Introduction to Data Integration

Lauren Vanderlinden

BIOS 6660

Spring 2019

Overview From Last Time

- Different databases
- How to retrieve data in R
- Gene signatures

NCBI

GEO Gene Expression Omnibus

EMBL-EBI

TCGA

GTEx Portal

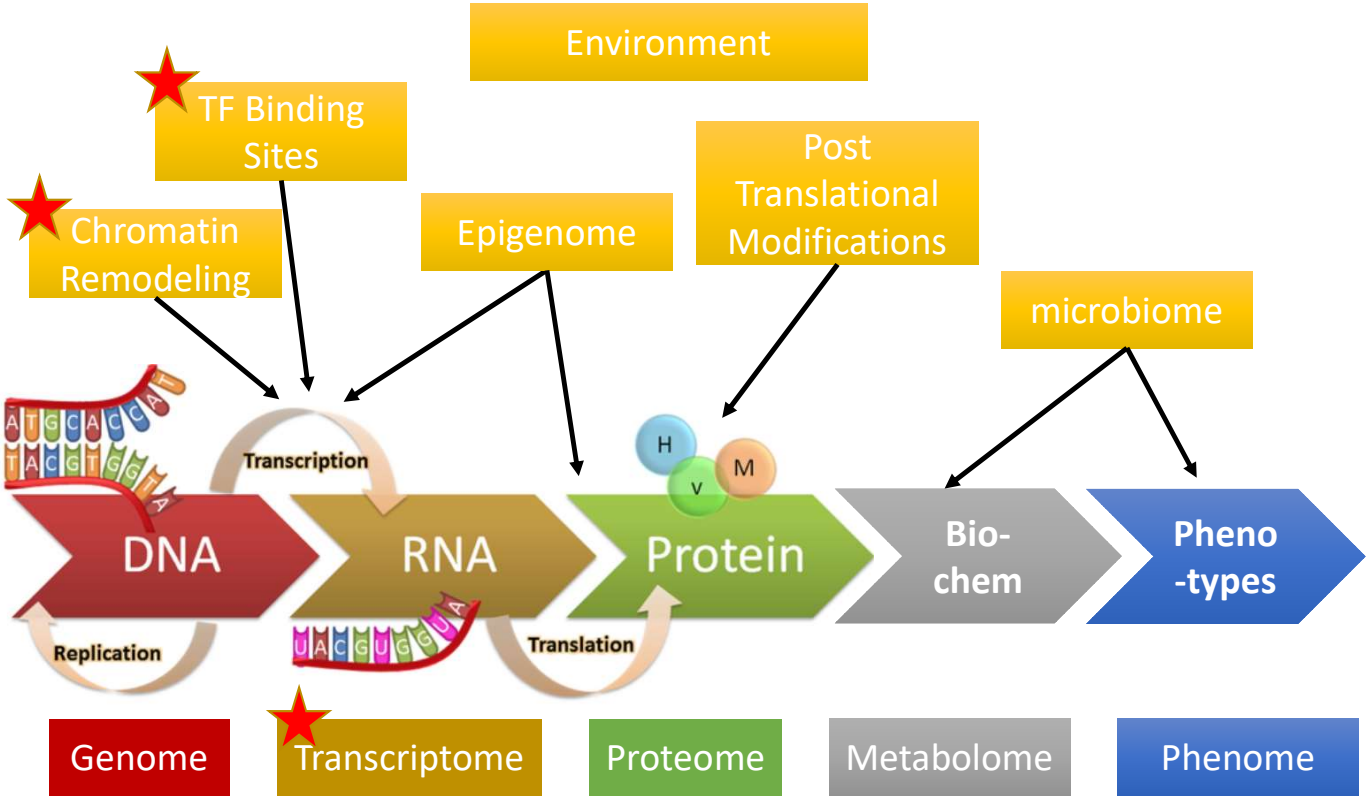
PhenoGen Informatics
The site for quantitative genetics of the transcriptome.

