Statistical Methods for High Dimensional Biology

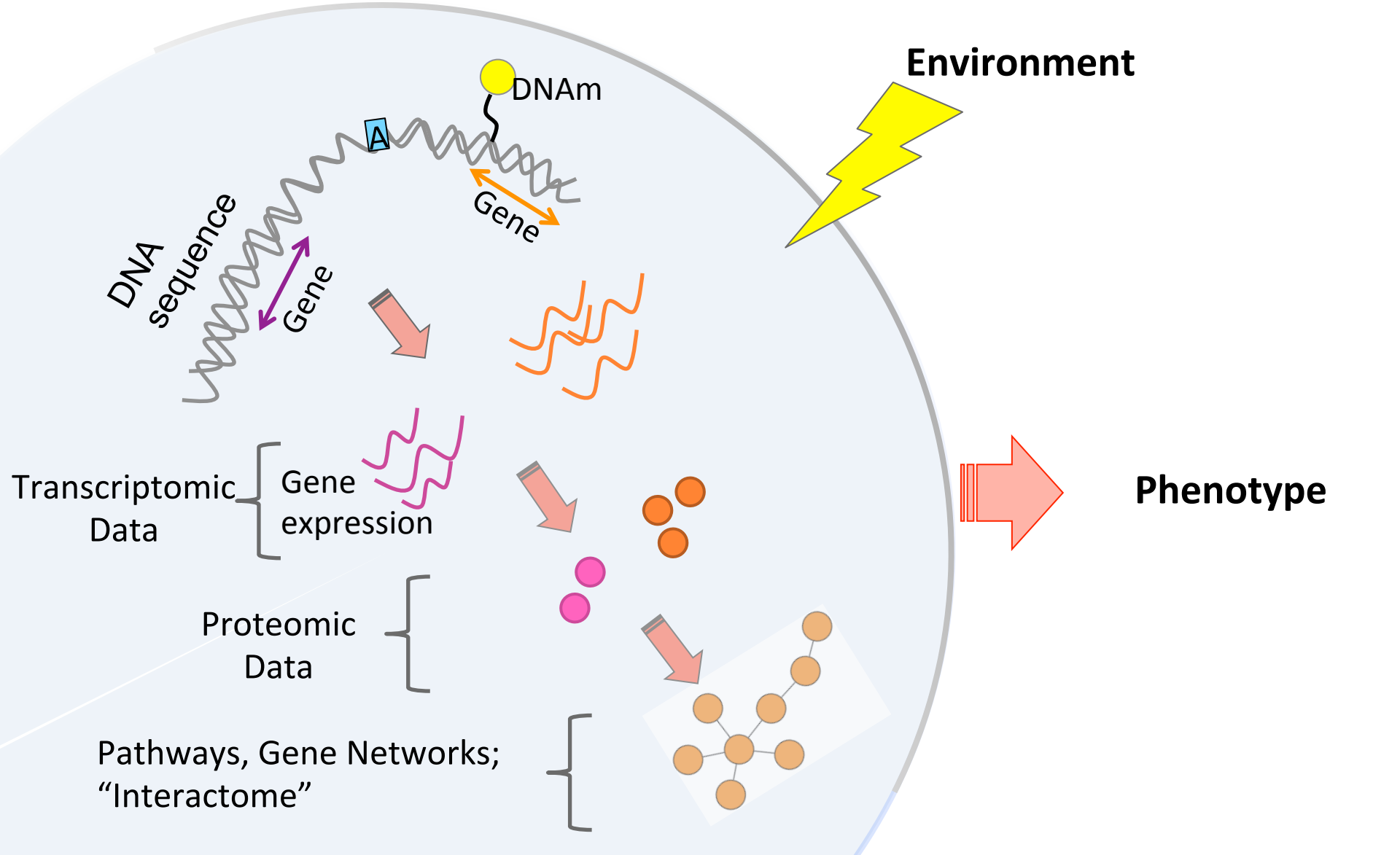
STAT/BIOF/GSAT 540

Lecture 22 – Multi-omics analysis

Sara Mostafavi

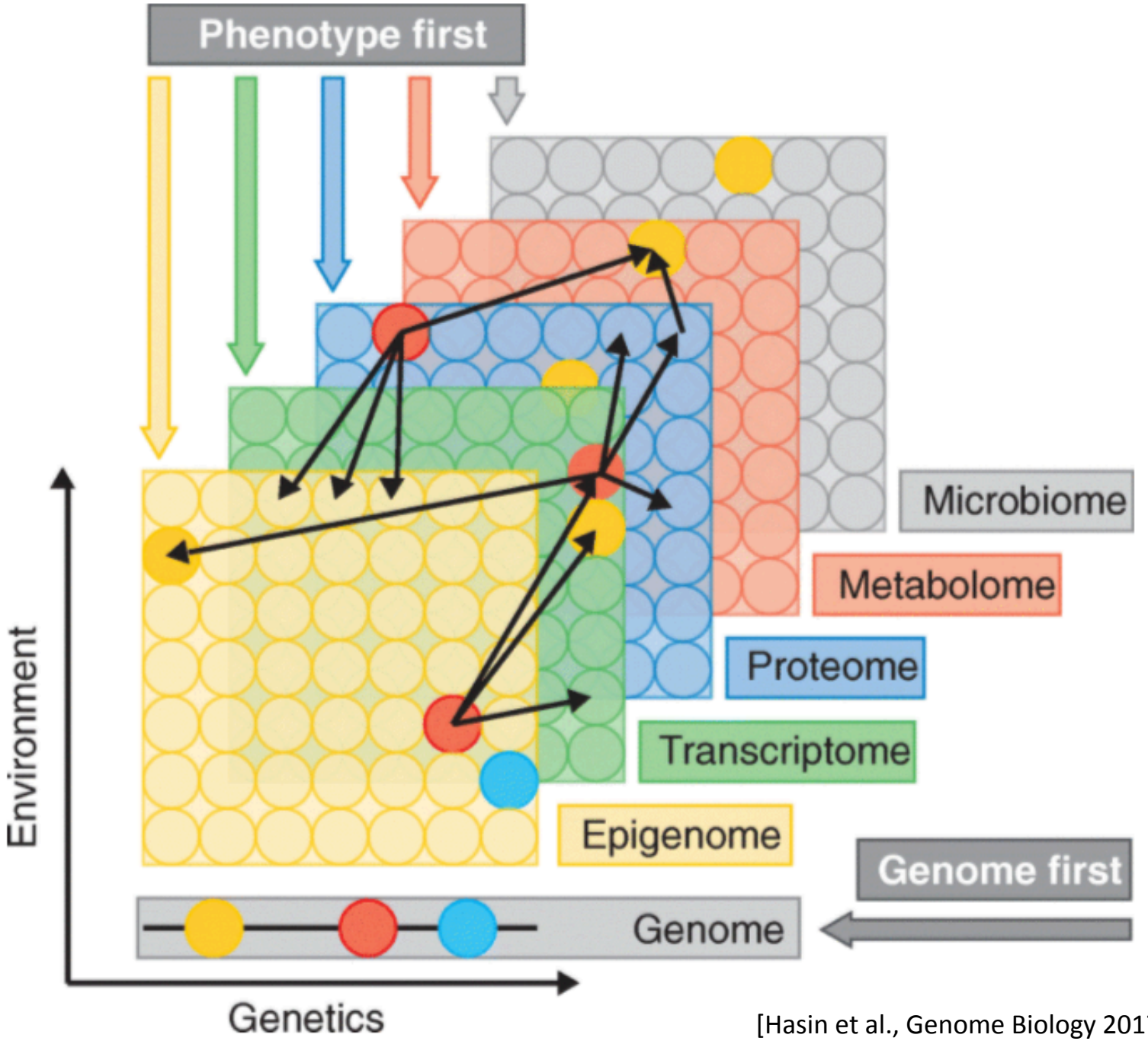
March 28,2018

Integration of multiple omics datasets--
Complementary omics data enable us to capture multiple views of complex biological systems



High-dimensional biological data

- Genomics (genetics): captures genetic variation across the genome
- Epigenomics: genome-wide characterization of DNA or histone modification
- Transcriptomics: examines genome-wide RNA levels (gene expression)
- Proteomics: examines genome-wide protein levels
- Metabolomics: examines intensity/activity of all measurable metabolites
- Microbiomics: examines all the microbiota presence in a sample from gut, skin, etc.



