Chromatin structure-based prediction of recurrent noncoding mutations in cancer

Journal club

10/25/2016

Aim of the work

To identify non-coding driver(recurrent) mutations that impact a cis-regulatory element of a gene via chromatin interactions.

Compare this model with respect to site specific recurrence model.

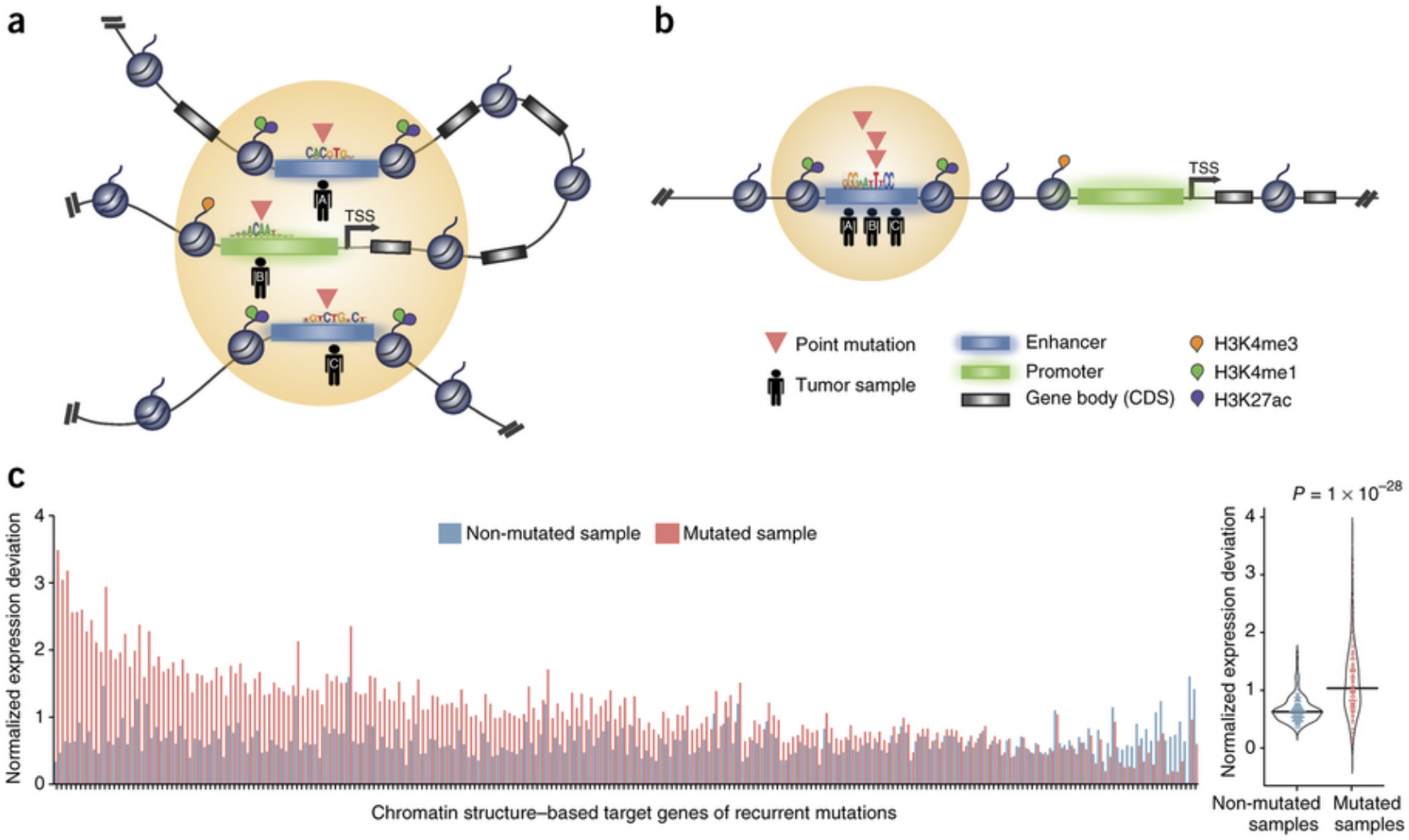
SNV datasets utilized

Cohort	Cancer type	Number of samples	Number of point mutations
Sanger	Breast	119	647,695
	Lung	24	1,446,336
GIS	Lung	30	1,763,572
TCGA	Breast	92	756,434
	Lung	90	3,594,840

Enhancer-promoter interaction data

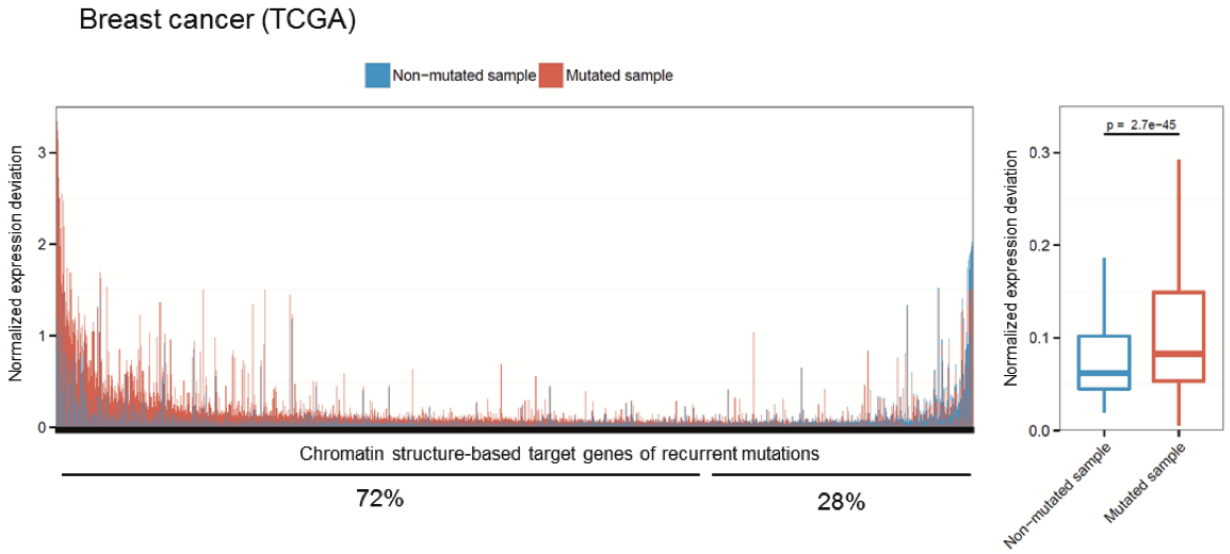
Cancer type	State	Cell line	Data source	Number of interaction pairs
Breast	Cancer	MCF7	ChIA-PET	24,501
		MCF7	IM-PET	25,746
		MCF7, T47D	DHS tag correlation	66,222
		MCF7, T47D	CAGE expression correlation	18,362
	Cell of origin	HMEC	IM-PET	20,254
		HMEC	DHS tag correlation	75,167
		HMEC	CAGE expression correlation	26,001
Lung	Cancer	A549	IM-PET*	27,435
		A549	DHS tag correlation	50,135
		A549	CAGE expression correlation	18,509
	Cell of origin	NHLF, IMR90	IM-PET	40,189
		AGO4550, HPF, NHLF, WI38	DHS tag correlation	70,785
		AGO4550, HPF, NHLF, WI38	CAGE expression correlation	25,400