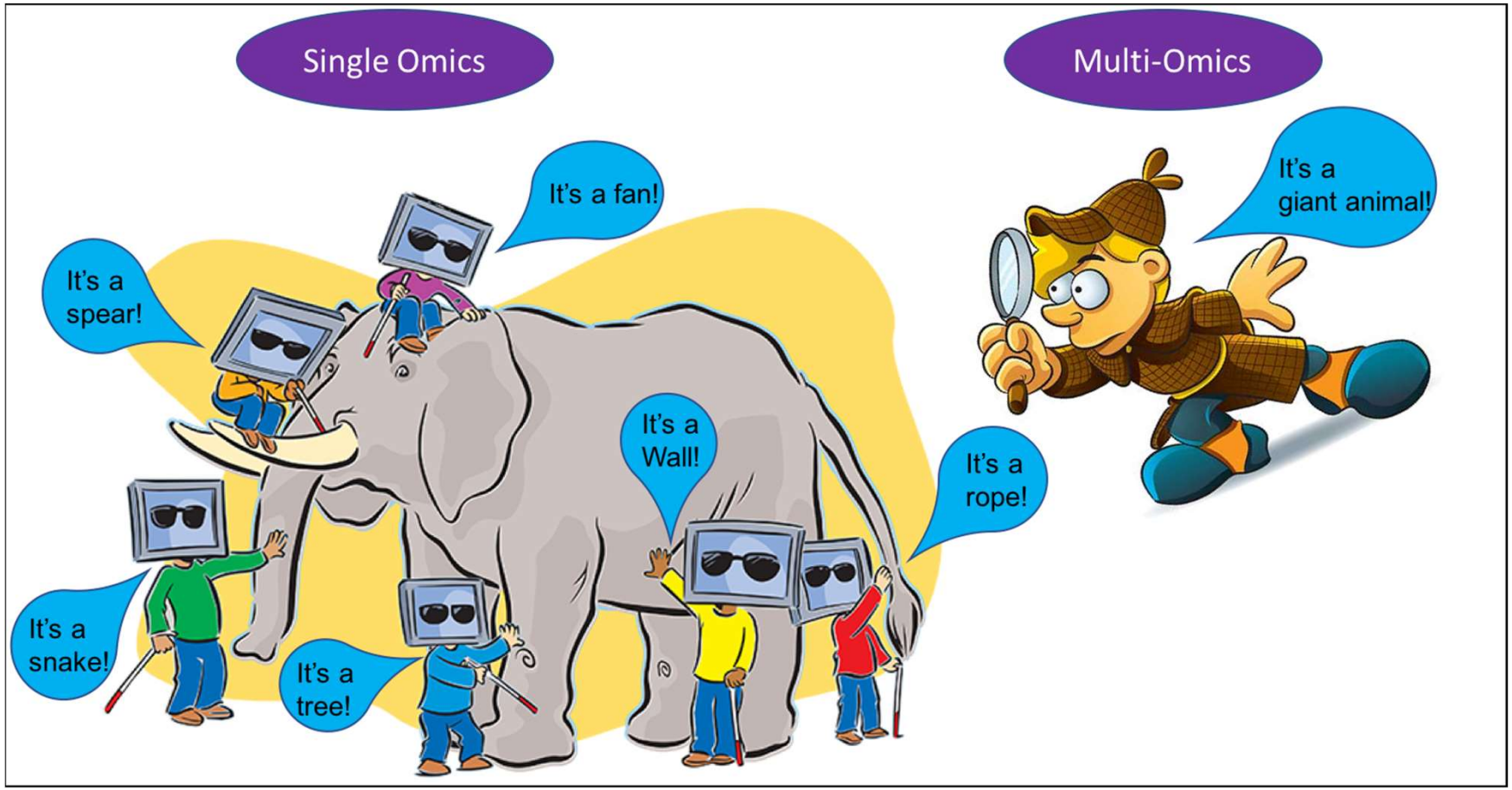
GeneNetwork
University of Tennessee: www.genenetwork.org