[Hasin et al., Genome Biology 2017]

Integration of multiple omics datasets

- Challenges:
  - We don't know how to map between measurements across different regulation layers:
    - We need a mapping between genes, probes, SNPs, etc
    - Solutions exists based on proximity: but essentially everything is collapsed to gene level
  - Different correlation structure at different regulation layers:
    - LD between SNPs, proximal correlation between probes.
    - Traditional statistical approaches are not well suited to the above.
  - Computational scalability:
    - Nuanced approaches are not scalable to genome-wide setting
- Overall goal: prediction vs understanding

Integration of multiple omics datasets

- Possible approaches:
  1. Post-hoc integration: e.g., meta-analysis using Fisher's and other approaches.

Integration of multiple omics datasets

- Possible approaches:
  1. Post-hoc integration: e.g., meta-analysis using Fisher's and other approaches.
  2. Integrate at feature level, ignore the correspondence between regulatory layers. ("data concatenation")
     - Ignores biology, results not very interpretable.

Integration of multiple omics datasets

- Possible approaches:
  1. Post-hoc integration: e.g., meta-analysis using Fisher's and other approaches.
  2. Integrate at feature level, ignore the correspondence between regulatory layers. ("data concatenation")
     - Ignores biology, results not very interpretable.
  3. "Genome first approach": Build careful "region based" models that include a biological "direction".
     - Mediation analysis, xQTL analysis, TWAS analysis
     - Currently not feasible to be applied genome-wide and off the shelf, active area of research.
     - Genome-wide flavor: Building a Bayesian Network as in PARADIGM (Vaske etal, Bioinformatics 2010).

Integration of multiple omics datasets

- Possible approaches:
  1. Post-hoc integration: e.g., meta-analysis using Fisher's and other approaches.
  2. Integrate at feature level, ignore the correspondence between regulatory layers. ("data concatenation")
     - Ignores biology, results not very interpretable.
  3. "Genome first approach": Build careful "region based" models that include a biological "direction".
     - Mediation analysis, xQTL analysis, TWAS analysis
     - Currently not feasible to be applied genome-wide and off the shelf, active area of research.
     - Genome-wide flavor: Building a Bayesian Network as in PARADIGM (Vaske etal, Bioinformatics 2010).
  4. "Phenome first approach": Kernel-based integration. Map each data to a "patient/sample" space first.
     - Doesn't naturally identify or reveal mechanisms.
     - Major confounding factor problems for confounded phenotypes.

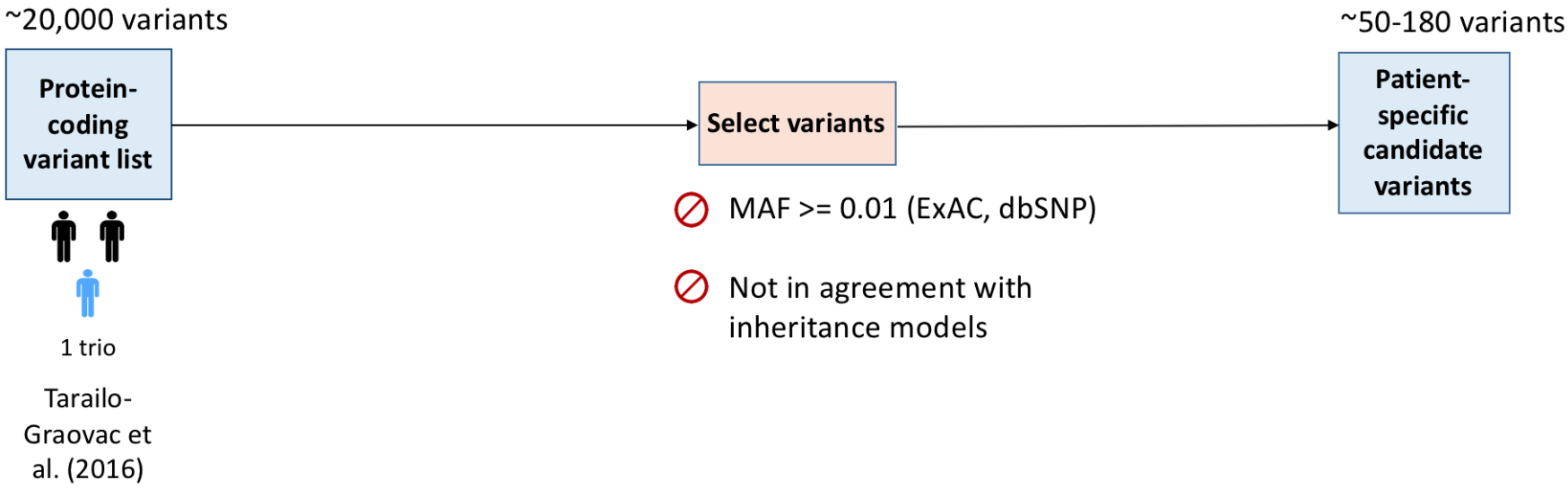
Goal of multiomics data generation depends on the complexity of the disease/trait etiology

- Simple Mendelian traits:
  - Eg., Rare inborn errors of metabolism
  - Validation or prioritization of potential causative variants
- Complex/multifactorial traits:
  - E.g., Alzheimer's disease, cancer
  - (1) causal inference ; (2) wider net ; (3) understanding patient heterogeneity ; (4) mechanistic insights

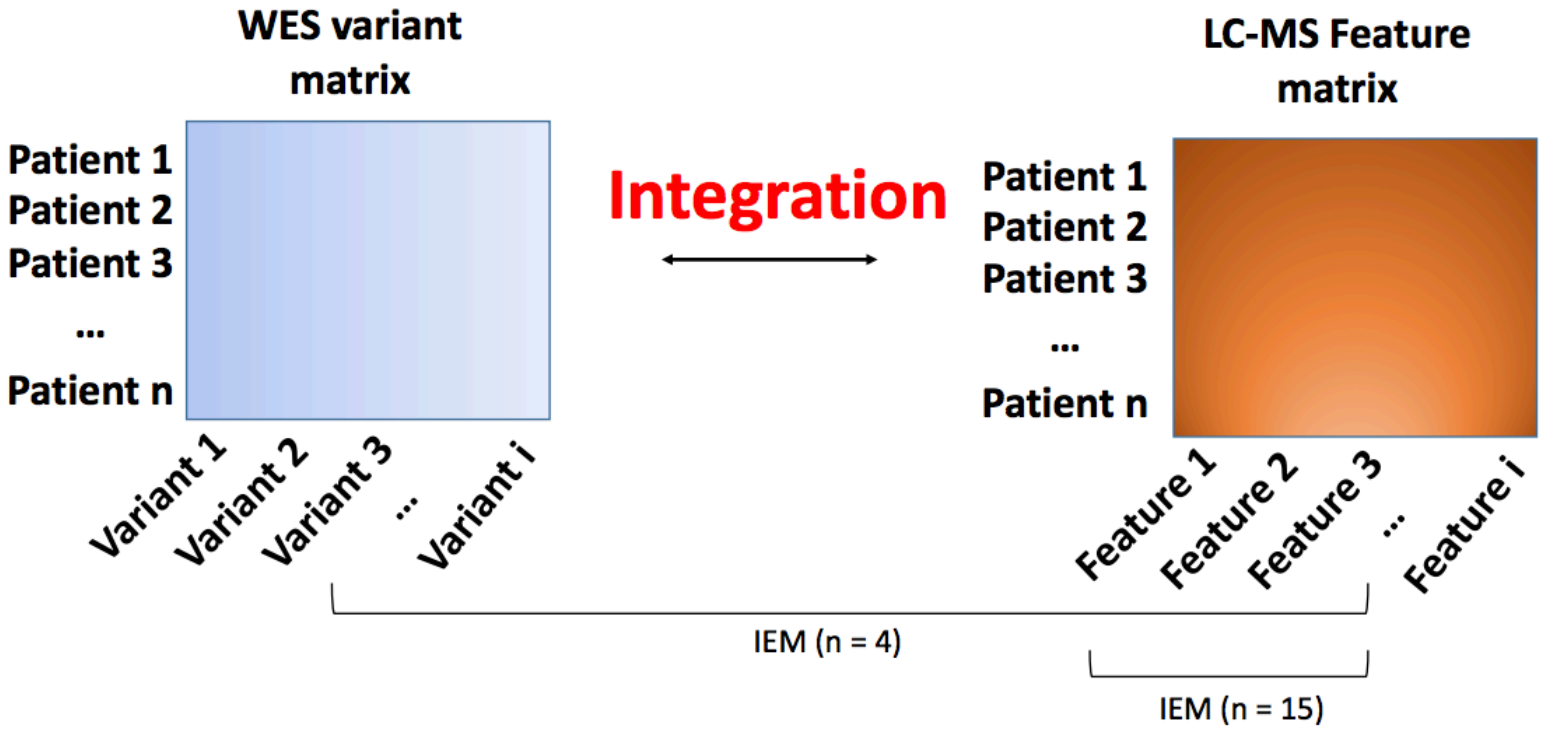
Simple trait example: IEMs

- TIDEx-multiomics project: improve diagnosis and treatment of inborn errors of metabolism (IEMs) that cause neurodevelopmental delay
  - Intellectual developmental disorders caused by defects in enzymes that help break down (metabolize) parts of food.
  - Characterized by movement, speech, and cognitive abnormalities
  - Early diagnosis is key so that treatment can be initiated
  - TIDEX clinic at BC Children's uses Whole Exome Sequencing to identify causative variants