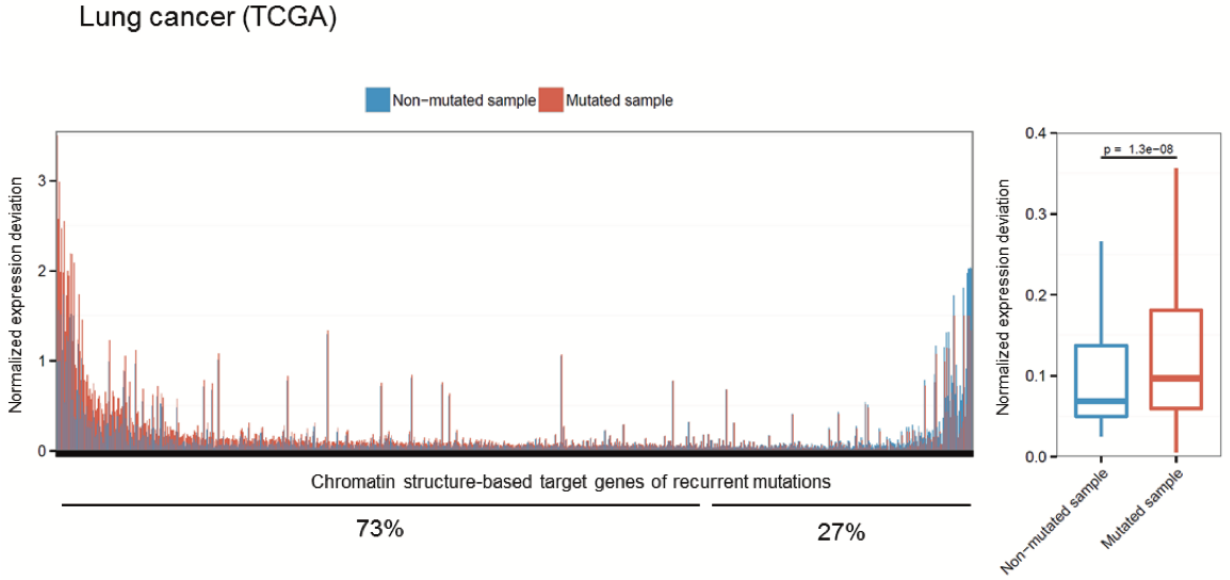
Chromatin-structure based recurrence model



Deviation in expression for genes with recurrent mutations tend to higher compared to non-mutated samples.

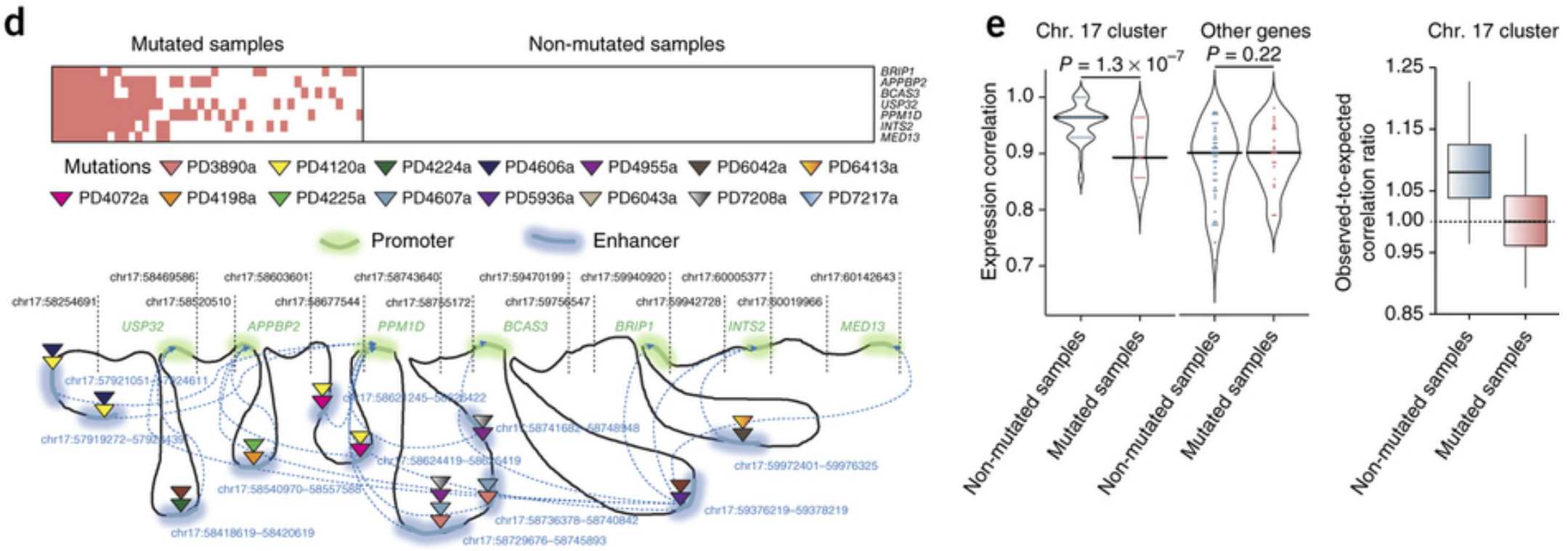
Gene expression deviation of genes influenced by recurrent mutations





