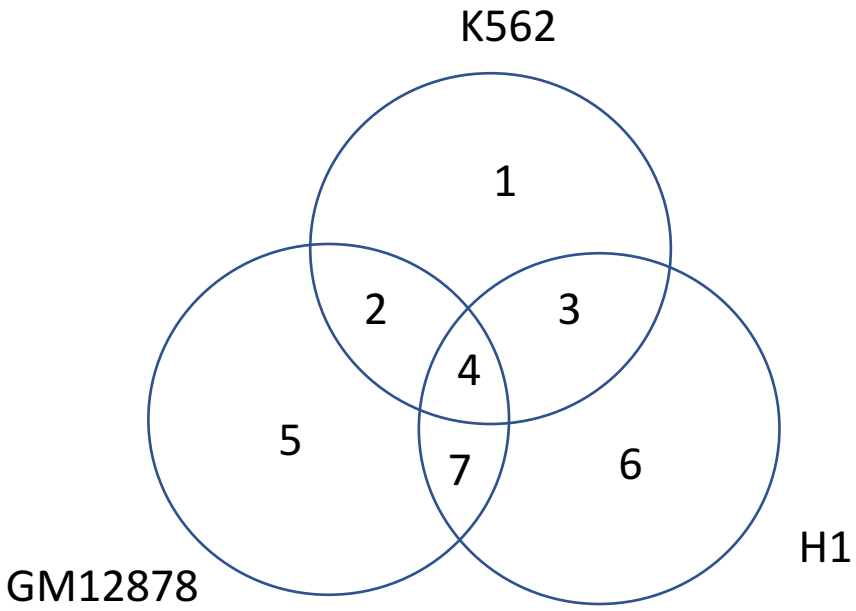
the Klein Lab

Michael Snyder Lab

Kevin White Lab

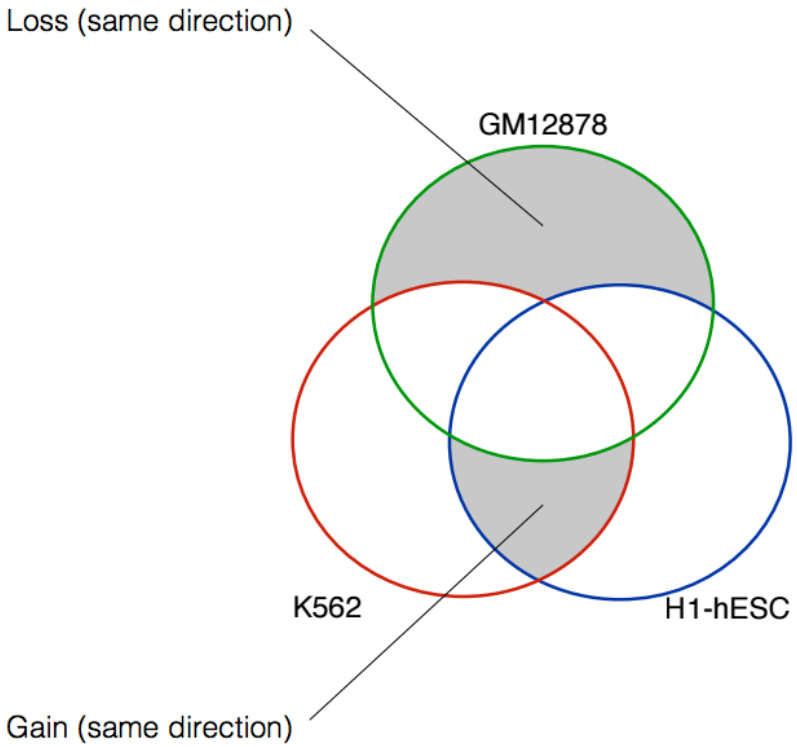
Discussion

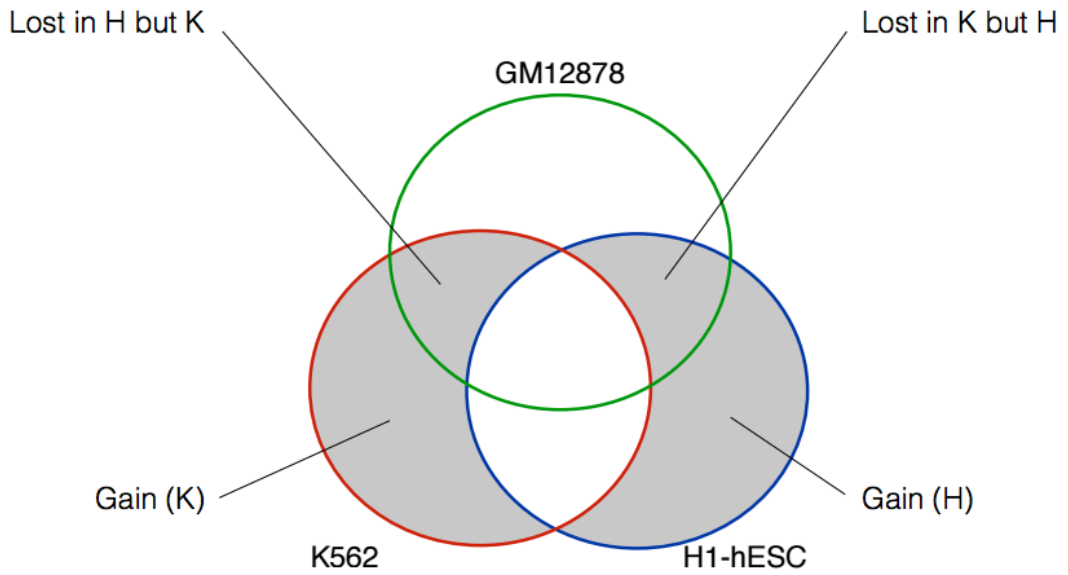
Ultimate question: a cancer cell is going through a more differentiated way or un-differentiated