Diagnosis of IEMs based on WES



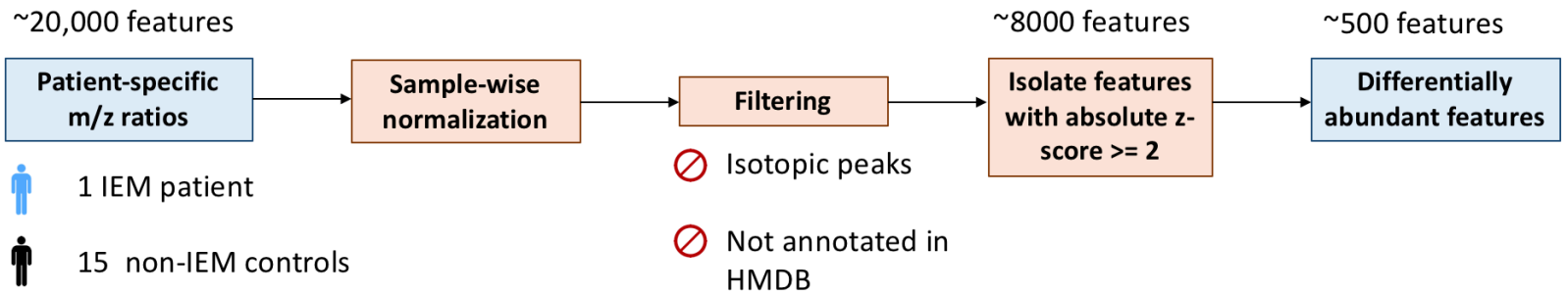
Overlap genetic variants with differential metabolites to prioritize the likely causal variants



Metabolomic data for prioritization of patient-specific causal variants

Metabolomic data:

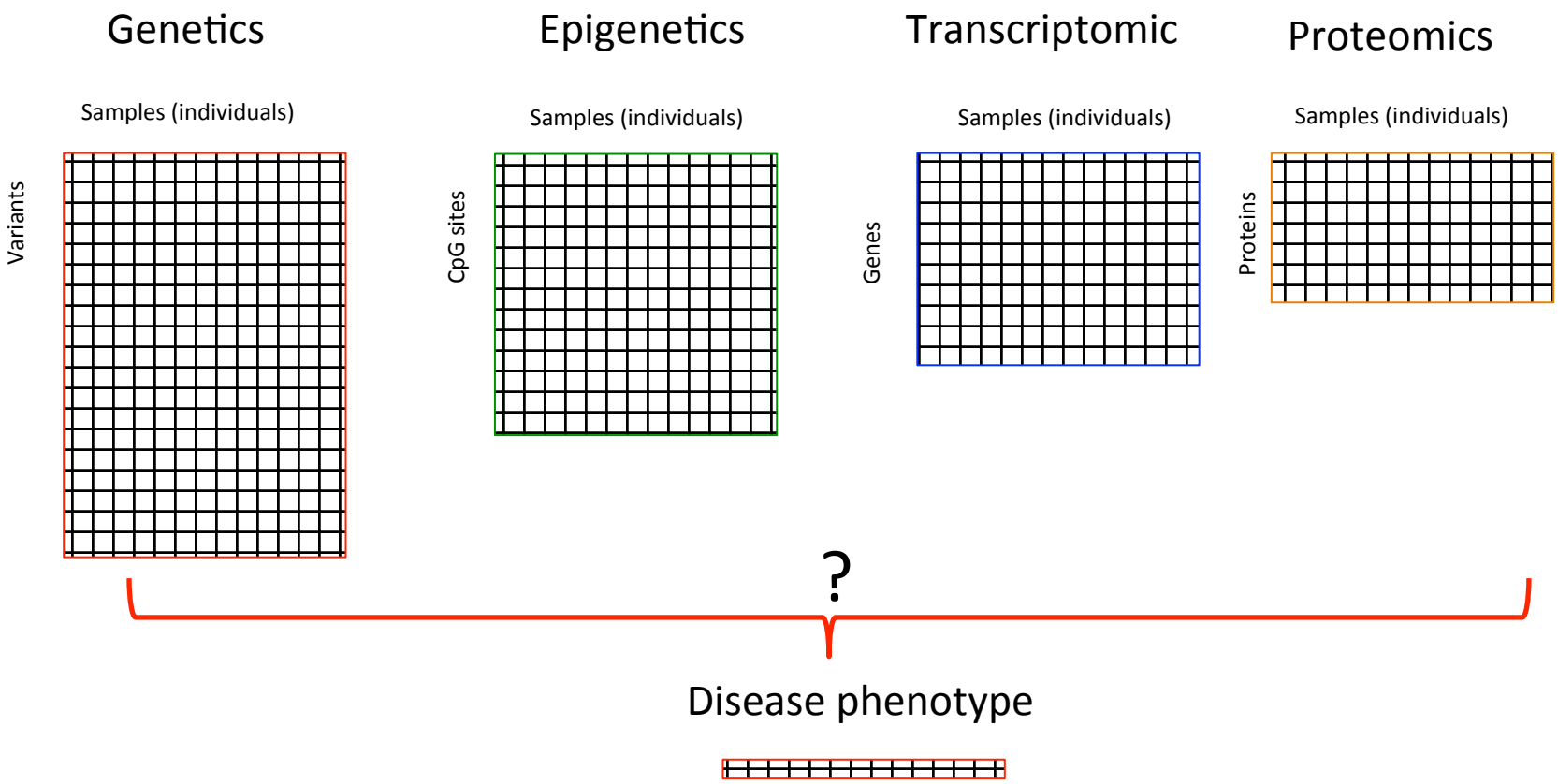
- Set of all small molecular metabolites (e.g., enzyme, signaling molecules, hormones) present within a tissue sample. Could be endogeneous or exogeneous.
- Provides an instantaneous snapshot of the physiology of the cell.



[Graham et al., JIMD 2018]

