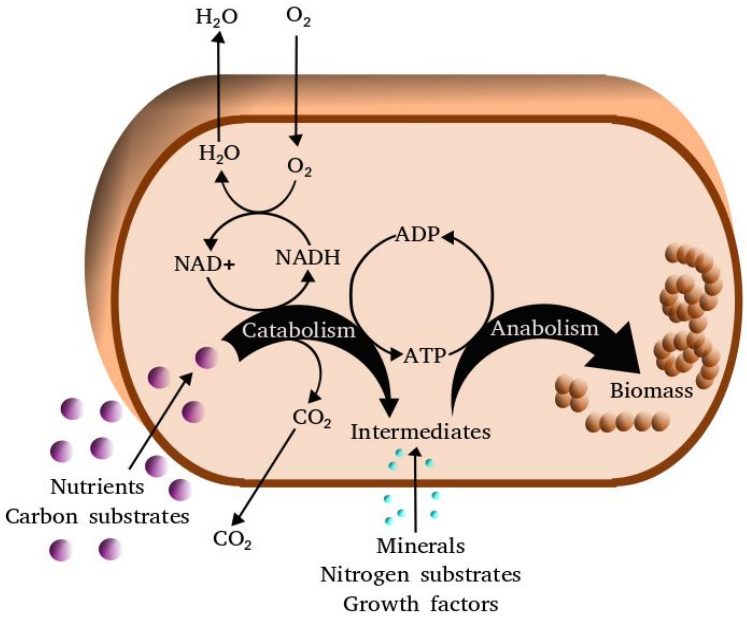
Gene expression is a poor predictor of the metabolite abundance in cancer cells

Huaping Li
School of Biomedical Sciences, HKU
2022.03

Background

Metabolism refers to all biochemical processes taking place in live cells and organisms.

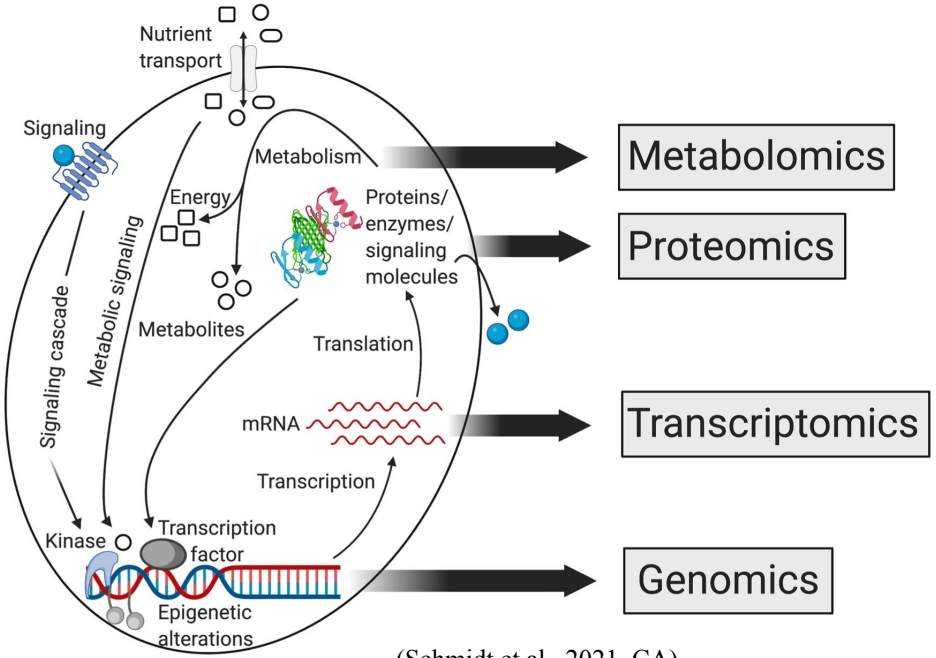
Metabolites are the by-products or products of cellular metabolic activities and are usually shown as small molecules, such as amino acids, lactic acid, lipids, nucleotide, vitamins, ethanol, and glycerol, etc.



Simplified view of the cellular metabolism

Background

- Cancer cells tend to modulate their metabolism to meet the high anabolic requirements of rapid proliferation, and therefore promotes cancer progression(DeBerardinis et al., 2007; Liberti & Locasale, 2016)
- The global changes in metabolite levels has been reported as a hallmark of cancer(Faubert et al., 2020).



(Schmidt et al., 2021, CA)

Research Aim

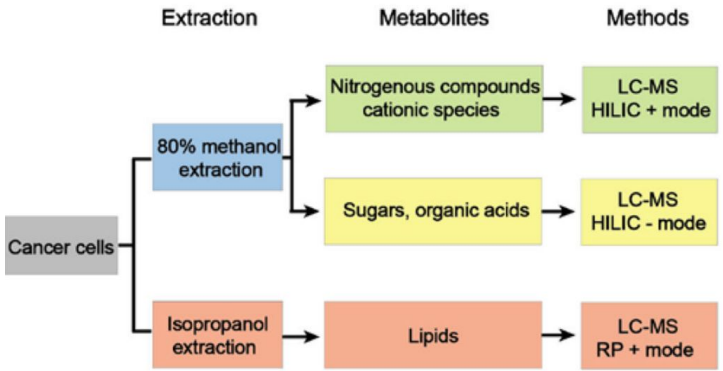
Metabolomics
less resources
hard to capture & measure

?

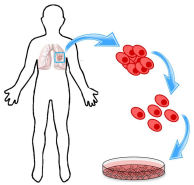
Transcriptomics
huge resources
easily detect & profile

Database

CCLE Cancer Cell Line Encyclopedia

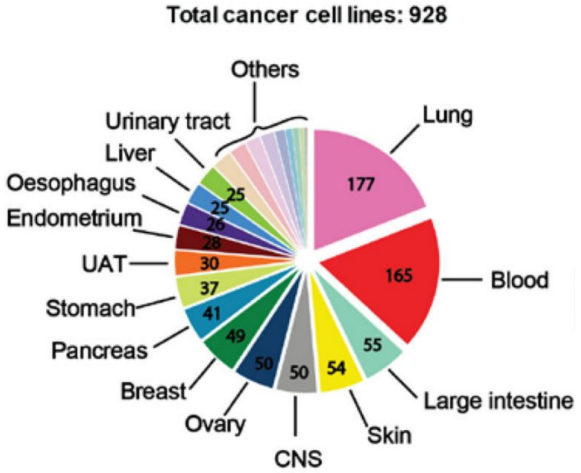


large-scale genetic characterization of ~1000 cancer cell lines



Metabolomics:
225 metabolites * 928 cell lines (29 cancer types)

Transcriptomics:
28k genes * 1019 cell lines



(Li et al., 2019, Nature Medicine)

