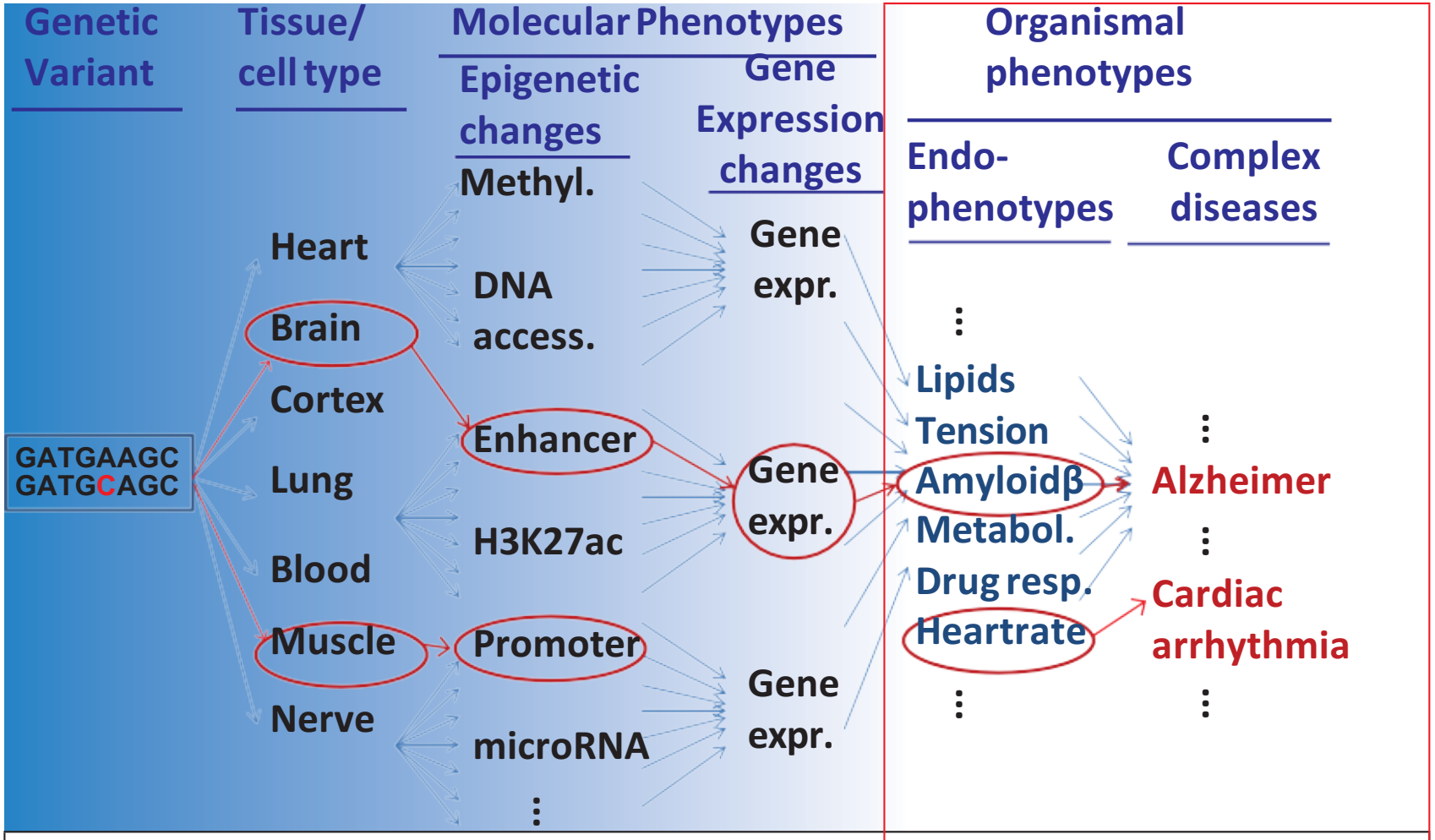
A Bayesian method to infer disease networks and expected phenotypes using electronic health records

Yue Li
Postdoctoral researcher
Kellis Lab
MIT

Dissecting regulatory circuitry of human complex diseases

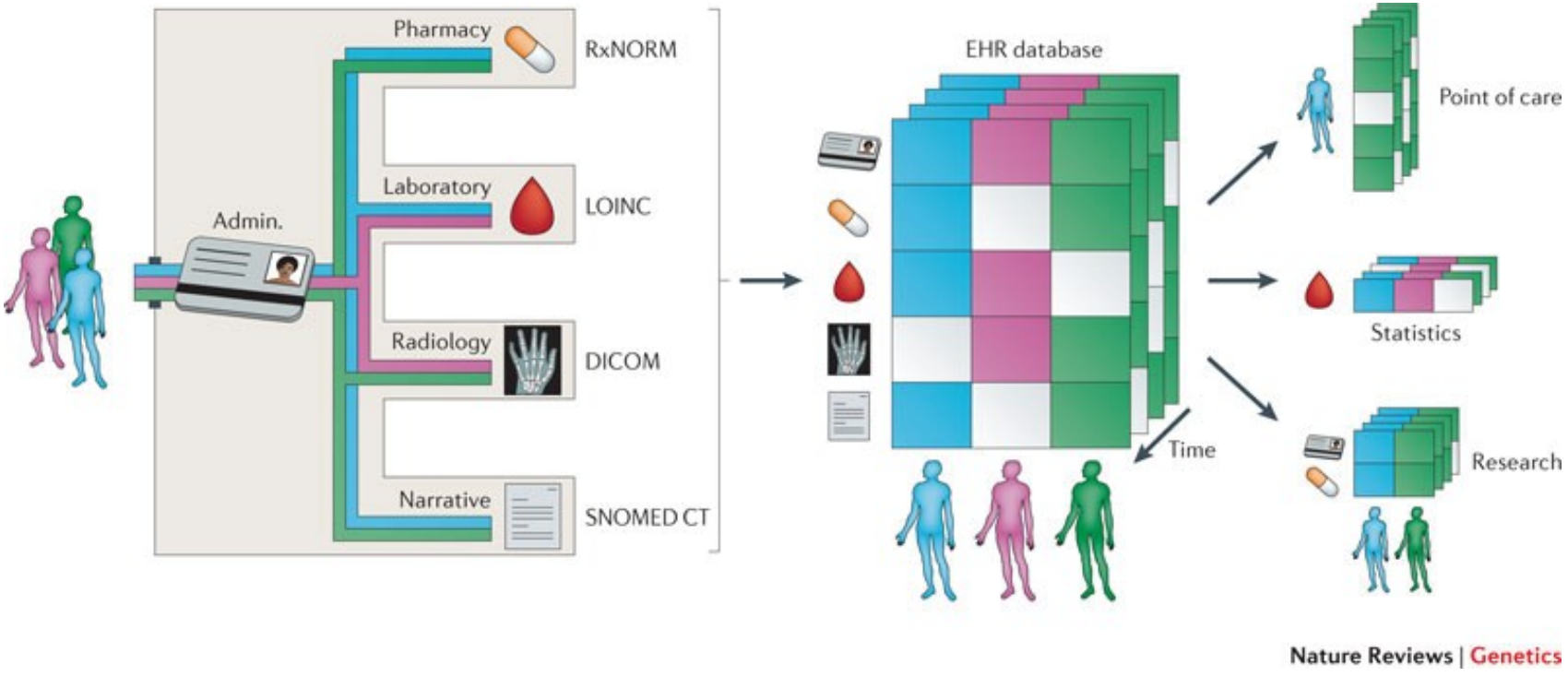


Research outline

Inferring regulatory variants in diseases

Disease networks using electronic health records

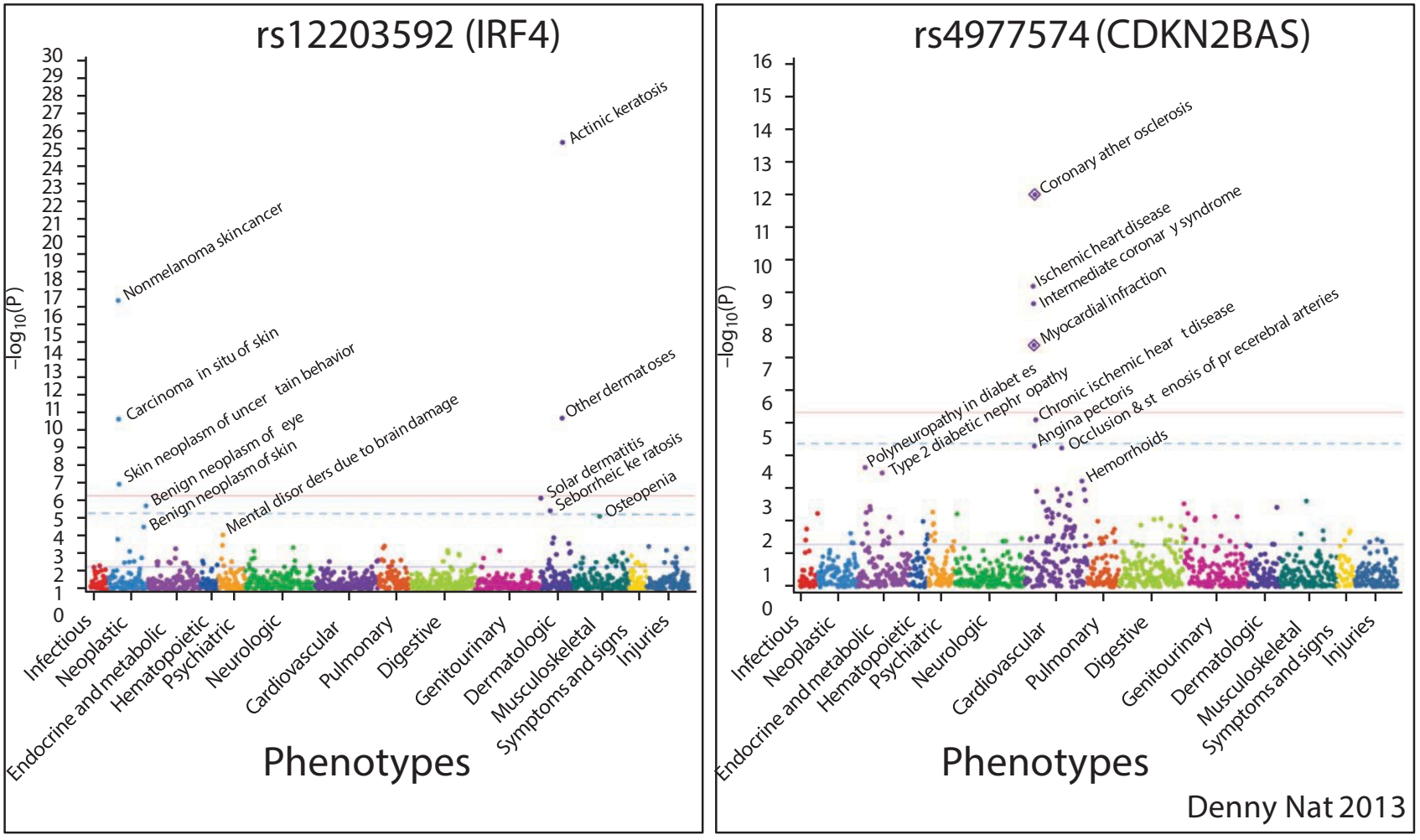
Electronic health records contain rich patient-level data



Jensen et al., Nature Rev. Gen. 2012

- Lab tests: Logical Obs. Identifiers Names & Codes (LOINC)
- Pharmaceutical: Prescription data (RxNorm)
- Imaging: Digital Imaging and Comm. in Medicine (DICOM)
- Phenotype: International Classification of Disease-9 (ICD-9)

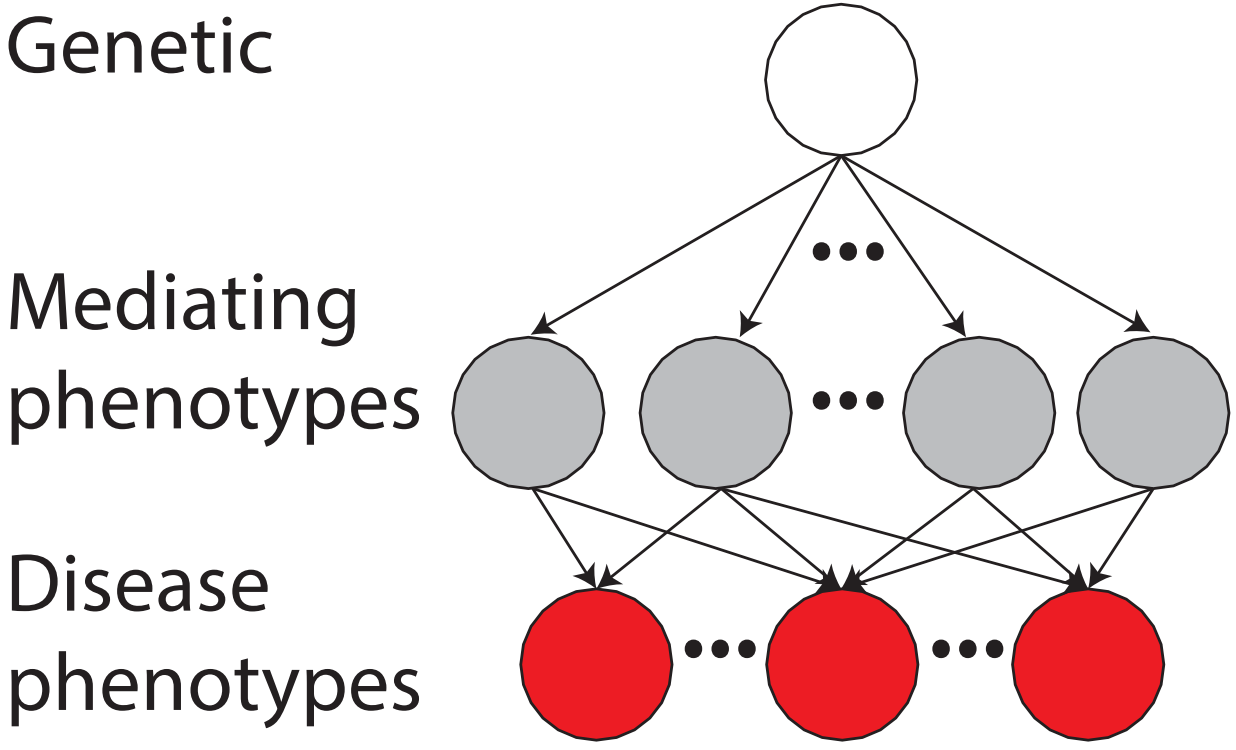
Phenome-wide association studies with genetic information



Pleiotropy: the same SNP is associated with multiple traits

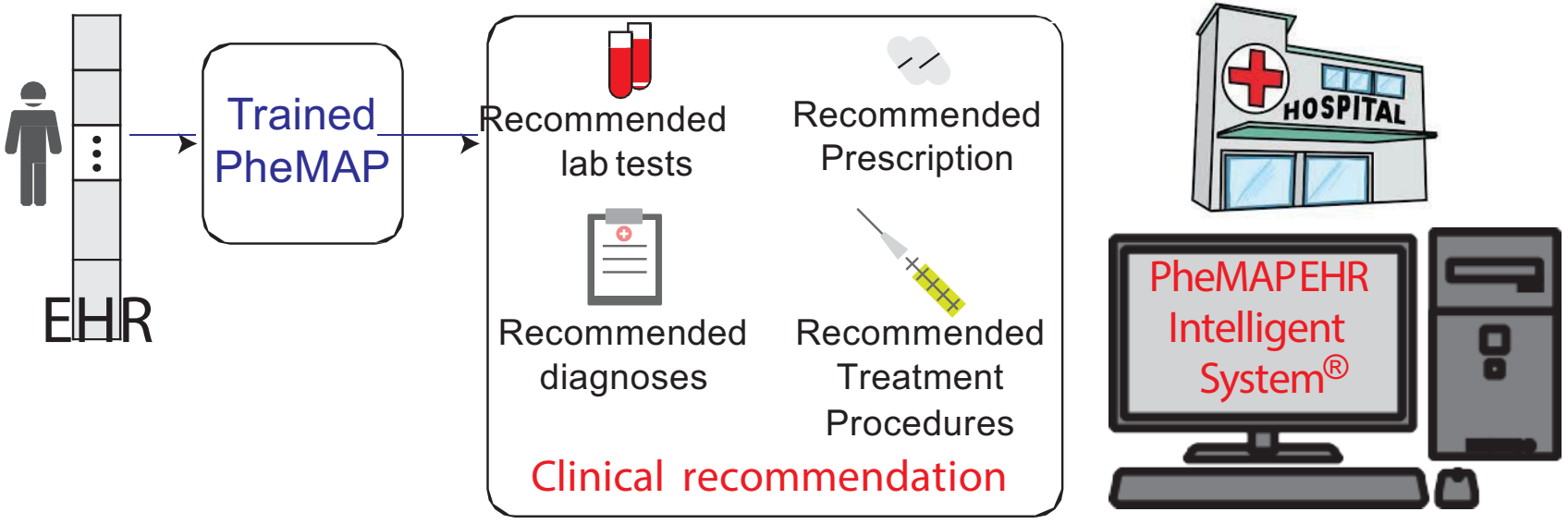
PheWAS without genetics information using EHR

- Genotype are often **not available** over large patient cohort
- Given the causal mediating phenotypes, **diseases of interest are conditionally independent of genotype**

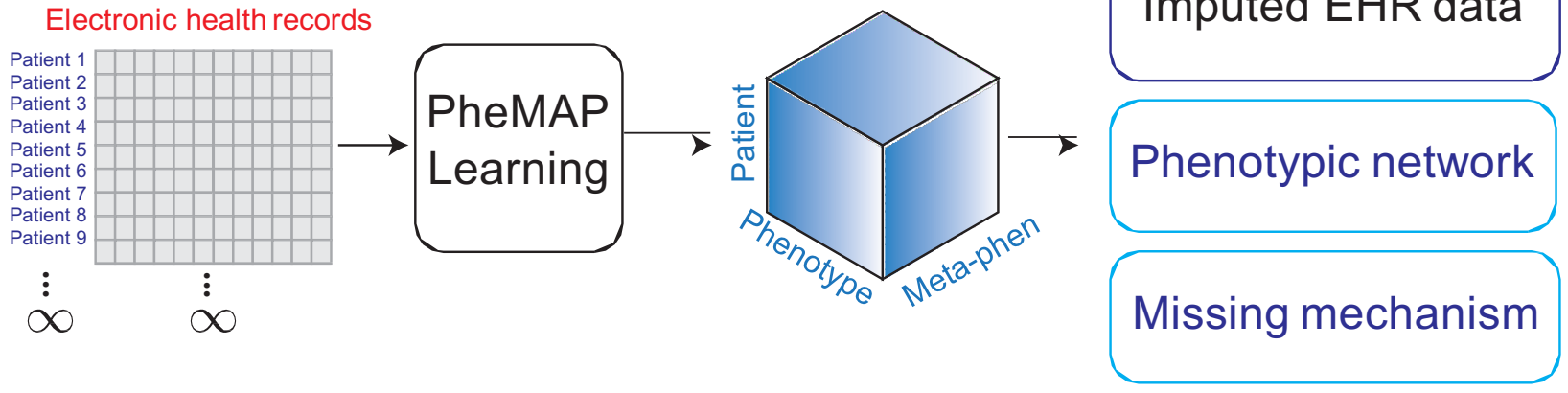


Overall goal:
building an intelligent medical recommendation system

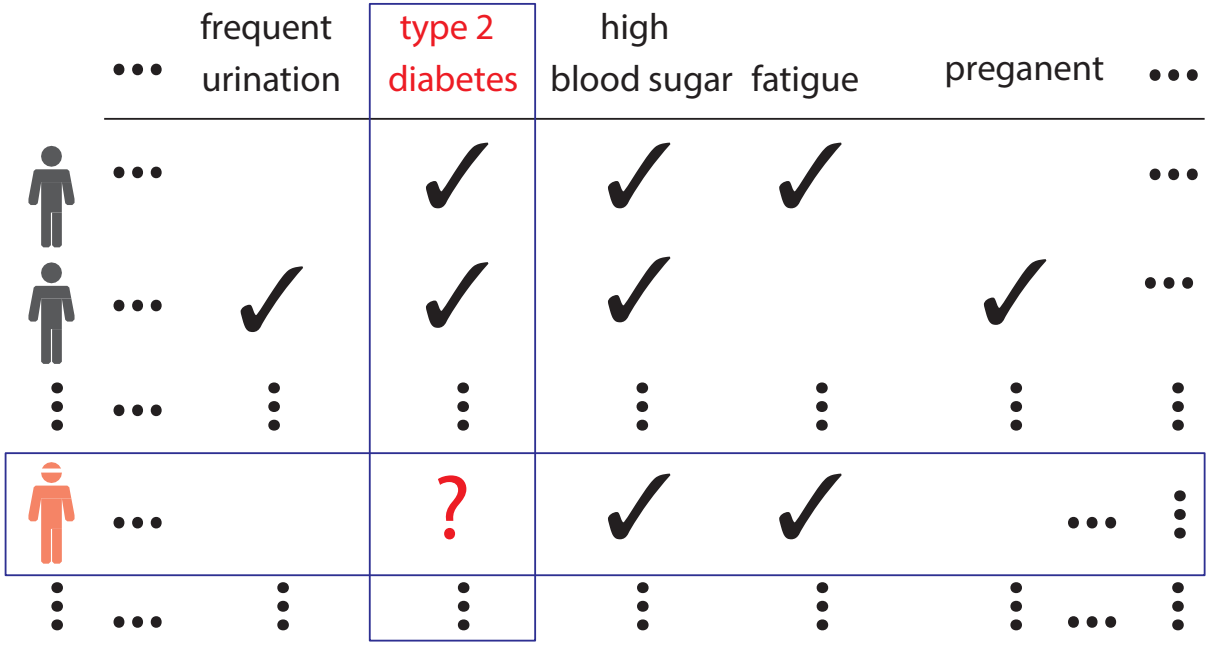
Diagnosis of new patients



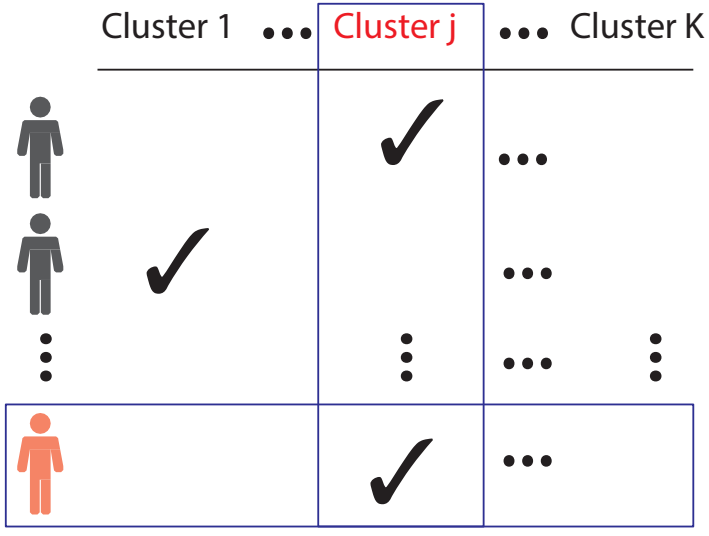
Phenotype MAtrix prediction (PheMAP)



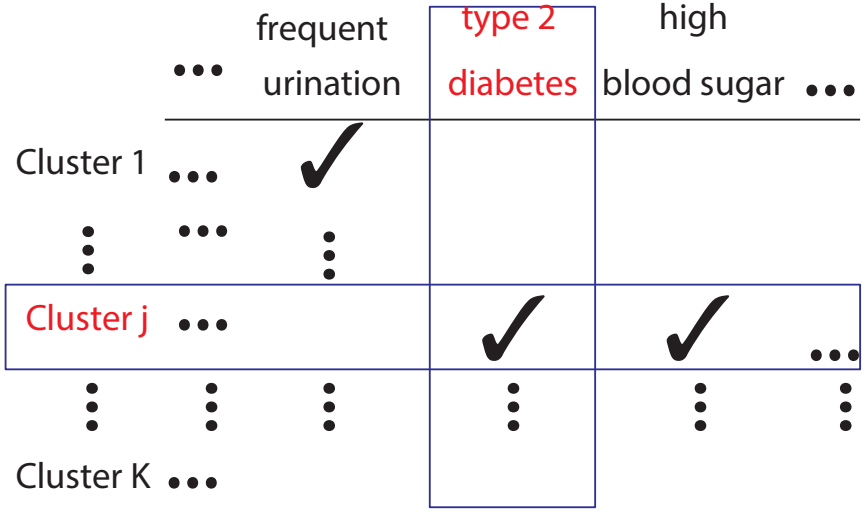
Intuition behind predicting phenotypes by factorization