Multiomics approaches for complex traits

- Complex traits: small effect size signal
- Requires very large sample sizes

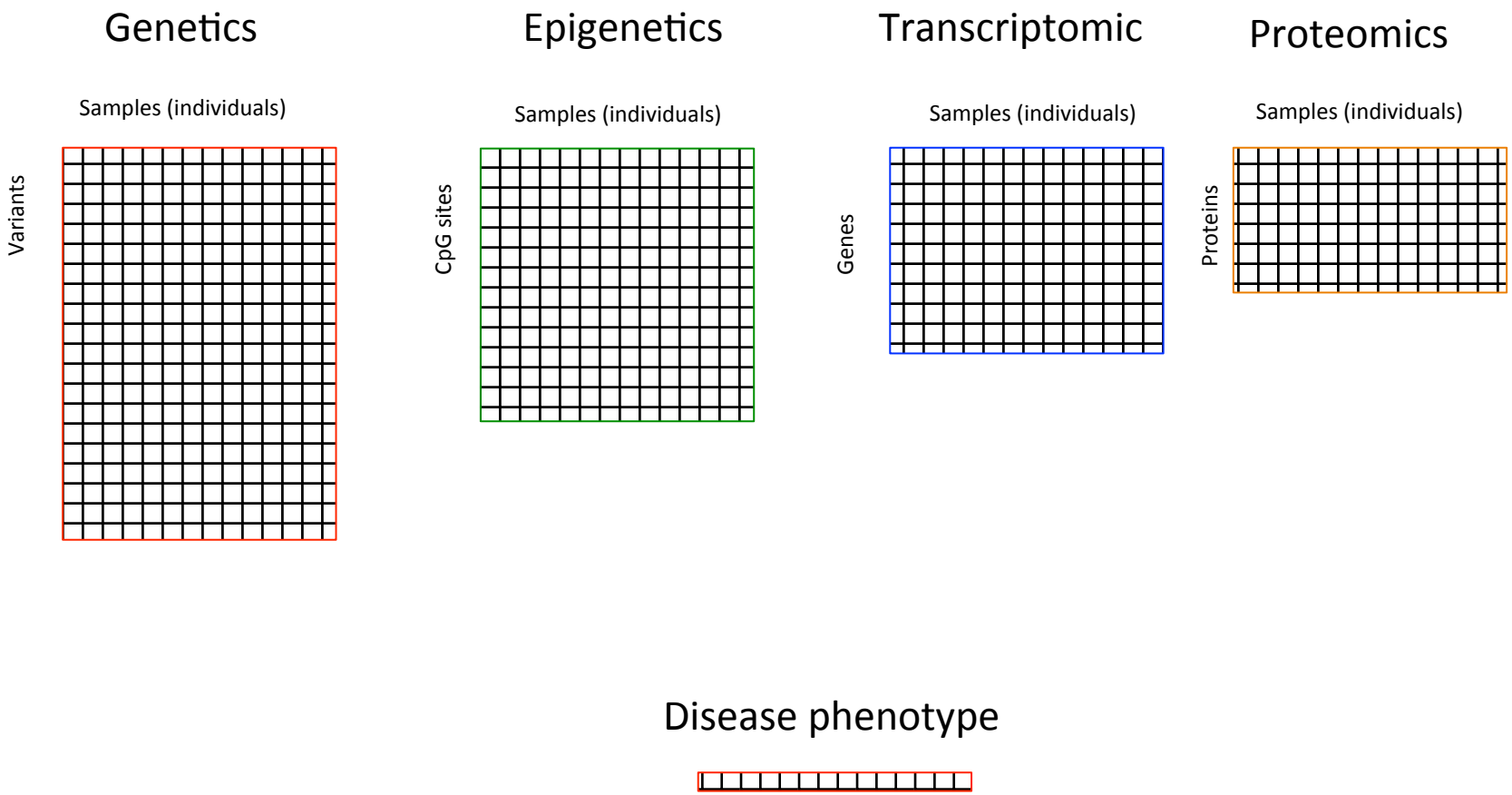


Multiomics approaches for complex traits

1) “genome first approach” or systems genetics: inference of causal gene pathway(s) from DNA sequence to phenotype

2) “phenome first approach”: prediction of disease outcome AND understanding patient phenotypic heterogenity

Multiomics approaches for complex traits

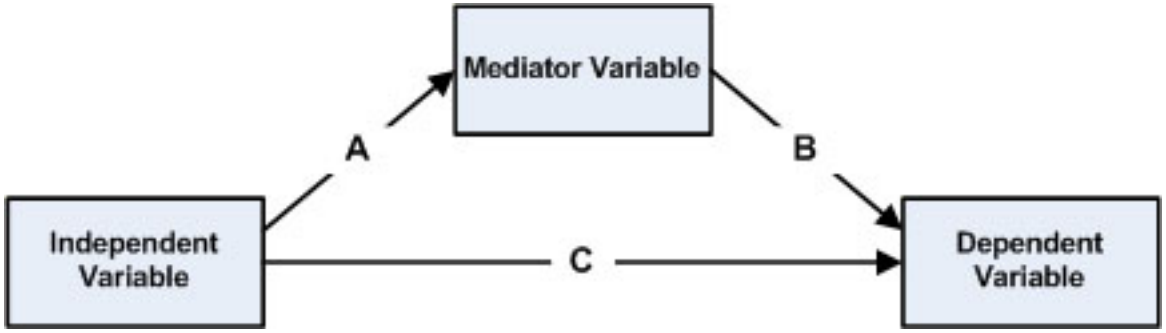


The genome first approach

- Deriving disease mechanism:
  - Perform a GWAS to find plausible variants/SNPs
  - Perform xQTL analysis to derive molecular/cellular factors associated with SNPs
  - Perform mediation analysis to establish the molecular mechanisms for the causal/upstream variants

Mediation analysis

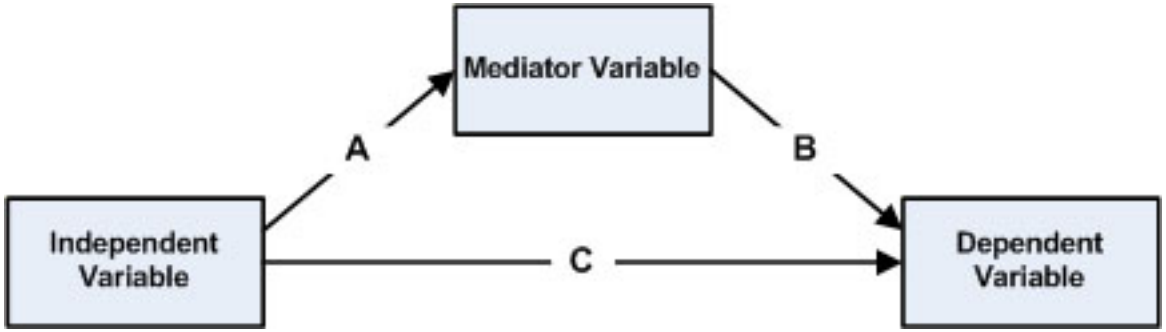
- Statistical definition: a model that seeks to explain the relationship between an independent variable and a dependent variable via the inclusion of a third intermediate variable.



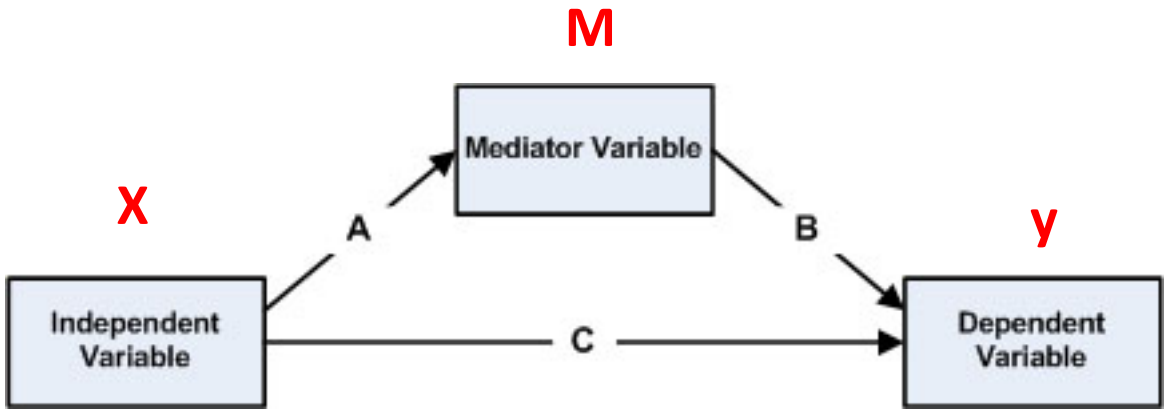
Mediation analysis

- Statistical definition: a model that seeks to explain the relationship between an independent variable and a dependent variable via the inclusion of a third intermediate variable.

- A mediation model proposes that the independent variable influences the (non-observable) mediator variable, which in turn influences the dependent variable

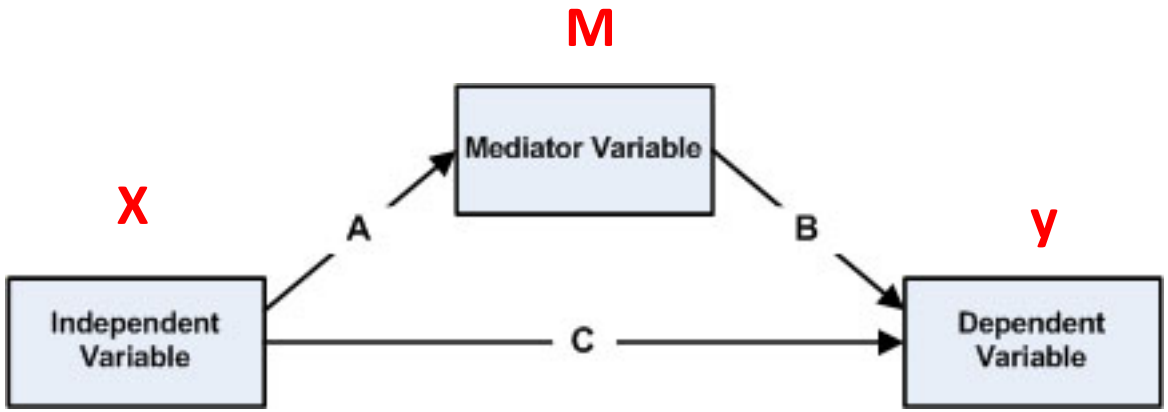


Mediation analysis: three regression steps



1) X is a predictor of Y:

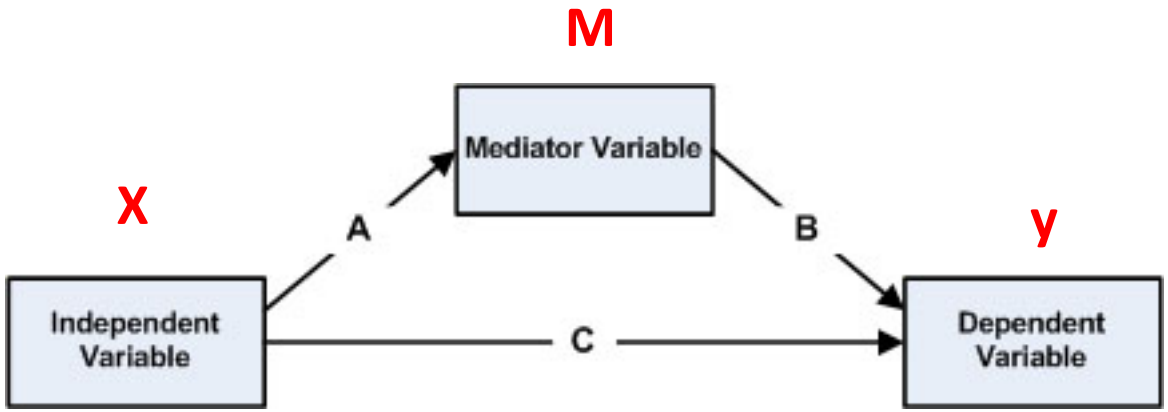
Mediation analysis: three regression steps



1) X is a predictor of Y:

$$y = \alpha_1 + \beta_1 x + \varepsilon$$

Mediation analysis: three regression steps

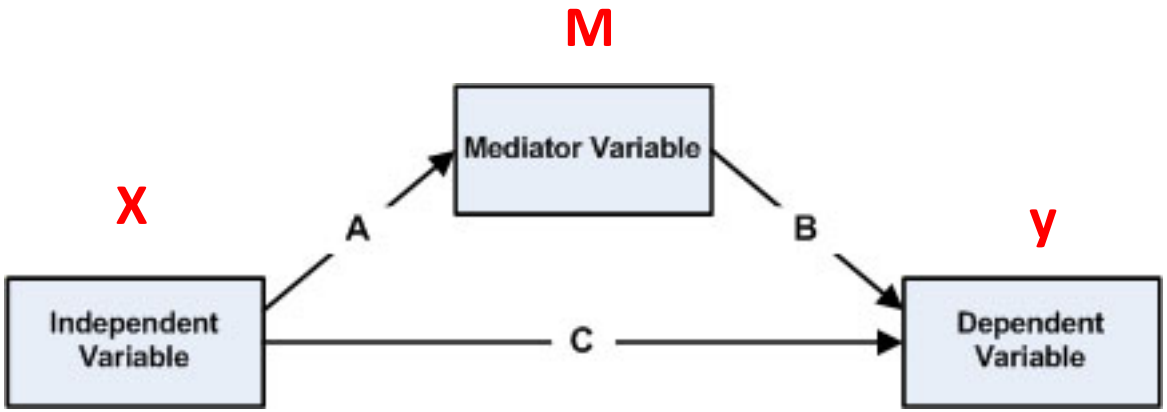


1) X is a predictor of Y:

$$y = \alpha_1 + \beta_1 x + \varepsilon$$

2) X is a predictor of M:

Mediation analysis: three regression steps



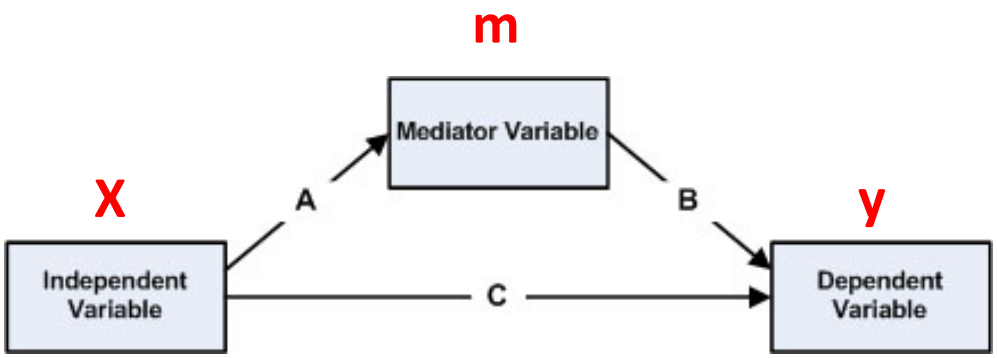
1) X is a predictor of Y:

$$y = \alpha_1 + \beta_1 x + \varepsilon$$

2) X is a predictor of M:

$$M = \alpha_2 + \beta_2 x + \varepsilon$$

Mediation analysis: three regression steps

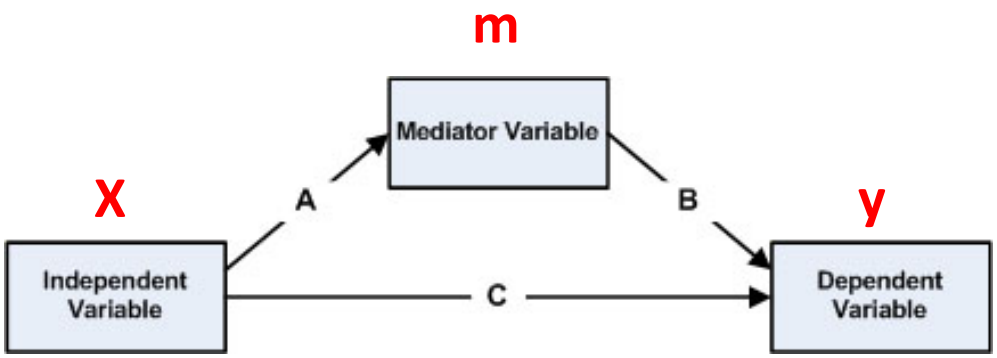


1) X is a predictor of Y: $y = \alpha_1 + \beta_1 x + \varepsilon$

2) X is a predictor of M: $m = \alpha_2 + \beta_2 x + \varepsilon$

3) M is a predictor of Y, conditioned on Y: $Y = \alpha_3 + \beta_3 x + \beta_4 m + \varepsilon$