Analysis pipeline

CCLE metabolic profiling
trend normalization

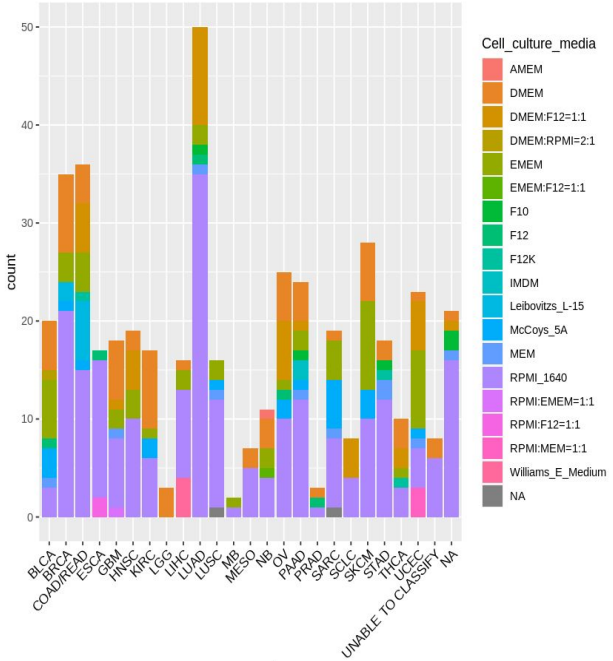
CCLE transcriptomic profiling

-keep coding region only
-remove features with > 10% missing/0 values
-log transfer
-quantile normalization
-(scale)

sample filtering
454 adherent-growth solid cancer human cell lines * 225 metabolites

merge

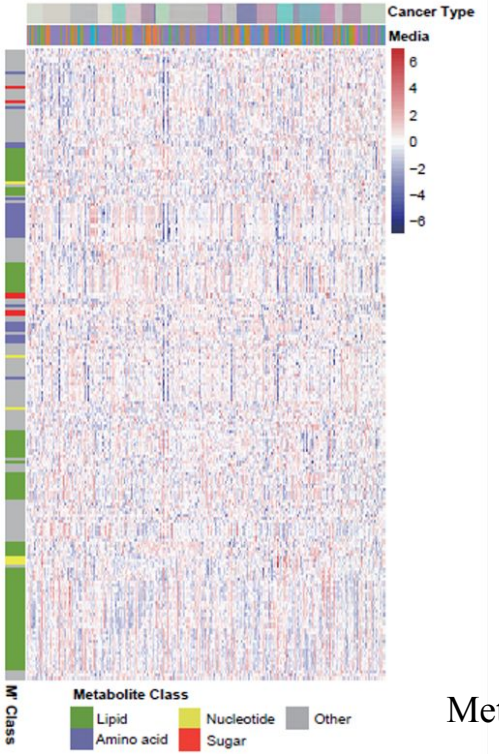
data preprocessing
454 cell lines * 13,403 coding genes expression profiles

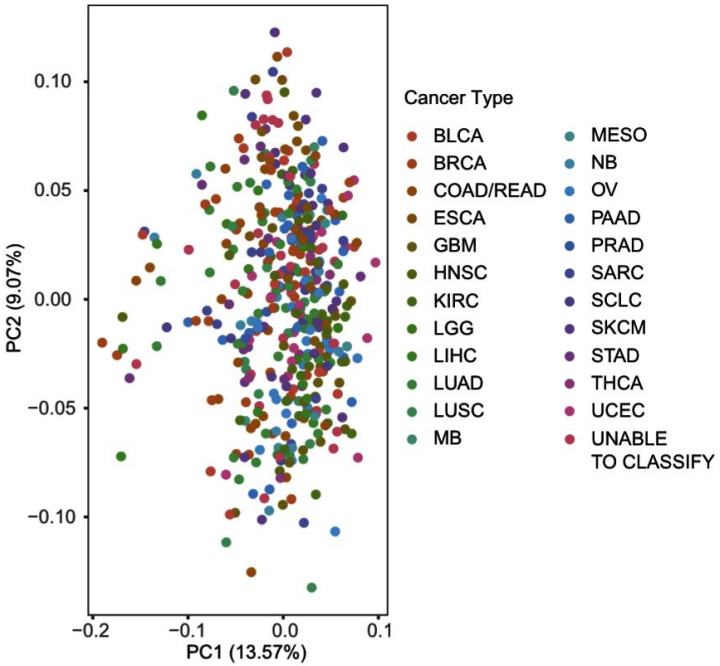


Basic statistics analysis:
PCA, Hierarchical clustering, CV

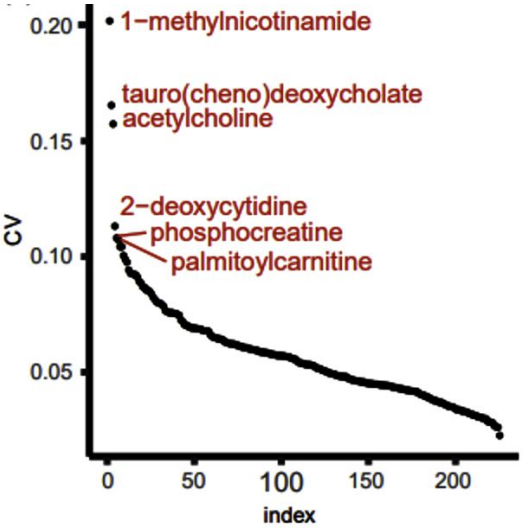
Metabolism heterogeneity

- “Metabolic phenotypes in tumors are both heterogeneous and flexible” (Kim & DeBerardinis, 2019)





Metabolism heterogeneity among cancer cell lines



Coefficient of variation(CV) of each metabolite.[CV = SD/mean]

Analysis pipeline

CCLE metabolic profiling
trend normalization

CCLE transcriptomic profiling

-keep coding region only
-remove features with > 10% missing/0 values
-log transfer
-quantile normalization
-(scale)

sample filtering
454 adherent-growth solid cancer human cell lines * 225 metabolites

merge

data preprocessing
454 cell lines * 13,403 coding genes expression profiles

(basic statistics analysis)

Models training

A) Covariates model
B) Global gene-based prediction model
C) Metabolism gene-based prediction model

meraged CCLE profiling (n=454)

media

	SampleType	tcga_code
Cell line Sample 1	tissue_cell	NB
Cell line Sample 2	tissue_cell	NA
Cell line Sample 3	tissue_cell	OV
Cell line Sample 4	tissue_cell	BLCA
Cell line Sample 5	tissue_cell	BRCA
Cell line Sample 6	tissue_cell	KIRC
Cell line Sample 7	tissue_cell	KIRC
Cell line Sample 8	tissue_cell	HNSC
Cell line Sample 9	tissue_cell	SKCM
Cell line Sample 10	tissue_cell	GBM