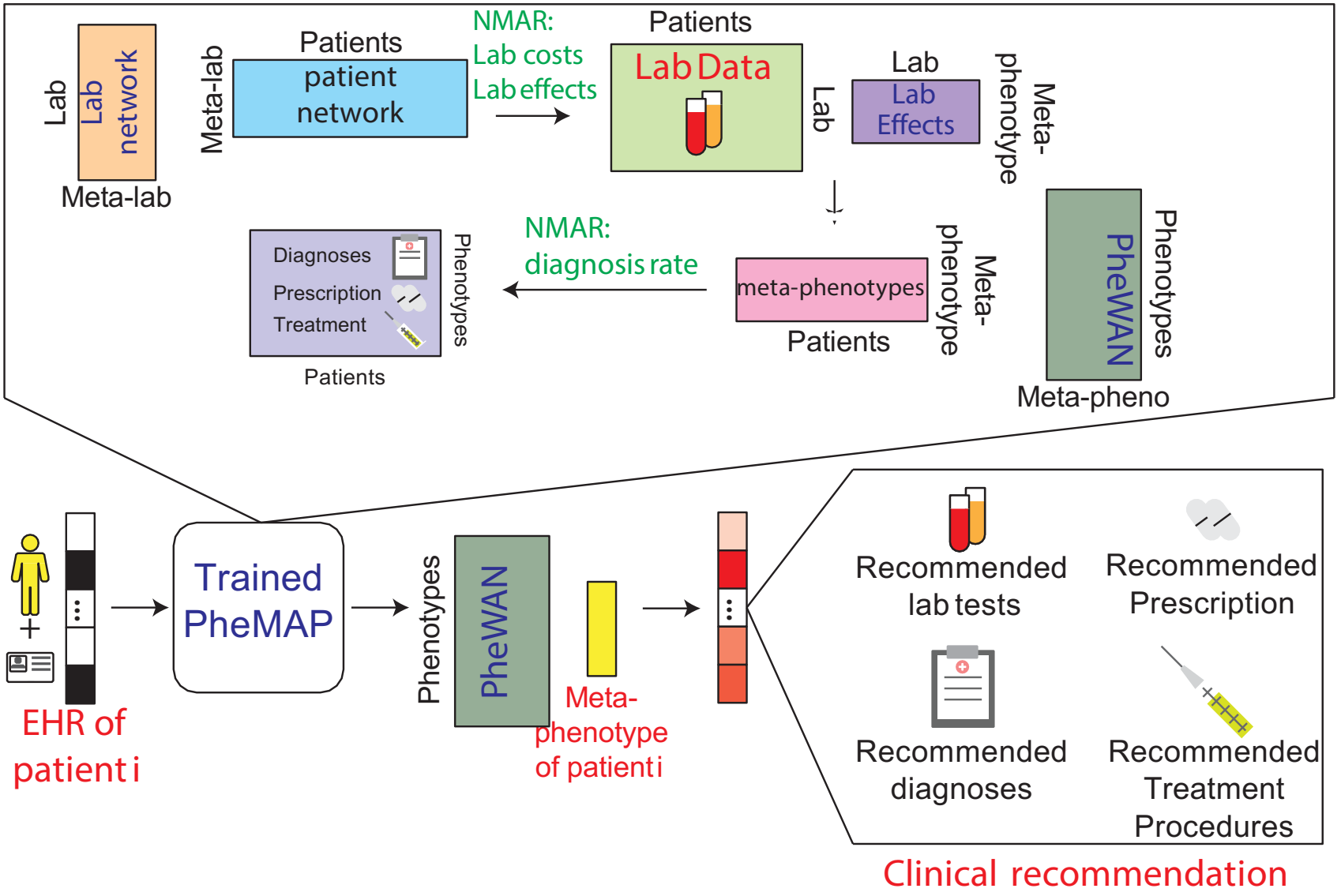
Patient similarity



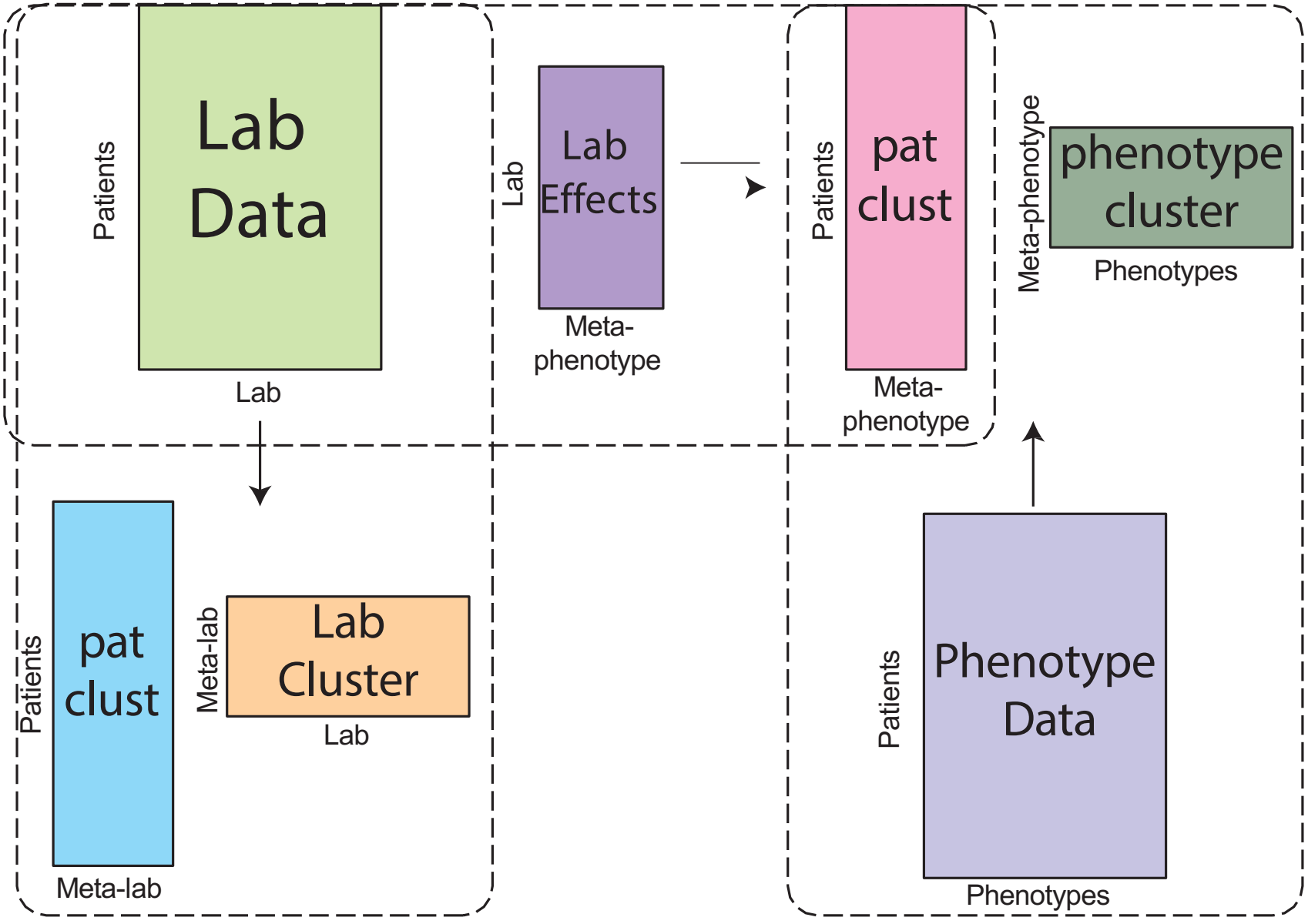
Phenotype similarity



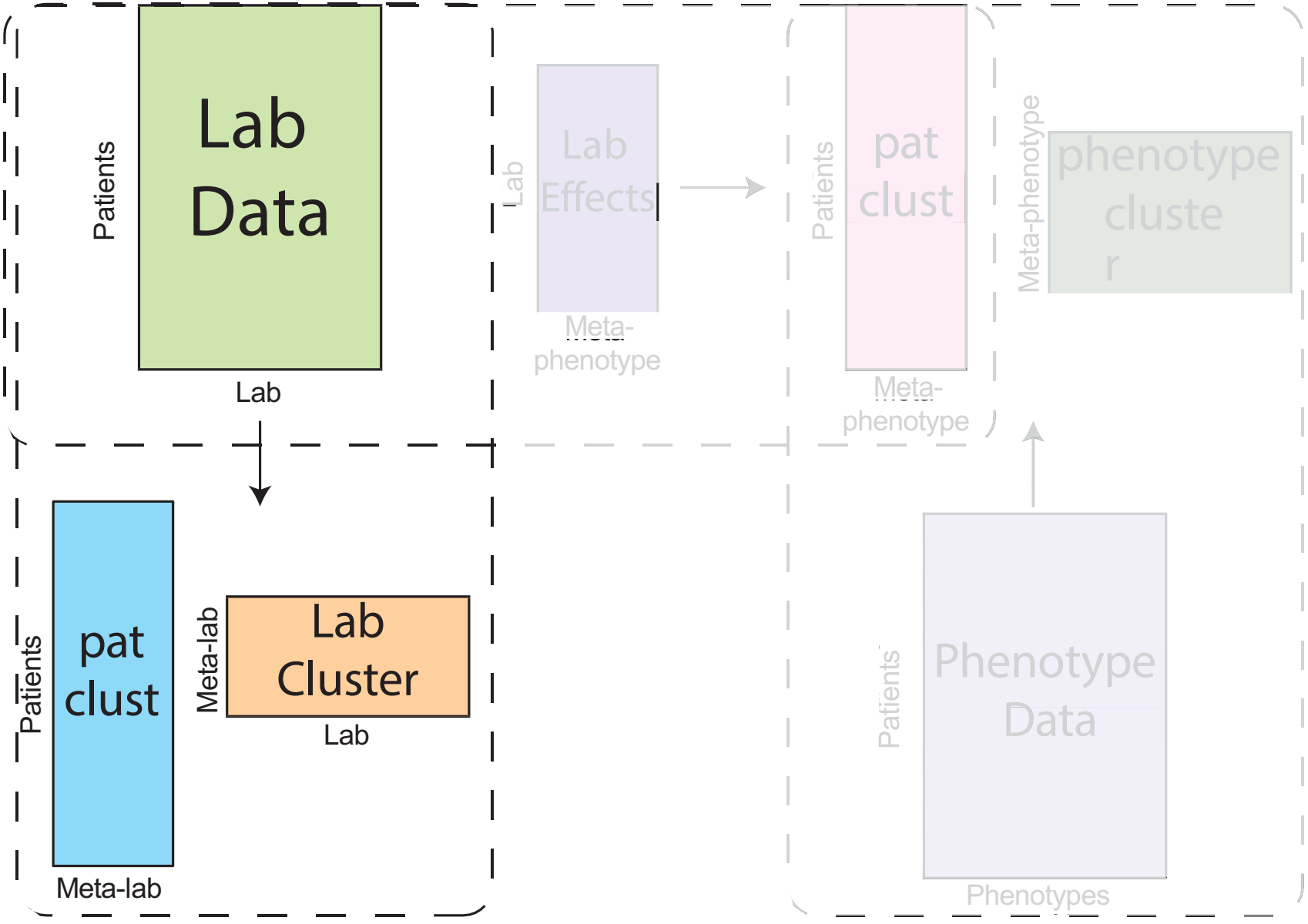
Model the EHR data as data generative process



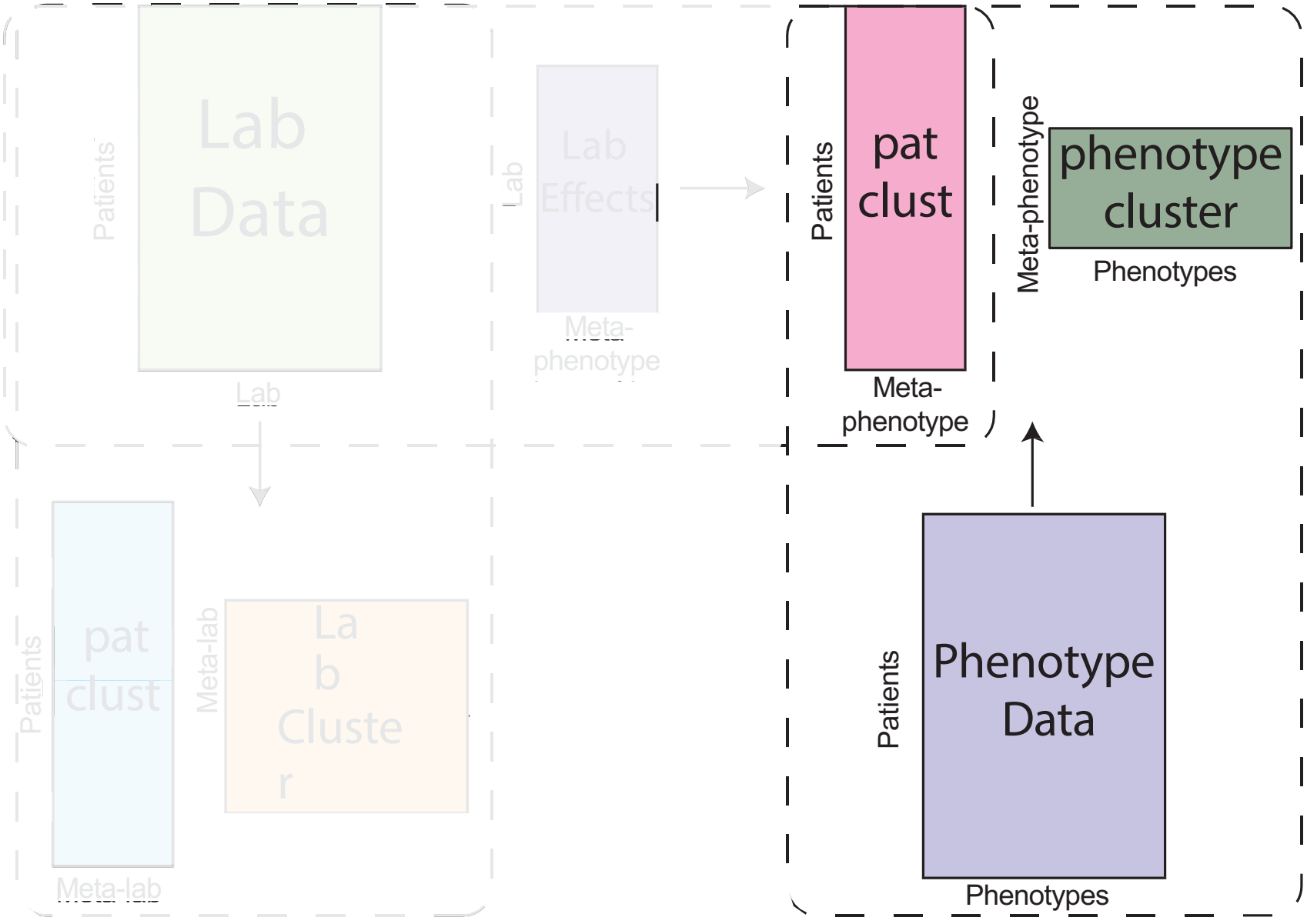
PheMAP is a probabilistic matrix factorization method



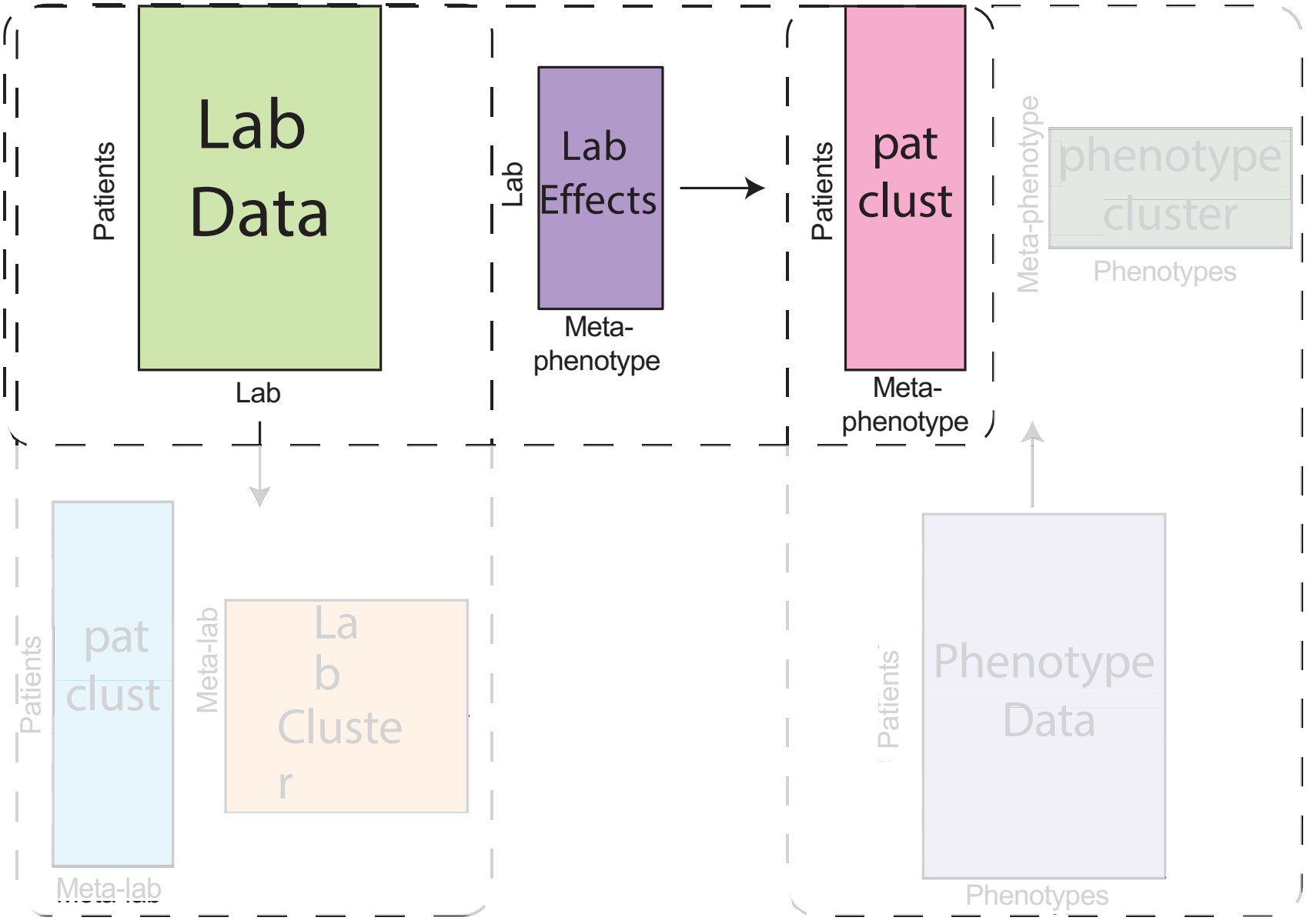
PheMAP is a probabilistic matrix factorization method



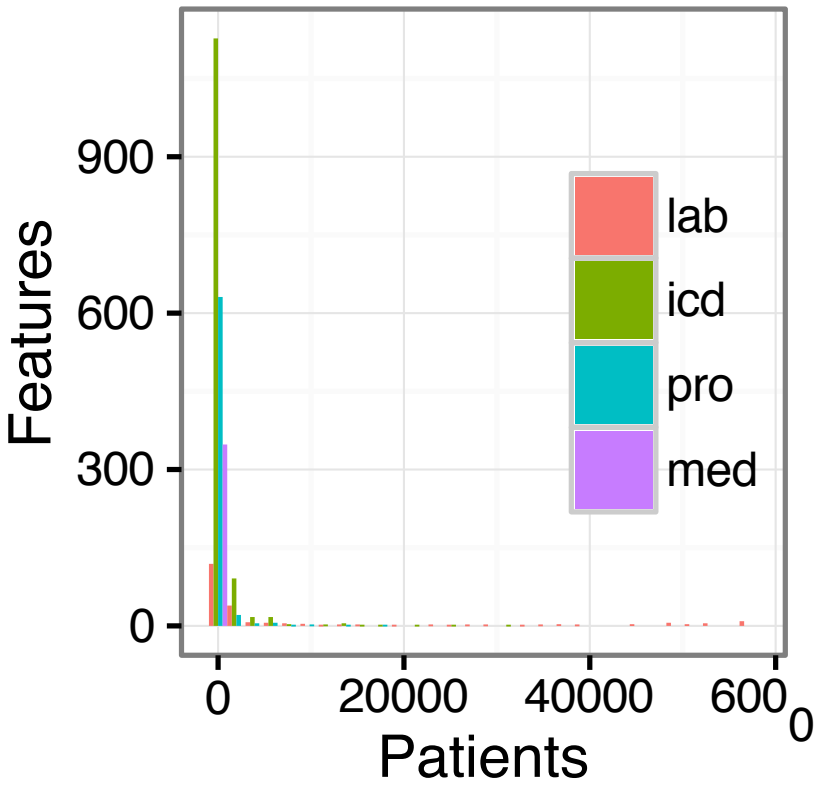
PheMAP is a probabilistic matrix factorization method

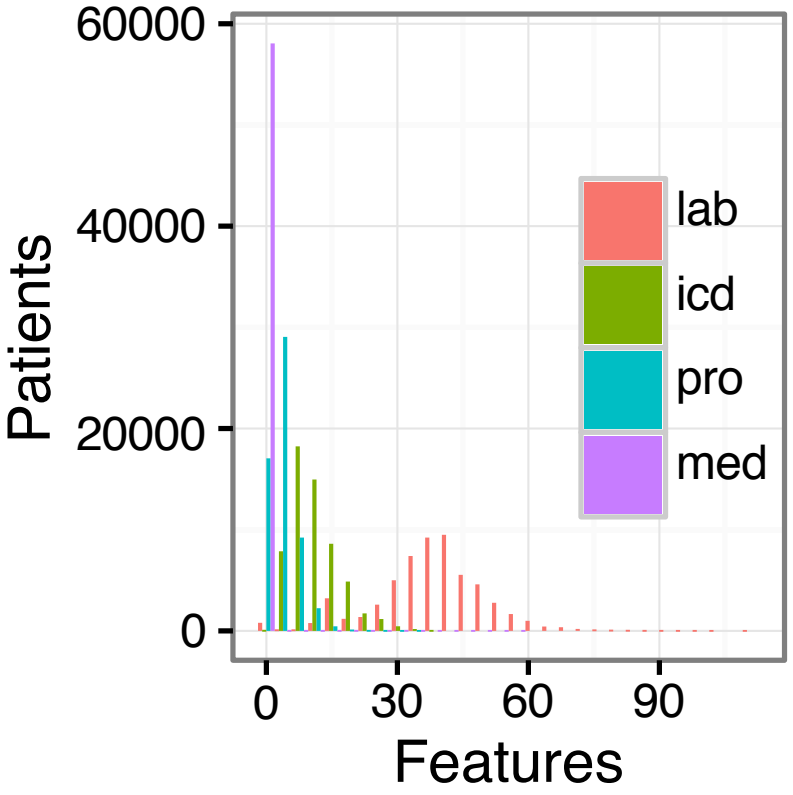


PheMAP is a probabilistic matrix factorization method

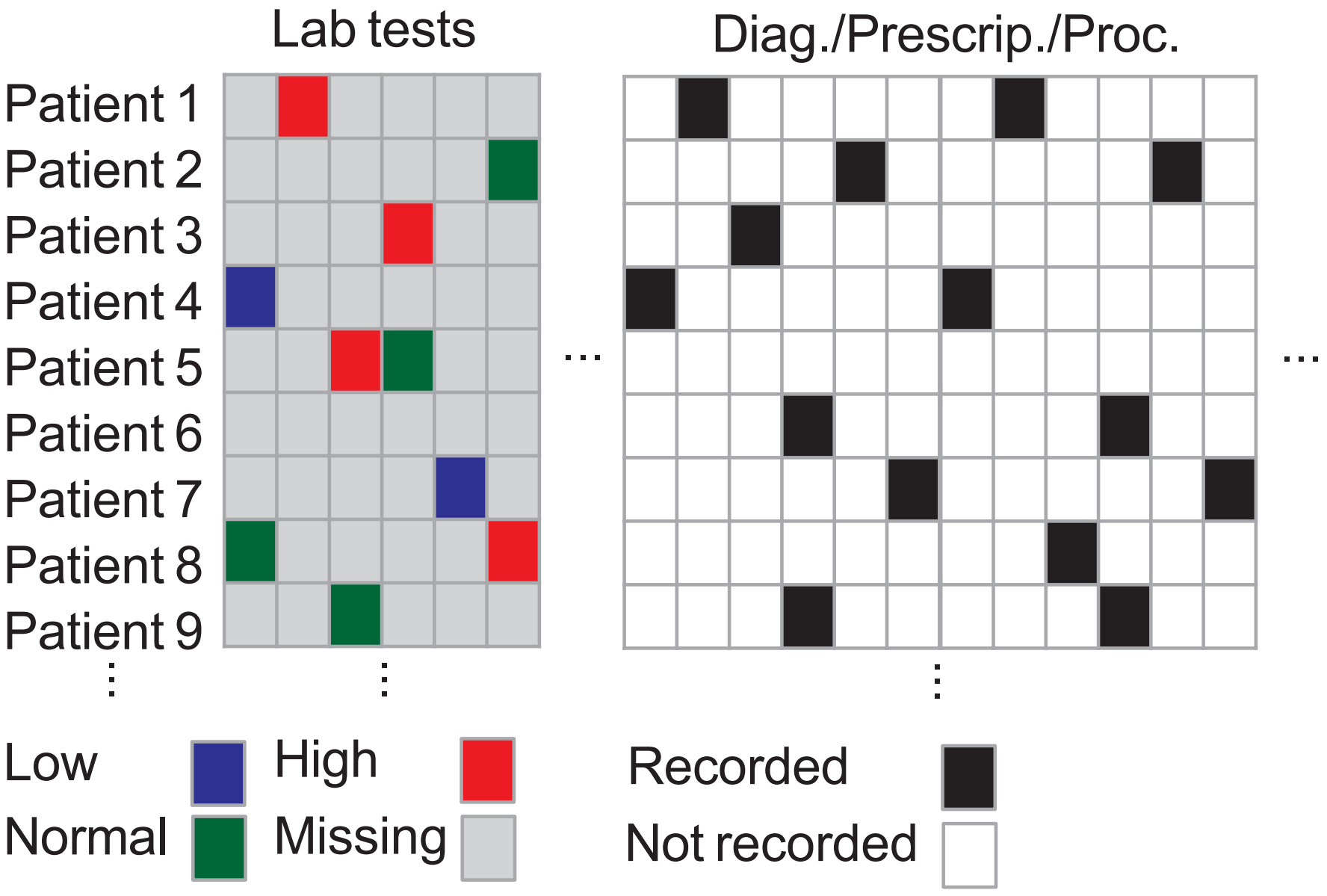


EHR data are extremely sparse

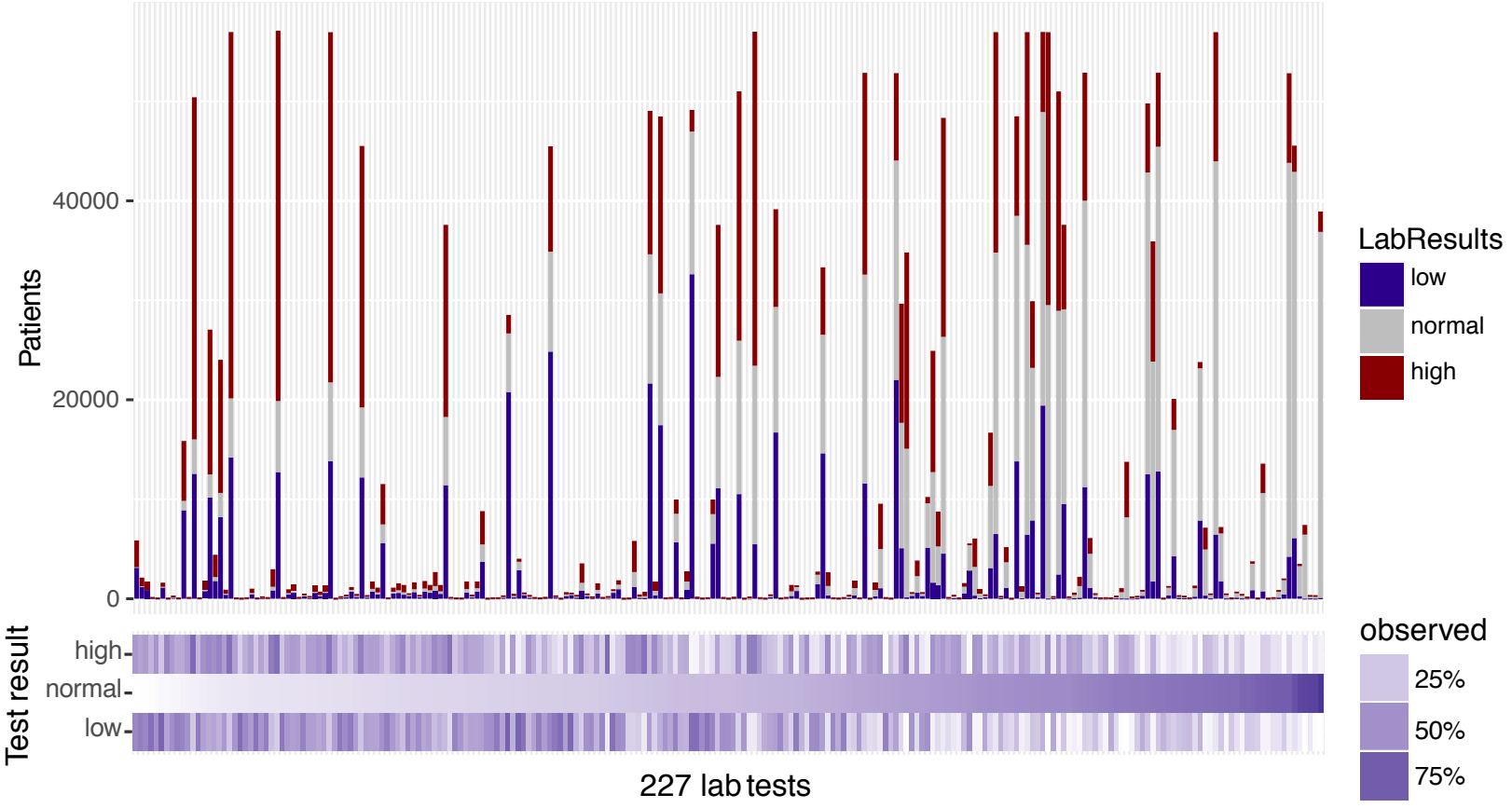




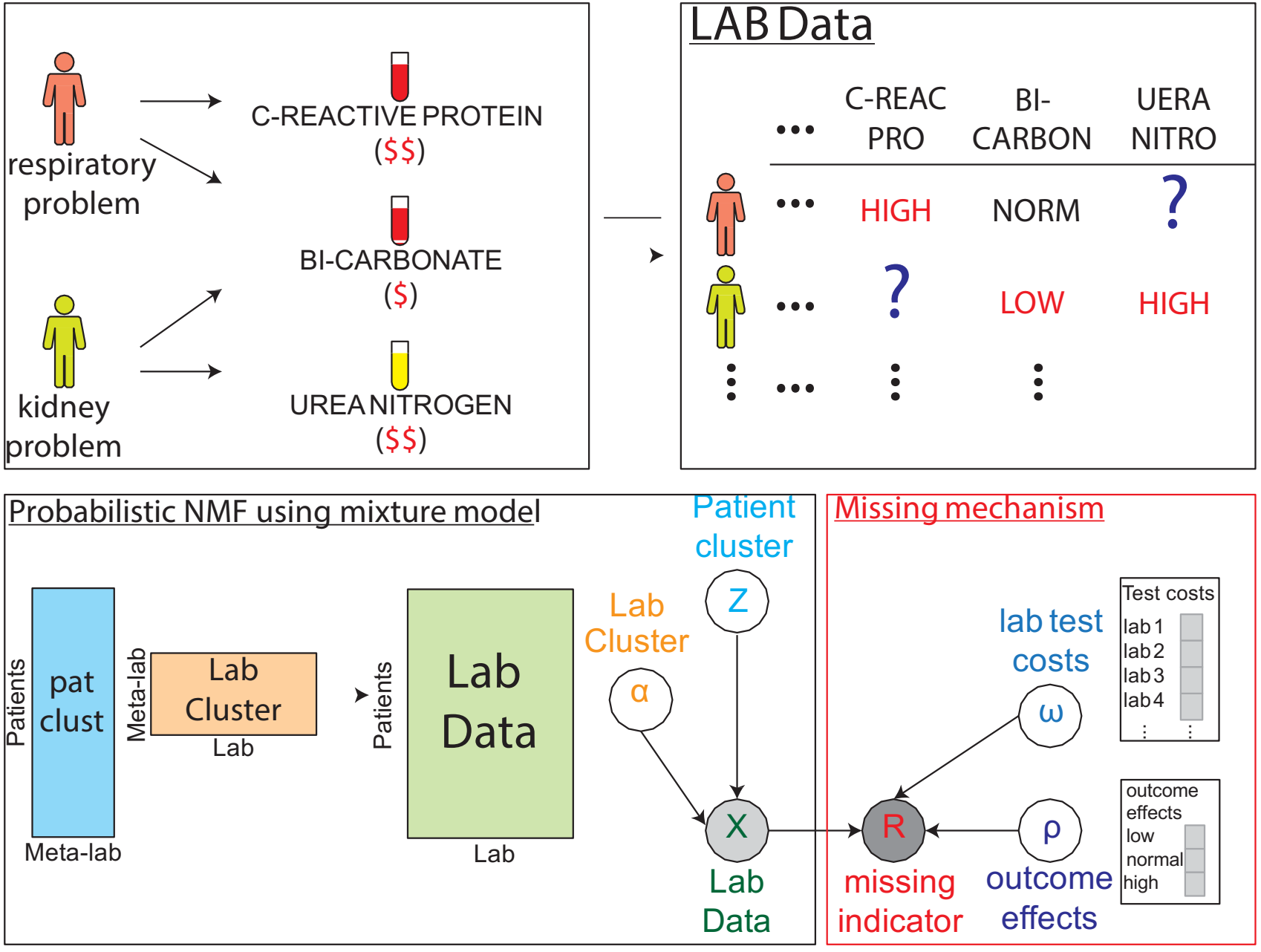
Two types of missing data in EHR



EHR data are not missing at random (NMAR)