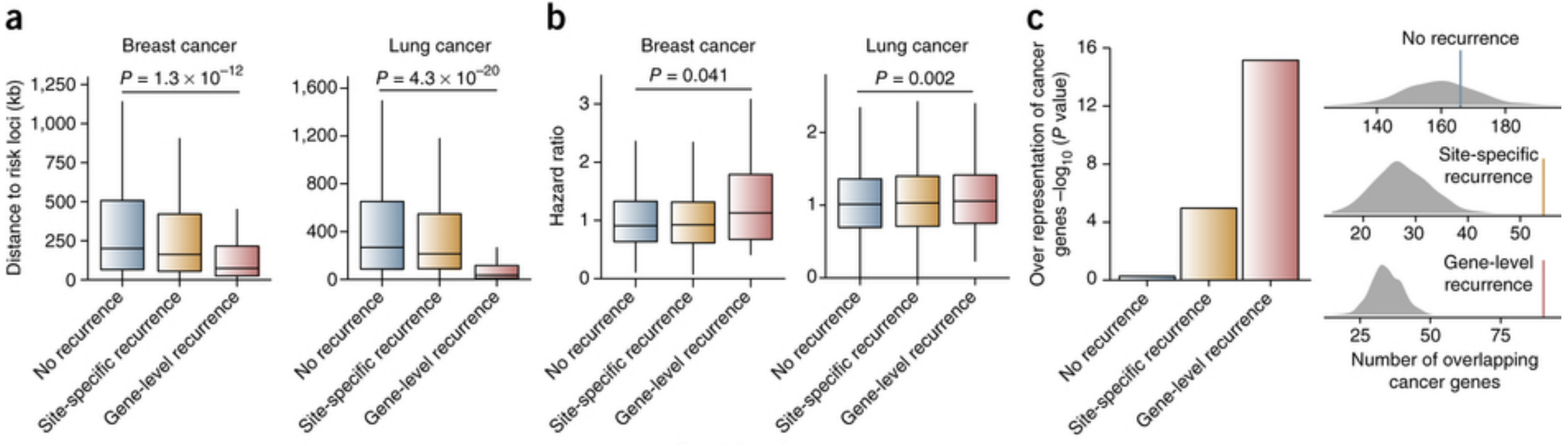
Approximately 72% and 73% of genes with recurrent mutations show higher level of gene expression deviation for the mutated Breast and Lung samples, respectively.

Mutation profile of genes with recurrent and clustered mutations



Cluster of recurrent mutations observed in various genes on the chromosome 17 for breast cancer cohort. These genes had similar recurrent mutation profile.

Oncogenic relevance of gene-wise recurrent mutations

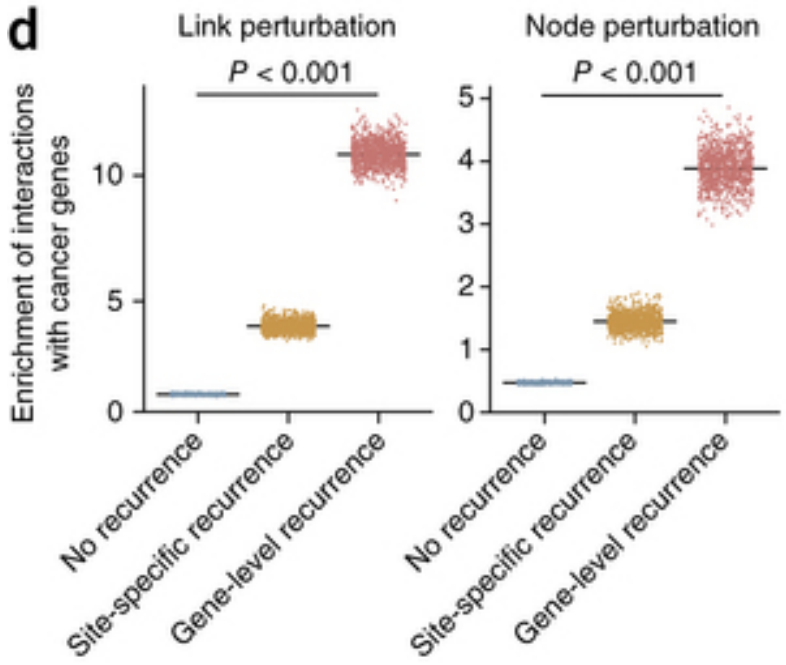


Recurrent mutations are in close proximity with cancer gene locus.

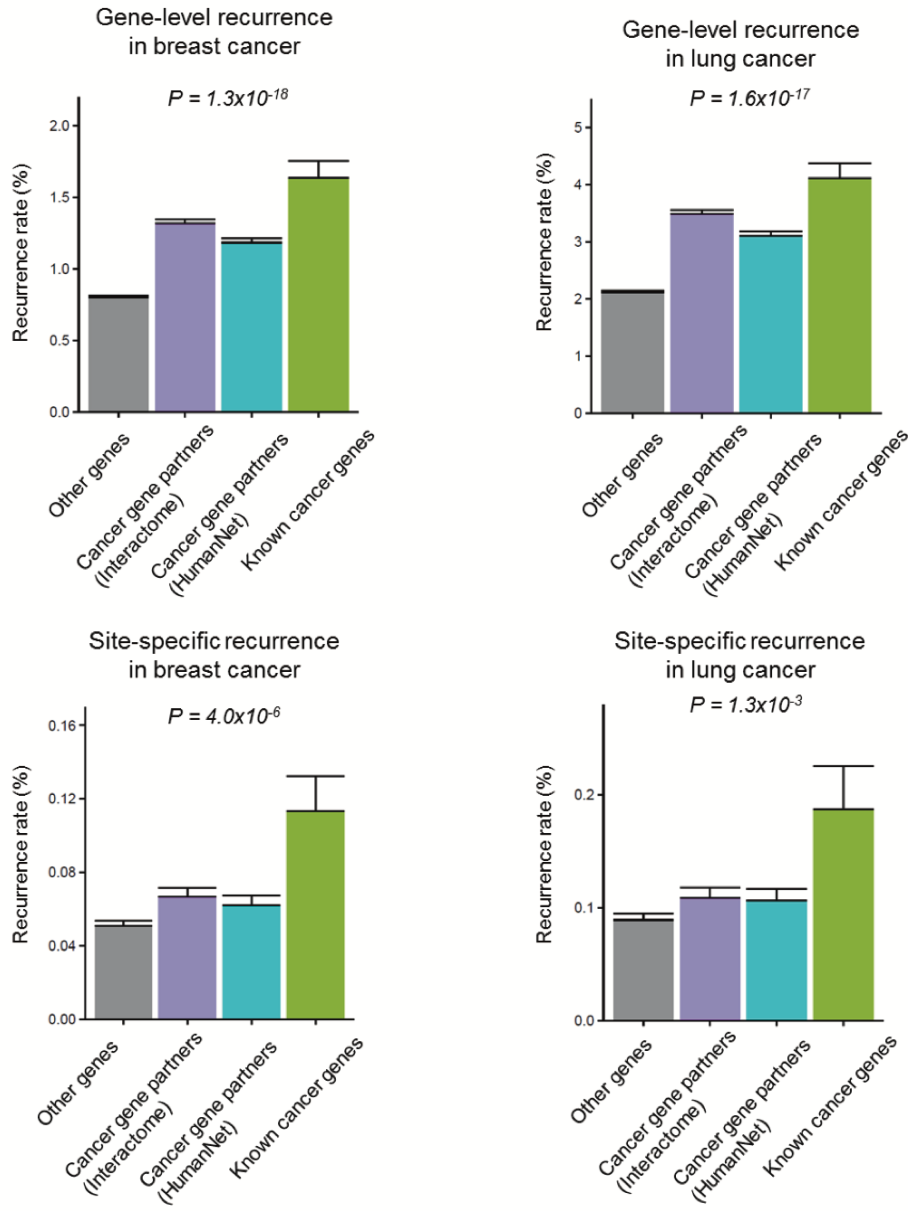
Expression of genes with recurrent mutations correlate better with patient survival.