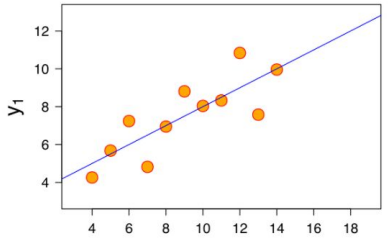
metabolite

aconitate	adenine	adipate	a-ketoglut	AMP	citrate
5.582619	5.857043	6.550616	5.919698	5.310705	6.162948
6.222526	6.218087	5.788157	5.764736	6.038368	6.241364
5.804038	5.245122	5.946702	5.826242	5.71312	5.836146
5.516183	5.245486	5.905076	6.010646	5.625985	5.657373
5.516094	5.140008	5.788156	5.936196	5.607163	5.689864
5.845154	5.574732	5.782307	6.440875	5.672406	5.782251
5.53609	5.661936	5.740511	6.107778	5.641792	5.490199
5.572837	5.771789	5.874517	5.800793	5.751495	5.77687
6.496666	5.814723	5.597575	6.257081	6.152208	6.405899
6.070379	5.713437	5.878509	5.917343	5.850285	5.977321
6.177149	6.09626	6.081746	5.911435	5.602765	6.027855
5.837388	5.89852	6.045058	6.029349	5.963923	5.930403

gene expression

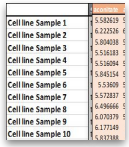
SCYL3	C1orf112	CFH	FUCA2	GCLC	NFYA	STPG1	NIPAL3	LAS1L	ENPP4
0.475868	0.951326	0.467636	0.581929	1.659403	1.913952	-0.41182	0.483405	1.847822	0.214526
0.336572	1.100636	-1.38864	1.721416	2.163875	0.996204	0.587333	1.072303	1.537776	0.624914
0.468839	1.080463	1.000494	1.862707	1.670757	1.294999	1.029017	1.352454	1.516014	0.518507
0.47769	1.30209	0.316076	1.765064	1.532204	1.187153	0.851332	1.229902	1.632266	-0.5067
0.508408	1.182788	-0.41014	1.849298	1.515505	1.156239	1.01043	1.492947	1.706607	1.193058
0.720225	1.343995	1.545451	2.13593	1.099995	1.296737	1.02181	1.418132	1.704572	0.580171
0.643239	1.096592	1.703187	1.960155	1.294478	1.280444	0.675187	0.978908	1.727506	0.866826
0.462219	0.779801	1.736512	1.525971	1.789479	1.305551	0.721949	1.685478	1.330817	-3.26826
0.459289	1.181995	-1.33928	1.879775	1.413074	1.233807	0.690548	1.532992	1.618933	-0.07702
0.484399	1.316908	-0.35525	1.711942	0.803481	1.190435	0.081424	0.725688	1.537202	-1.80014
0.566625	0.806211	-0.44789	1.747438	1.383525	0.993533	0.630013	1.105358	1.89864	0.651833
0.556308	0.822657	-2.30386	1.575902	1.31228	1.248692	0.501592	1.052845	1.653969	0.786244
0.291977	0.916635	1.744566	2.005164	1.131656	1.238653	0.570494	0.785322	1.506695	-4.32645
0.275262	0.782733	0.976961	1.804753	0.917639	1.04776	0.535819	0.912507	1.505699	-0.02874
0.341553	0.961941	1.061196	1.857764	1.337054	1.461578	0.57148	0.911518	1.85624	0.80519

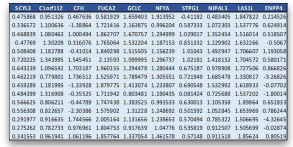
Linear & Lasso regression



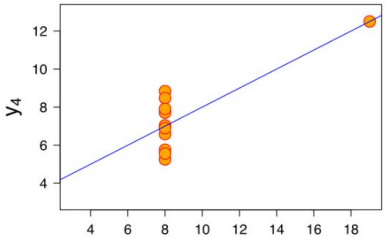
"In statistics, **linear regression** is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables"

$$y \sim w1x1 + w2x2 + \ldots + wnxn$$



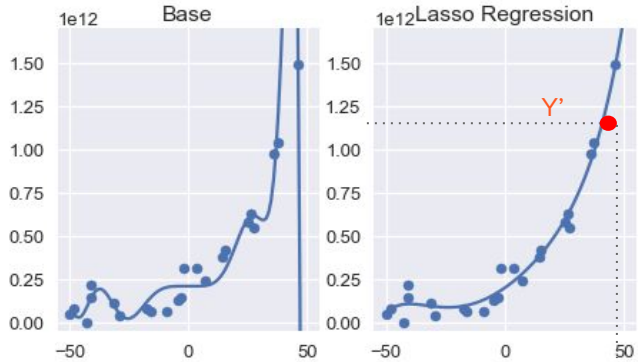


multicollinearity (eg.x1=logx2)
feature bias



" **Lasso** is a modification of linear regression, where the model is penalized for the sum of absolute values of the weights. It performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model."

$$\frac{1}{2m}\sum_{i=1}^{m}(y - Xw)^2 + alpha\sum_{j=1}^{p}|w_j|$$ penalize

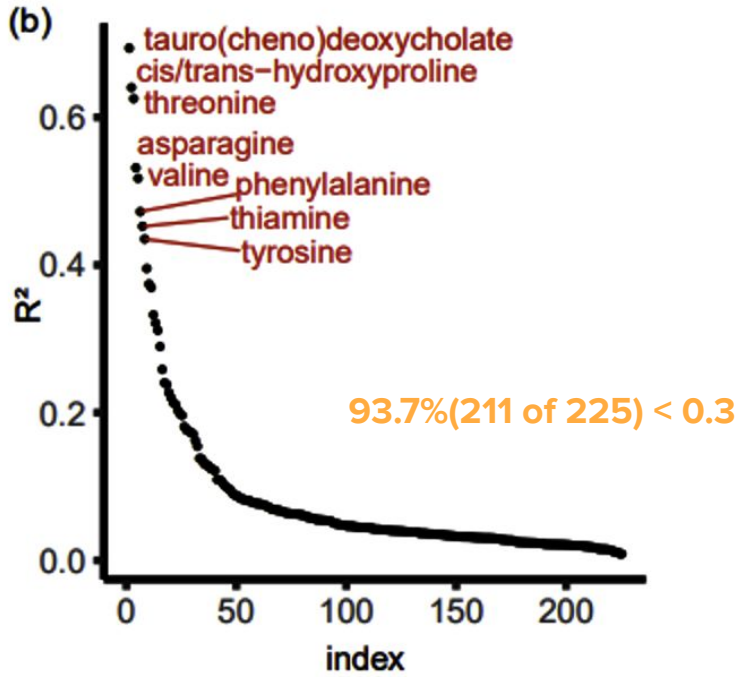


Covariates analysis

10-fold cv, step-wise, linear regression models to check the influence from media and clinical traits

lm: y ~ 55 medium componunts + clinical information

(inferred ethnicity, patient age, pathology, cancer type, mutation rates and doubling time (hr))

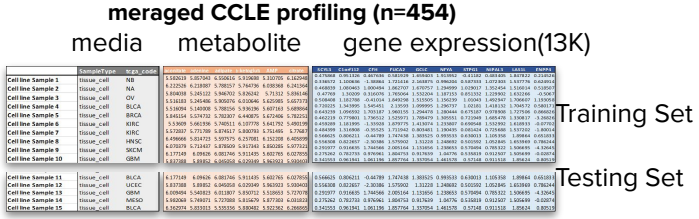


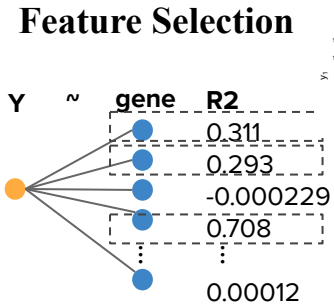
Metabolite	R2
taurodeoxycholate/taurochenodeoxycholate	0.693
cis/trans-hydroxyproline	0.64
threonine	0.625
asparagine	0.531
valine	0.517
phenylalanine	0.472
thiamine	0.452
tyrosine	0.435
lysine	0.395
histidine	0.374
isoleucine	0.369
arginine	0.332
niacinamide	0.321
tryptophan	0.311