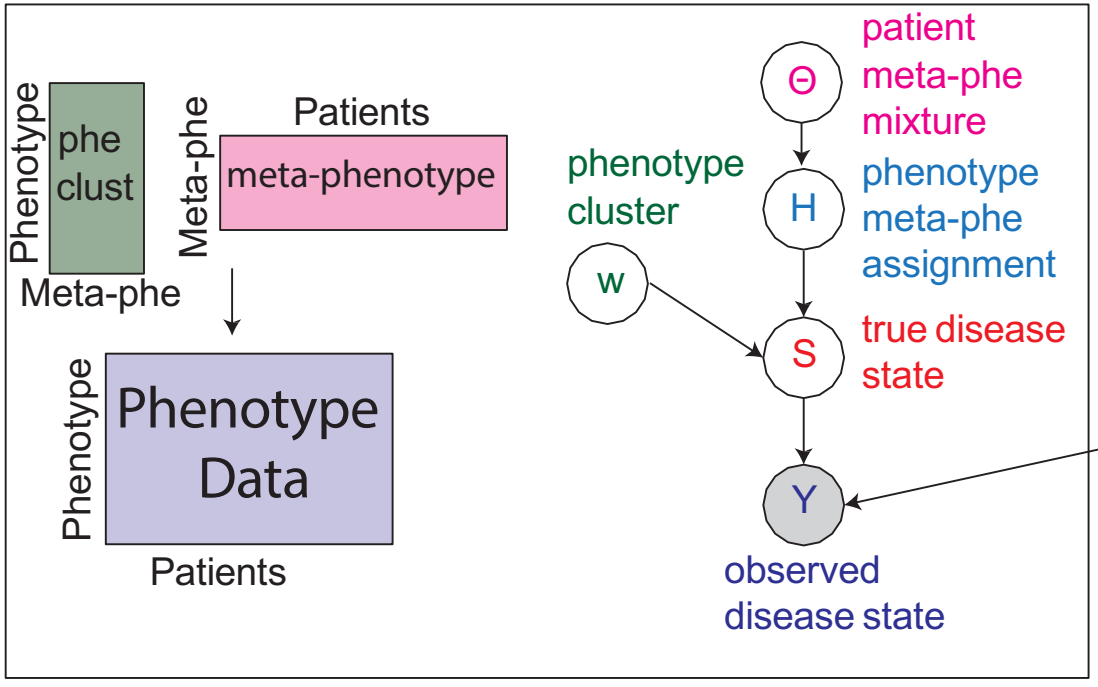
Modeling missing mechanism in lab test

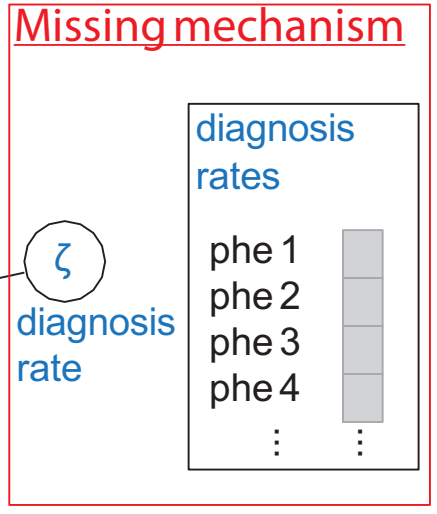


Modeling missing mechanism in phenotype data

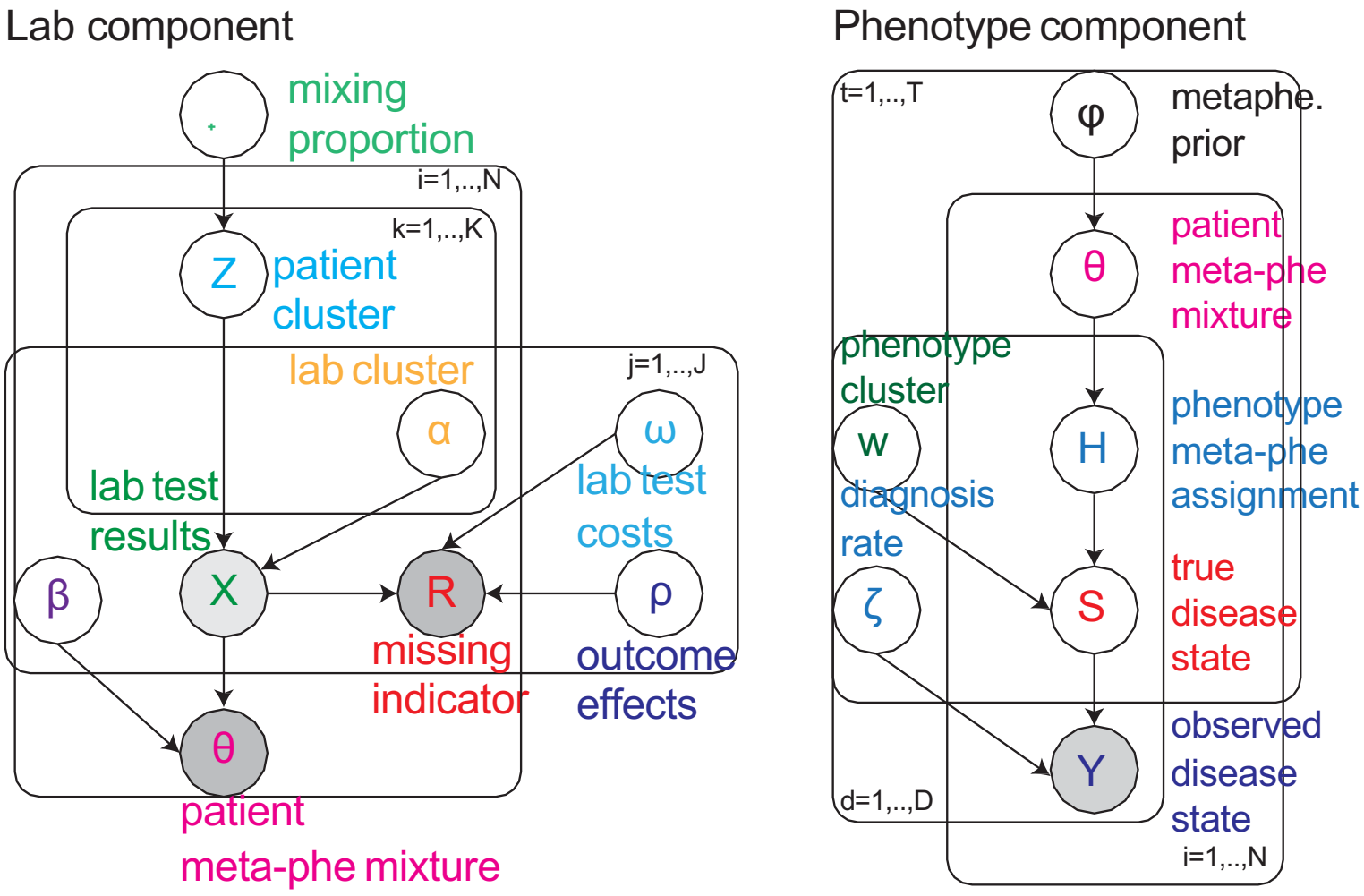
Phenotype Data

	...	Asthma	Influenza	Pneumonia	Malignant neoplasm of lung
patients with respiratory problems	...	✓	NA	NA	NA
	...	NA	NA	✓	NA
	...	⋮	⋮	⋮	⋮



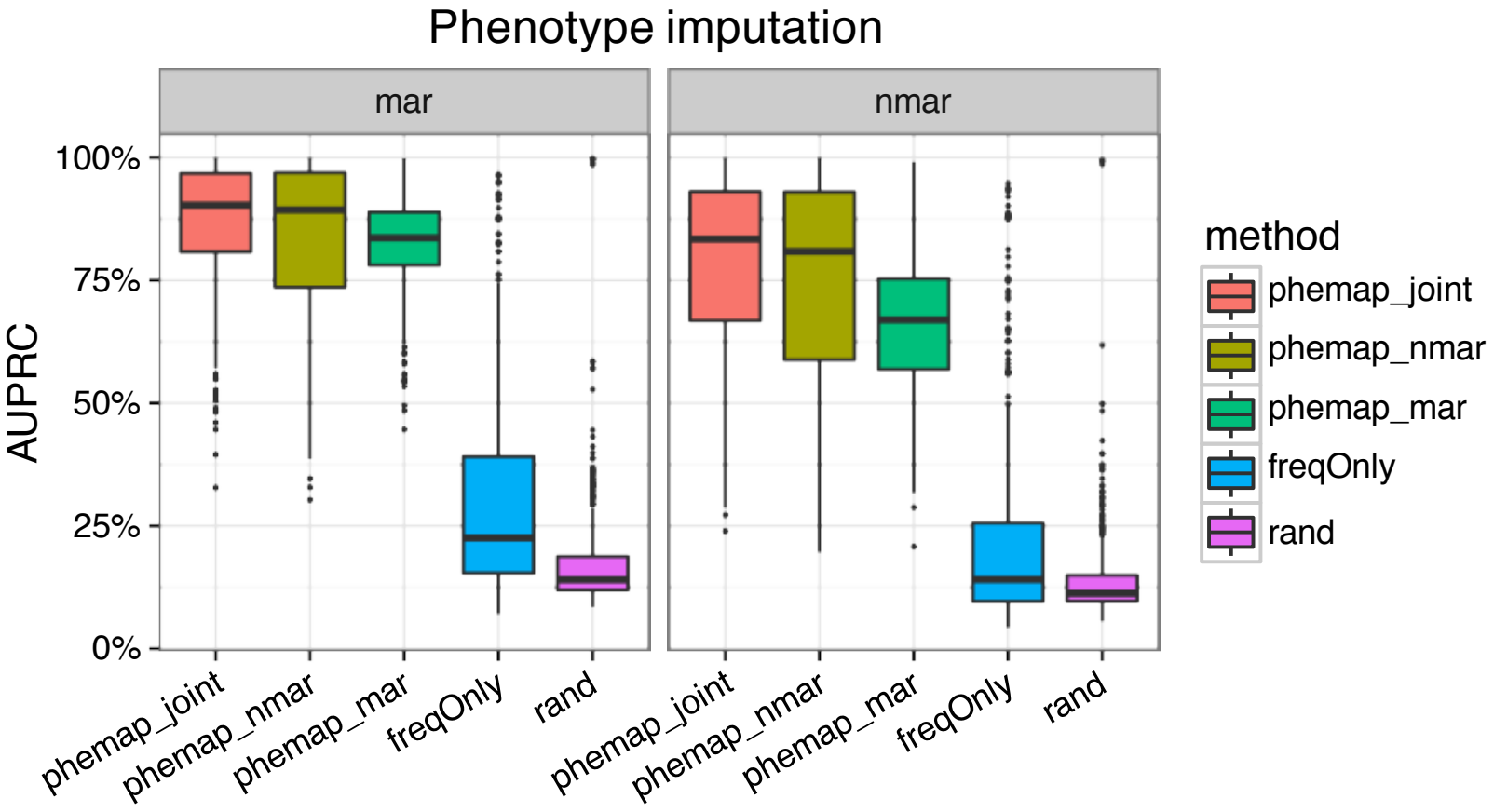


PheMAP joint inference over multiple data types

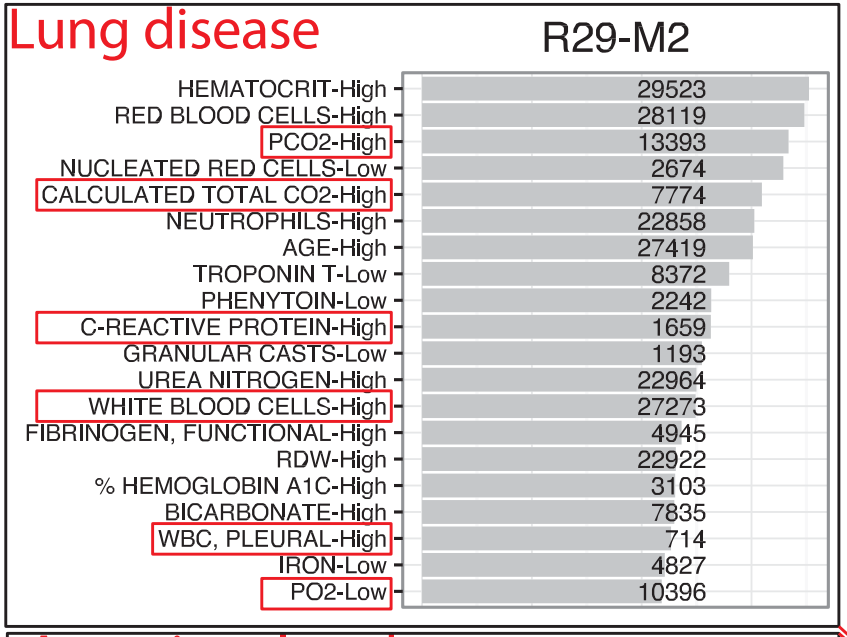


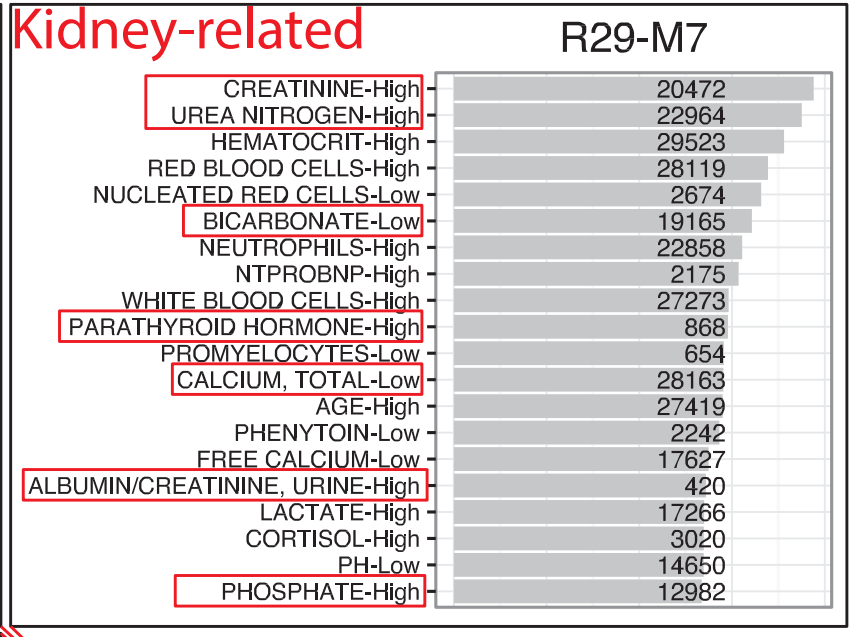
Meta-phenotype likelihood follows Dirichlet distribution

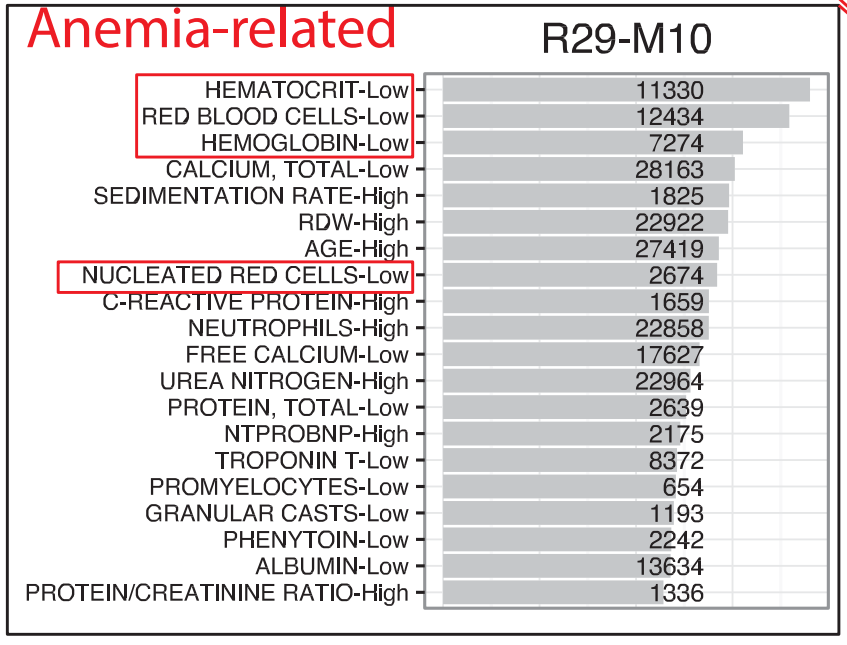
PheMAP achieves promising imputation accuracy

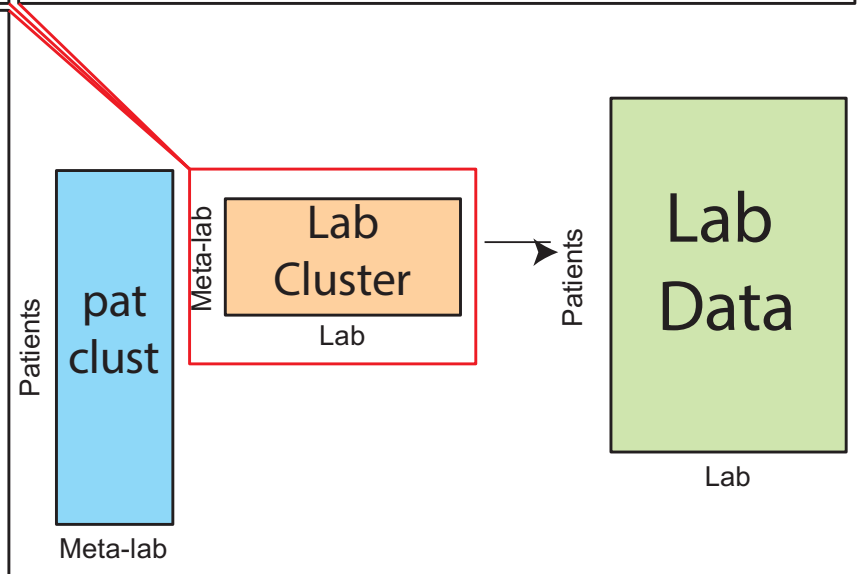


Many meta-lab clusters are biologically meaningful



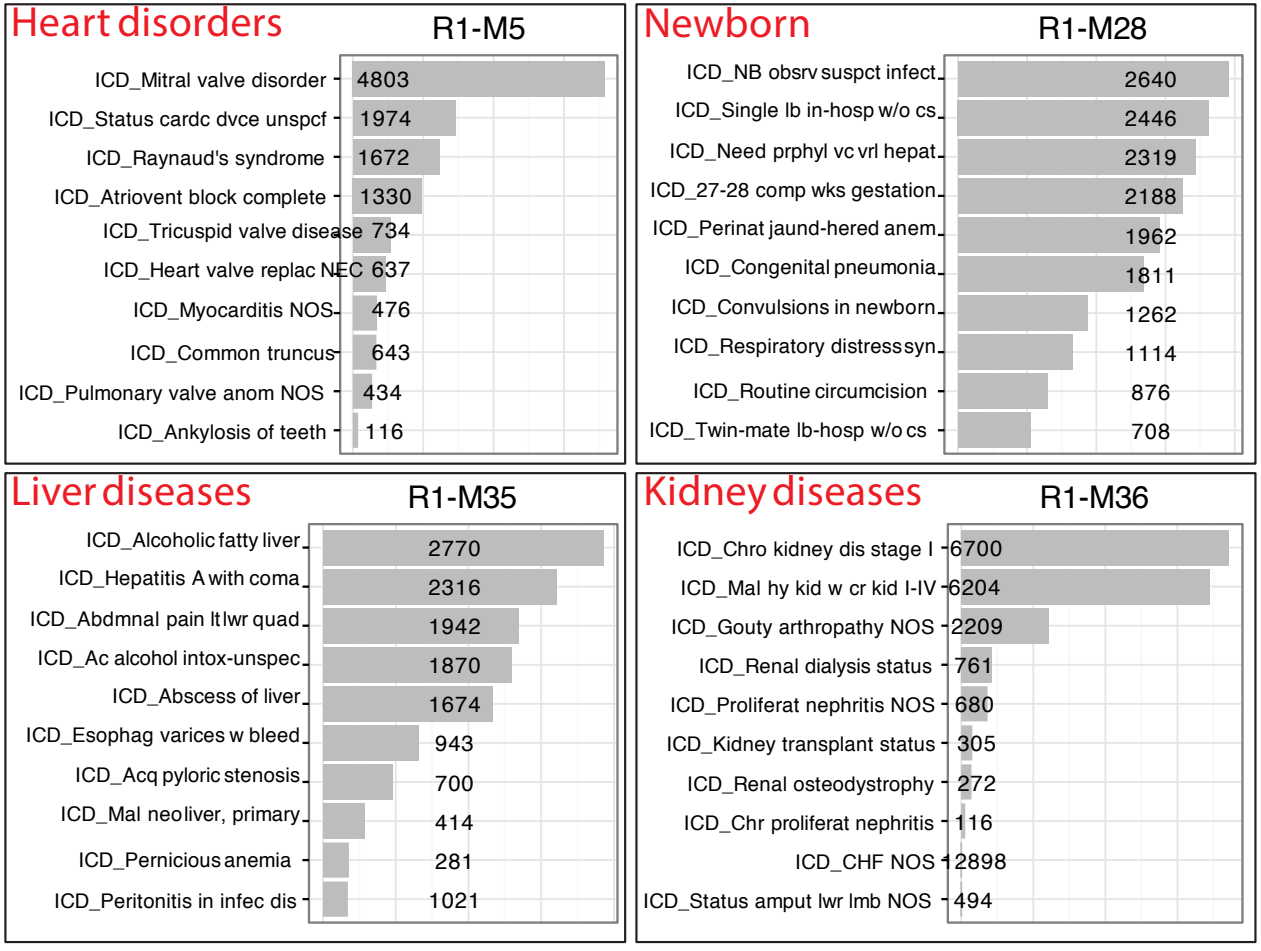


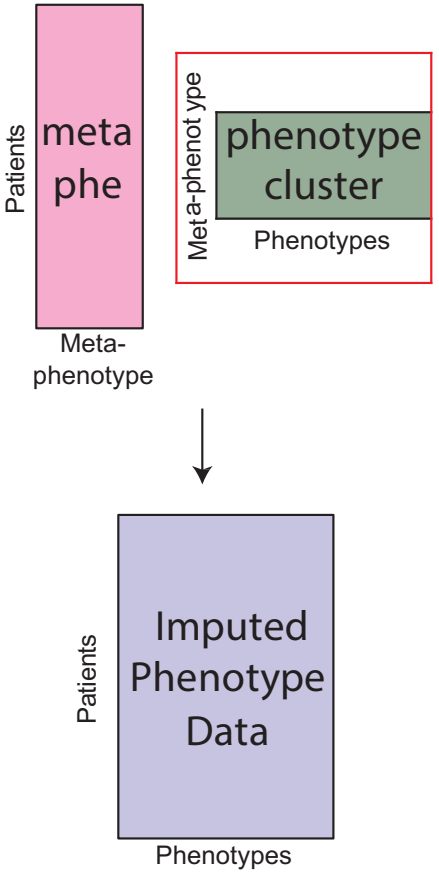




Thanks Yuri Anjura (MD student at HMS) for help interpreting the meta-lab!

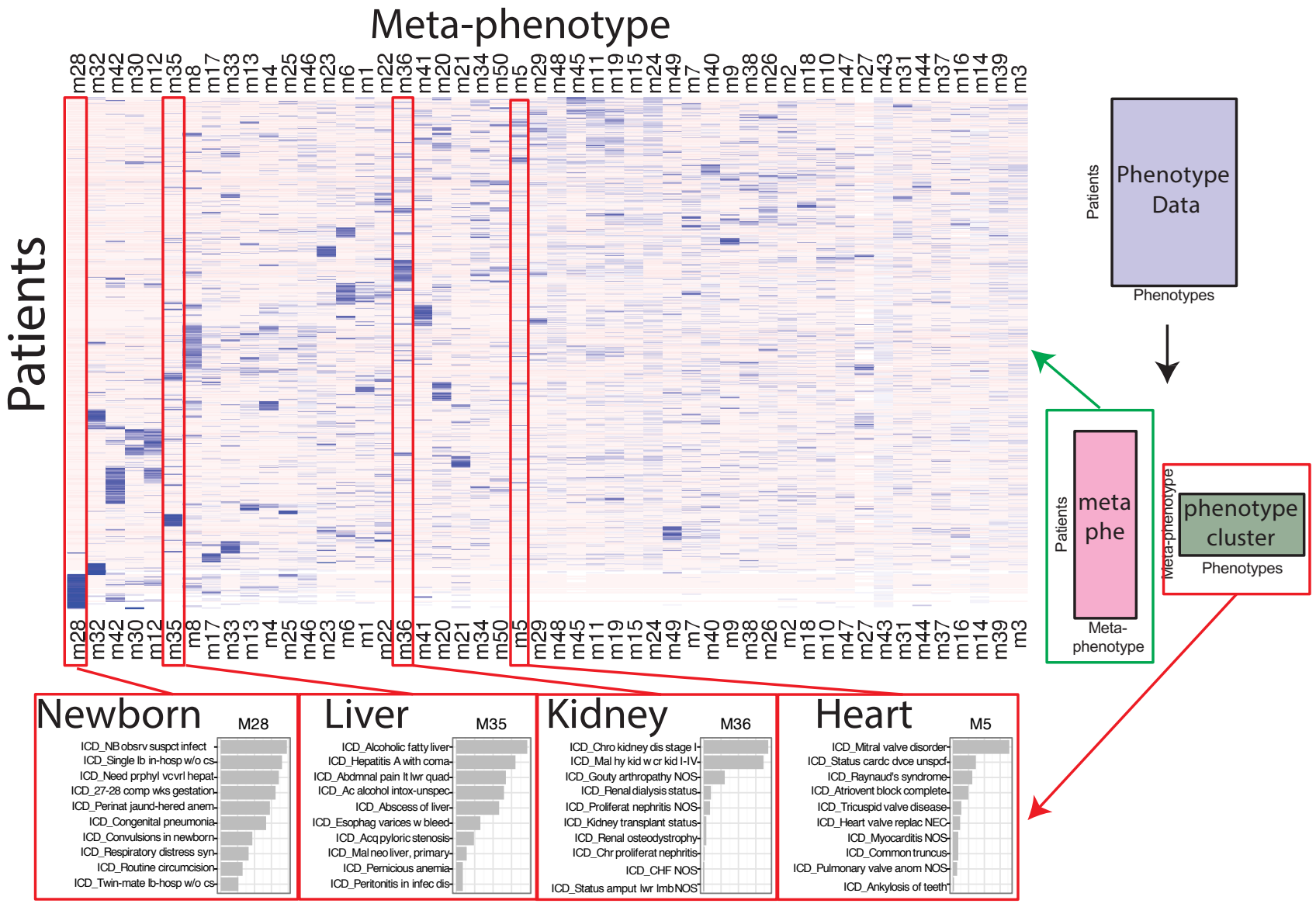
Many meta-phenotypes clusters are biologically meaningful



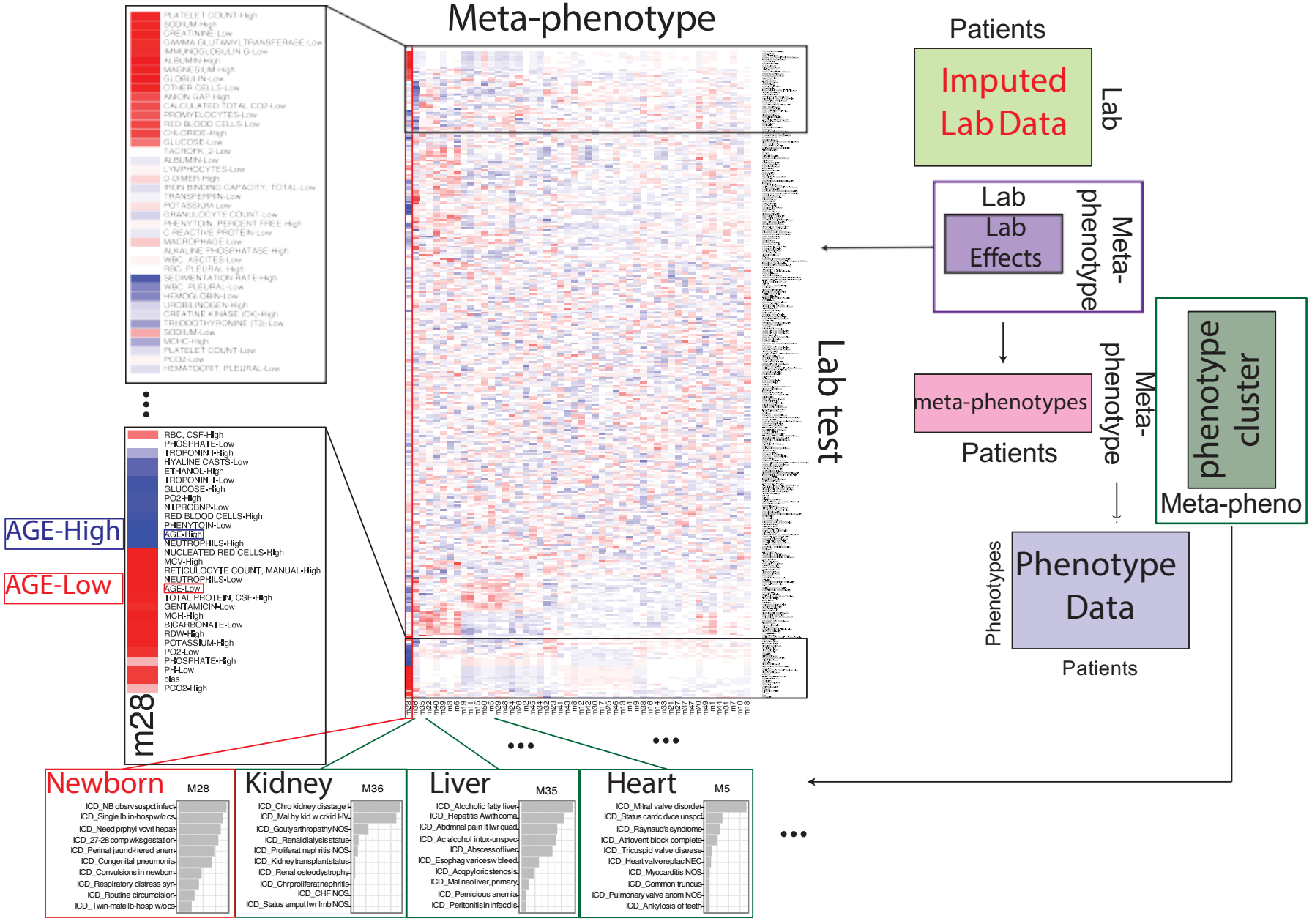


Thanks Brad Ruzicka (MD at McLean Hospital) for help interpreting the meta-phenotypes!