Known cancer genes were over-represented among genes affected with recurrent mutations.

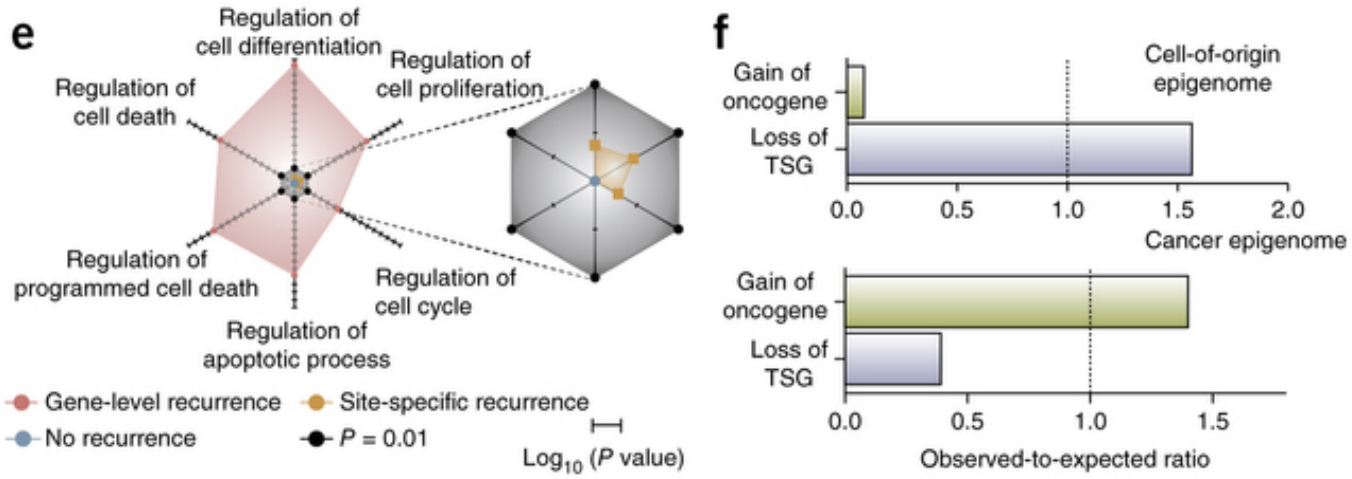
Interactions of gene-wise recurrent mutations with known cancer genes

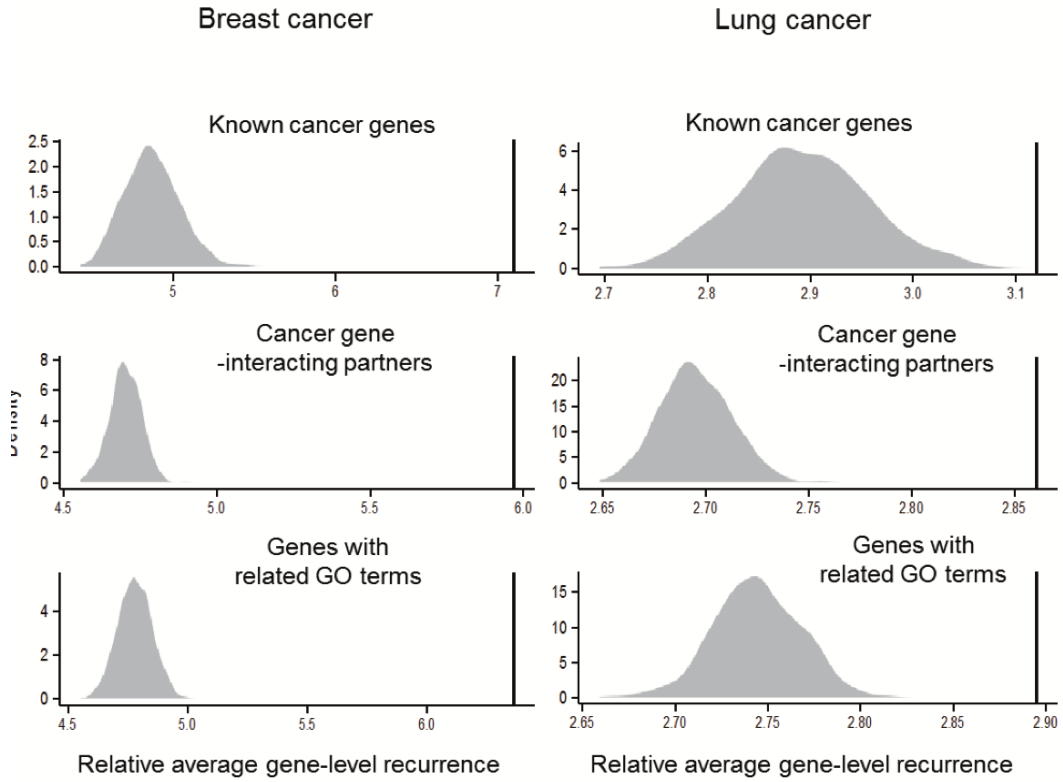


Enrichment of interactions involving recurrently mutated genes with known cancer genes.