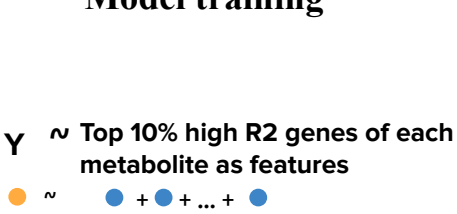
Global gene-based prediction models

linear model: y ~ cancer type + media type + gene

Lasso model: y ~ selected genes







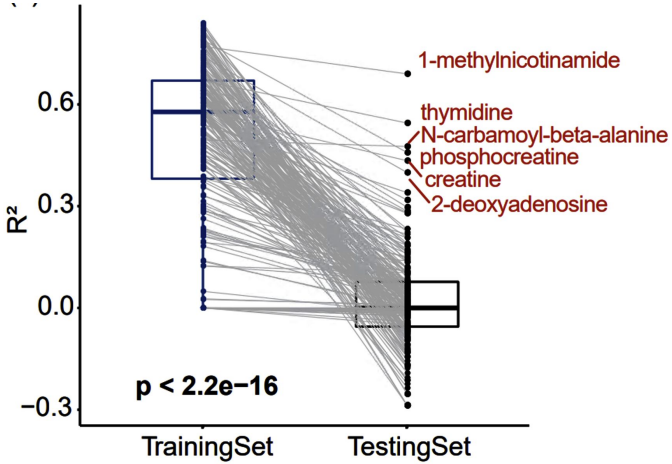
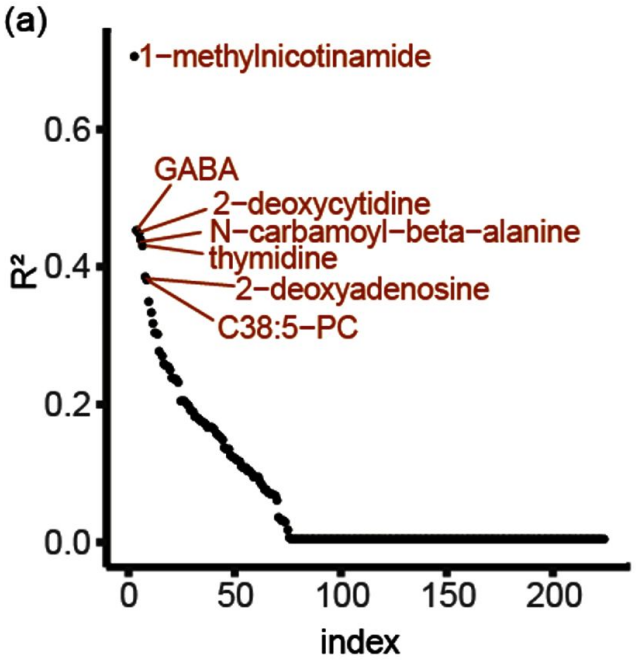


Table 1. Metabolites with high R-squared(R^2) from the prediction models.

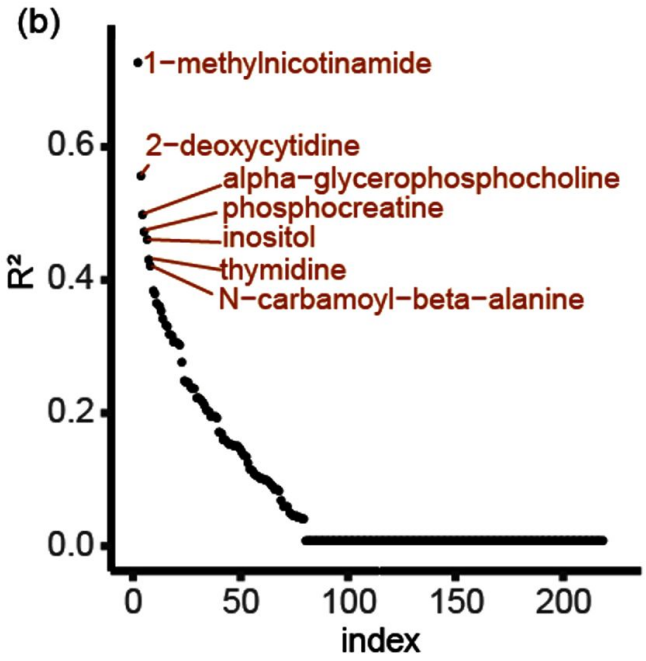
	Global gene-based LASSO model		covariates linear model^
Metabolite	**TrainingSet_R^2**	**TestingSet_R^2**	**R^2**
1-methylnicotinamide	0.771	0.69	0.0612
thymidine	0.67	0.545	0.0622
N-carbamoyl-beta-alanine	0.495	0.476	0.0625
Phosphocreatine	0.79	0.458	0.0217
creatine	0.575	0.434	0.0479
2-deoxyadenosine	0.741	0.399	0.0395

Metabolism gene-based prediction models

Lasso regression: Y ~ all metabolism genes



554 metabolism genes from **SMPDB**



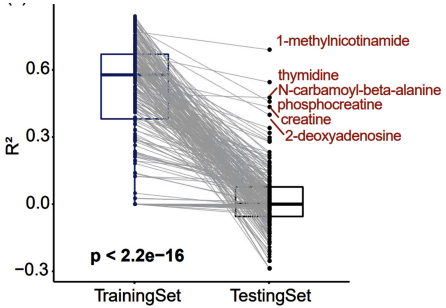
1701 metabolism genes **Reactome**

Summary

Gene expression alone cannot be used to estimate metabolism activity

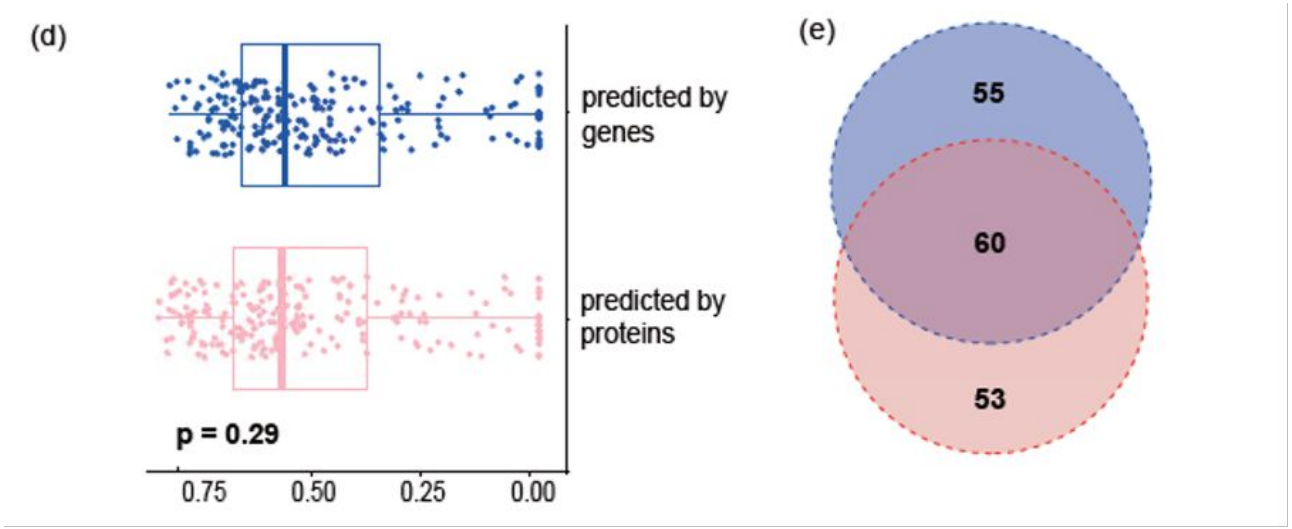
- Even when accounting for cell culture conditions or cell lineage in the model, few metabolites could be accurately predicted.