Associating patients by the inferred meta-phenotypes



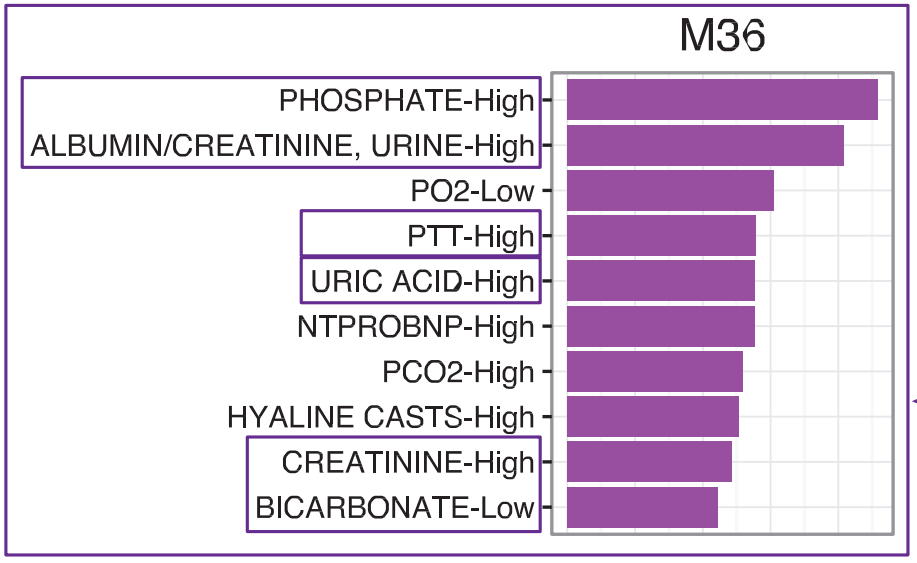
Linking lab tests to phenotypes via meta-phenotypes

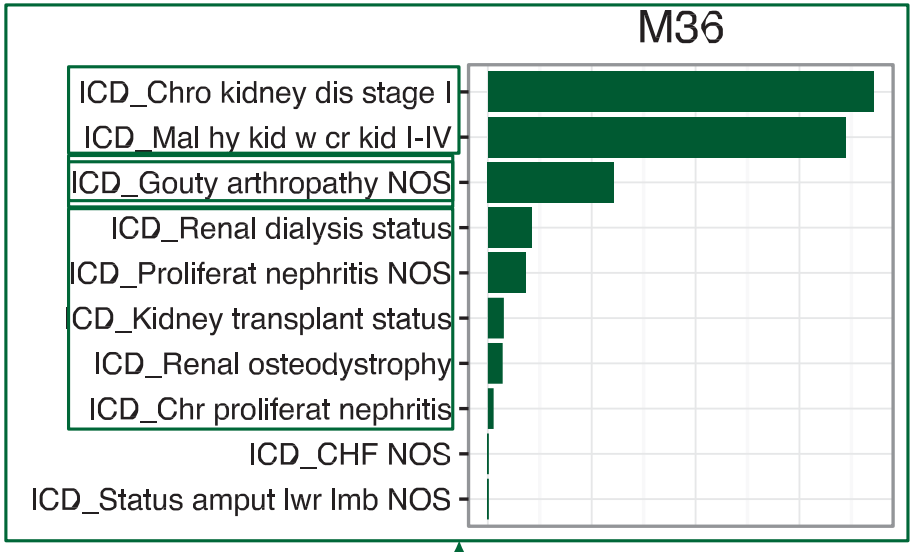


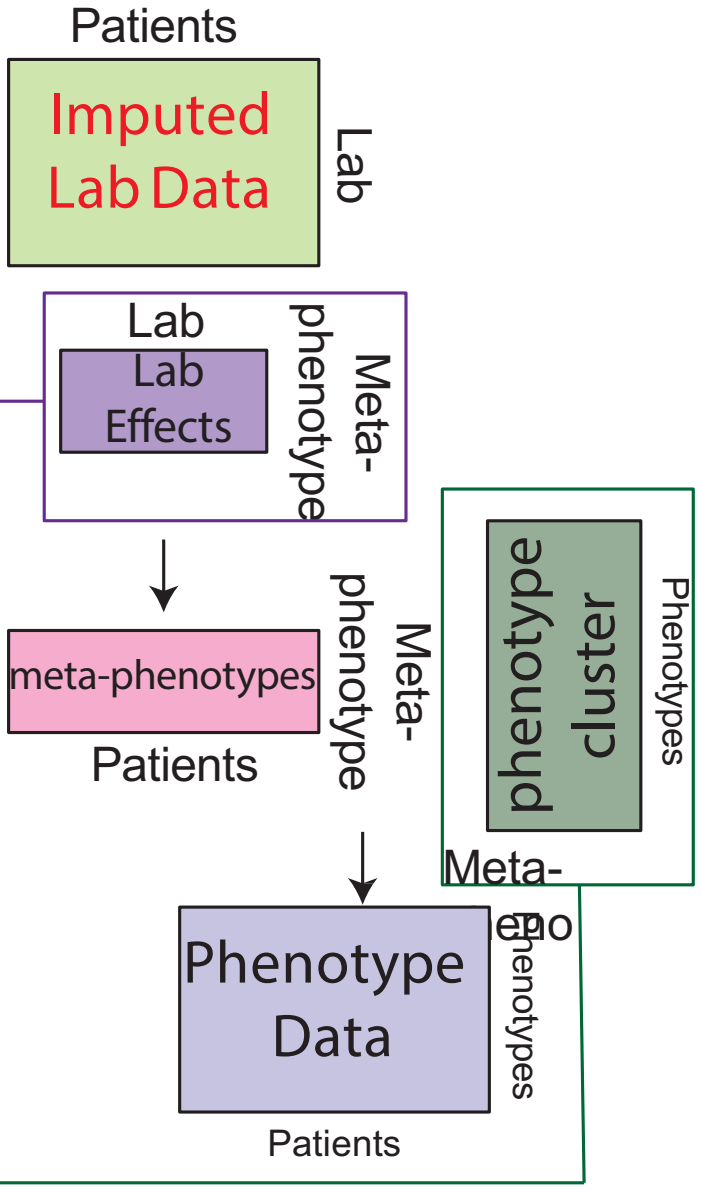
"Newborn" is an "outlier" meta-phenotype based on the lab tests

Linking lab tests to phenotypes via meta-phenotypes

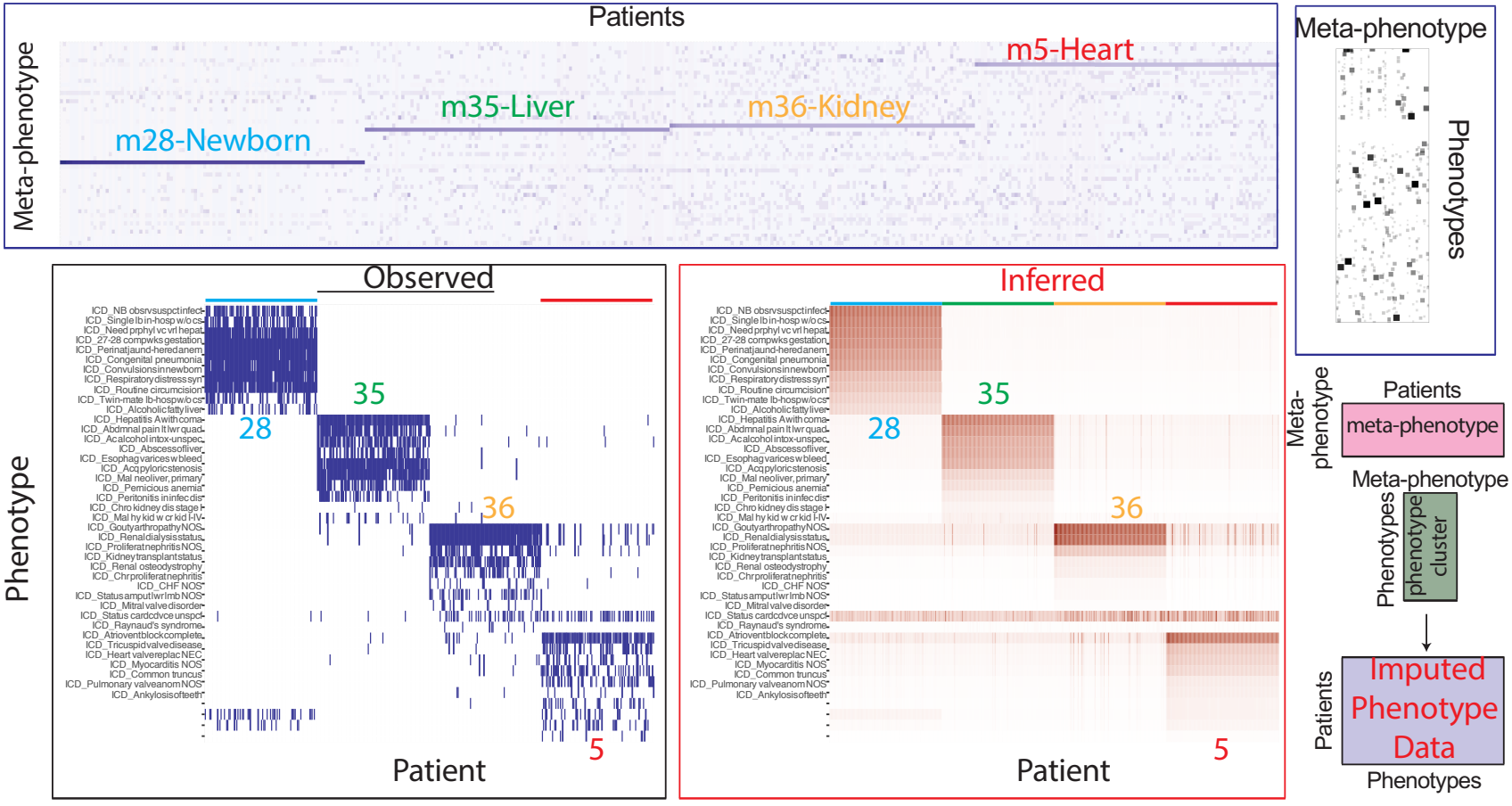
Top lab tests on kidney disease module





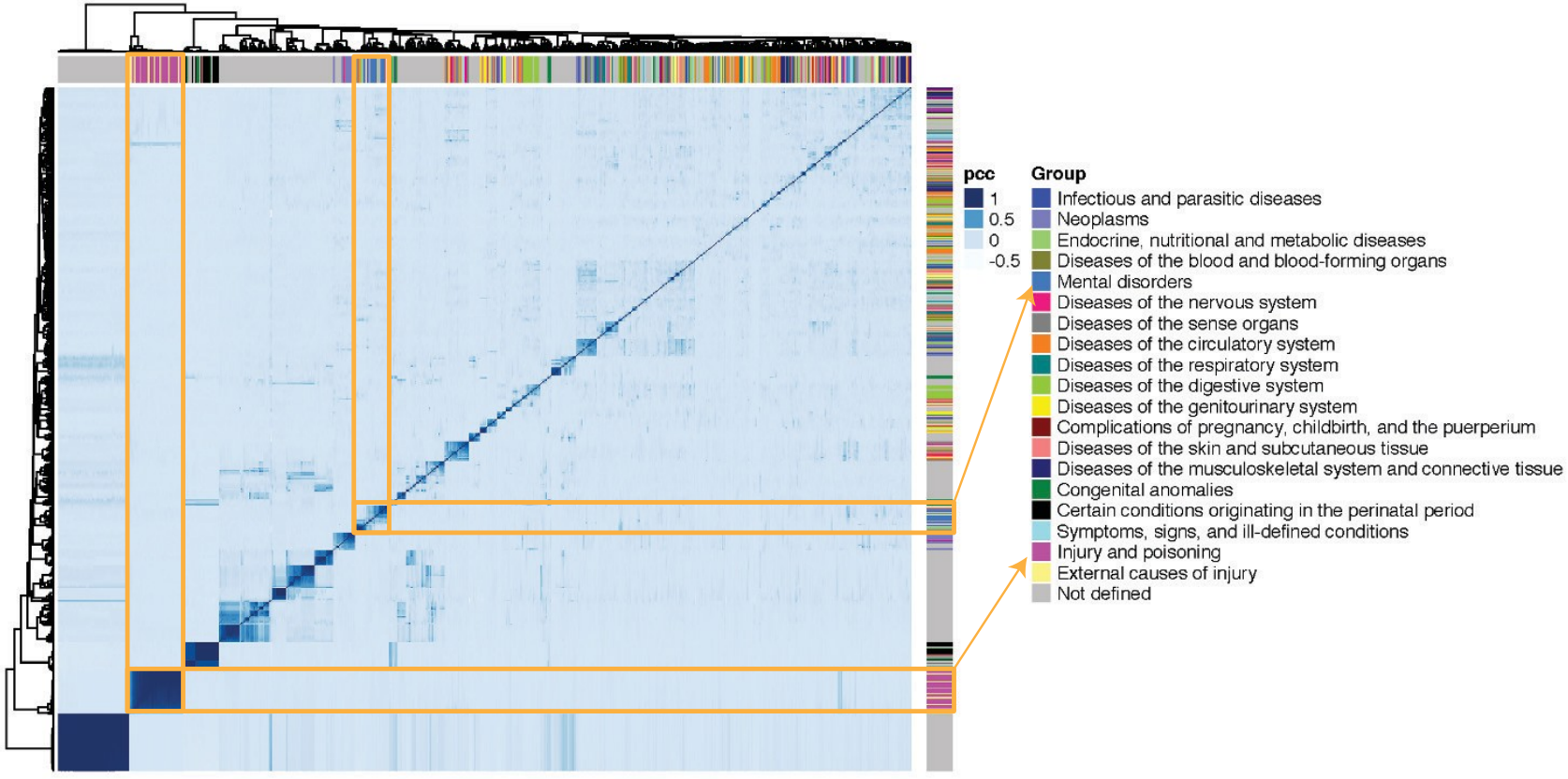


Imputing phenotypes by meta-phenotype associations



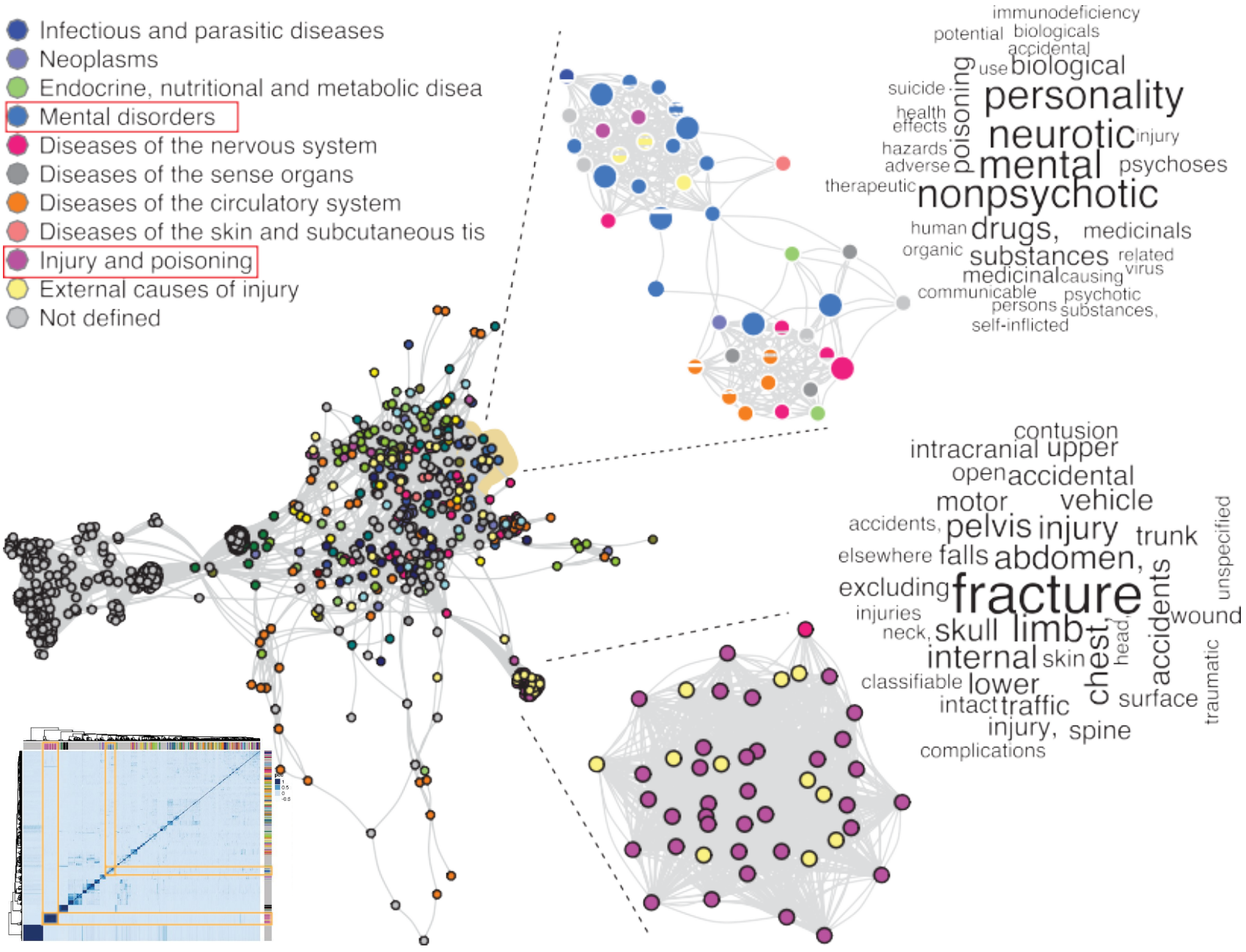
- Patients and phenotypes are sorted in decreasing order of their probabilistic associations with each meta-phenotype
- For each meta-phenotype, the top 100 patients and top 10 phenotypes are selected

Correlation across meta-phenotypes reveal high modularity

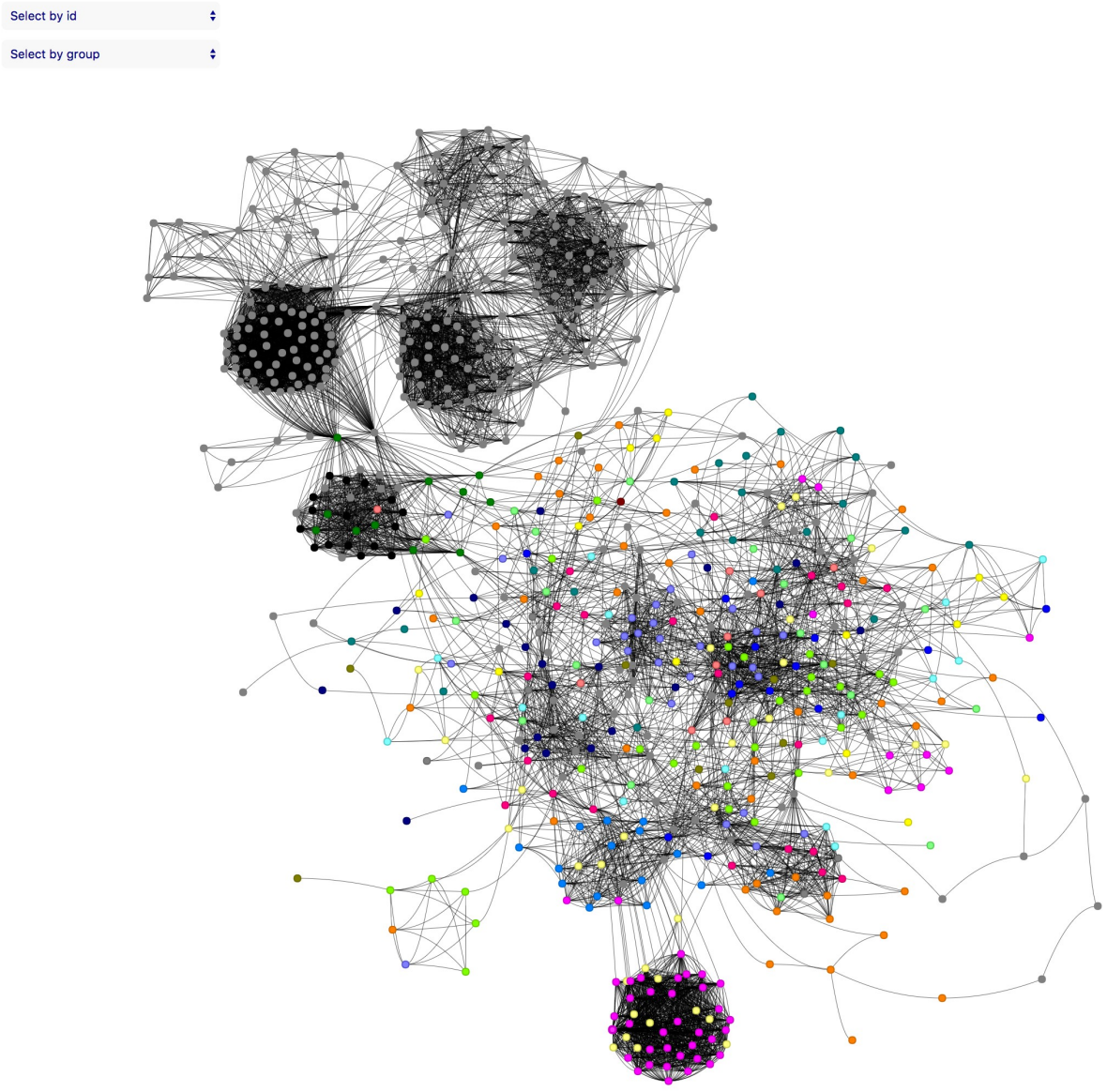


Many modules are highly enriched for common disease categories defined by ICD-9 system

Visualizing PheWAN by correlation across meta-phenotypes



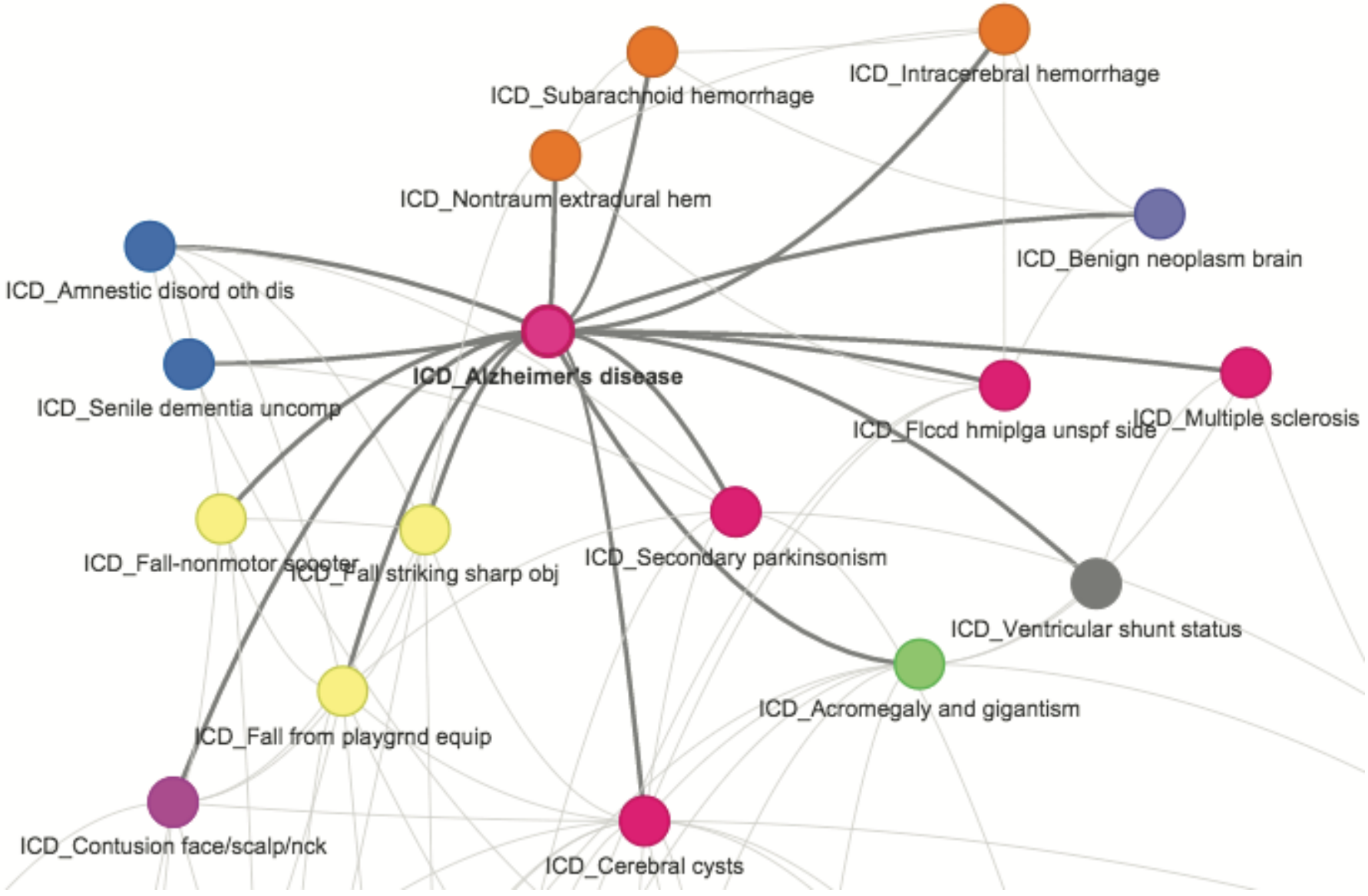
Online visualization portal of disease network



http://people.csail.mit.edu/yueli/phewan/mimic/CompleteNetAnnotated.html

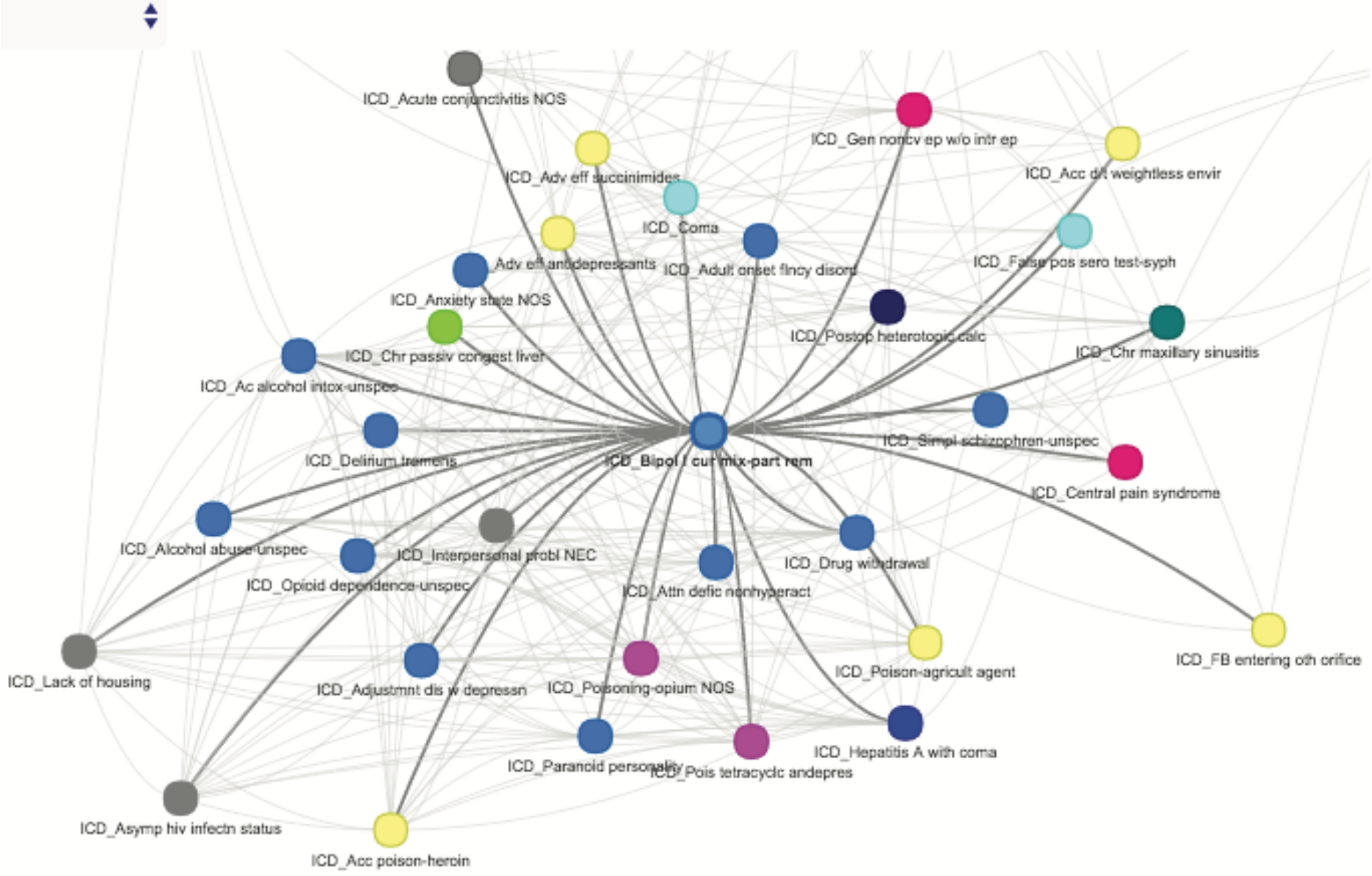
Collaborating with postdoc Jose Davila on the visualization portal