Functional enrichment analysis of gene-wise recurrent mutations



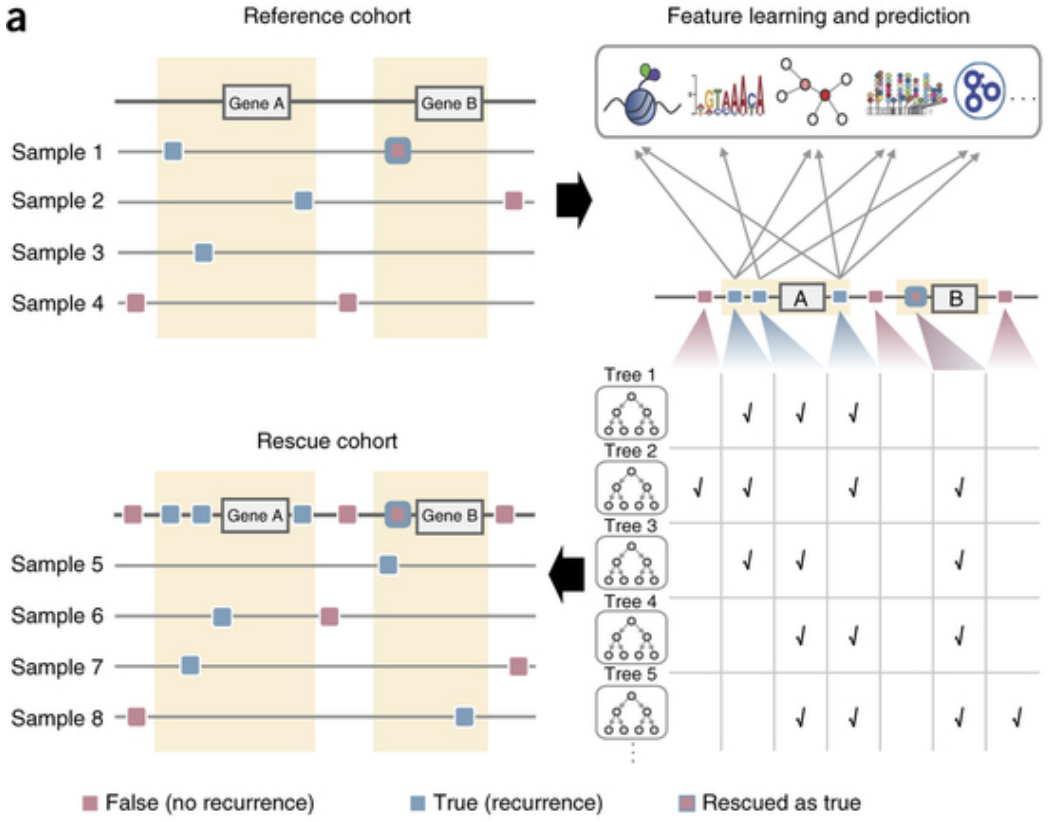


Recurrently mutated genes interact with genes enriched with biological function related to cancer genes.

Recurrently mutation among TSG enriched in enhancer active in the cell-of-origin epigenome.

For oncogenes, enrichment observed in enhancers active in cancer epigenome.

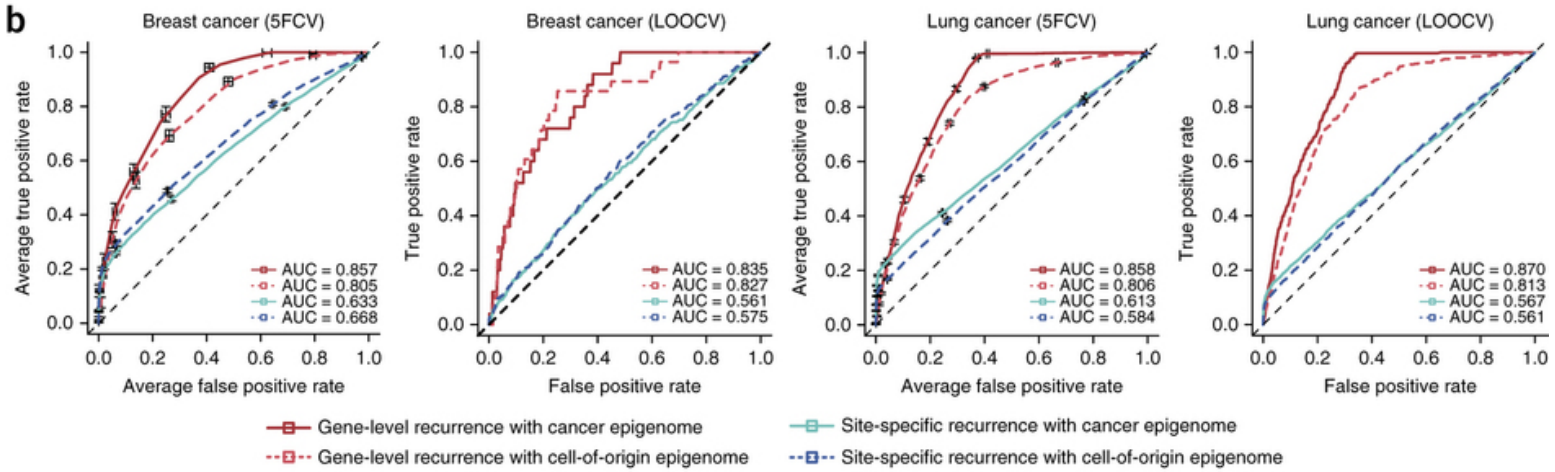
Classifier to predict recurrent mutations