Differentiated direction: $\frac{3}{1+3+6}$

Un-Differentiated direction: $\frac{7}{5+7+6}$

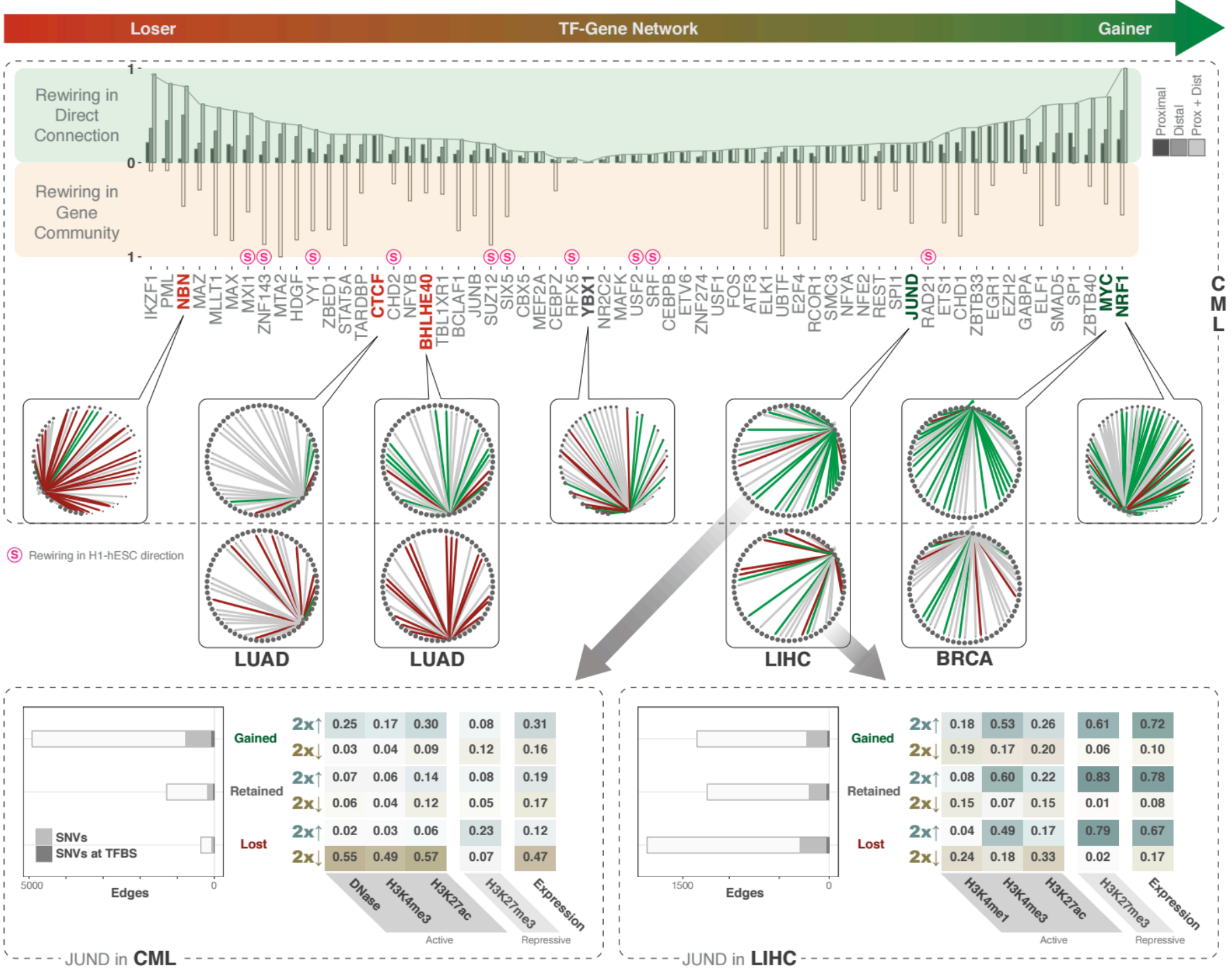




- Edges in the same direction vs. edges in the opposite direction
- TIP+enhancers for “rScore”, TSS only for “H1 similarity”
- More edges in G+K
- Gain dominant

	n.edges
GM12878	302,295
K562	572,944
H1-hESC	131,330

TF	RFX5	SRF	CHD2	ZNF143	MXI1	USF2	SUZ12	SIX5	YY1	RAD21	CTCF	USF1	CEBPB	SP1	ATF3	EZH2	MAFK	CHD1	GABPA	JUND	EGR1	MAX	MYC	REST	NRF1
same.dir	1353	1541	2015	2491	2478	1157	1325	716	1914	1193	1221	1053	1167	1249	883	337	663	982	930	1248	985	1055	917	832	313
same.dir x2	2706	3082	4030	4982	4956	2314	2650	1432	3828	2386	2442	2106	2334	2498	1766	674	1326	1964	1860	2496	1970	2110	1834	1664	626
opposite.dir	705	1073	2081	2578	2583	1264	1539	963	3017	2281	2640	2411	3085	3369	2569	1027	2178	3285	3598	4891	5143	5617	4916	5836	5022
overall.dir	2001	2009	1949	2404	2373	1050	1111	469	811	105	-198	-305	-751	-871	-803	-353	-852	-1321	-1738	-2395	-3173	-3507	-3082	-4172	-4396
% same dir	0.79	0.74	0.66	0.66	0.66	0.65	0.63	0.60	0.56	0.51	0.48	0.47	0.43	0.43	0.41	0.40	0.38	0.37	0.34	0.34	0.28	0.27	0.27	0.22	0.11