Alzheimer's disease subnetwork module



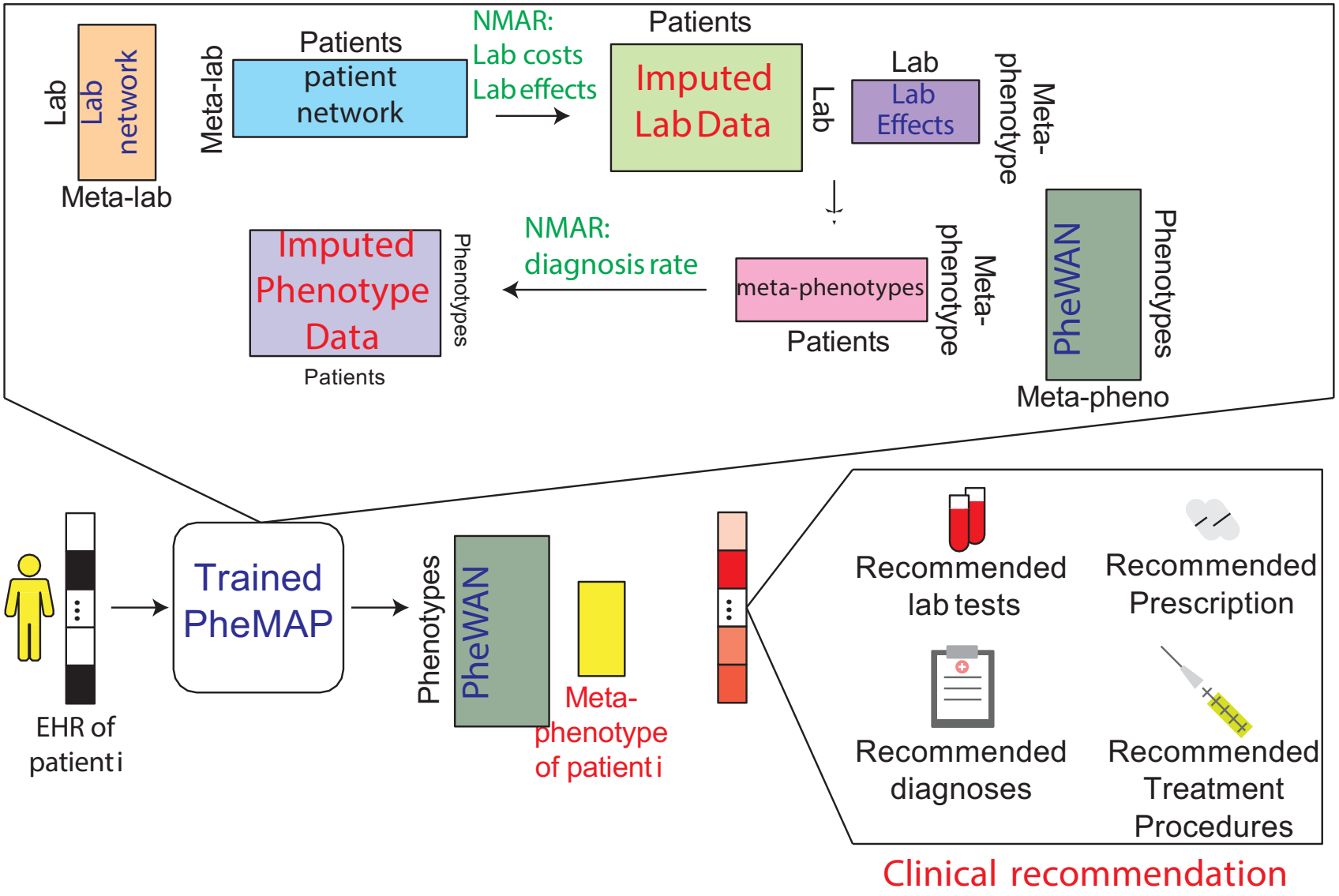
http://people.csail.mit.edu/yueli/phewan/mimic/NewMentalNetInt.html

Bipolar disorder subnetwork module



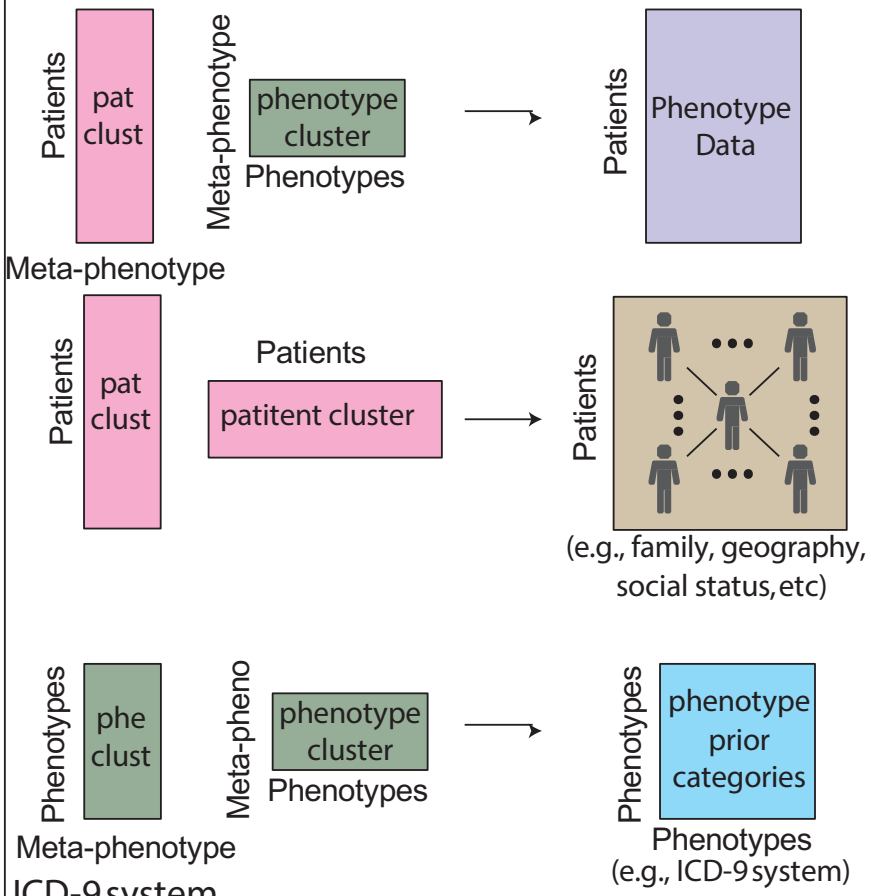
http://people.csail.mit.edu/yueli/phewan/mimic/NewMentalNetInt.html

Summary of the EHR PheMAP model

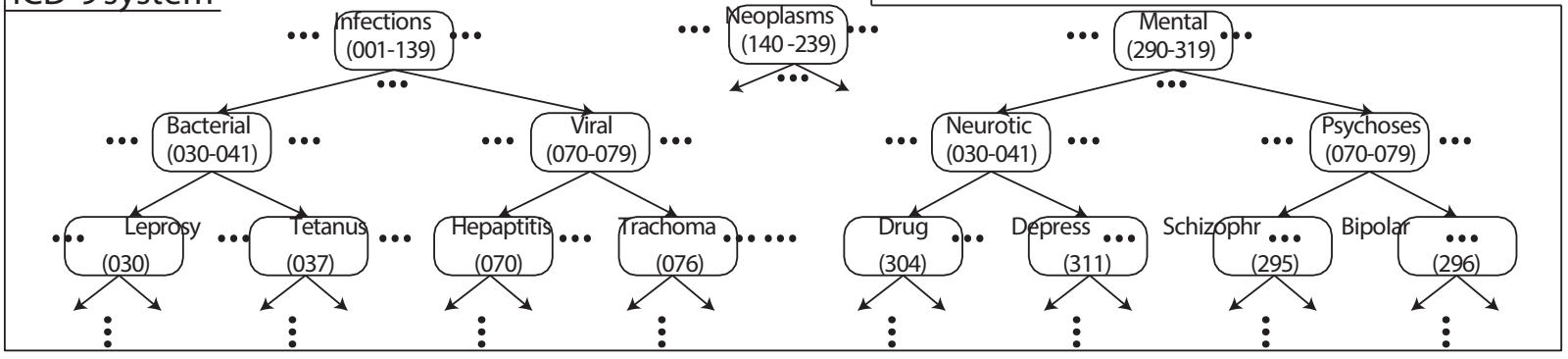


Future works

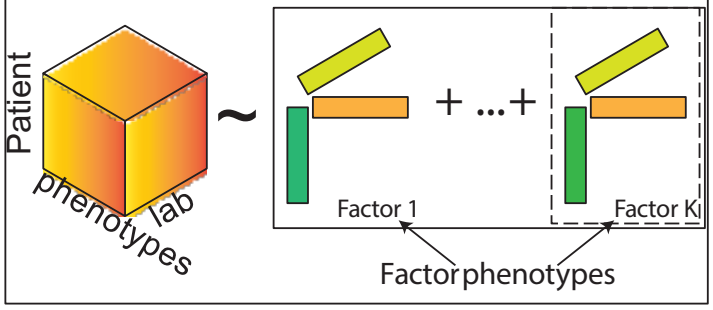
1. Integrating prior phe/pat networks



ICD-9 system

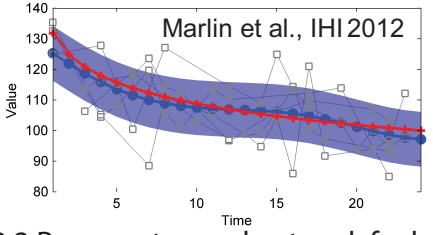


2. Tensor decomposition

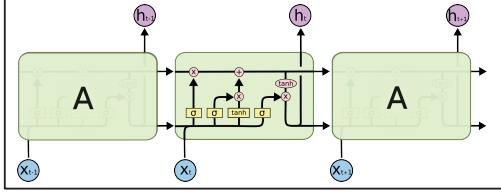


3. Temporal measurements

3.1 Gaussian kernel (Marlin et al., IHI 2012)
or Kalman filter (Qian, Osgood, & Stanley, 2014)



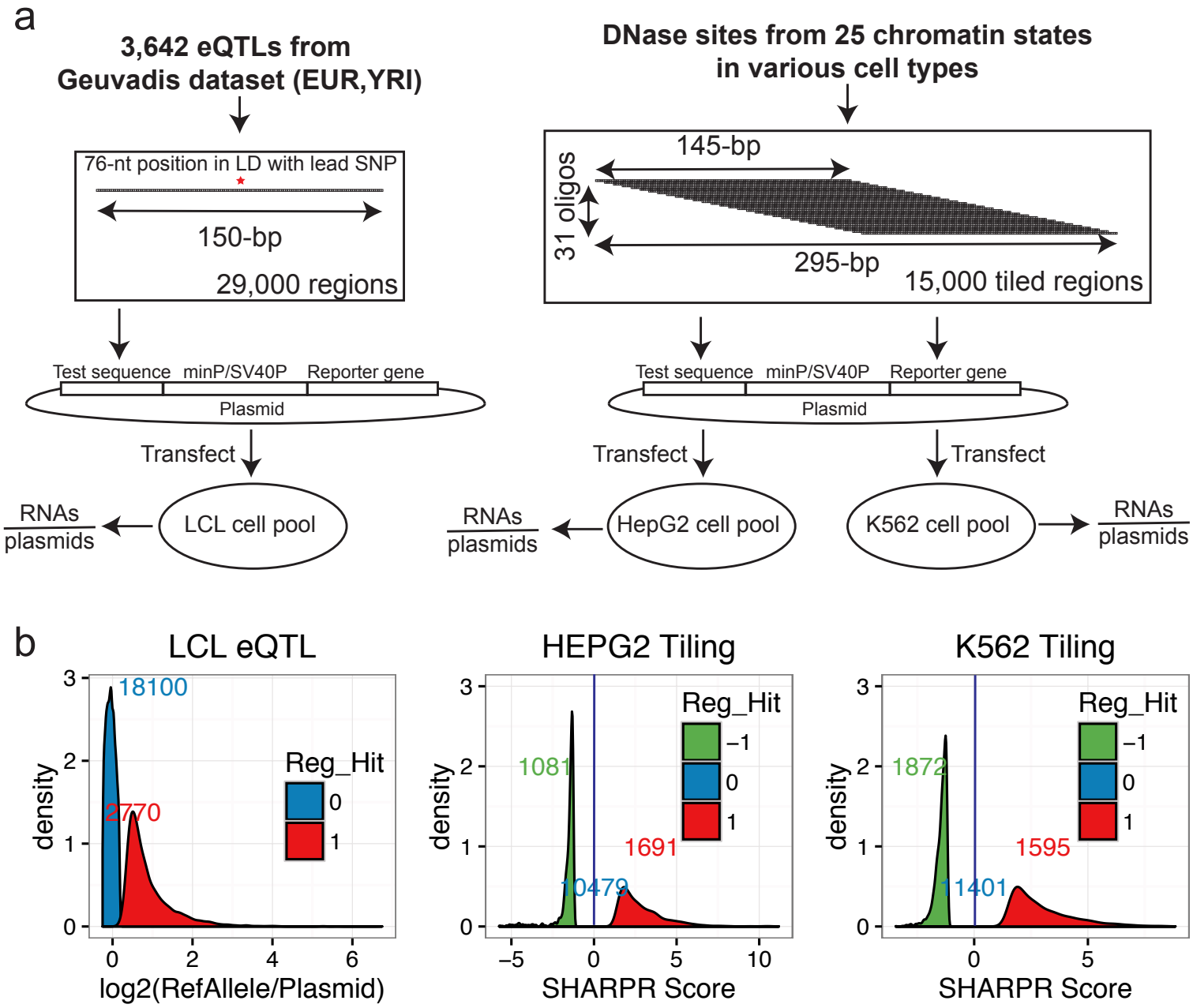
3.2 Recurrent neural network for long time trajectory



MPRA analysis

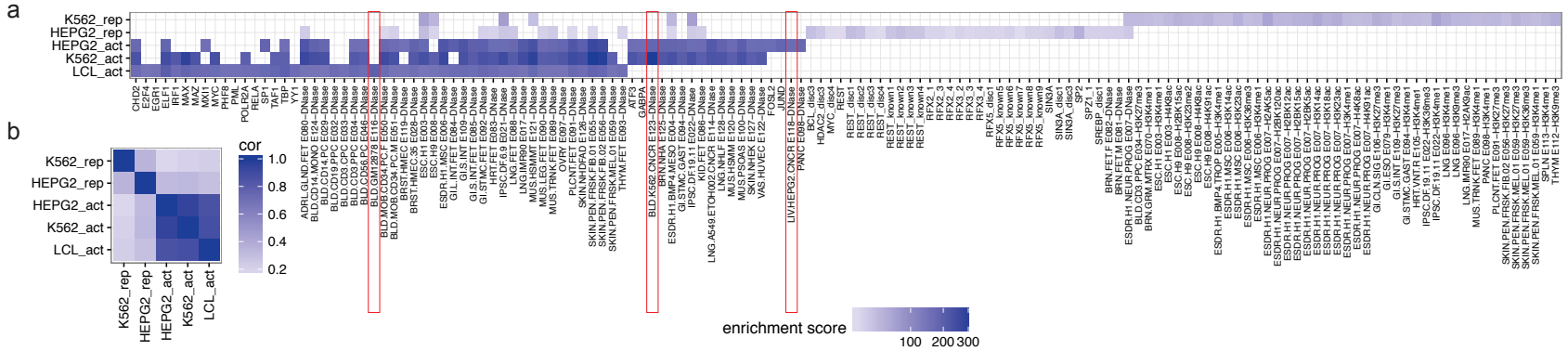
MPRA training data

Figure 1



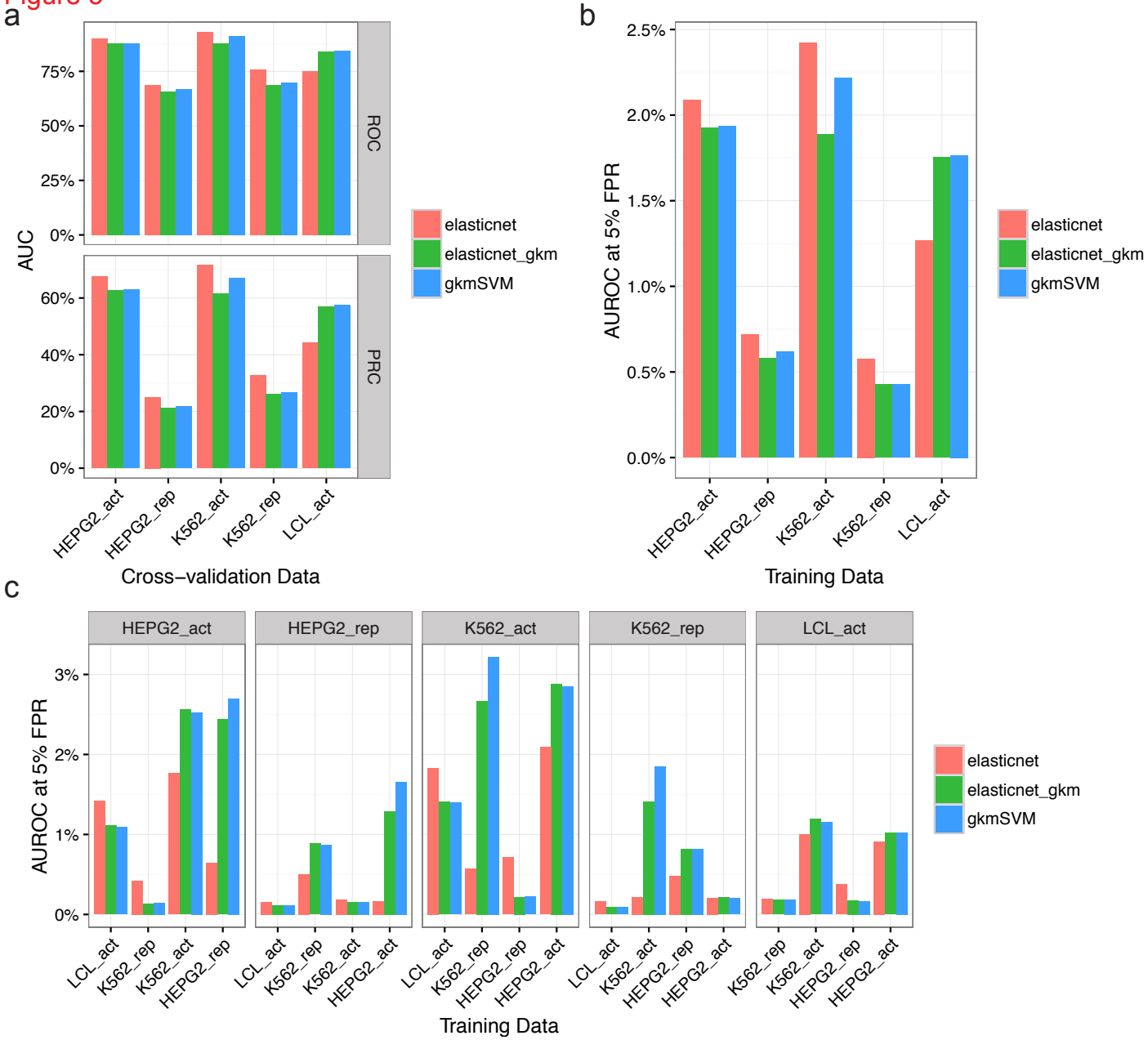
MPRA features enrichments

Figure 2



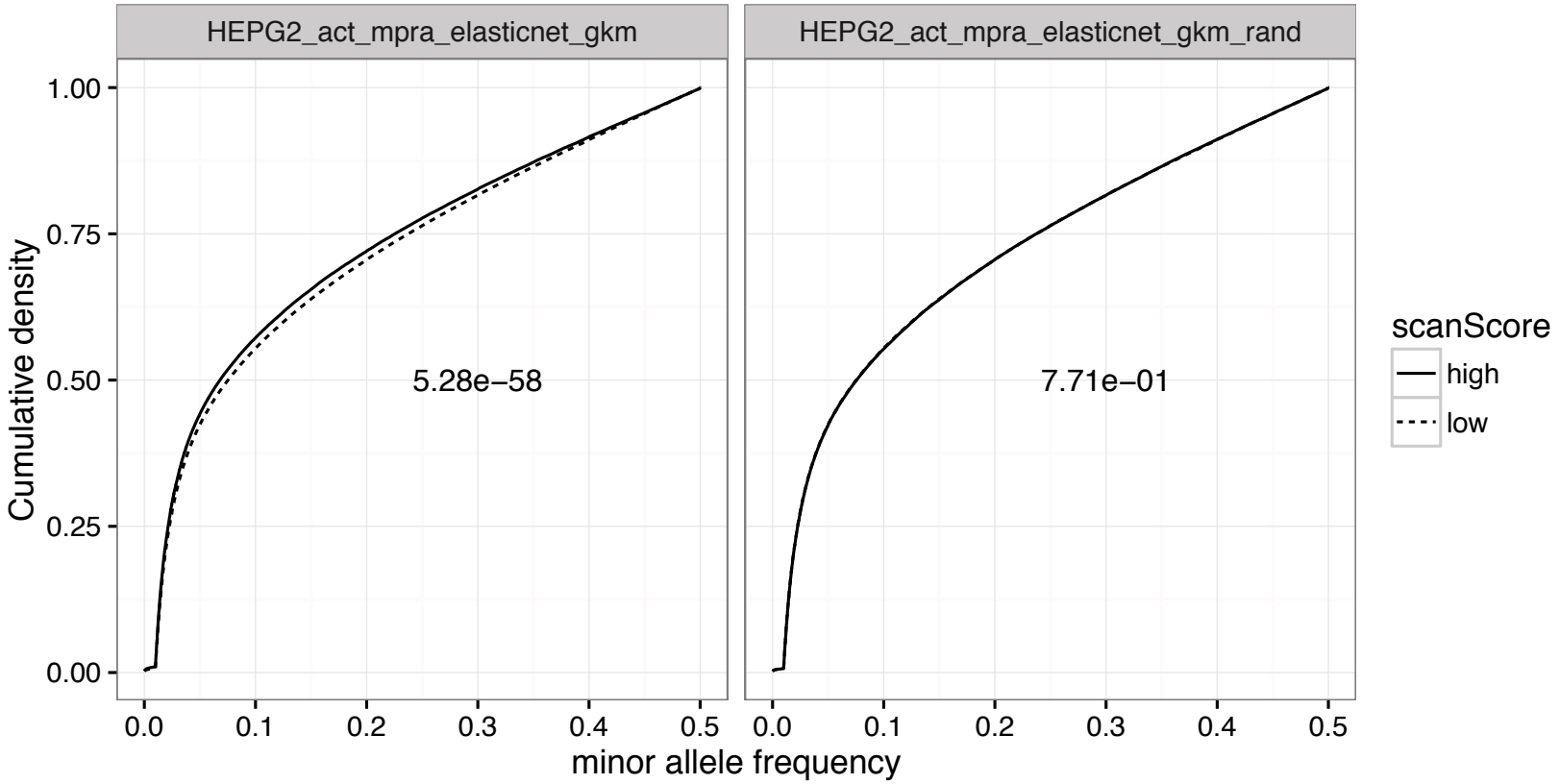
MPRA predictions

Figure 3



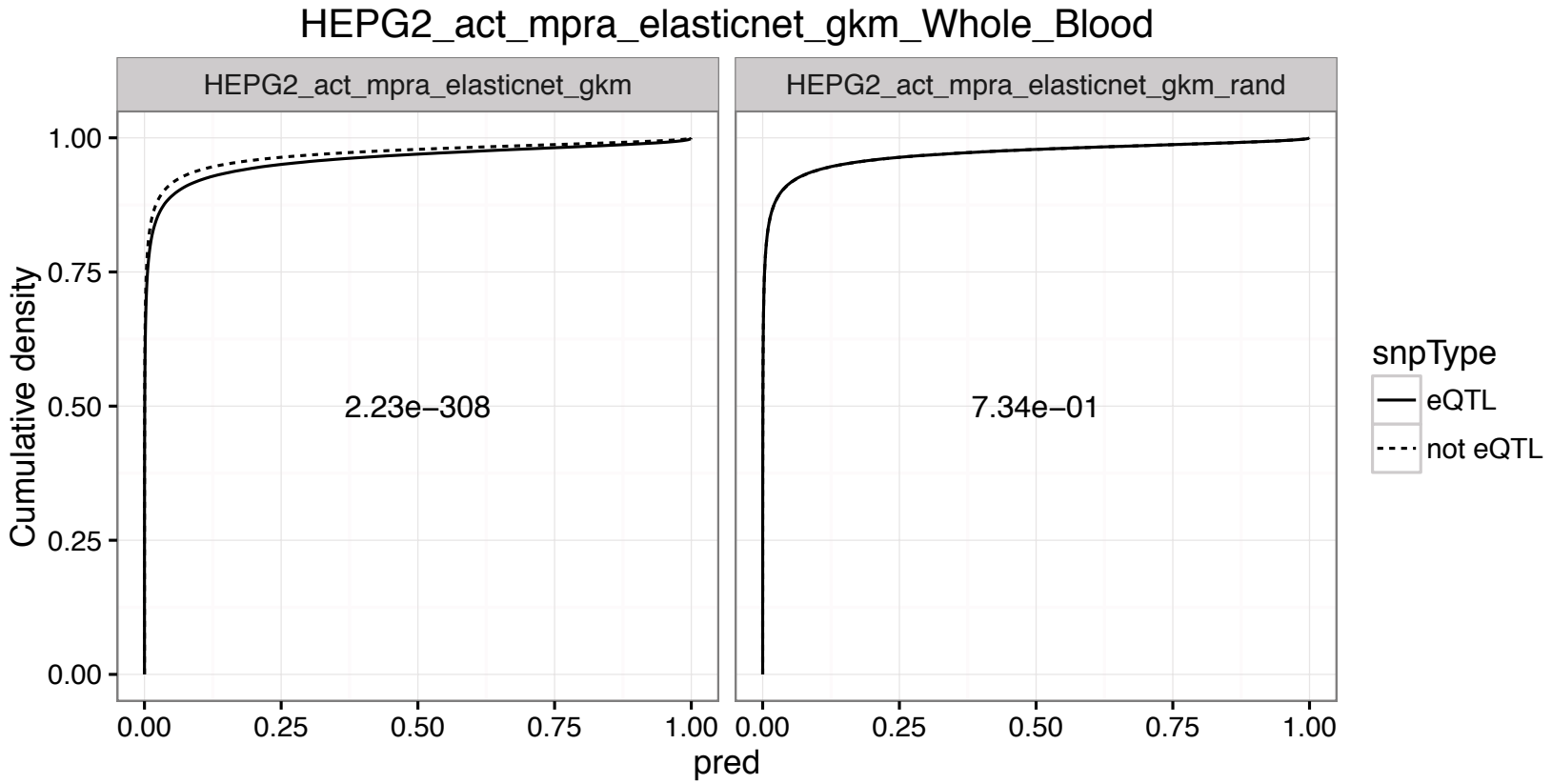
Common variants with high predicted scores exhibit lower MAF

Figure 4



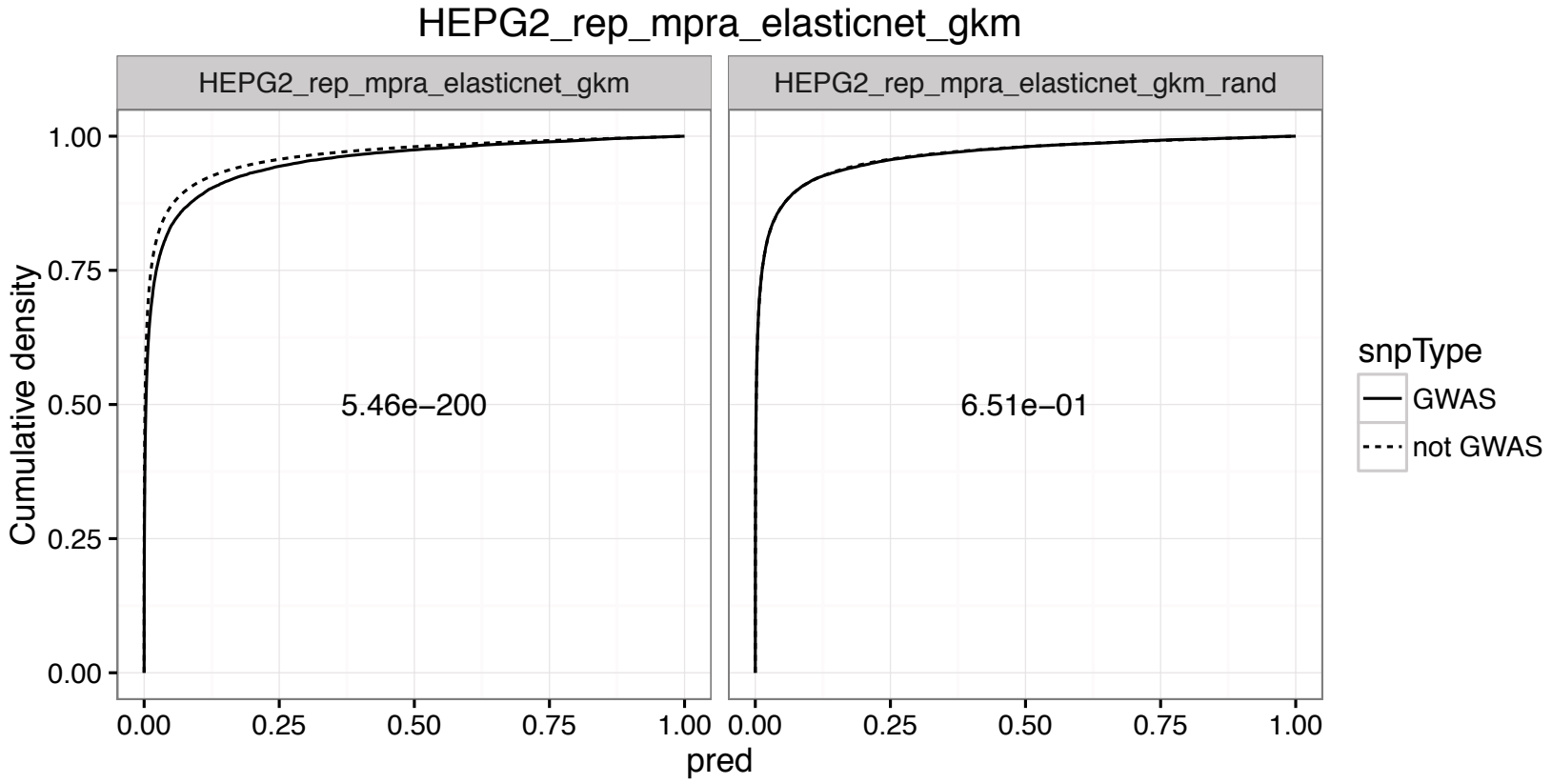
Common variants in eQTL exhibit higher predicted scores