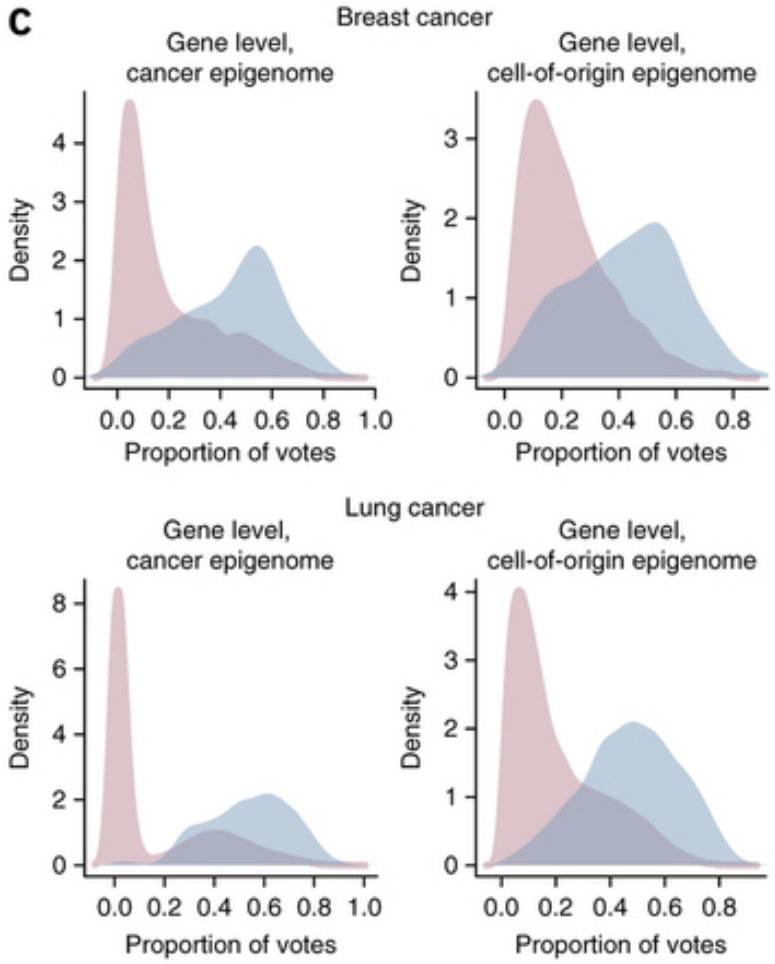


Random forest classifier to predict gene level recurrent functional mutations with inclusion of additional samples.

	Feature ID	Description
Target gene features	Target_Cancer_Gene	Whether the target gene is a known cancer gene
	Target_HumanNet.CGS_L1	Number of one-hop interactions of the target gene with known cancer genes in HumanNet
	Target_HumanNet.CGS_L2	Number of two-hop interactions of the target gene with known cancer genes in HumanNet
	Target_Interactome.CGS_L1	Number of one-hop interactions of the target gene with known cancer genes in Interactome
	Target_Interactome.CGS_L2	Number of two-hop interactions of the target gene with known cancer genes in Interactome
	Target_GO_cell.death	Whether the target gene belongs to a Gene Ontology term related to cell death
	Target_GO_cell.differentiation	Whether the target gene belongs to a Gene Ontology term related to cell differentation
	Target_GO_cell.proliferation	Whether the target gene belongs to a Gene Ontology term related to cell proliferation
	Target_GO_mitototic.cell.cycle	Whether the target gene belongs to a Gene Ontology term related to mitotic cell cycle
	Target_DEG_score	Differential expression of the target gene between cancer and normal tissues*
	Target_Expression	Expression level of the target gene in cancer
	Target_CNA	Copy number alteration of the target region
Binding TF features	TF_Cancer_Gene	Whether the regulator is a known cancer gene
	TF_HumanNet.CGS_L1	Number of one-hop interactions of the regulator with known cancer genes in HumanNet
	TF_HumanNet.CGS_L2	Number of two-hop interactions of the regulator with known cancer genes in HumanNet
	TF_Interactome.CGS_L1	Number of one-hop interactions of the regulator with known cancer genes in Interactome
	TF_Interactome.CGS_L2	Number of two-hop interactions of the regulator with known cancer genes in Interactome
	TF_GO_cell.death	Whether the regulator belongs to a Gene Ontology term related to cell death
	TF_GO_cell.differentiation	Whether the regulator belongs to a Gene Ontology term related to cell differentation
	TF_GO_cell.proliferation	Whether the regulator belongs to a Gene Ontology term related to cell proliferation
	TF_GO_mitototic.cell.cycle	Whether the regulator belongs to a Gene Ontology term related to mitotic cell cycle
	TF_DEG_score	Differential expression of the regulator gene between cancer and normal tissues*
TFBS features	diff_Pval	Change in the P value of the motif score by the mutation
	avg_Pval	Average of the P values of the motif scores with and without the mutation
	TF_sign	Gain or loss of the TF binding site by the mutation
Genomic features	Dist_GWAS_2D	Chromosomal distance from the mutation to the closest cancer GWAS locus
	Early.to.late_Rate	Replication timing represented as the early-to-late ratio, (G1B+S1) /(S4+G2)**
	PhastCons	Evolutionary conservation of the underlying sequences as meaured by PhastCons
Cancer epigenome features	Dnase	DNase peak score in the cancer cell line
	H3k20me1	H3K20me1 modification peak score in the cancer cell line
	H3k27ac	H3K27ac modification peak score in the cancer cell line
	H3k27me3	H3K27me3 modification peak score in the cancer cell line
	H3k36me3	H3K36me3 modification peak score in the cancer cell line
	H3k4me1	H3K4me1 modification peak score in the cancer cell line
	H3k4me2	H3K4me2 modification peak score in the cancer cell line
	H3k4me3	H3K4me3 modification peak score in the cancer cell line
	H3k79me2	H3K79me2 modification peak score in the cancer cell line
	H3k9ac	H3K9ac modification peak score in the cancer cell line
	H3k9me3	H3K9me3 modification peak score in the cancer cell line
Cell-of-origin epigenome features	Orig_Dnase	DNase peak score in the cells of origin
	Orig_H3k20me1	H3K20me1 modification peak score in the cells of origin
	Orig_H3k27ac	H3K27ac modification peak score in the cells of origin
	Orig_H3k27me3	H3K27me3 modification peak score in the cells of origin
	Orig_H3k36me3	H3K36me3 modification peak score in the cells of origin
	Orig_H3k4me1	H3K4me1 modification peak score in the cells of origin
	Orig_H3k4me2	H3K4me2 modification peak score in the cells of origin
	Orig_H3k4me3	H3K4me3 modification peak score in the cells of origin
	Orig_H3k79me2	H3K79me2 modification peak score in the cells of origin
	Orig_H3k9ac	H3K9ac modification peak score in the cells of origin
	Orig_H3k9me3	H3K9me3 modification peak score in the cells of origin