Validated in proteomics level



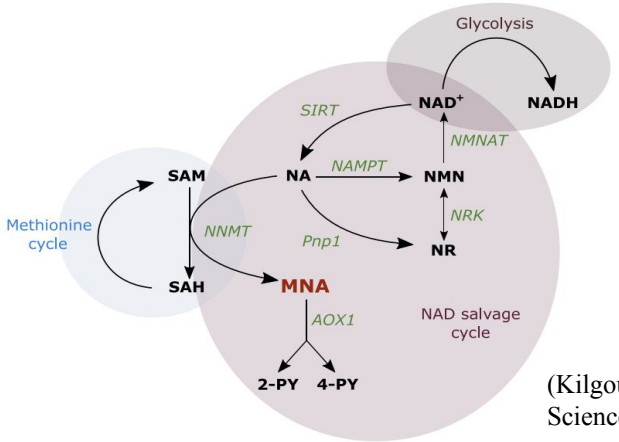
Global gene-based prediction models

Proteomics-based prediction models

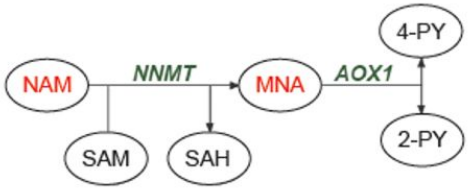


d) R2 distribution of global gene-based Lasso models (blue) and proteomics-based Lasso models (red). e) Venn diagram of common predictable metabolites (Top 50% high R2 metabolites in training set) between global gene-based (blue) and proteomics-based (red) Lasso models.

Using associated metabolites in prediction



(Kilgour et al., 2021, Science Advances)



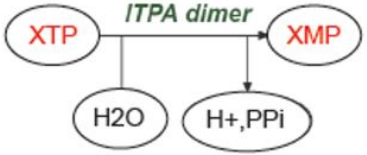


Table 2. Prediction ability change with reactant metabolite level

Prediction Model	R^2 in testing set	Reactant Coefficient(rank)
MNA ~ selected genes	0.683	\
MNA ~ selected genes + NAM	0.683	\
XMP ~ selected genes	-0.422	\
XMP ~ selected genes + XTP	0.376	0.57(1)
phosphocreatine ~ selected genes	0.516	\
phosphocreatine ~ selected genes + creatine	0.646	0.902(1)
thymidine ~ selected genes	0.525	\
thymidine ~ selected genes +thymine	0.649	0.399(1)

MNA: 1-methylnicotinamide; NAM: Nicotinamide; XMP: xanthosine; XTP: xanthine.

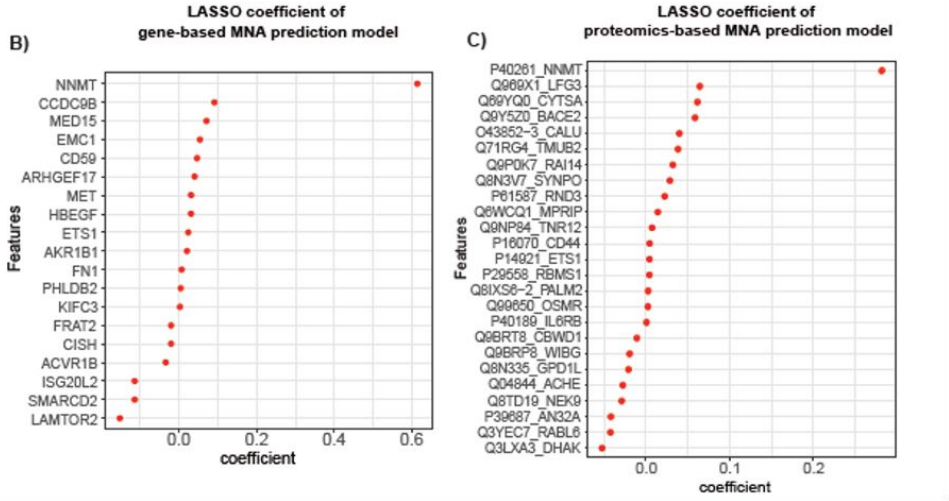
Summary

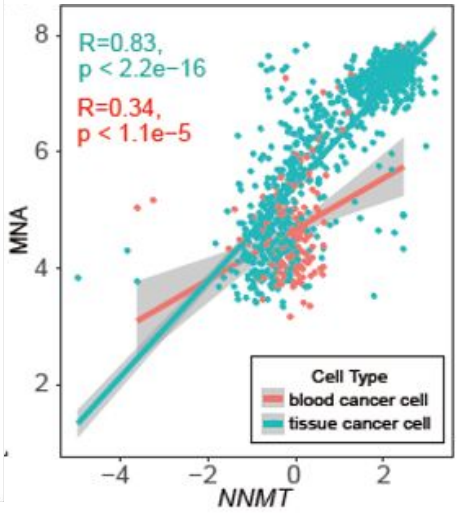
Gene expression alone cannot be used to estimate metabolism activity

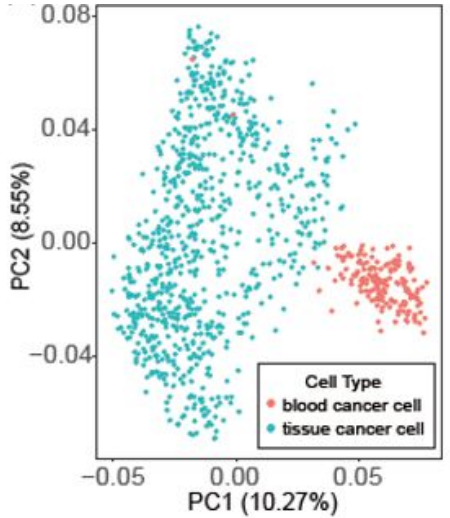
- Even when accounting for cell culture conditions or cell lineage in the model, few metabolites could be accurately predicted.
- In some cases, the inclusion of the upstream and downstream metabolites are needed in prediction

NNMT correlated well with MNA in CCLE cohort

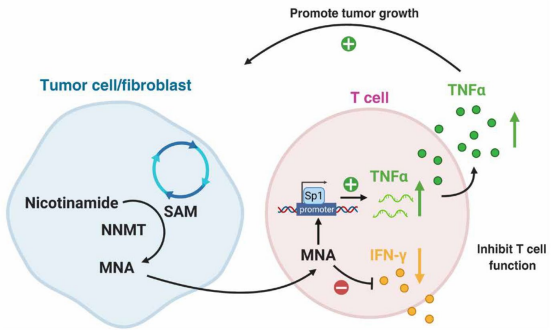
(928 cell lines)







1. Solid cancer cells and blood cancer cells show substantially different metabolic profiles
2. Some metabolites maybe cancer-type specific - considering cell type and even cancer purity in metabolites prediction.