Figure 5

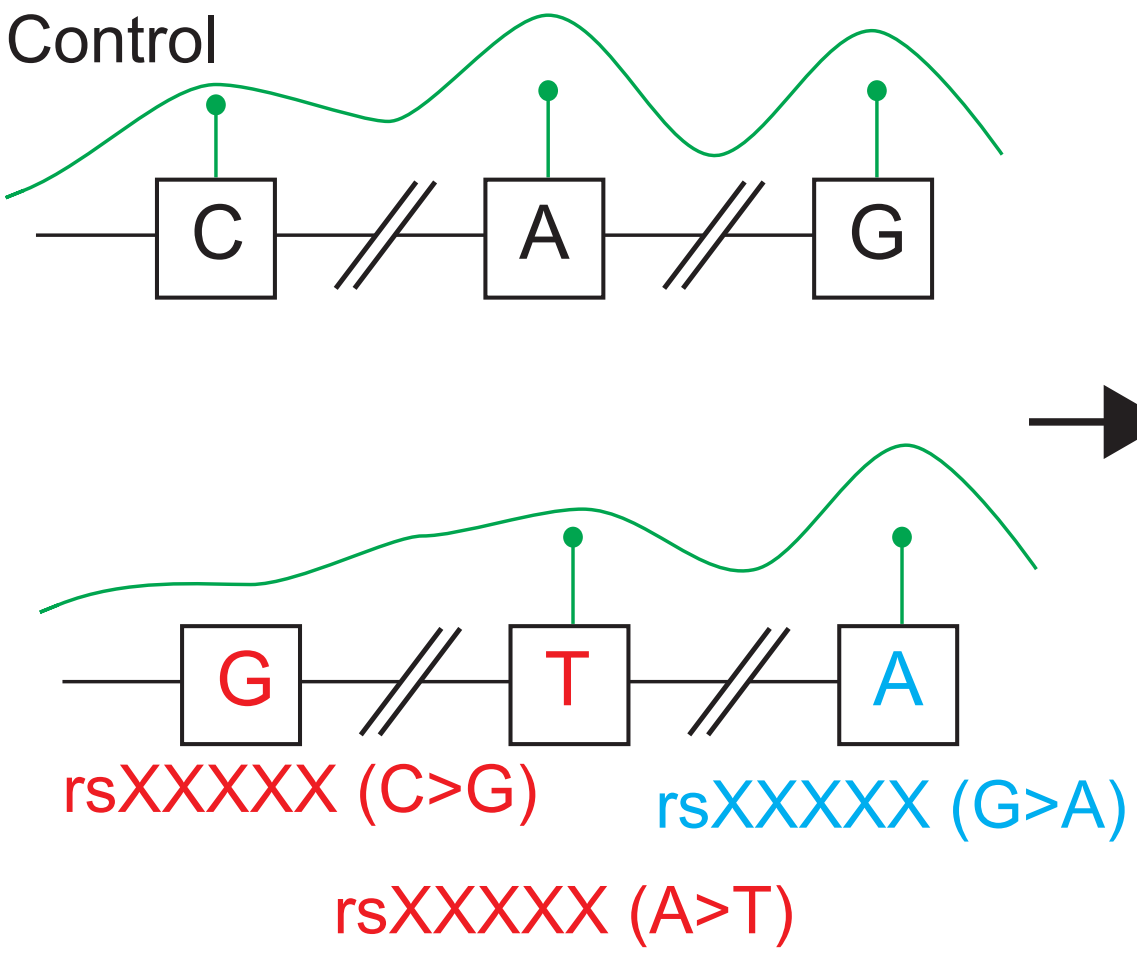


Common variants in GWAS catalog exhibit higher predicted scores

Figure 6



Incorporation of CNN model trained MPRA as prior model into the fine-mapping model



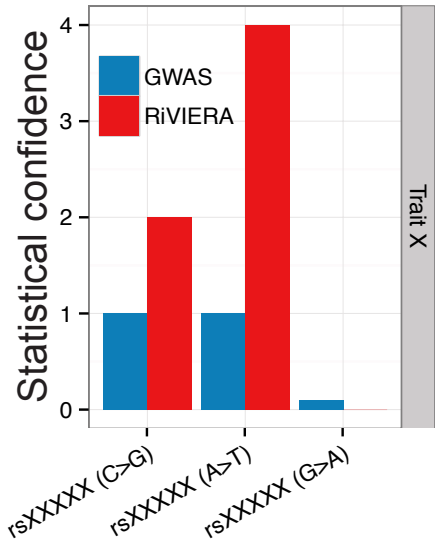


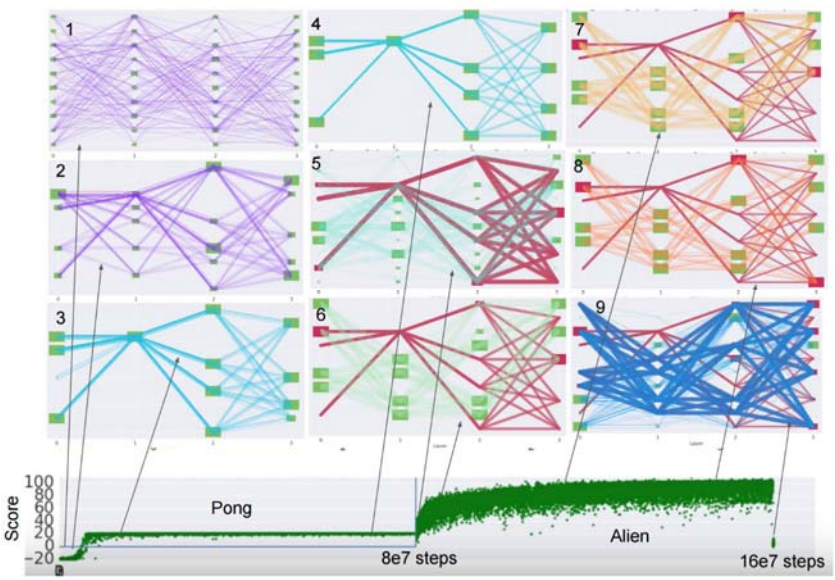
Illustration only

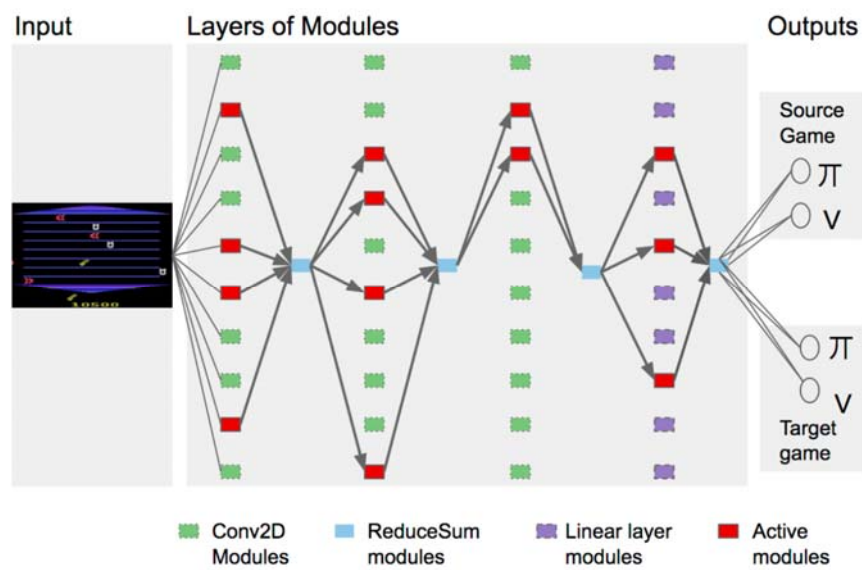
Transfer learning CNNs

Alvin Shi
Yue Li
Manolis Kellis
3/26/2017

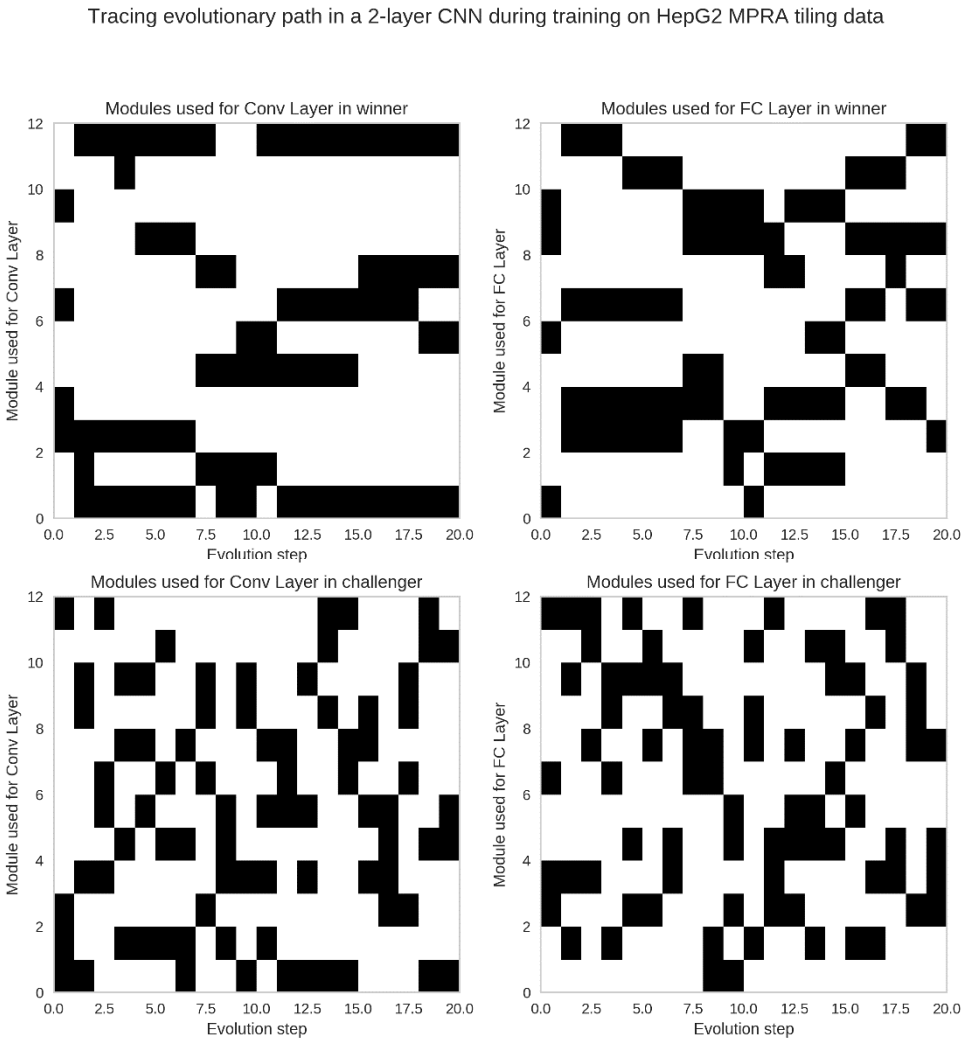
Evolutionary training algorithm and transfer learning

- **Training phase**
  - Initialize modules for each layer and randomly pick two subset as active
    - Example: We can initialize 10 modules, each containing 25 convolutional filters for a total of 250 total filters.
  - Designate the set of active modules a “path” or “genotype”
  - Train both networks until convergence and compare costs/performance metrics
    - Keep the “winning” path, with a small chance to mutate the winning path.
    - Reinitialize the “losing” path randomly
  - Repeat until desired number of iterations have concluded
- **Transfer from task 1 to task 2**
  - Network weights from best path from task 1 is frozen and remaining modules are reinitialized. Initialize the “winning” path for task 2 from the wining path from task 1.
  - Repeat training process until convergence for task 2.
- **Motivations for PathNet**
  - Generalizes the idea of dropout to modular sections of a neural net – prevents overfitting.
  - Prevents overfitting when training large networks when transferring from a larger training task to a small training task. Furthermore, decreases training time/cost when transferring between related tasks.





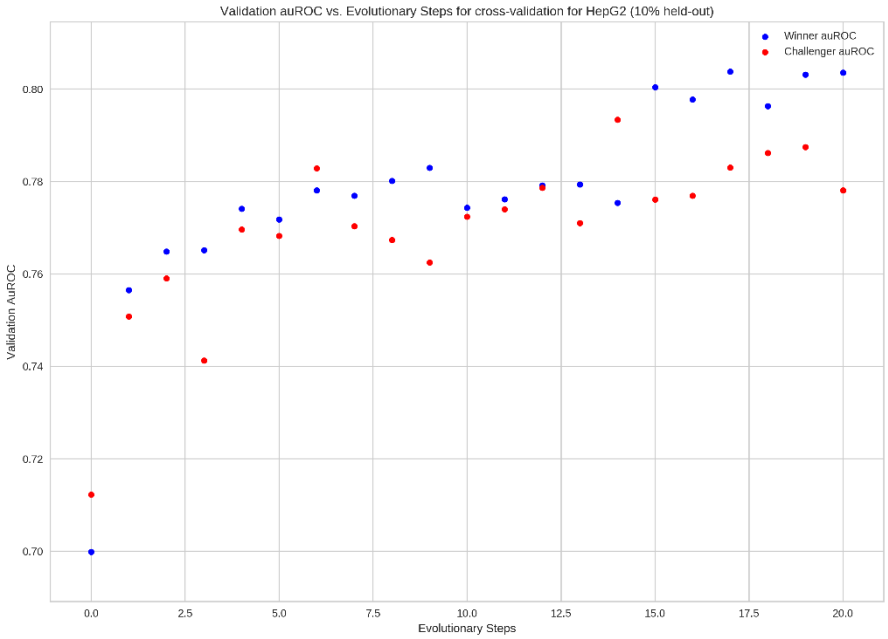
PathNet in action: Task 1

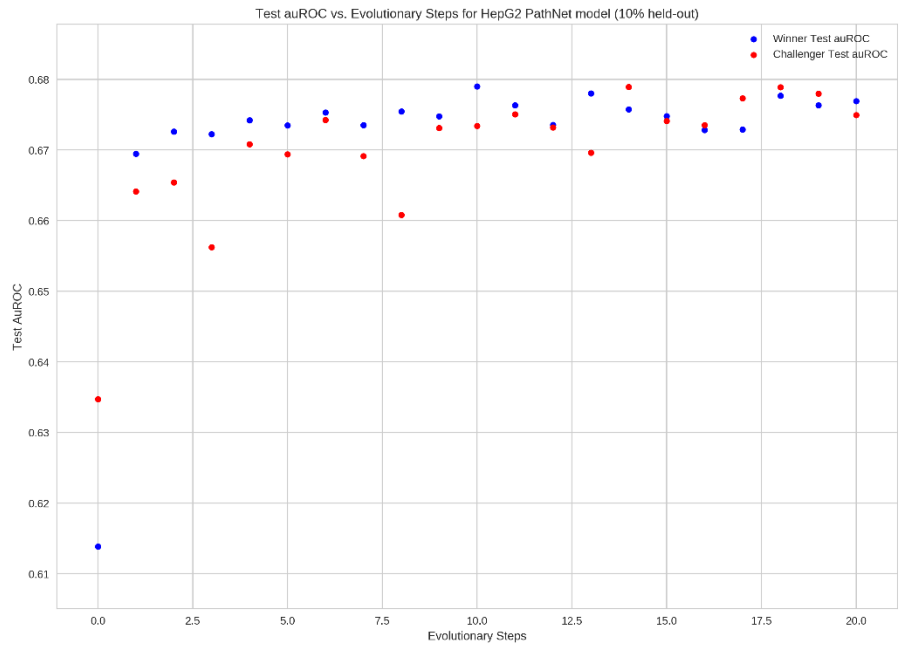


Layer 1: CNN

Layer 2: Fully-Connected

Shaded cells denote active modules

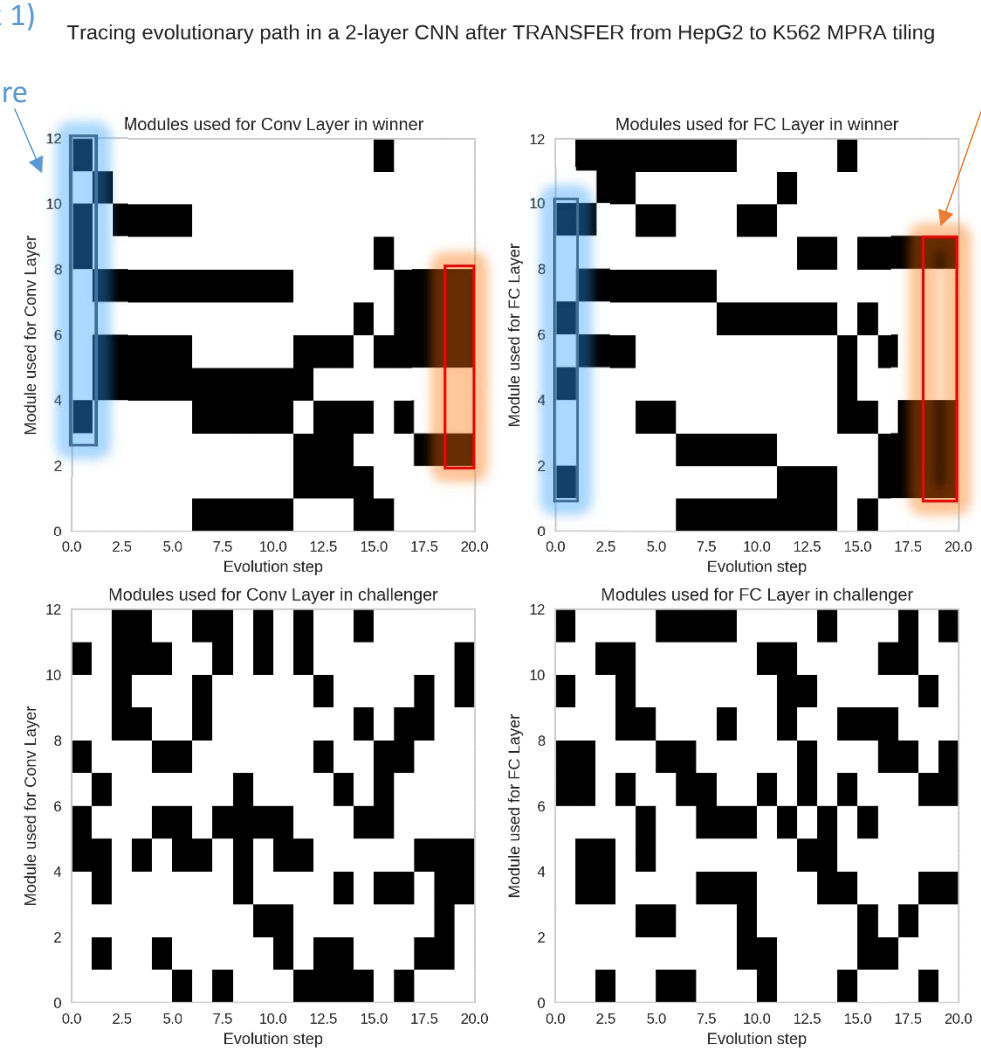




PathNet in action: Transfer from Task 1 to Task 2

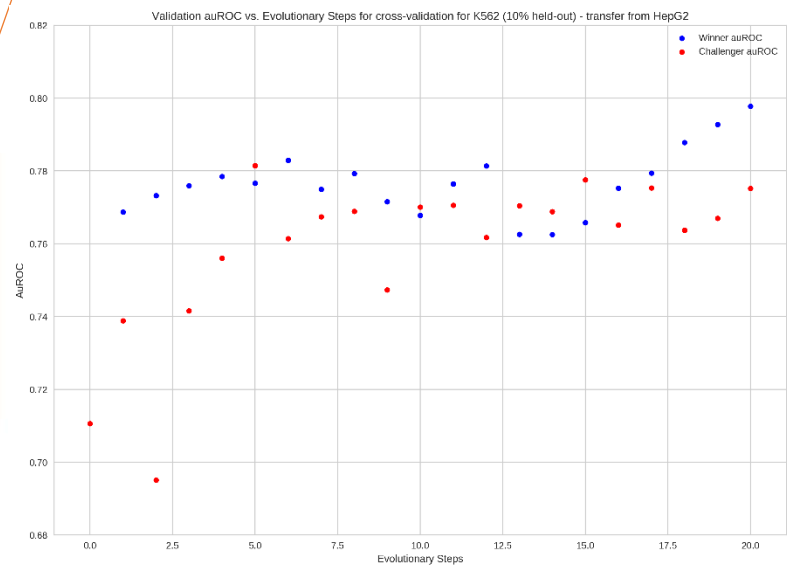
Transferred weights (from Task 1) for these modules are fixed

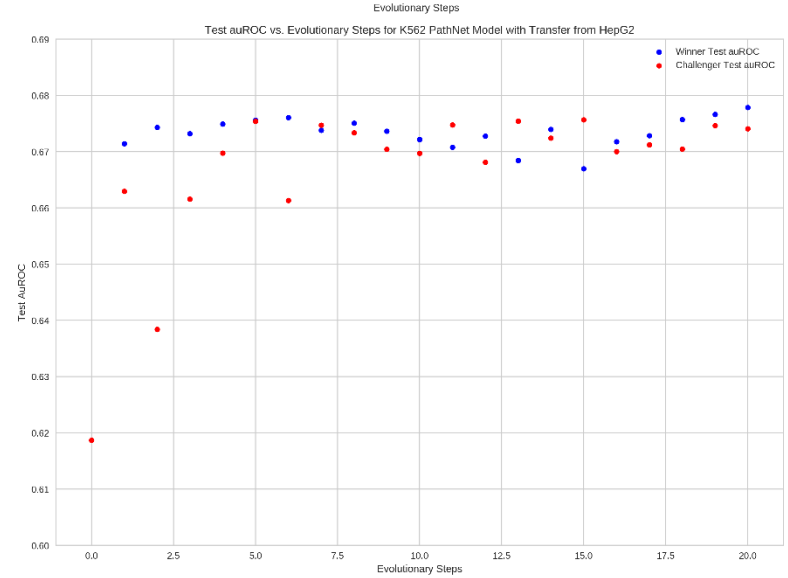
Depending on the task, only a fraction of the transferred modules will be used



Layer 1: CNN

Layer 2: Fully-Connected





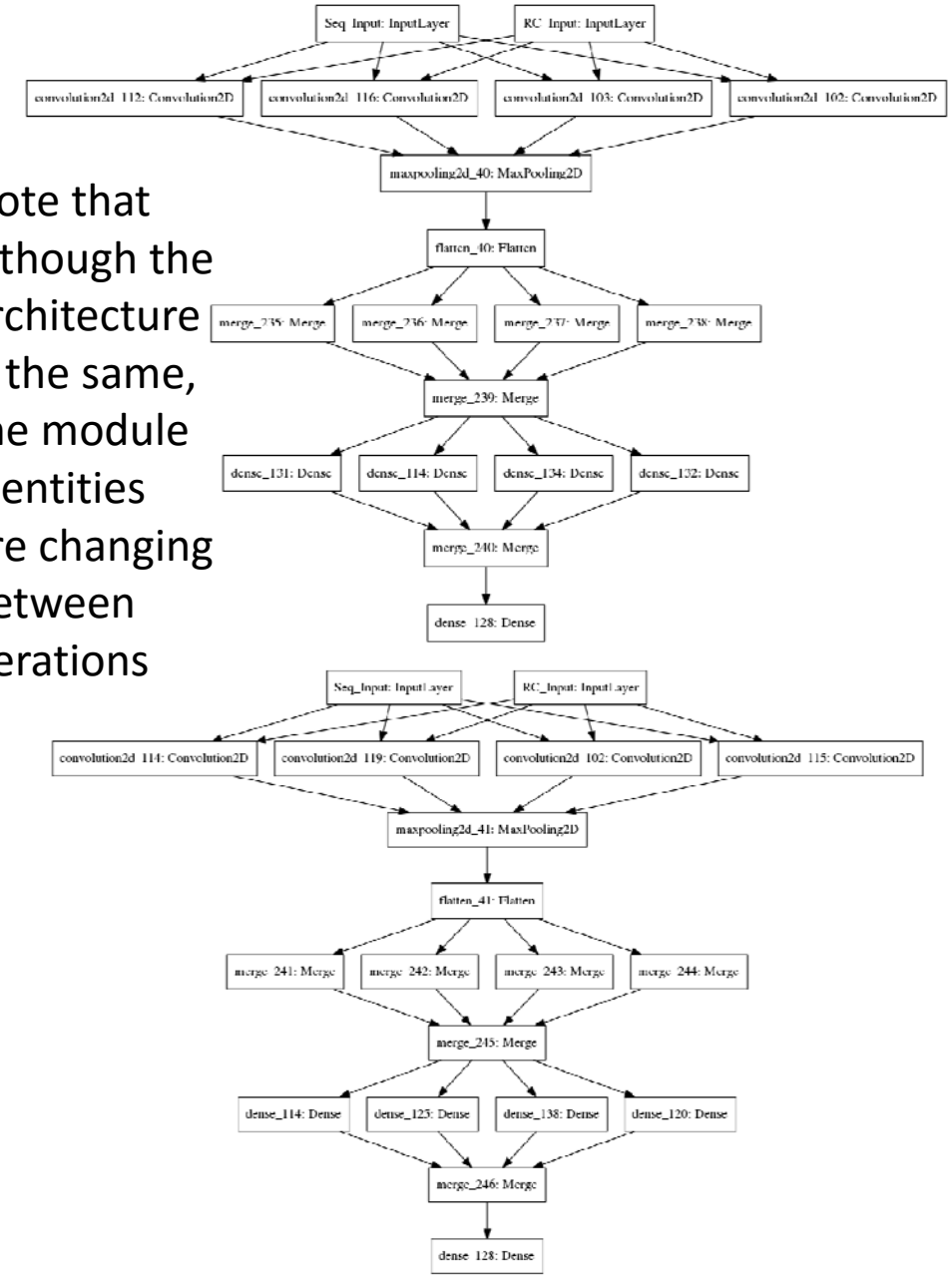
Pathnet Model

Layer (type)	Output Shape	Param #	Connected to
Seq_Input (InputLayer)	(None, 1, 4, 150)	0	
RC_Input (InputLayer)	(None, 1, 4, 150)	0	
convolution2d_19 (Convolution2D)	(None, 40, 1, 133)	2920	Seq_Input[0][0] RC_Input[0][0]
convolution2d_12 (Convolution2D)	(None, 40, 1, 133)	2920	Seq_Input[0][0] RC_Input[0][0]
convolution2d_13 (Convolution2D)	(None, 40, 1, 133)	2920	Seq_Input[0][0] RC_Input[0][0]
convolution2d_17 (Convolution2D)	(None, 40, 1, 133)	2920	Seq_Input[0][0] RC_Input[0][0]
maxpooling2d_82 (MaxPooling2D)	(None, 40, 1, 33)	0	convolution2d_19[34][0] convolution2d_12[38][0] convolution2d_13[24][0] convolution2d_17[28][0] convolution2d_19[35][0] convolution2d_12[39][0] convolution2d_13[25][0] convolution2d_17[29][0]
flatten_81 (Flatten)	(None, 1320)	0	maxpooling2d_82[0][0] maxpooling2d_82[1][0] maxpooling2d_82[2][0] maxpooling2d_82[3][0] maxpooling2d_82[4][0] maxpooling2d_82[5][0] maxpooling2d_82[6][0] maxpooling2d_82[7][0]
merge_476 (Merge)	(None, 1320)	0	flatten_81[0][0] flatten_81[4][0]
merge_477 (Merge)	(None, 1320)	0	flatten_81[1][0] flatten_81[5][0]
merge_478 (Merge)	(None, 1320)	0	flatten_81[2][0] flatten_81[6][0]
merge_479 (Merge)	(None, 1320)	0	flatten_81[3][0] flatten_81[7][0]
merge_480 (Merge)	(None, 1320)	0	merge_476[0][0] merge_477[0][0] merge_478[0][0] merge_479[0][0]
dense_23 (Dense)	(None, 4)	5284	merge_480[0][0]
dense_16 (Dense)	(None, 4)	5284	merge_480[0][0]
dense_22 (Dense)	(None, 4)	5284	merge_480[0][0]
dense_18 (Dense)	(None, 4)	5284	merge_480[0][0]
merge_481 (Merge)	(None, 4)	0	dense_23[14][0] dense_16[16][0] dense_22[23][0] dense_18[3][0]
dense_26 (Dense)	(None, 1)	5	merge_481[0][0]

Total params: 32,821
Trainable params: 32,821
Non-trainable params: 0

Pathnet model at two evolutionary time steps:

Note that although the architecture is the same, the module identities are changing between iterations



Testing PathNet: MPRA Transfer Learning

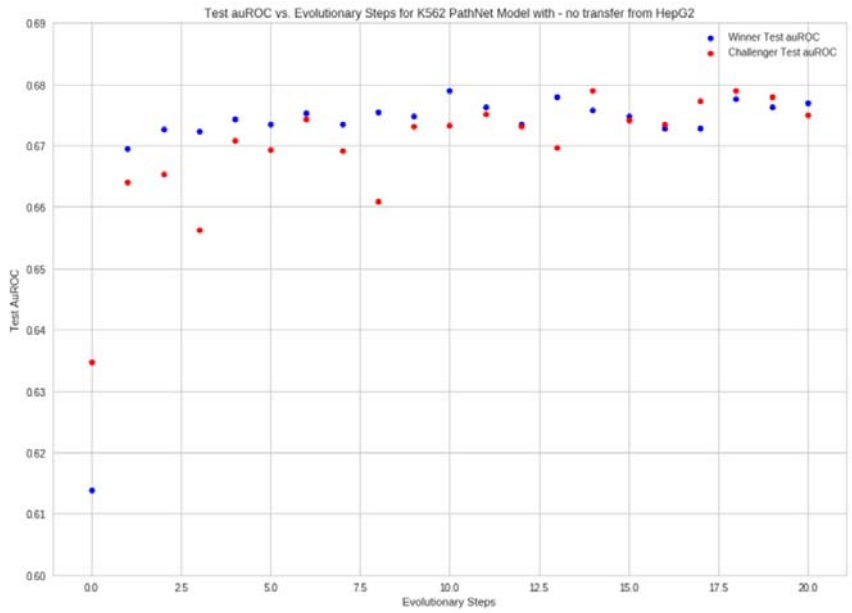
- Training Data from Ernst et al.[1]
  - Task 1: Binarized HepG2 MPRA tiling data (10% validation)
  - Task 2: Binarized K562 MPRA tiling data (10% validation)
- Testing Data
  - Testing dataset: Binarized LCL MPRA data
- Evaluate against matched CNN
  - Same architecture, hyperparameter settings, total number of weights