Features contributing to the efficacy of recurrence classifiers

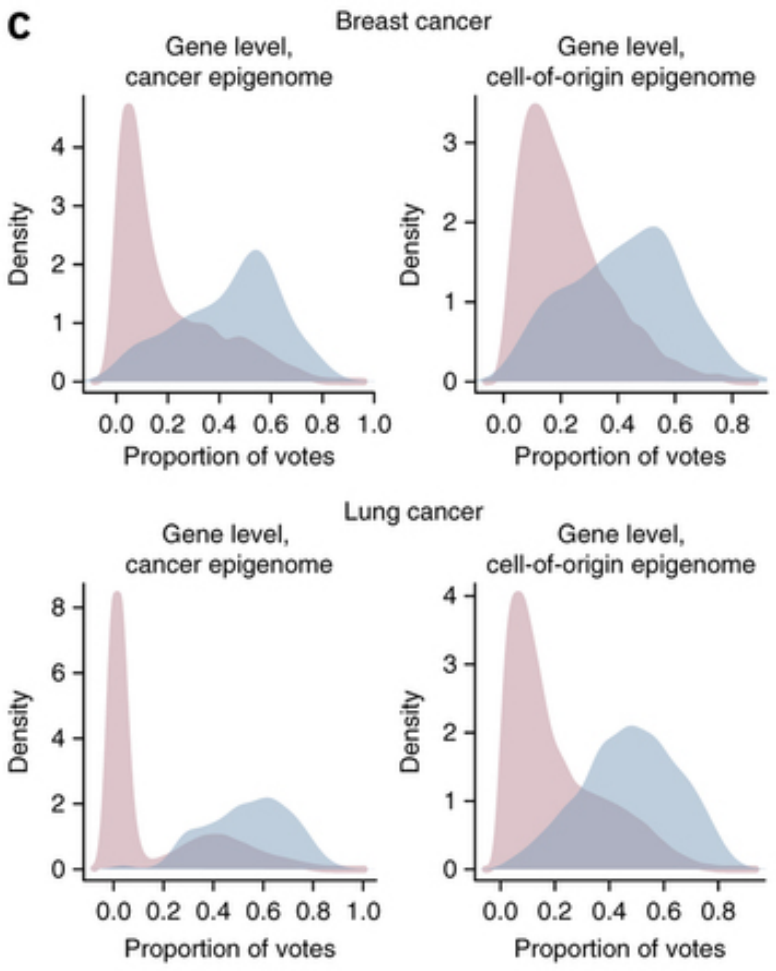


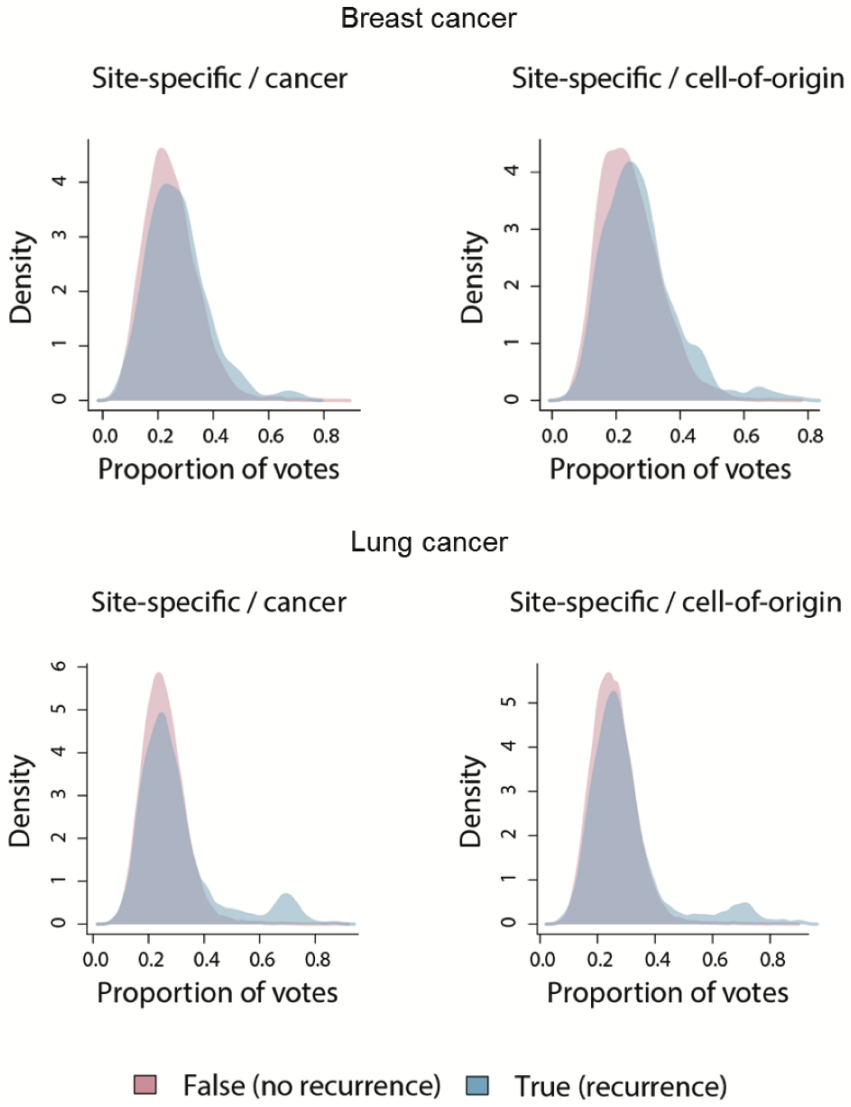


Gene-level recurrence model outperform site-specific model.

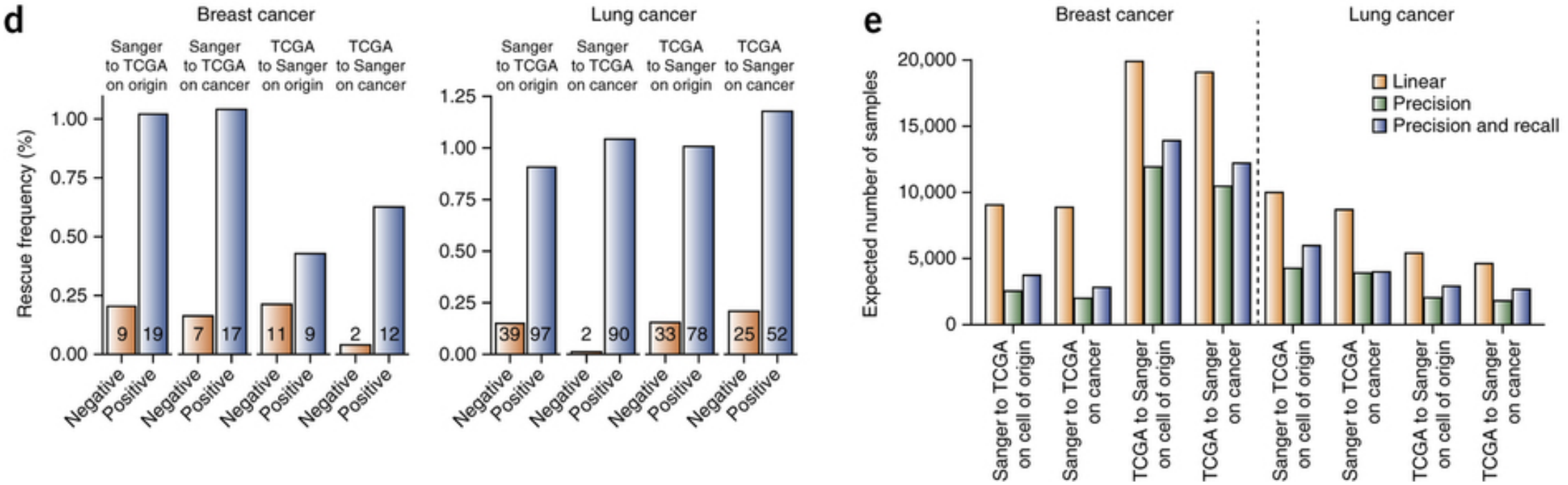
Important predictive features include cancer/cell-of-origin epigenome and TFBS.