RGD

Multiple 'Omics Datasets

- Want to get a bigger picture on what is happening
- Multiple 'omics datasets

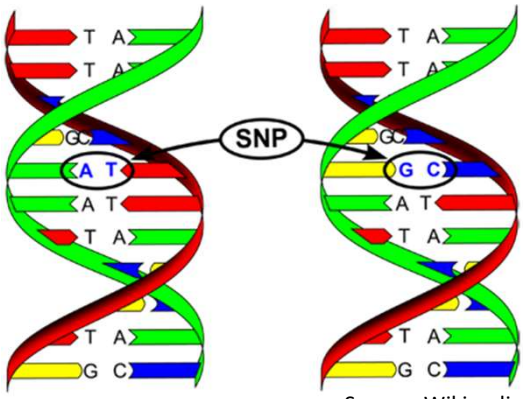




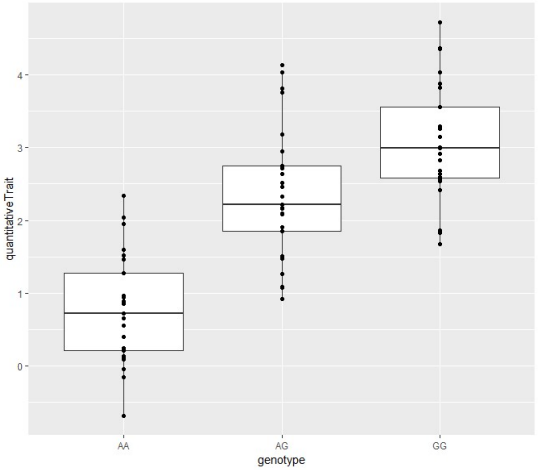
Source: Melgen

Integration with DNA

- Most common form of data integration
- DNA marker set of Single Nucleotide Polymorphisms (SNPs)
  - 2 copies of chromosomes
  - Homozygous major allele (AA)
  - Heterozygous (AG)
  - Homogzygous minor allele (GG)
- 99.9% genome same among individuals
- Easy to integrate with other ‘omics datasets tied to genome
  - mRNA
  - Proteomics
  - DNA methylation
- Look at physically closest SNP or candidate marker

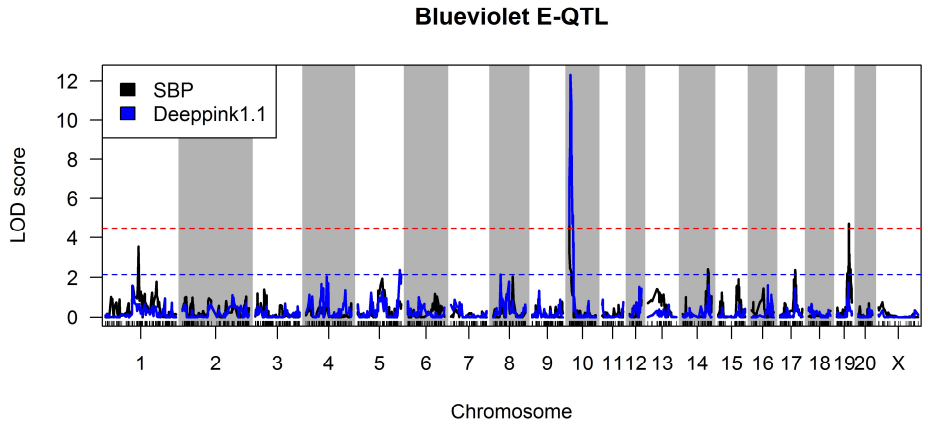


Source: Wikipedia



Quantitative Trait Loci (QTL) Mapping

- Attempt to explain the genetic basis of variation in complex traits
- Outcome is a continuous measure
  - Phenotype (pQTL)
  - Gene/Transcript Expression (eQTL)
- Predictor is SNP marker
- ANOVA at marker loci (marker regression):
  - Outcome = # Alleles (0, 1 or 2 value)
  - Outcome = Dominant Allele (0 or 1 value)
  - Repeat for each marker