Mediation analysis: three regression steps



Full vs partial mediation: $\beta_3 = 0$ vs reduced reduction of VE by *X*

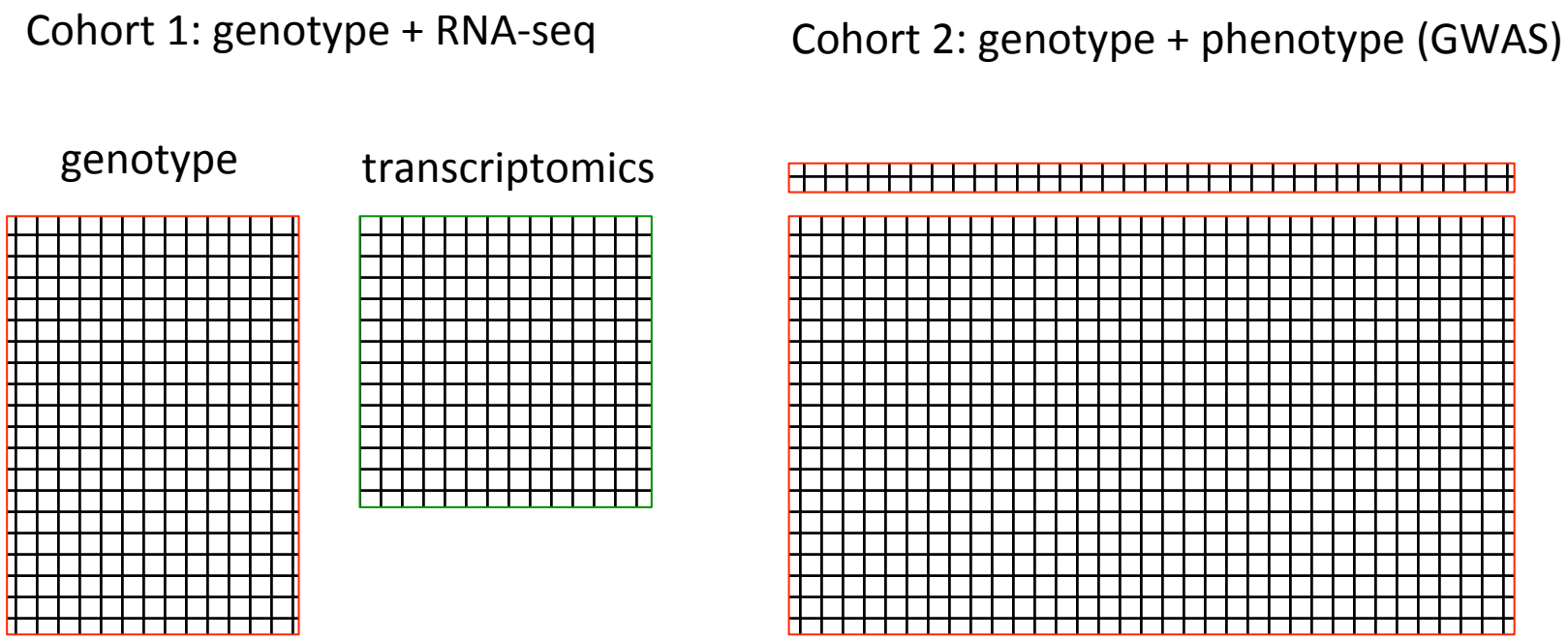
1) X is a predictor of Y: $y = \alpha_1 + \beta_1 x + \varepsilon$

2) X is a predictor of M: $m = \alpha_2 + \beta_2 x + \varepsilon$

3) M is a predictor of Y, conditioned on Y: $Y = \alpha_3 + \beta_3 x + \beta_4 m + \varepsilon$

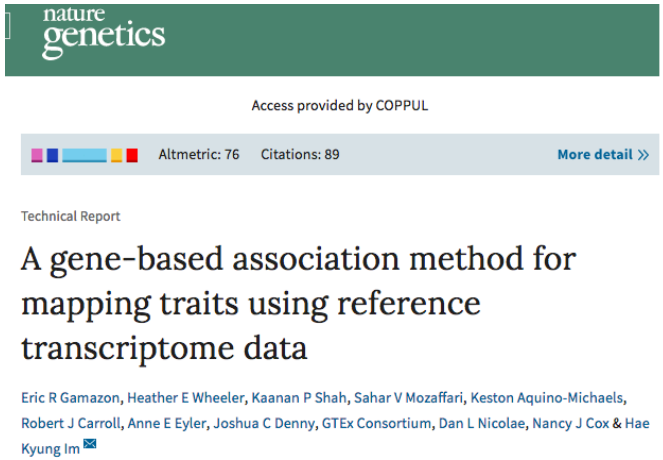
Genome first approach with disparate samples

- E.g., Combining huge GWAS studies with smaller transcriptomics + genetic studies.



Genome first approach with disparate samples

- TWAS: e.g., PrediXcan



Three step approach:

1. Build a “gene expression predictor” from genotype: call this M1
2. Use M1 model to predict gene expression levels in GWAS
3. Use predicted gene expression levels to perform association analysis in the GWAS sample.

Genome first approach with disparate samples

- TWAS: PrediXcan, MetaXcan, SMR,

Three step approach:

1. Build a "gene expression predictor" from genotype: call this M1

Model the expression level of gene as a function of SNP alleles

$$y = \alpha_1 + X\mathbf{b} + \varepsilon$$

SNP data for nearby SNPs

Gene expression level for gene 1

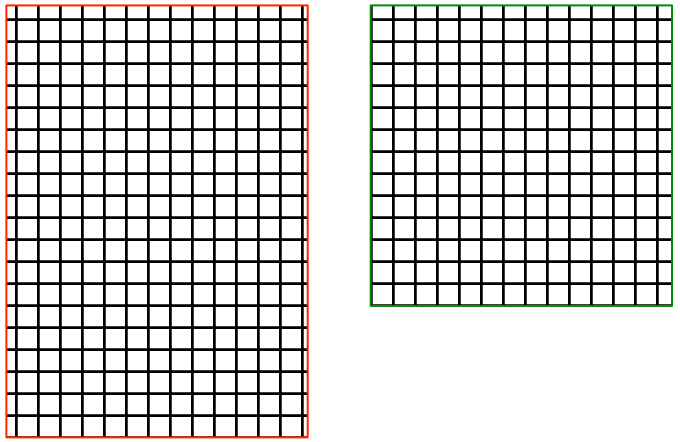
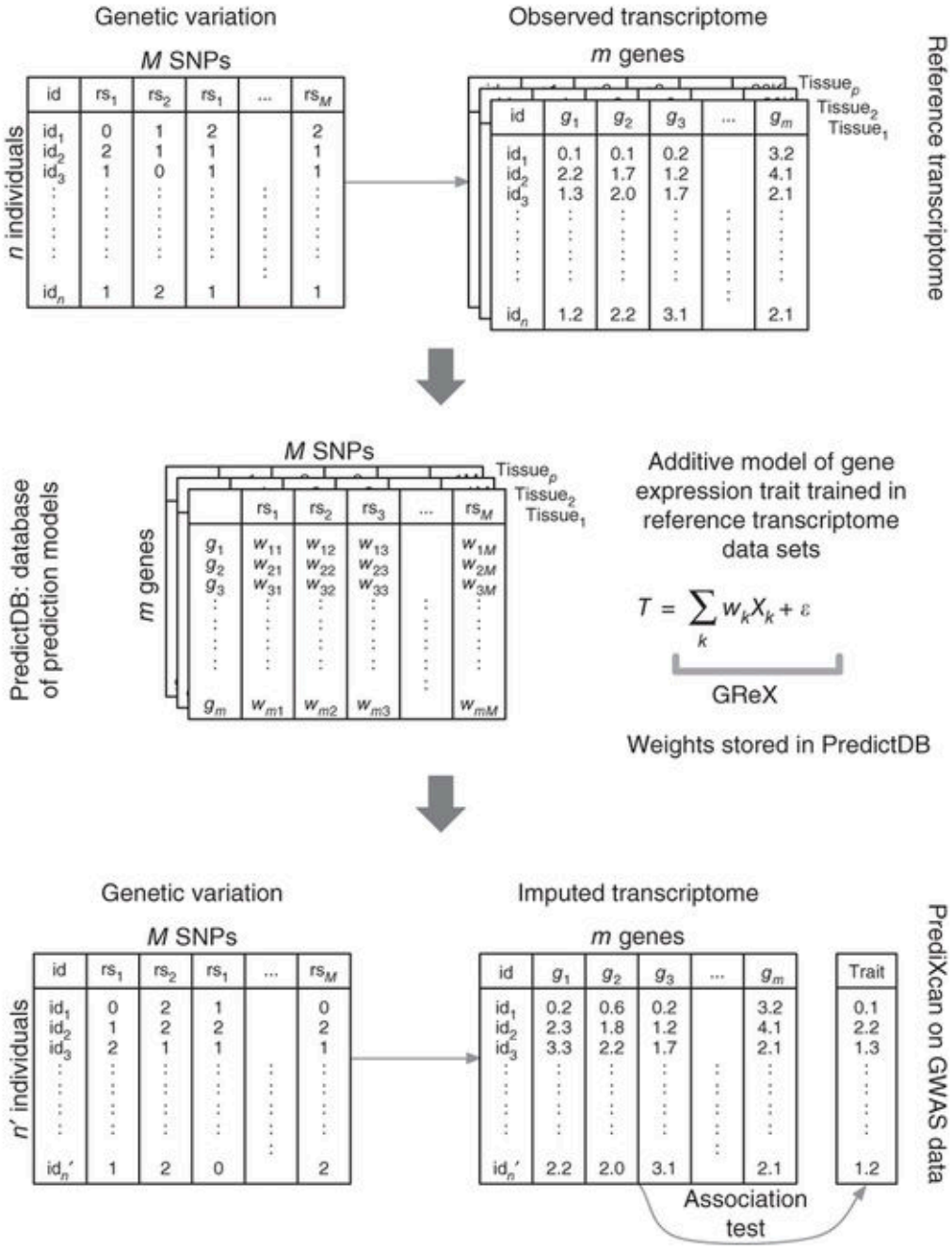


Figure 2: PrediXcan framework.