(Kilgour et al., 2021, Science Advances)

Summary

Gene expression alone cannot be used to estimate metabolism activity

- Even when accounting for cell culture conditions or cell lineage in the model, few metabolites could be accurately predicted.
- In some cases, the inclusion of the upstream and downstream metabolites are needed in prediction
- The metabolism heterogeneity and sample purity enlarge the challenge of metabolite prediction.

Summary

Gene expression alone cannot be used to estimate metabolism activity

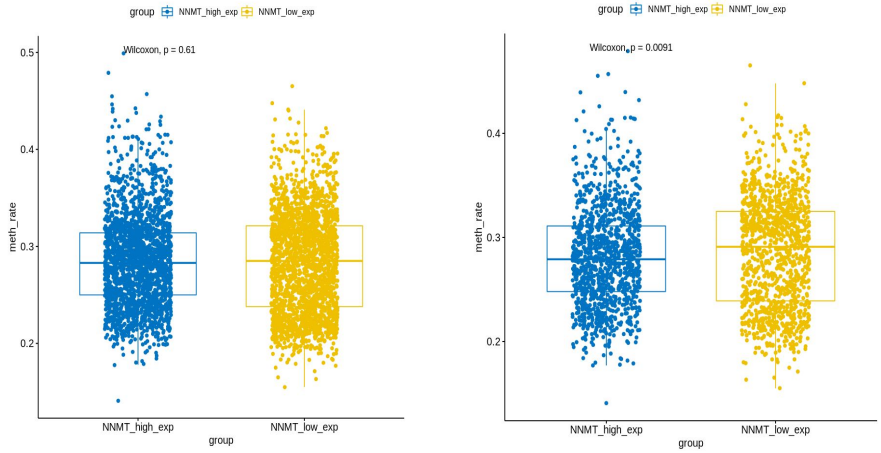
- Even when accounting for cell culture conditions or cell lineage in the model, few metabolites could be accurately predicted.
- In some cases, the inclusion of the upstream and downstream metabolites are needed in prediction
- The metabolism heterogeneity and sample purity enlarge the challenge of metabolite prediction.

Why do most of the prediction models don't perform well?

- Gene expression may not reflect the activity of some metabolism associated enzymes. The enzyme activities could be regulated allosterically and post-translationally;
- The bi-directional effect of metabolism and gene expression, making the metabolite level much more unpredictable.
- There are many flexible transcriptional solutions to regulate the metabolism activities.

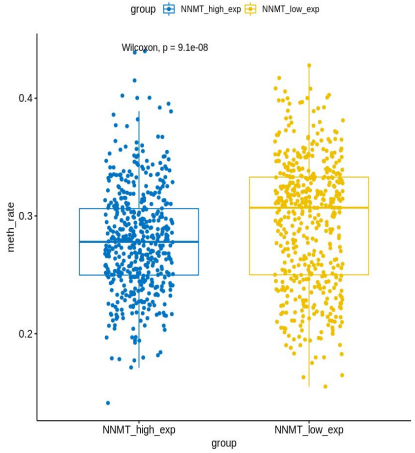
TCGA bulk samples - NNMT vs methy rate,purity

150k detected “Island” methylation sites

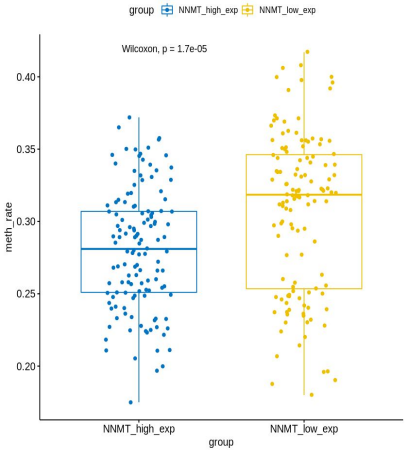


purity = 0.8,n=4k+

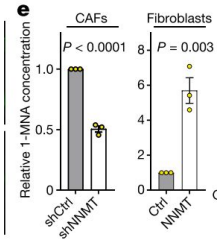
purity = 0.90,n=2k+

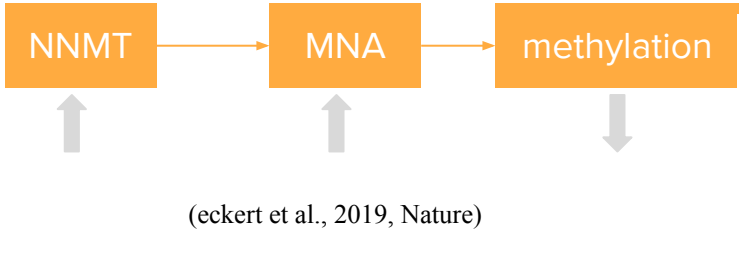


purity = 0.95,n=1k+

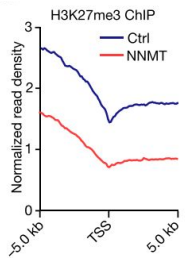


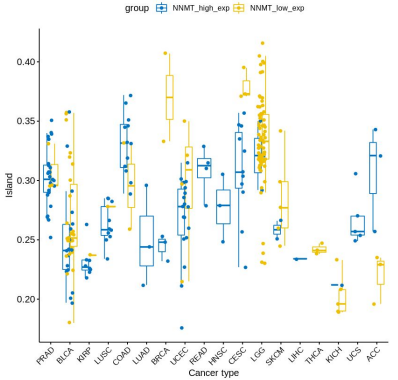
purity = 0.98,n=200+





(eckert et al., 2019, Nature)





Future Plans

1. Study the possible mechanism of methylation regulation by MNA and NNMT in tumors.

2. See if there are any metabolism signals associated with anti-cancer drug sensitivity.

Pharmacologic data:
4686 drugs * 579 cell lines (Corsello et al., 2020, Nat Cancer)
264 anti-cancer drugs * 1001 cell lines (Iorio et al., 2016, Cell)