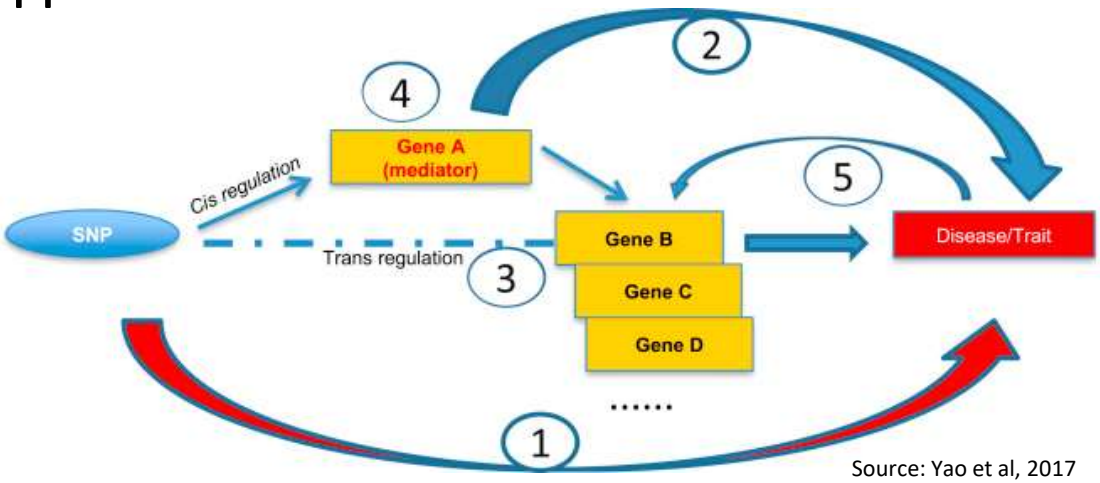
Example QTL for both a phenotype (systolic blood pressure, black trace) and a candidate eigengene expression (from WGCNA, blue trace). Red and blue dotted lines show the genome-wide significant and suggestive thresholds for the eQTL.

QTL Continued

- Logarithm of odds "LOD" Score:
  - log10 likelihood ratio comparing hypothesis of a QTL at position λ versus that of no QTL
  - $LOD(\lambda) = log_{10}\left\{\frac{\mathrm{P}(y|QTL\ at\ \lambda)}{P(y|no\ QTL)}\right\}$
- P-values also reported and plotted (Manhattan plot)
- This is common among animals models
- Limited to population you are looking at
  - 1K -10K number of markers
- Gives you an idea on genomic region, not specific SNP in particular
  - Low resolution, high statistical power
  - Genome wide significance 10^{-5}
- Estimated at least 30% of gene transcripts are substantially influenced by eQTL (Romanoski et. al, 2010)

Cis vs Trans Regulation

- Cis is genetic control from some SNP close to gene
- Trans is genetic control from SNP far away from gene
- You do see these eQTL "hotspots"

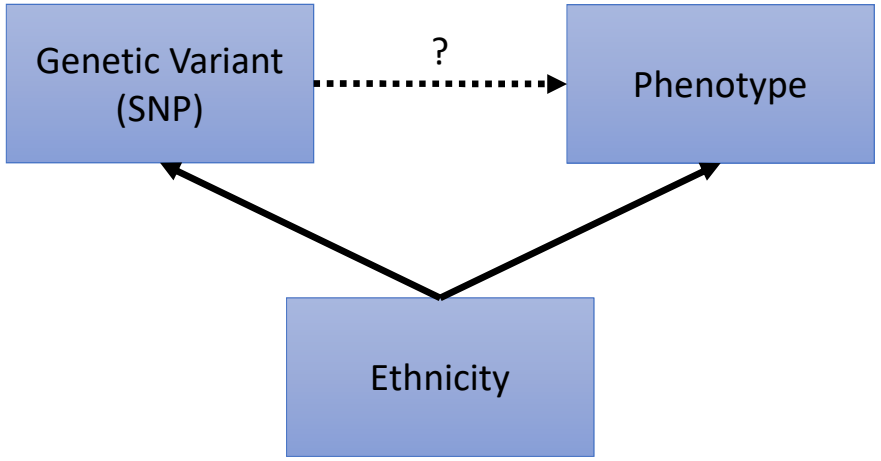


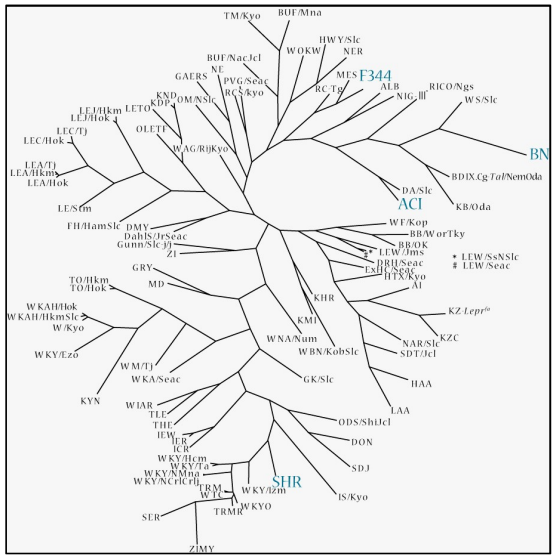
Source: Yao et al, 2017

(1) missense SNP affects protein structure/function
(2) non-coding SNP affects gene expression (*cis*)
(3) non-coding SNP affects remote (*trans*) gene expression directly or by
(4) cis-eGene mediation of the *trans*-eQTL-*trans*-eGene association; or
(5) reverse causality (trait has feedback effect on gene expression).