Loser
TF-Gene Network
Gainer
Rewiring in Direct Connection
Rewiring in Gene Community
Proximal
Distal
Prox + Dist
IKZF1 PML NBN MAZ MLLT1 MAX MXI1 ZNF143 MTA2 HDGF YY1 ZBED1 STAT5A TARDBP CTCF CHD2 NFYB BHLHE40 TBL1XR1 BCLAF1 JUNB SUZ12 SIX5 CBX5 MEF2A CEBPZ RFX5 YBX1 NR2C2 MAFK USF2 SRF CEBPB ETV6 ZNF274 USF1 FOS ATF3 ELK1 UBTF E2F4 RCOR1 SMC3 NFYA NFE2 REST SPI1 JUND RAD21 ETS1 CHD1 ZBTB33 EGR1 EZH2 GABPA ELF1 SMAD5 SP1 ZBTB40 MYC NRF1
CML
Rewiring in H1-hESC direction
LUAD
LUAD
LIHC
BRCA
SNVs
SNVs at TFBS
5000
Edges
0
Gained
Retained
Lost
2x↑ 0.25 0.17 0.30 0.08 0.31
2x↓ 0.03 0.04 0.09 0.12 0.16
2x↑ 0.07 0.06 0.14 0.08 0.19
2x↓ 0.06 0.04 0.12 0.05 0.17
2x↑ 0.02 0.03 0.06 0.23 0.12
2x↓ 0.55 0.49 0.57 0.07 0.47
DNase
H3K4me3
H3K27ac
H3K27me3
Expression
Active
Repressive
JUND in CML
1500
Edges
0
Gained
Retained
Lost
2x↑ 0.18 0.53 0.26 0.61 0.72
2x↓ 0.19 0.17 0.20 0.06 0.10
2x↑ 0.08 0.60 0.22 0.83 0.78
2x↓ 0.15 0.07 0.15 0.01 0.08
2x↑ 0.04 0.49 0.17 0.79 0.67
2x↓ 0.24 0.18 0.33 0.02 0.17
H3K4me1
H3K4me3
H3K27ac
H3K27me3
Expression
Active
Repressive
JUND in LIHC