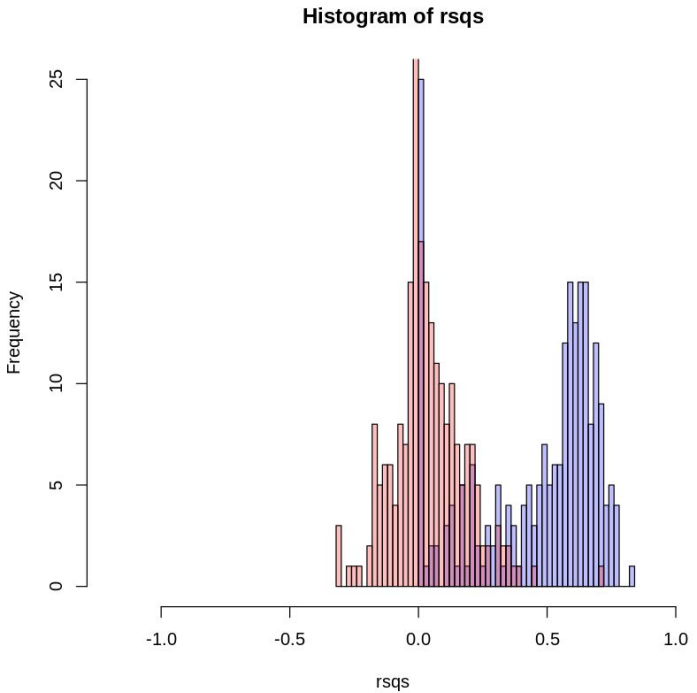
THANKS

bk pages

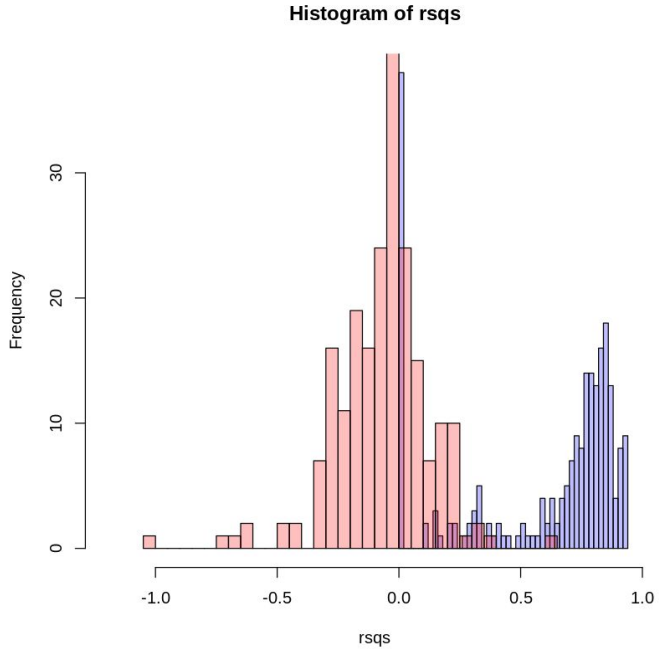
double check

majority of the metaboilites are not
influenced by common cell culture media.

multiple media , normed
cancer cell lines (**n=454**)



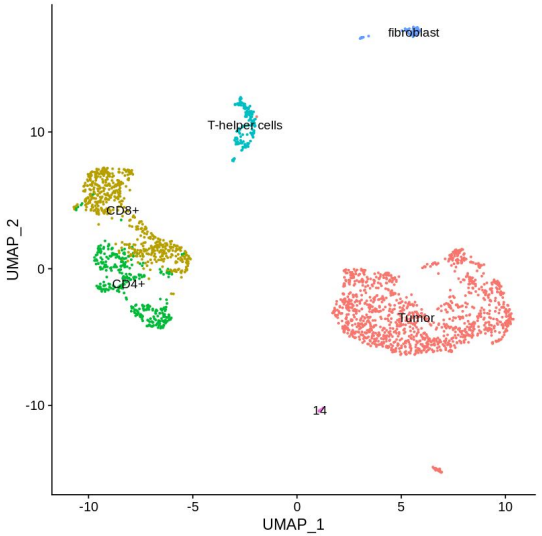
RPMI_1640 ,normed
cancer cell lines (**n=226**)

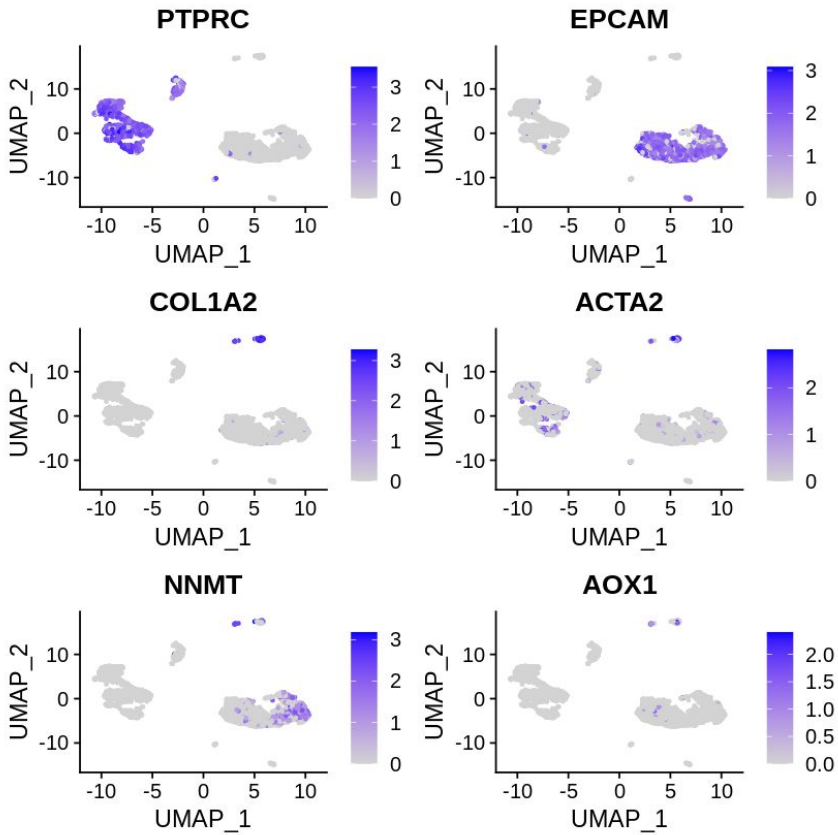


ovarian Single cell cohort - NNMT distribution in cells

"EpCAM+ (tumor) cells"
EPCAM (pan-epithelial cell marker),
COL1A2 and ACTA2 (fibroblast markers)
PTPRC (CD45+ ,immune cell mariker)

blue: T-helper cells
yellow: CD8+
green: CD4+
red: Tumor

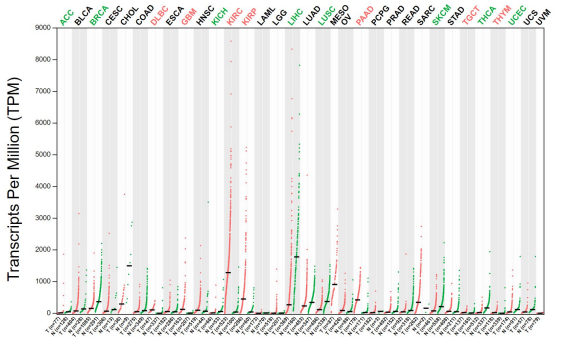




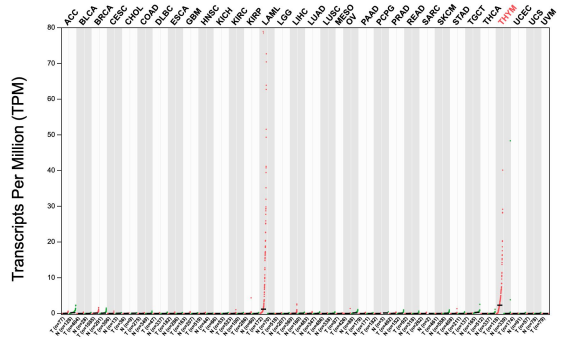
metabolism pathway - gene and metabolite distribution among cancer types

1. NNMT expressed in tumor cell, while most AOX1 not
2. AOX1 highly expressed in LAML tummor cells
3. MNA is not predictable in LAML samples