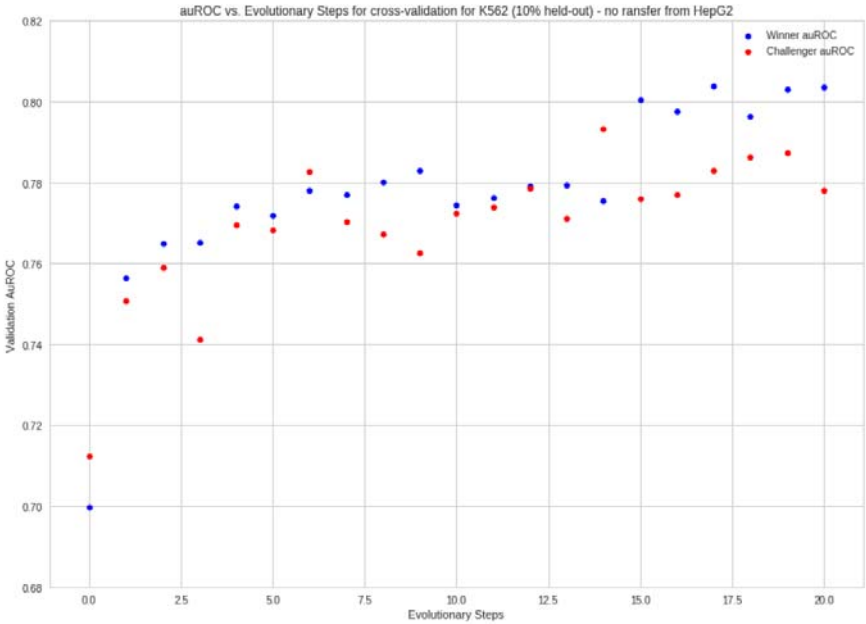
Task 1: HepG2	Validation auROC	Test auROC
PathNet	0.735	0.67
Matched CNN	0.726	0.65

Task 2: K562	Validation auROC	Test auROC
PathNet + Transfer	0.80	0.68
PathNet	0.80	0.67
Matched CNN	0.73	0.64

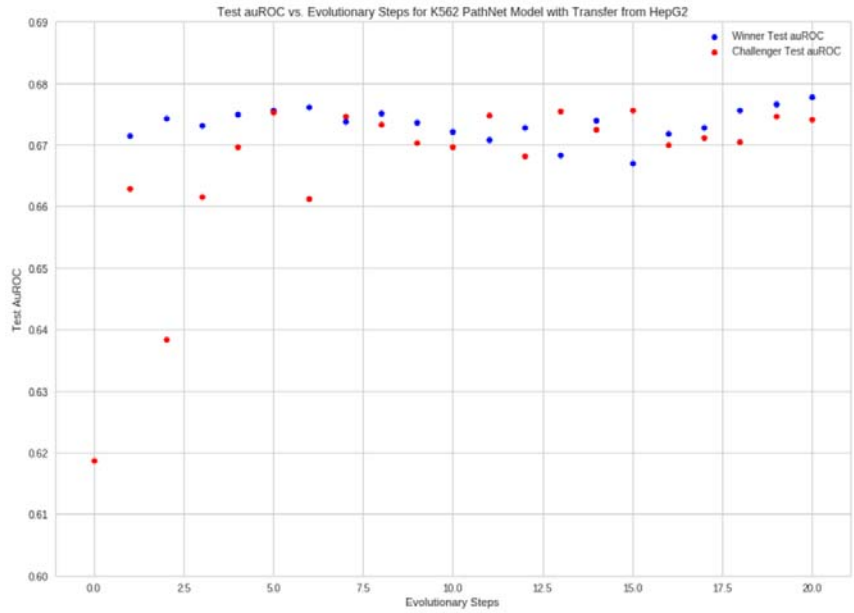
1 - Ernst, J., Melnikov, A., Zhang, X., Wang, L., Rogov, P., Mikkelsen, T. S., & Kellis, M. (2016). Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nature Biotechnology*, (October 2015). http://doi.org/10.1038/nbt.3678
2 - Tewhey, R., Kotliar, D., Park, D. S., Liu, B., Winnicki, S., Reilly, S. K., … Sabeti, P. C. (2016). Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*, *165*(6), 1519–1529. http://doi.org/10.1016/j.cell.2016.04.027

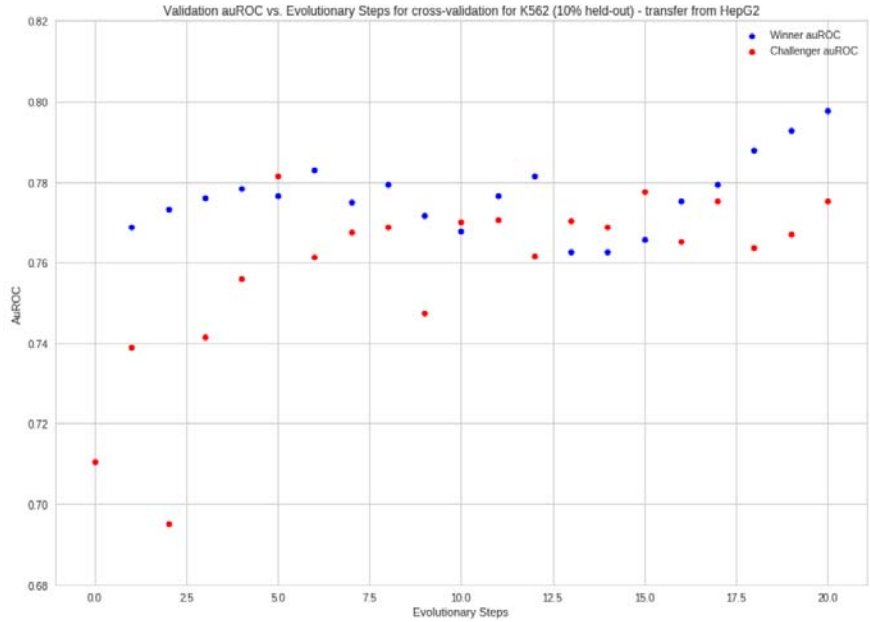
PathNet – No Transfer





PathNet – With Transfer





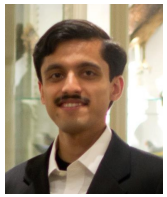
Conclusion and future directions

- On our initial tests, PathNet outperforms generic CNNs in single-task prediction.
- Current implementation of transfer learning produces no tangible evidence of faster training on MPRA testing/training datasets.
  - Continue to evaluate other transfer methods and other relevant prediction tasks.
  - Implement and test alternative transfer learning method (freeze both weights and path in task 2 – thereby expanding the total number of utilized modules).

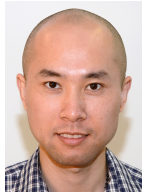
Decomposition and interpretation of Alzheimer's disease GWAS statistics from transcriptomic and epigenomic regulatory programs

Yongjin Park,
MIT CSAIL / Broad Institute


Yongjin Park


Abhishek Sarkar


Benjamin Iriarte


Liang He


Manolis Kellis


Kunal Bhutani

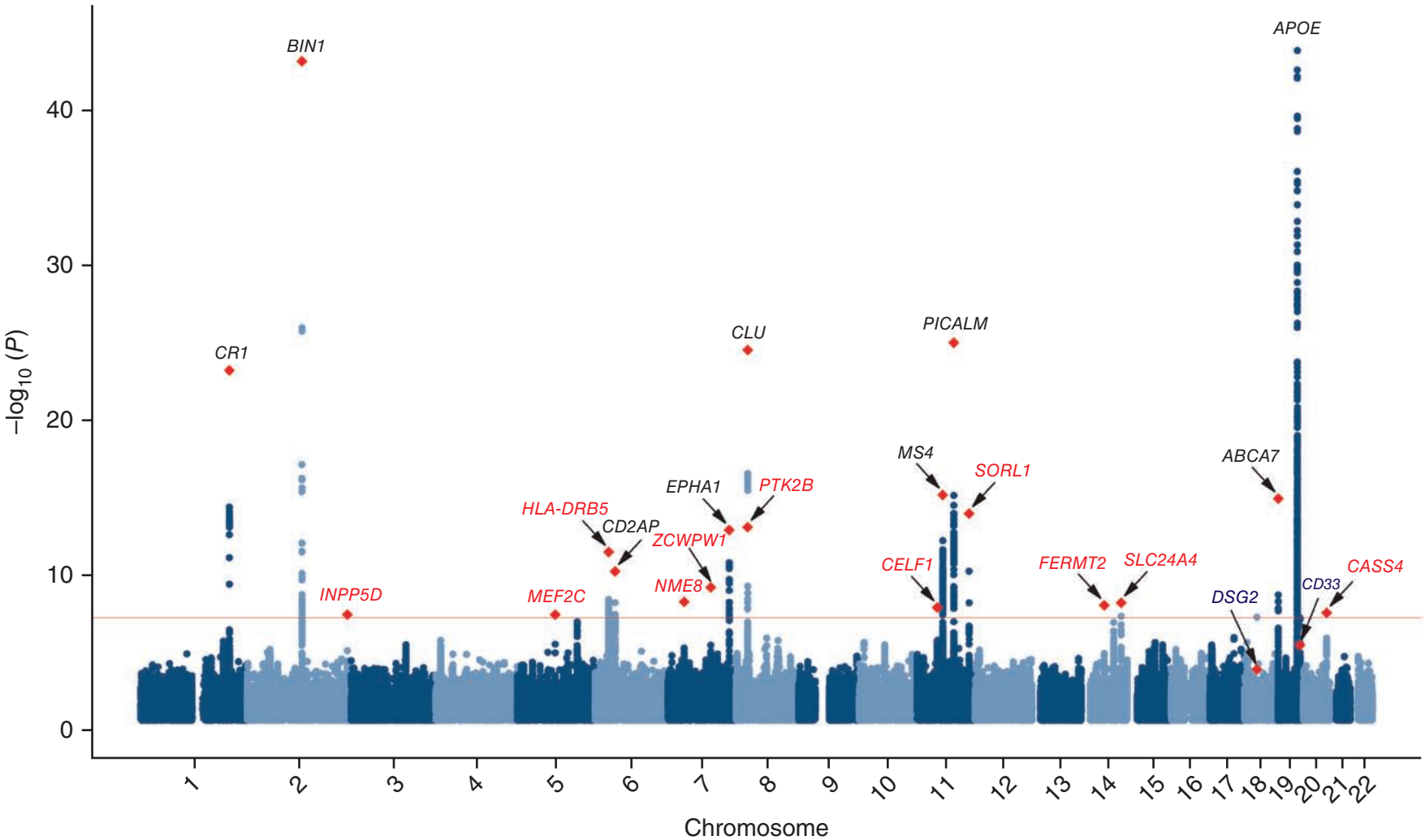

Bogdan Pasaniuc


Nick Mancuso


Alexander Gusev

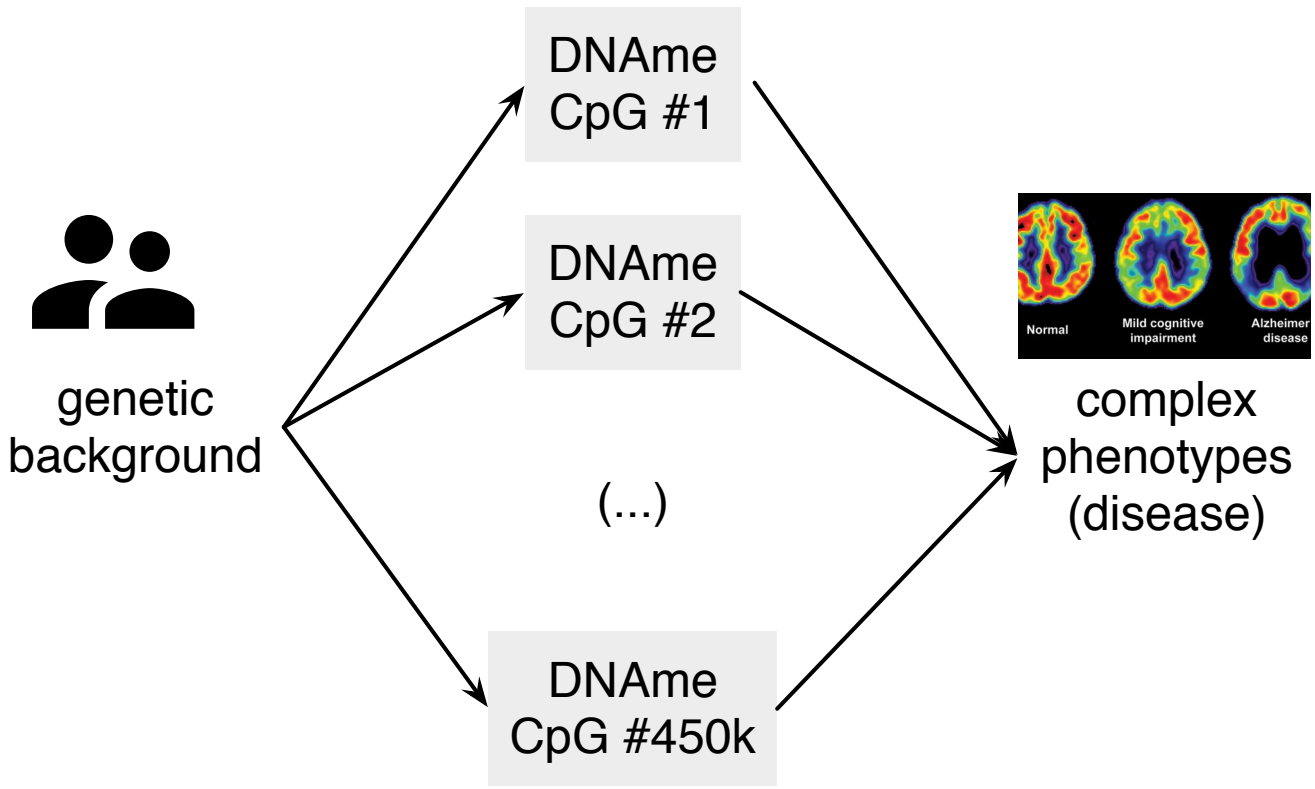

Philip De Jager

GWAS implicates existence of molecular & cellular mechanisms in Alzheimer's disease

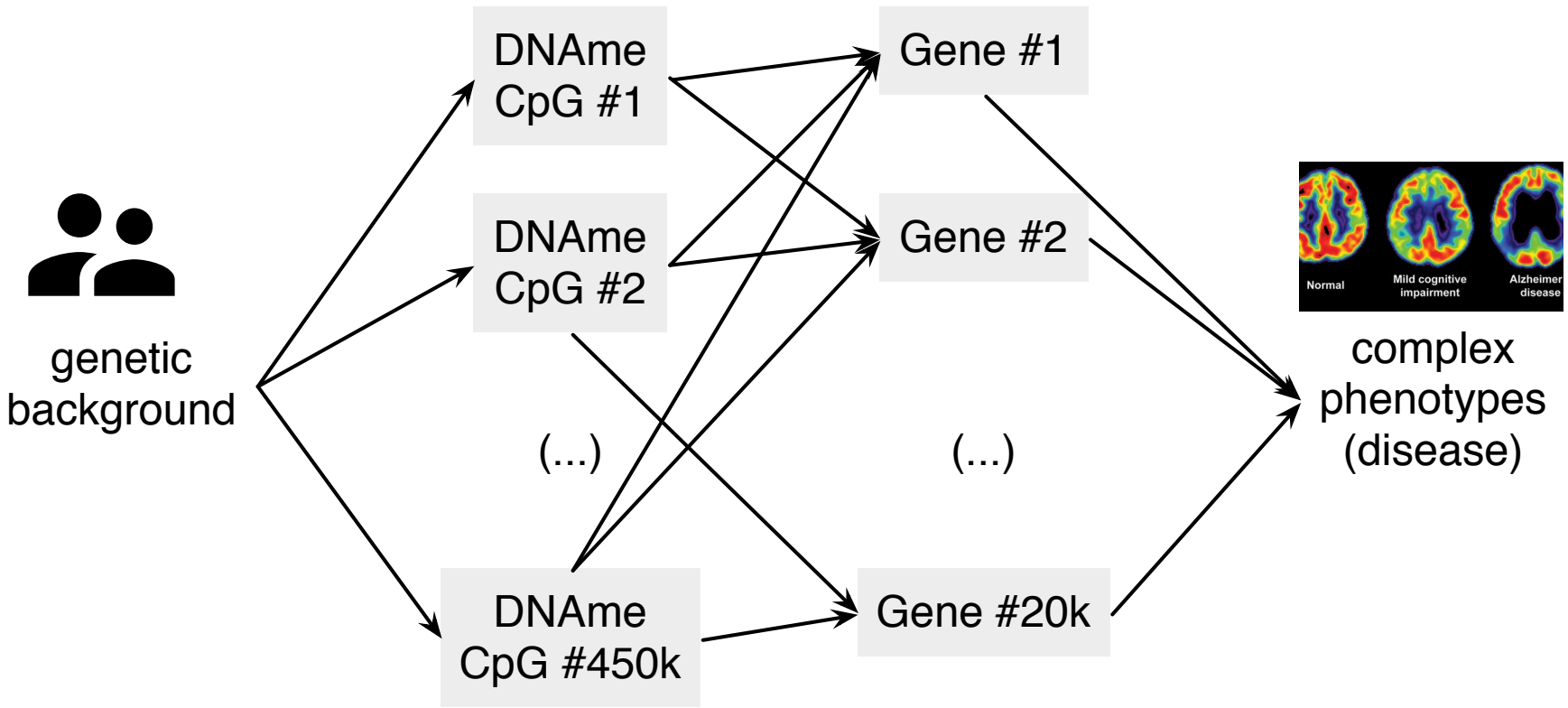


Lambert *et al.* Nat. Gen. (2013)

Goal: Identify embedded associations / causation of regulatory elements in between genetics and phenotype associations



Deeper knowledge of GWAS: Identify multiple types regulatory programs using multi-omics data



Association of phenotypic variability with imputed regulatory signals

Imputed *T*WAS

(1) Train a linear model of gene expression on reference cohort

$$GE_{ref} \approx X_{ref}\, \theta_{qtl}$$

(2) Impute individual-level gene expressions

$$GE_{pred} \leftarrow X_{gwas}\, \theta_{qtl}$$

(3) Measure correlation between the predicted expr. and observed phenotypes

$$Pheno \sim GE_{pred}$$

Gamazon *et al.* Nat. Gen. (2015)

Association of phenotypic variability with imputed regulatory signals

Imputed *TWAS*

(1) A linear model of gene expression on reference cohort

$$GE_{ref} \approx X_{ref}\, \theta_{qtl}$$

(2) Imputed gene expressions

$$GE_{pred} \approx X_{gwas}\, \theta_{qtl}$$

What if we don't have access to individual genotype data? or *n* is too small?

But we could have access to well-powered summary SNP-level (marginal) effect sizes!

Gamazon *et al.* Nat. Gen. (2015)

Summary-based *TWAS*

(1) Reference cohort QTL model

$$GE_{ref} \approx X_{ref}\, \theta_{qtl}$$

(2) Skip imp. & find a walk-round sol'n

Goal: $\phi \sim GE_{pred}$?

Assume: $E[\phi] = X\, \theta_{gwas}$

Test stat. $T := GE_{pred}^{\top} \phi_{pred} / n$

$$E[T|gwas] \approx (X\theta_{qtl})^{\top} X\theta_{gwas} / n$$
$$\approx \theta_{qtl}^{\top}\, LD\, \theta_{gwas}$$
$$\approx \theta_{qtl}^{\top}\, z_{gwas}$$

$$V[T|gwas] \approx \theta_{qtl}^{\top}\, LD\, \theta_{qtl}$$

Gusev *et al.* Nat. Gen. (2016)

Fine-mapped identification of causal SNPs by co-localization of eQTL and GWAS

Summary-based test

Test stat. $T := GE_{pred}^{\top}\phi_{pred}/n$

$$E[T|gwas] \approx (X\theta_{qtl})^{\top} X\theta_{gwas}/n$$
$$\approx \theta_{qtl}^{\top}(LD\ \theta_{gwas})$$

$$V[T|gwas] \approx \theta_{qtl}^{\top} LD\ \theta_{qtl}$$

Co-localization of eQTL + GWAS

$$z_{gwas} \sim N(\lambda_g\ LD\ \theta_{shared}, LD)$$

$$z_{qtl} \sim N(\lambda_q\ LD\ \theta_{shared}, LD)$$

Aggregation of multiple signals within *cis*-region (causal + passenger)

Find credible set of SNPs driving both GWAS and QTL z-scores.

Hormozdiari .. Eskin, bioRxiv (2016)

Contributions of our work

Improving regulatory programs

- Accurately model types of data (DNAme arrays, RNA-seq, Chip-seq)
- Aggregating related information (tissue axis or multiple gene axis)
- Spike-slab type of sparse regression (reduce generalization errors; parsimonious model)
- Multiple levels of regulatory models

Summary-based *N*WAS

(1) Reference cohort QTL model

$$Reg_{ref} \approx X_{ref}\, \theta_{qtl}$$

(2) Test regulatory association

Goal: $\phi \sim Reg_{pred}$?

Contributions of our work

Improving regulatory programs

- Accurately model types of data (DNAme arrays, RNA-seq, Chip-seq)
- Aggregating related tissues (tissue axis or multiple gene axis)
- Spike-slab type of sparse regression (reduce generalization errors; parsimonious model)

Summary-based *N*WAS

(1) Reference cohort QTL model

$$Reg_{ref} \approx X_{ref}\, \theta_{qtl}$$

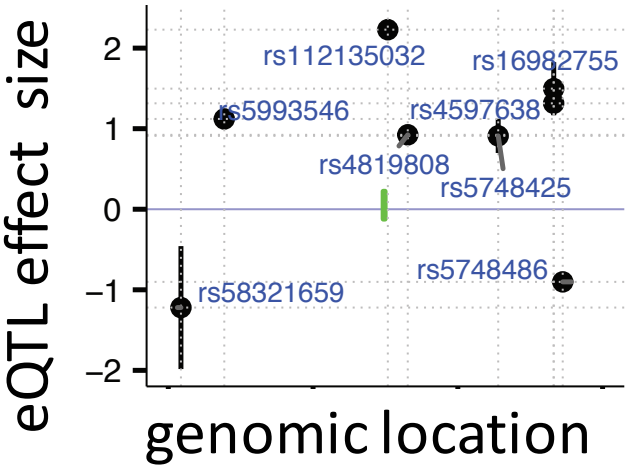
(2) Test regulatory association

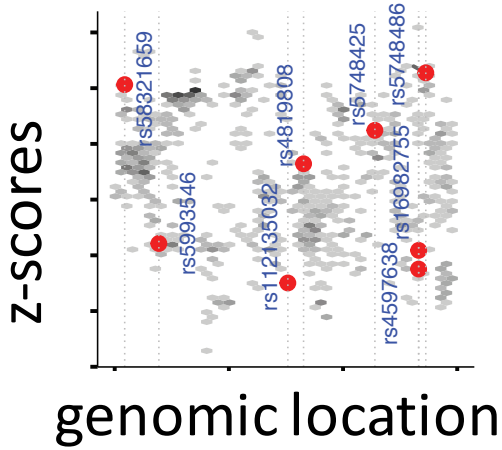
Goal: $\phi \sim Reg_{pred}$?

Distinguish sources of information

- Correct reverse-causation using observed phenotypes / proxy
- Account for direct effects in summary-based models

TWAS reveals target genes with tissue and cellular context by aggregating multivariate effects



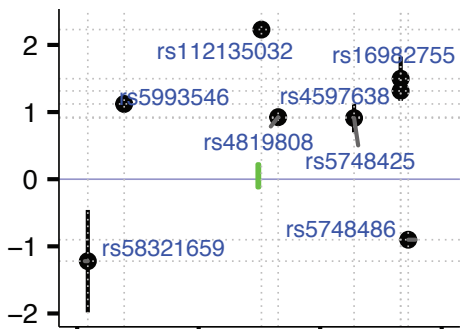


Reference cohort with regulatory contexts (GTEx tissues)

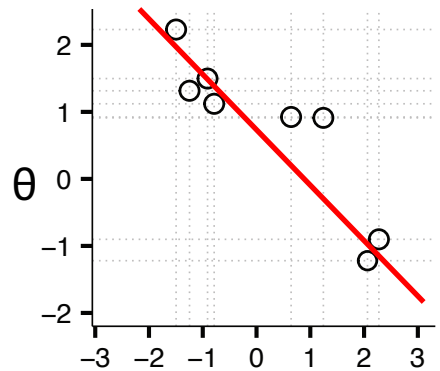
GWAS (well powered, but no context)

TWAS reveals target genes with tissue and cellular context by aggregating multivariate effects

Multi-SNP eQTL effect θ



genomic location



z-scores

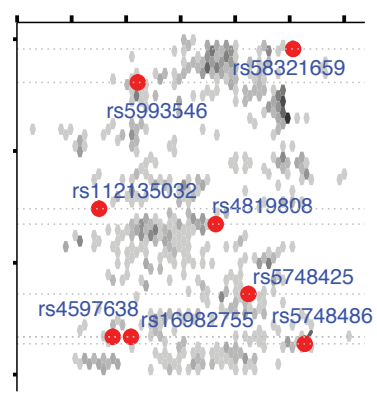
sTWAS statistic:

$$T = \theta^\top z$$

Standard error:

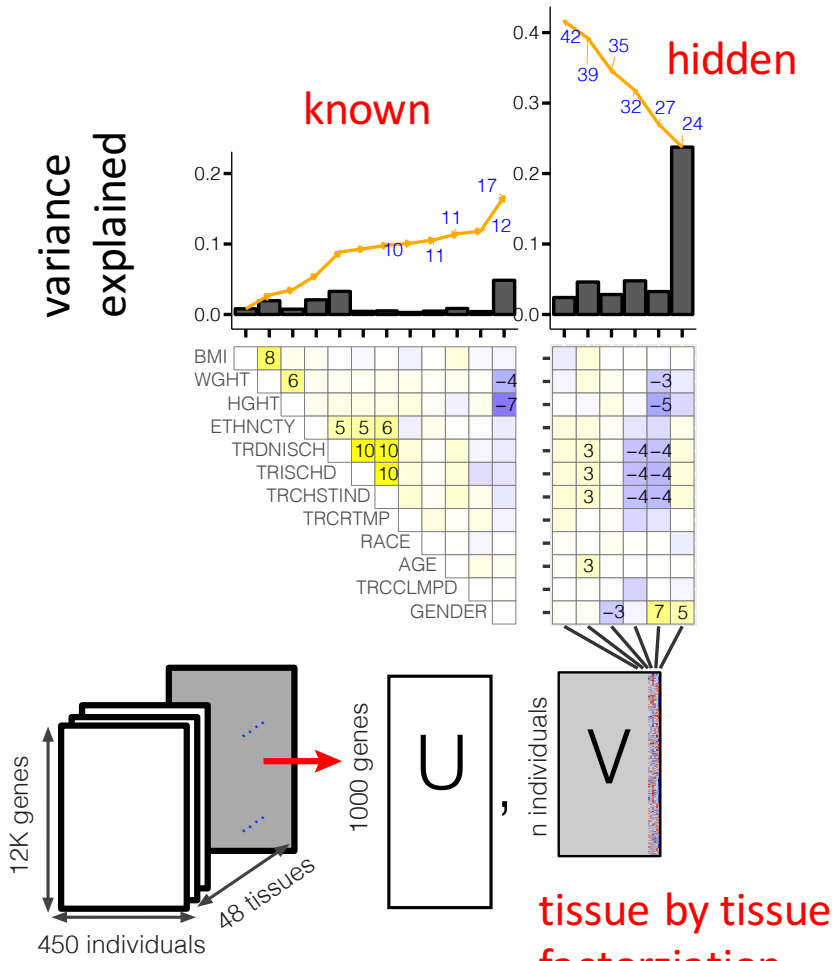
$$(\theta^\top LD\ \theta)^{1/2}$$

Reference cohort with regulatory contexts



GWAS (well powered, but no context)

Removing non-genetic sources of variability using low-ranked matrix factorization model



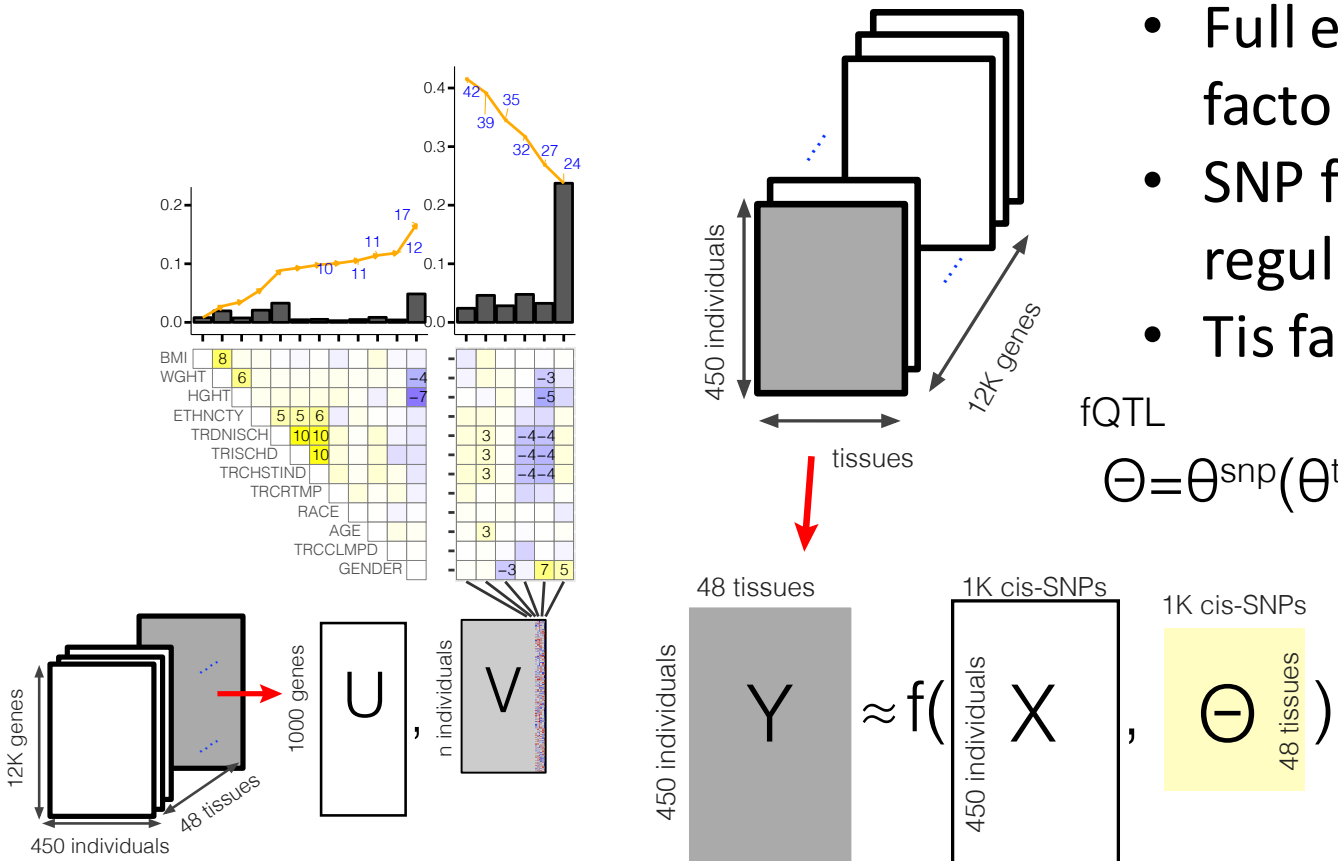
Factor analysis on most variable 1,000 genes in tissue and correct hidden & known confounders

tissue by tissue factorziation in GTEx

- Matrix factorization with known covariates (including demographic, technical confounders & common variants within 1Mb *cis*-regulatory regions of each gene)
- Automatic identification of ranks using generalized spike-slab prior on columns of latent factors; resolve #dimensions by posterior probability > .5)