Multiomics approaches for complex traits

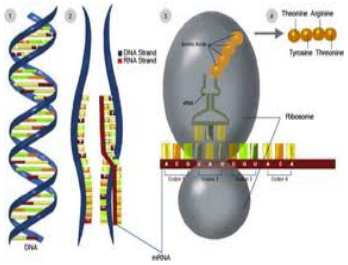
1) “genome first approach” or systems genetics: inference of causal gene pathway(s) from DNA sequence to phenotype

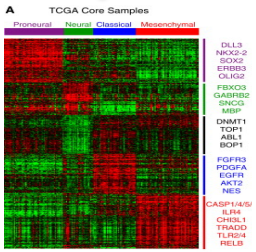
2) “phenome first approach”:

- predicting disease phenotype
- understanding patient phenotypic heterogeneity

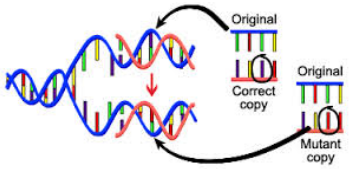
“Phenome first”: Independent analysis of each data for primary goal of prediction

mRNA



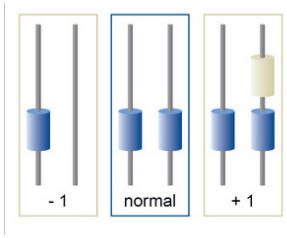


mutations



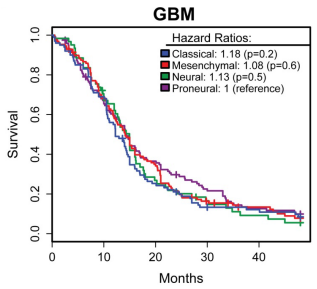
more genes

CNV



more genes

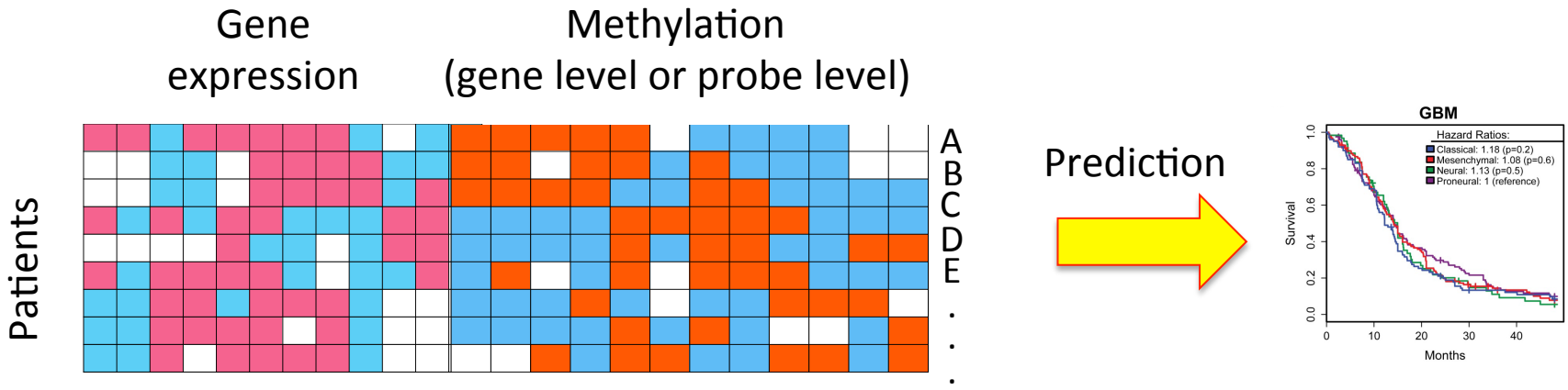
clinical



p-value = {0.2,0.6,0.5}

(Verhaak et al, Cancer Cell, 2010)

Concatenation



Phenome first: understanding patient phenotypic heterogeneity

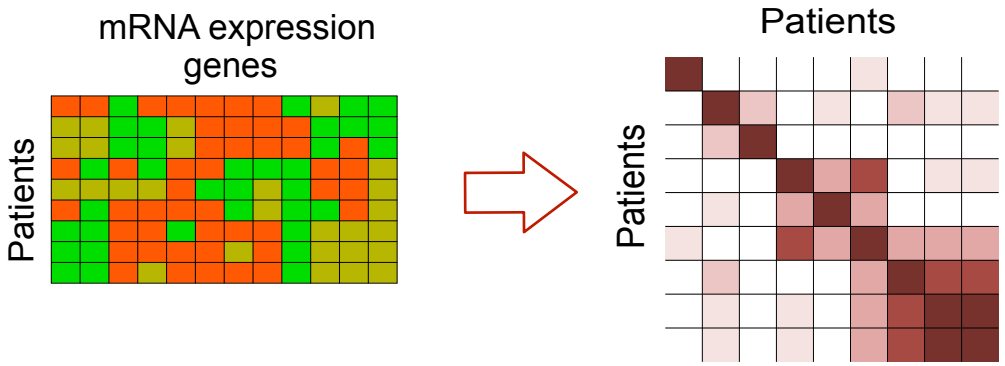
1. Concatenate and cluster (commonly used in TCGA analysis)
2. iCluster (Shen et al, 2009)
3. SNF (Wang et al, 2014)

Similarity Network Fusion (Wang et al, 2014)