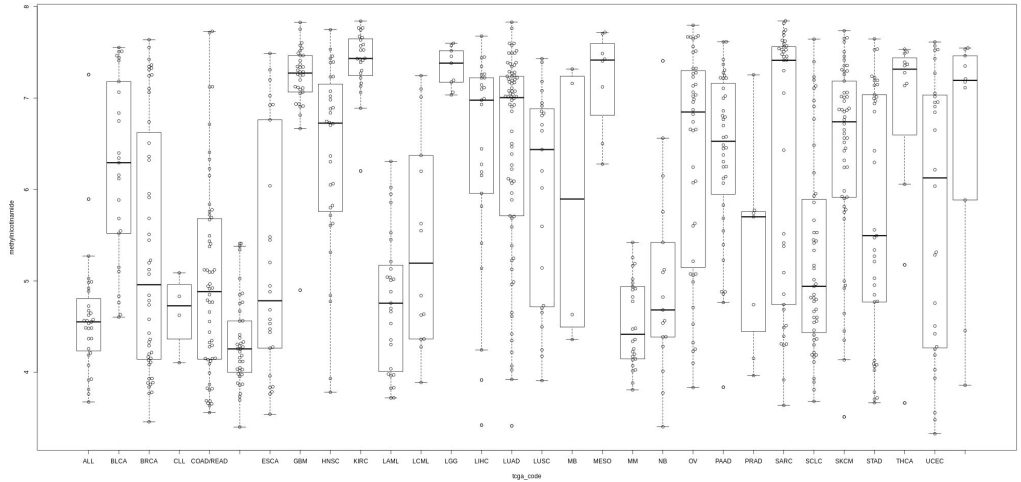
NNMT



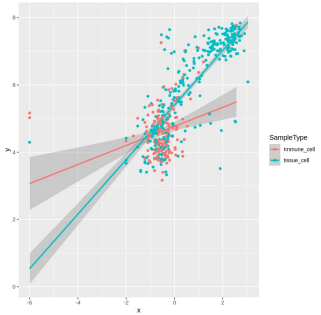
AOX1

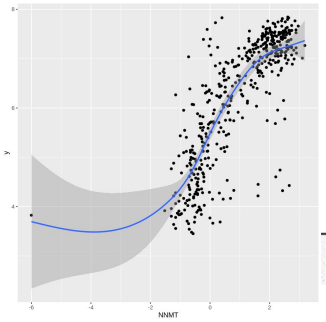


MNA



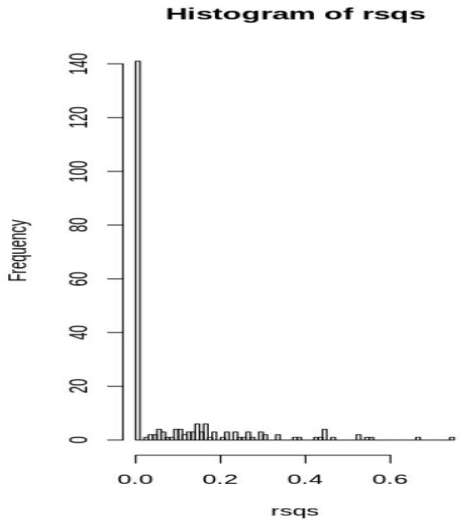
Permutation Test



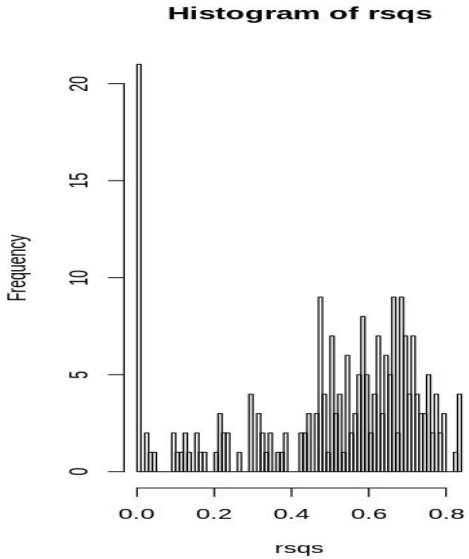


RMSE	R2
<dbl>	<dbl>
0.7009144	0.684

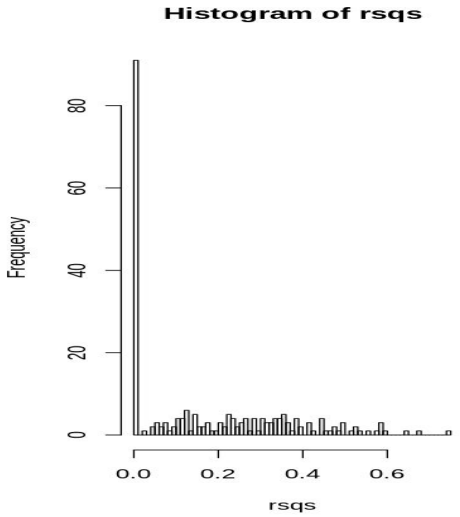
global LASSO y~ 13k genes



lm + LASSO y~selected genes

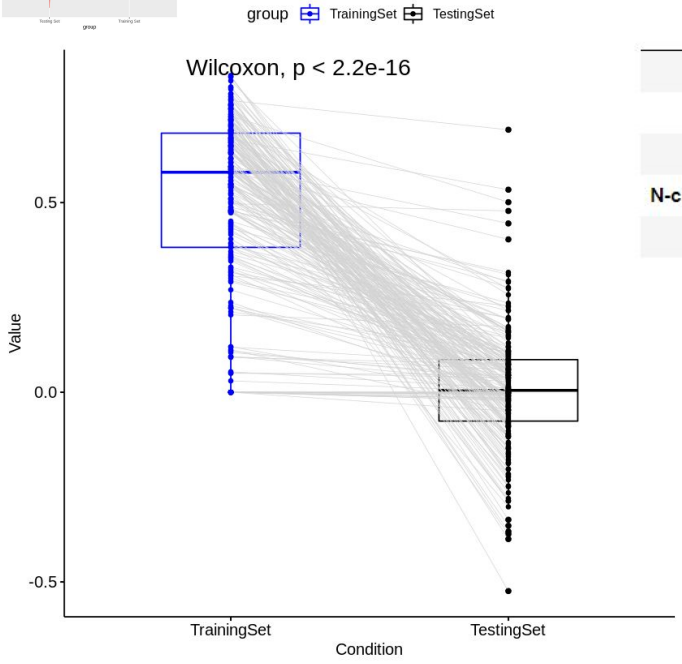


GAM + LASSO y~selected genes



LASSO models training

lm + LASSO y~selected genes



	R2	group
	<dbl>	<chr>
1-methylnicotinamide	0.769	TrainingSet
thymidine	0.666	TrainingSet
phosphocreatine	0.728	TrainingSet
N-carbamoyl-beta-alanine	0.497	TrainingSet
creatine	0.595	TrainingSet
2-deoxyadenosine	0.581	TrainingSet

	R2	group
	<dbl>	<chr>
1-methylnicotinamide	0.745	Training Set
N-carbamoyl-beta-alanine	0.484	Training Set
thymidine	0.497	Training Set
phosphocreatine	0.455	Training Set
2-deoxyadenosine	0.497	Training Set
inositol	0.387	Training Set

GAM + LASSO y~selected genes