Park & Sarkar *et al*. bioRxiv (2017)

Joint training of 48 GTEx tissues on shared genotype matrix with factored regression model



Factor analysis on most variable 1,000 genes in tissue and correct hidden & known confounders

Fit multi-tissue & polygenic regression with factored regression coefficients

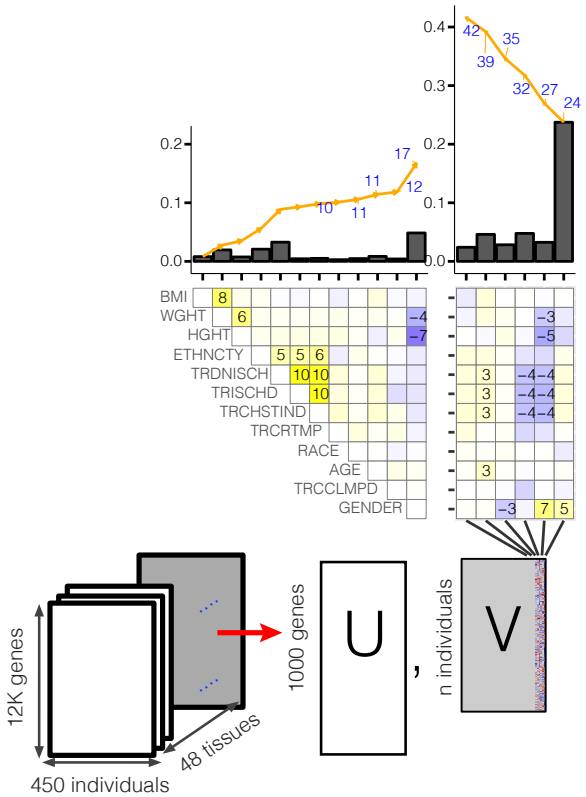
- Full effect size = SNP factor × Tissue factor
- SNP factor = shared regulatory motifs
- Tis factor = activity

fQTL

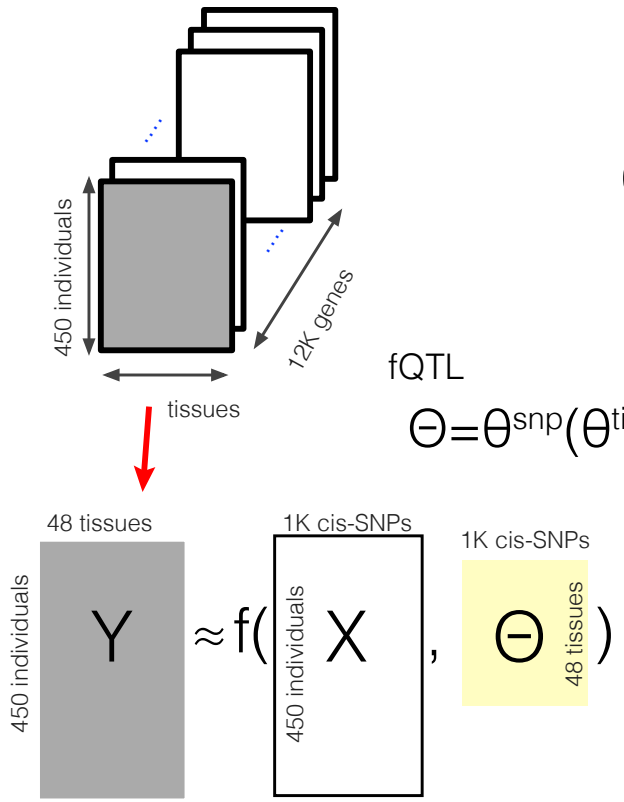
$$\Theta = \theta^{snp}(\theta^{tis})^T$$

Park & Sarkar *et al*. bioRxiv (2017)

Test gene-level association with AD GWAS z-scores in a tissue-specific or pan-tissue manner



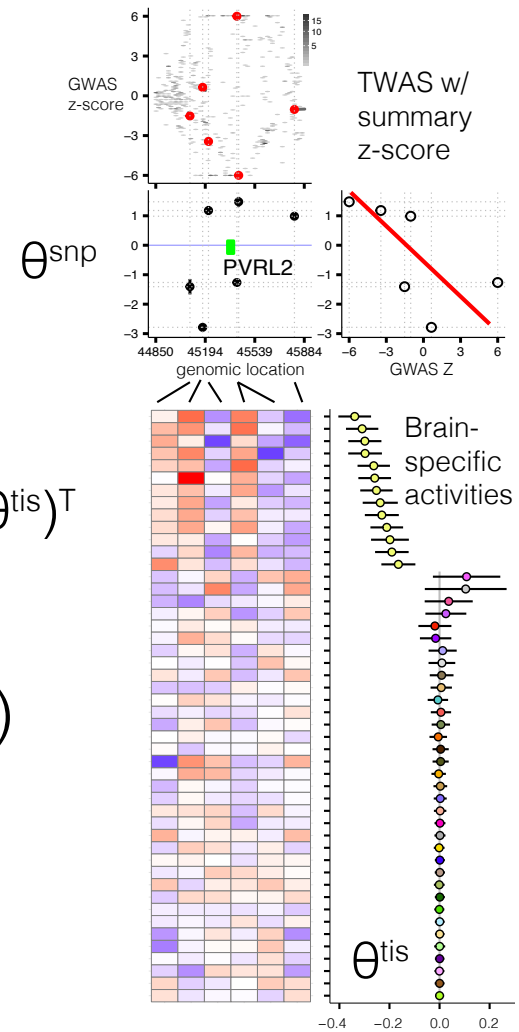
Factor analysis on most variable 1,000 genes in tissue and correct hidden & known confounders



fQTL

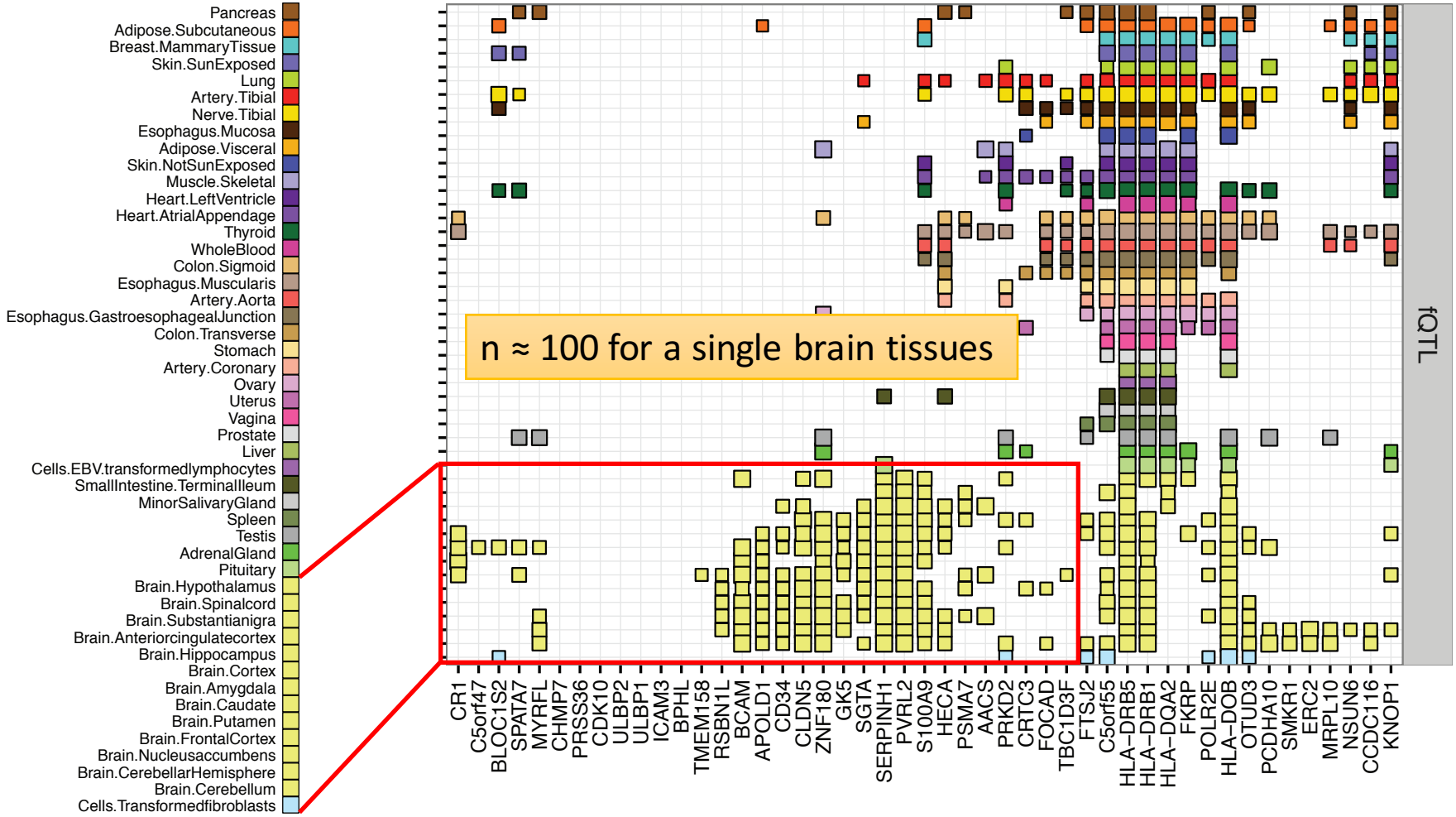
$$\Theta = \Theta^{snp}(\Theta^{tis})^T$$

Fit multi-tissue & polygenic regression with factored regression coefficients



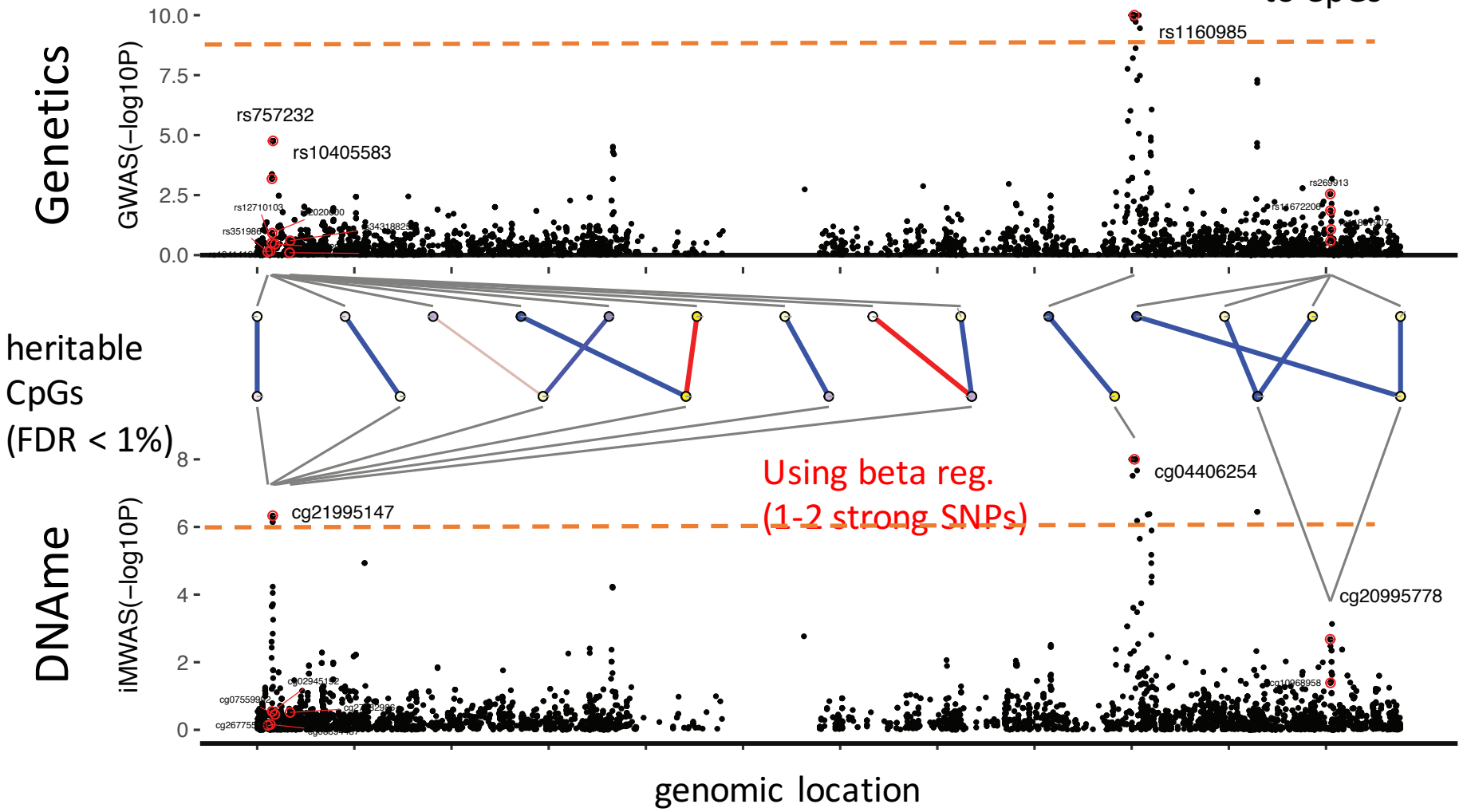
Park & Sarkar *et al*. bioRxiv (2017)

Brain-specific AD genes are only discovered by factored QTL models

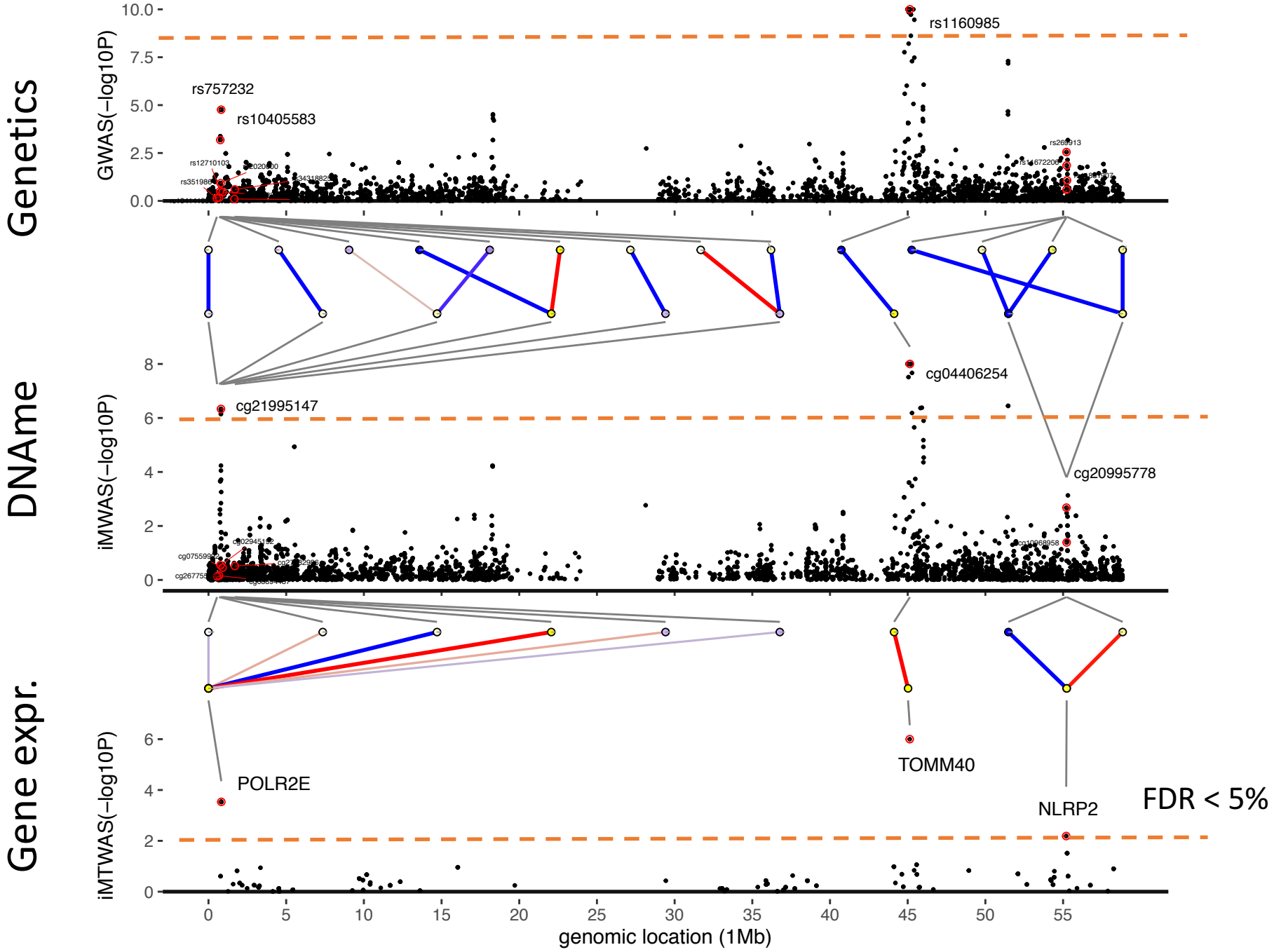


Alzheimer's disease sMWAS on Chr19

only show SNPs linked to CpGs

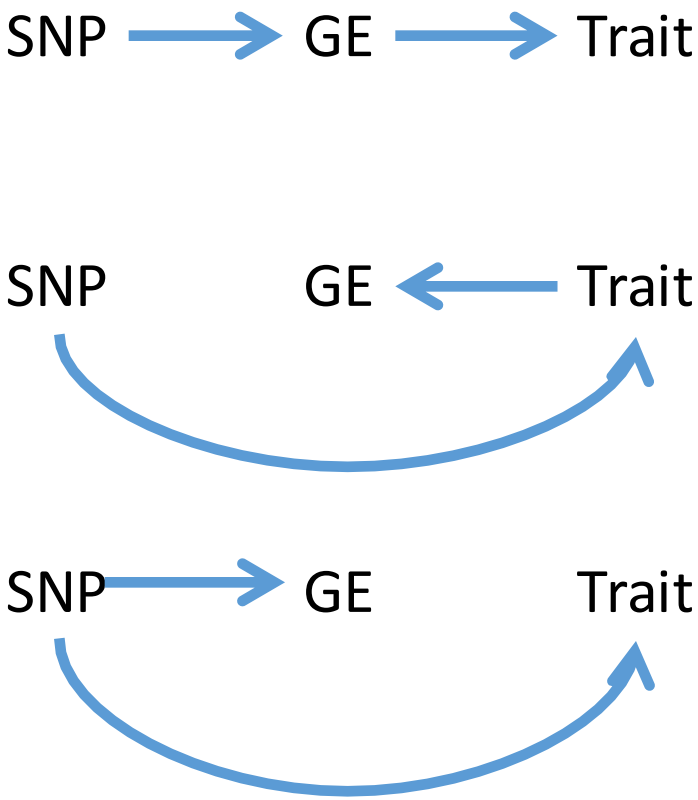


Park, *et al.*, *in preparation*


Genetics
DNAme
Gene expr.
GWAS(−log10P)
iMWAS(−log10P)
iMTWAS(−log10P)
rs1160985
rs757232
rs10405583
cg04406254
cg21995147
cg20995778
POLR2E
TOMM40
NLRP2
FDR < 5%
genomic location (1Mb)

Three types of association patterns iTWAS, sTWAS, co-localization can identify

(But TWAS cannot distinguish them from each other)



Mediation

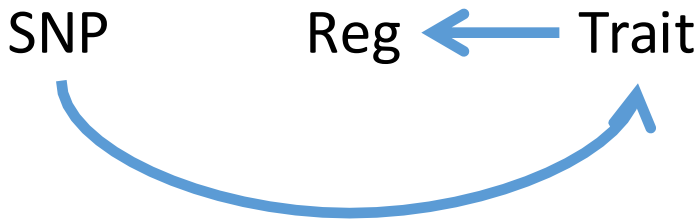
Reverse causation

Pleiotropy

Gusev *et al.* Nat. Gen. (2016)

Remove reverse causation in AD sTWAS, sMWAS

(What TWAS cannot distinguish from each other)



Reverse causation

DNAme ~ SNP (within ±1Mb) + Trait

Gene expression ~ SNP (within ±1Mb) + Trait

In the ROS-MAP cohort (with observed Aβ, NFτ, cognitive decline slope) pathological variables can be used as a surrogate of AD phenotype.

Park, *et al.*, *in preparation*

Distinguish mediation and pleiotropy by including direct effects in summary-based analysis

Pleiotropy model:

$$GE \approx X\,\theta_{qtl}$$

$$\phi \approx X\,\theta_{gwas}$$

vs

Mediation model (individual level data):

$$GE \approx X\,\theta_{qtl}$$

$$\phi \approx X\,\theta_{qtl}\,\theta_{mediation} + X\,\theta_{direct}$$

Ask: Can direct effect explain away mediation?

Without individual level data (apply summary-based regression; Zhu & Stephens, bioRxiv, 2016)

$$X^{\top}\phi \approx X^{\top}X\,(\theta_{qtl}\,\theta_{mediation} + \theta_{direct})$$

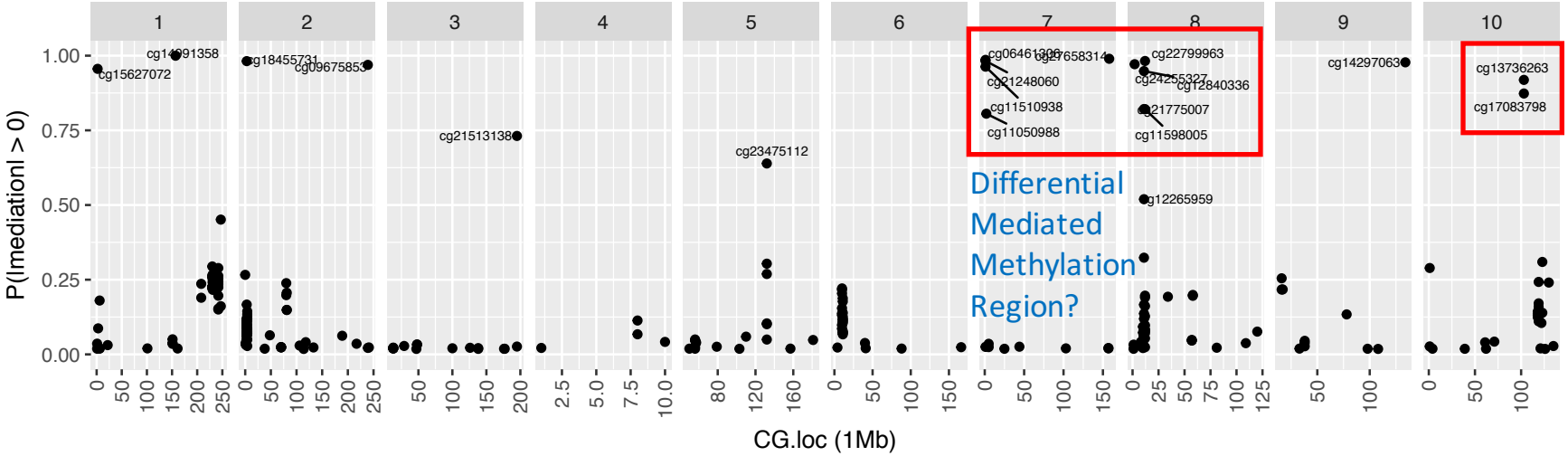
Or through fine-mapping model (Hormozdiari & Eskin)

$$z \sim N(\lambda LD(\theta_{qtl}\theta_{med} + \theta_{dir}), LD)$$

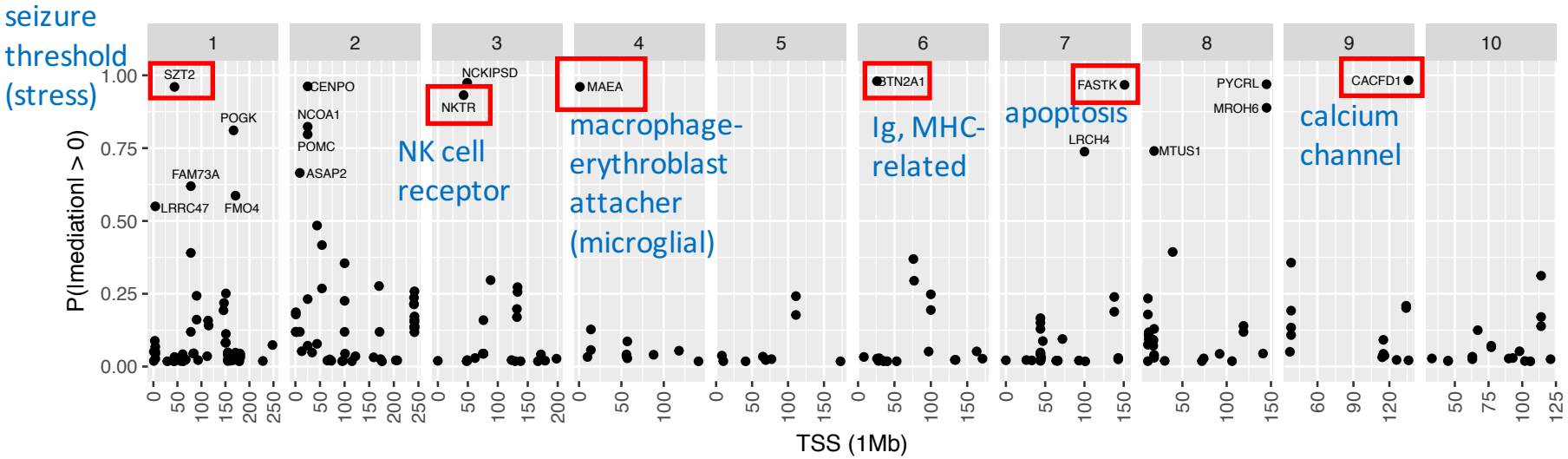
Estimate posterior distribution of spike-slab mediation effects using spectral transformation (Park, Sarkar, Kellis, *in preparation)*

AD GWAS causal mediation effects on Chr 1 - 10

MWAS mediation

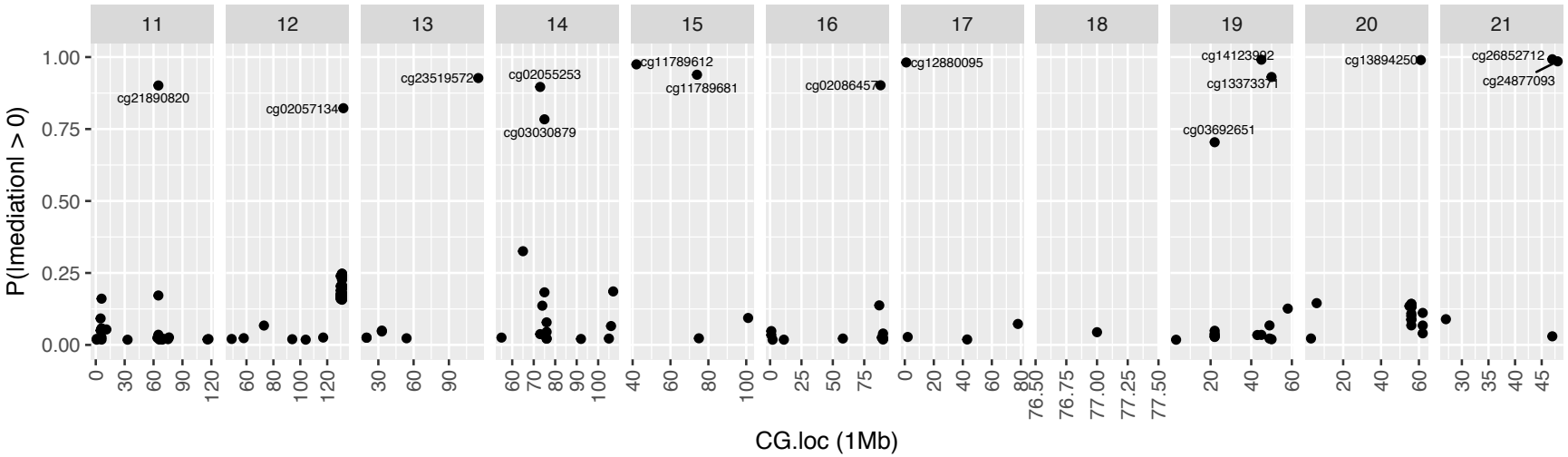


TWAS mediation



AD GWAS causal mediation effects on Chr 11 - 22

MWAS mediation



TWAS mediation