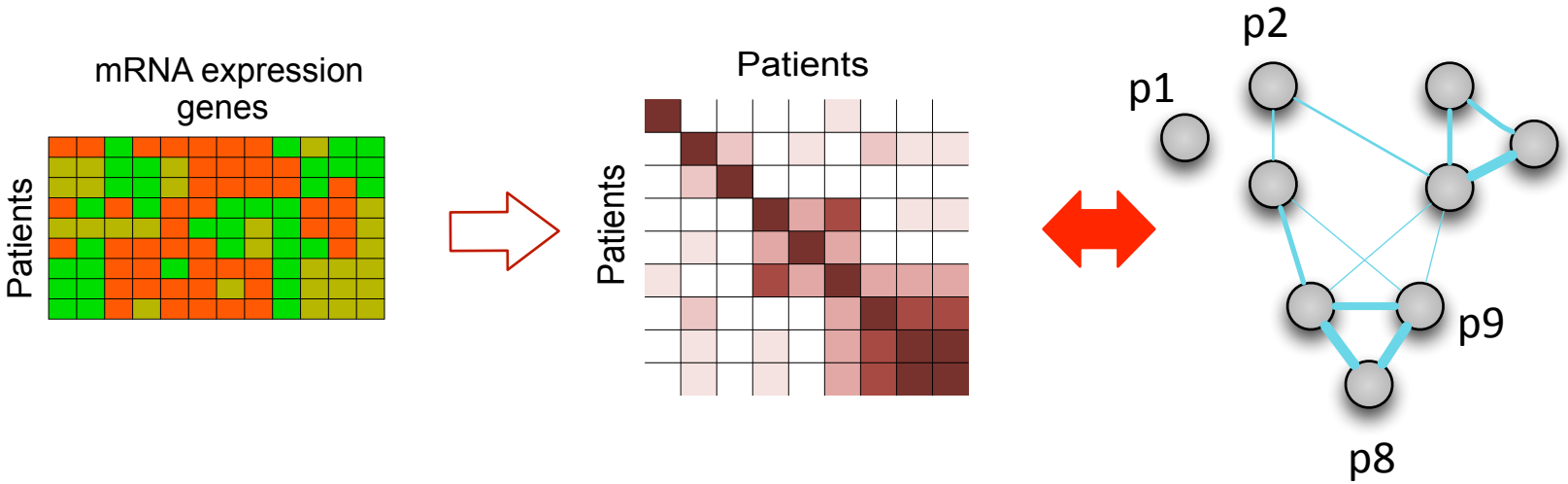
- Integrate data in the patient space

1. Construct patient similarity matrix
2. Fuse multiple matrices

1. Construct similarity networks

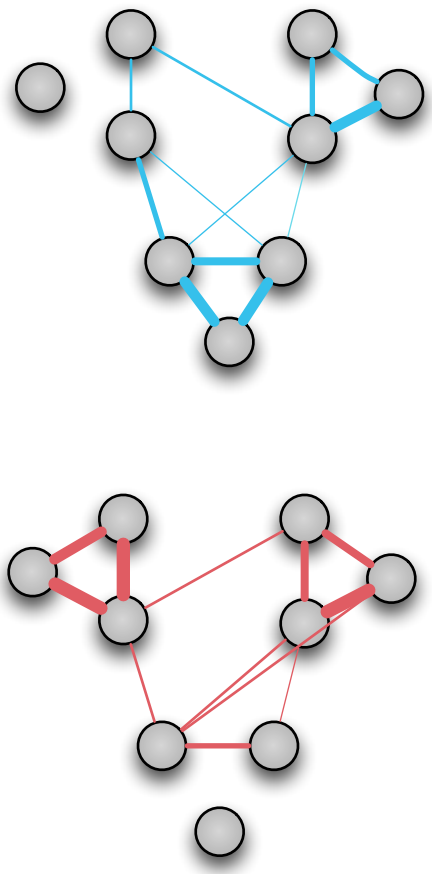


1. Construct similarity networks

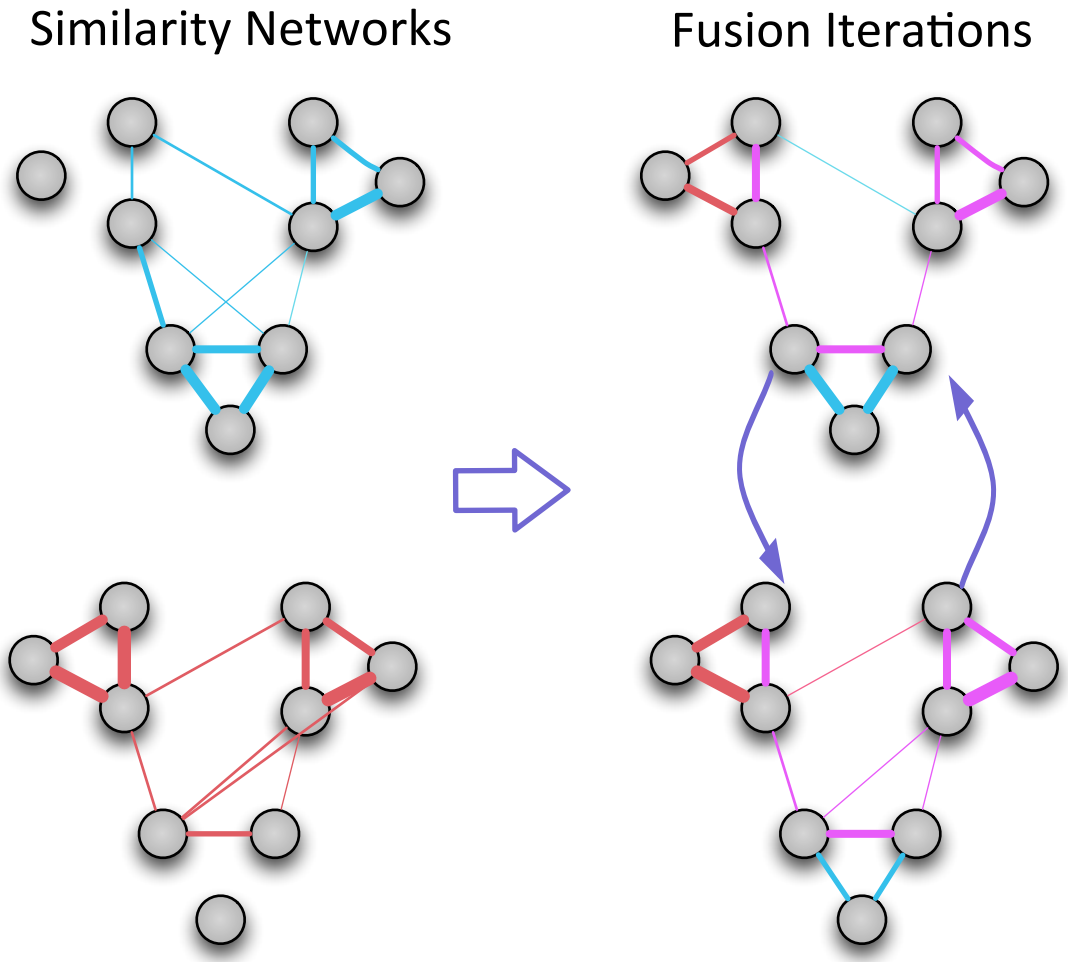


2. Combine networks

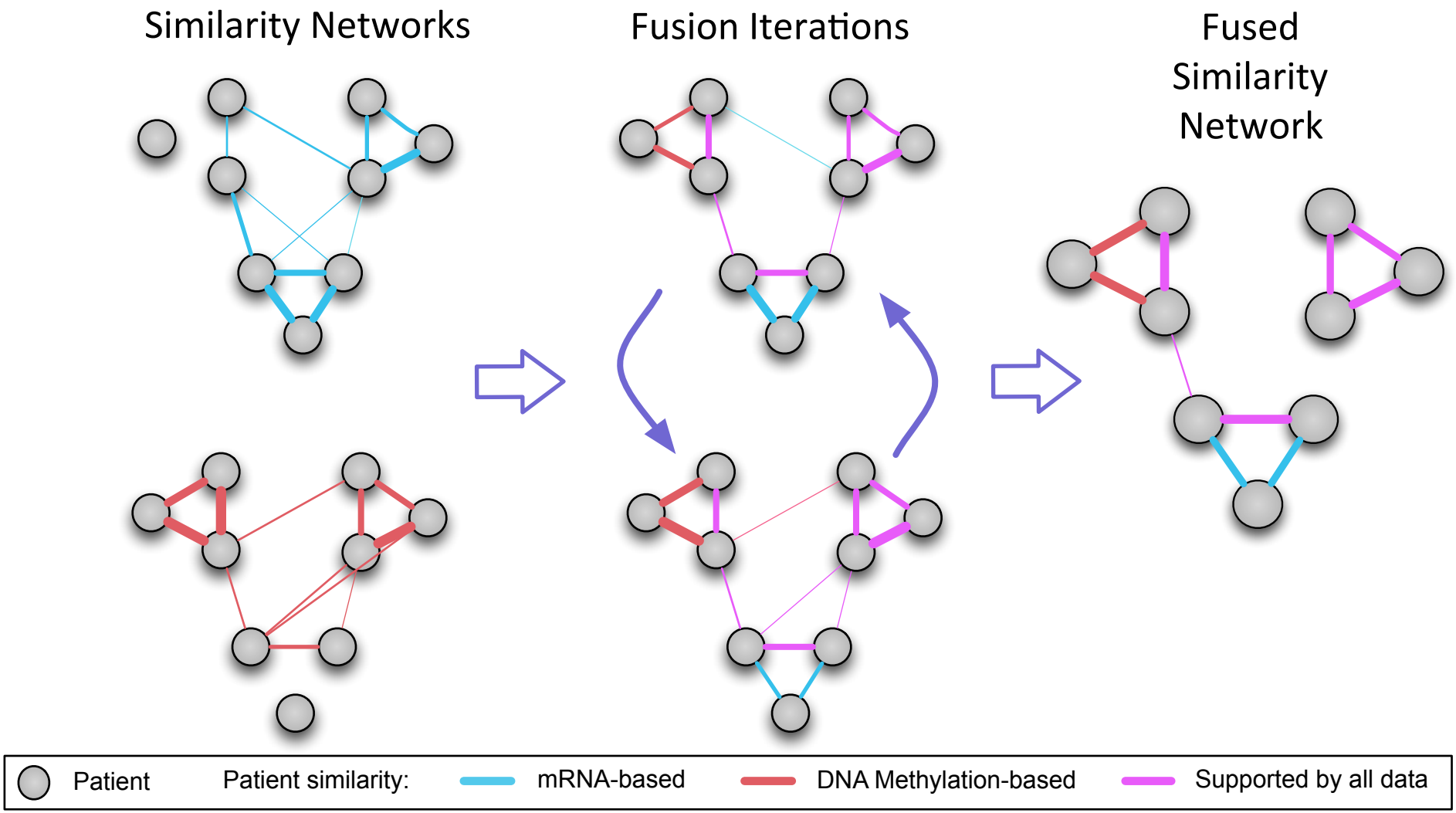
Similarity Networks



2. Combine networks



2. Combine networks



Summary

- Data integration has a long way to go:
  - Very little done to figure out complex mechanisms that link complex cellular systems to phenotypes.
- Inferring causality from observational data still under intense debate.
  - Best we can do is to eliminate potentially non-causal associations, and hope that we enrich for those that are causal.
- We need to distinguish between two goals in modeling data:
  - Prediction vs learning mechanisms