Features contributing to the efficacy of recurrence classifiers





Rescue frequency and power analysis to identify recurrence



1.0 Rescue fraction with addition of 100 samples.

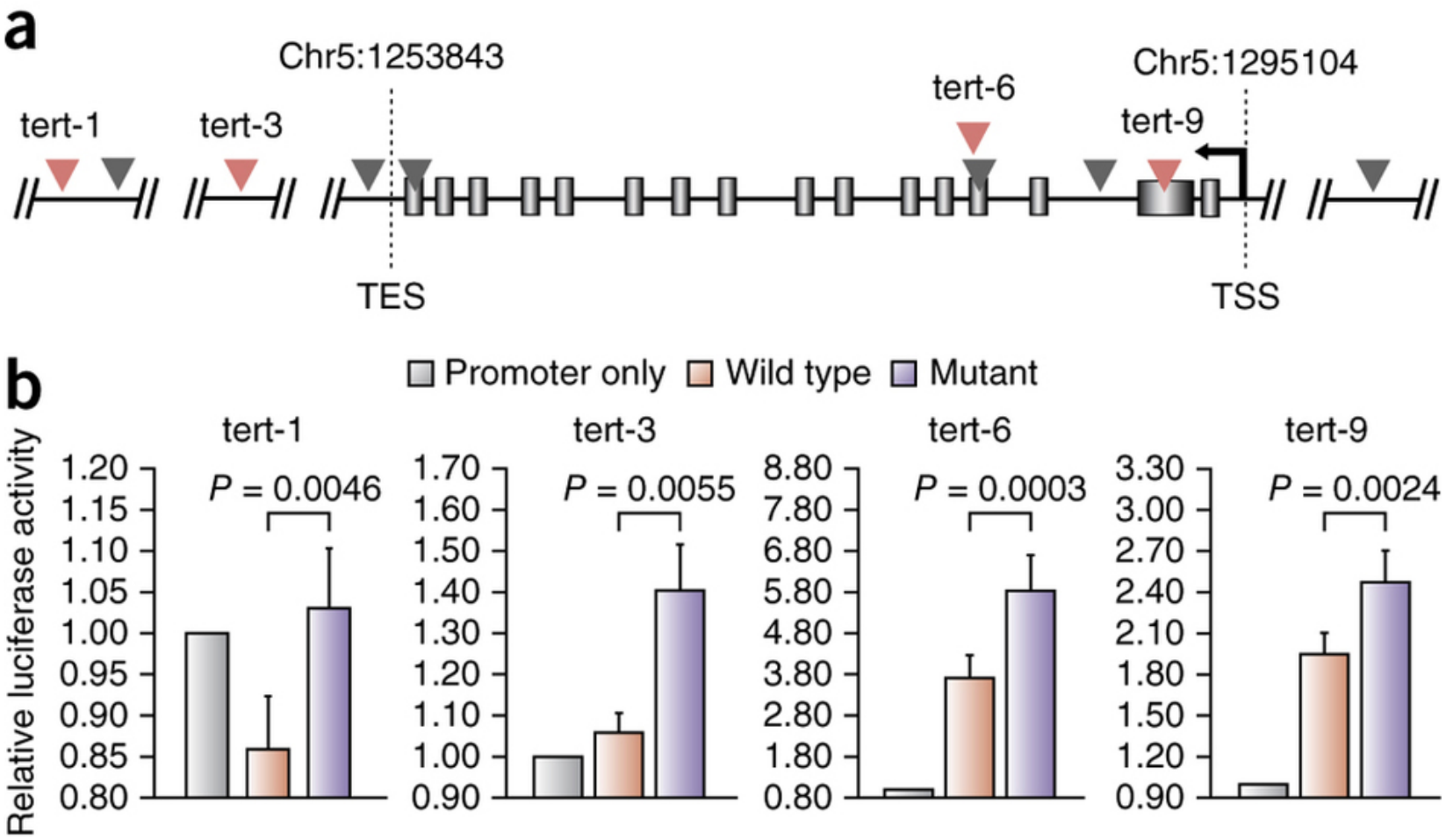
2k-15K samples per tumor type are required to fully rescue recurrent mutations.

Correlation of gene-level mutation burden and expression level

Survival gene	Survival P value (expression-survival correlation)	Correlation between mutation burden and expression level	Correlation P value
SNORA48	0.027	0.418	5.59E-05
SNORA14B	0.014	0.386	2.19E-04
LAMP1	0.005	0.34	1.28E-03
MAPK12	0.048	0.332	1.70E-03
PRSS54	0.044	0.329	1.83E-03
PITPNM2	0.033	0.328	1.92E-03
LOC100129935	0.044	0.328	1.61E-03
APCS	0.035	0.314	3.06E-03
VCX3B	0.029	0.311	2.86E-03
CARS2	0.011	0.304	4.16E-03
RPS6KA1	0.007	0.304	4.20E-03

Mutation burden of survival-associated genes correlates with gene expression level.

Functional validation of predicted noncoding TERT mutations



Luciferase assay based validation for 10 out of 18 TERT diver mutation.

4 mutations showed significant increase in the luciferase activity.

Discussion

This method is heavily reliant on chromatin interaction data.

Need to have accurate interaction data

Absence of complete chromatin-interaction data

The predictive power is still limited similar to site specific driver identification approaches.

Need more samples to identify recurrent mutations

Rescue frequency poor for False negative mutations