Population Stratification

- Presence of systematic differences in allele frequencies between subpopulations in a population
- Confounding by ethnicity
- In marker regression, assuming samples are independent
- Non-random mating between groups
  - Different relationships between each combination of strain (admixture)
    - True for the HRDP
  - In humans, physical separation (say African vs European vs Asian descent)
  - Genetic drift of allele frequencies in each group

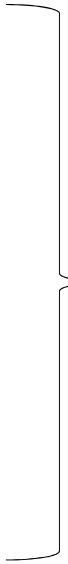




http://www.anim.med.kyoto-u.ac.jp/NBR/Images/phylogenetic_tree_132_200_2.png

QTL Mapping Software

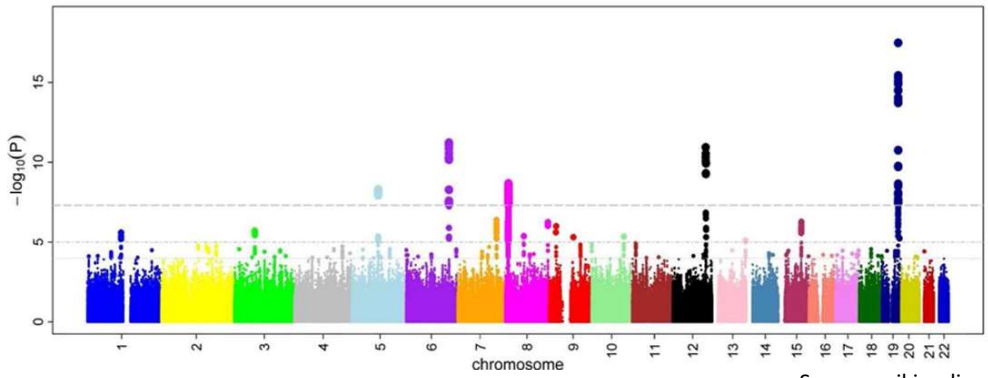
- R/qtl
- R/emma (Efficient mixed-model association)
- R/dlmap (Detection localization mapping for QTL)
- R/mppR (Multi-parent population QTL analysis)
- GEMMA (Genome-wide efficient mixed-model association)



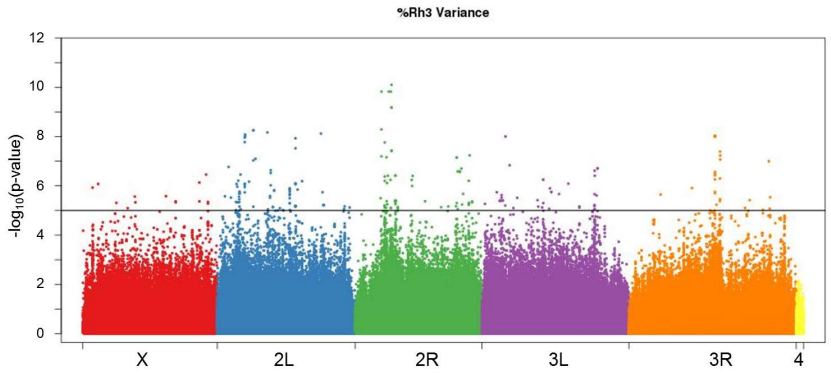
Adjusts for population structure

Genome-Wide Association Study (GWAS)

- Can do this with continuous qualitative traits (like QTLs) or classification traits
- Statistical model same, testing if SNP predicts outcome
- Difference is a much more comprehensive set of markers
  - Millions of markers
- High resolution, but lower statistical power
  - Genome wide significance 10^{-8}
- Software: plink



Source: wikipedia



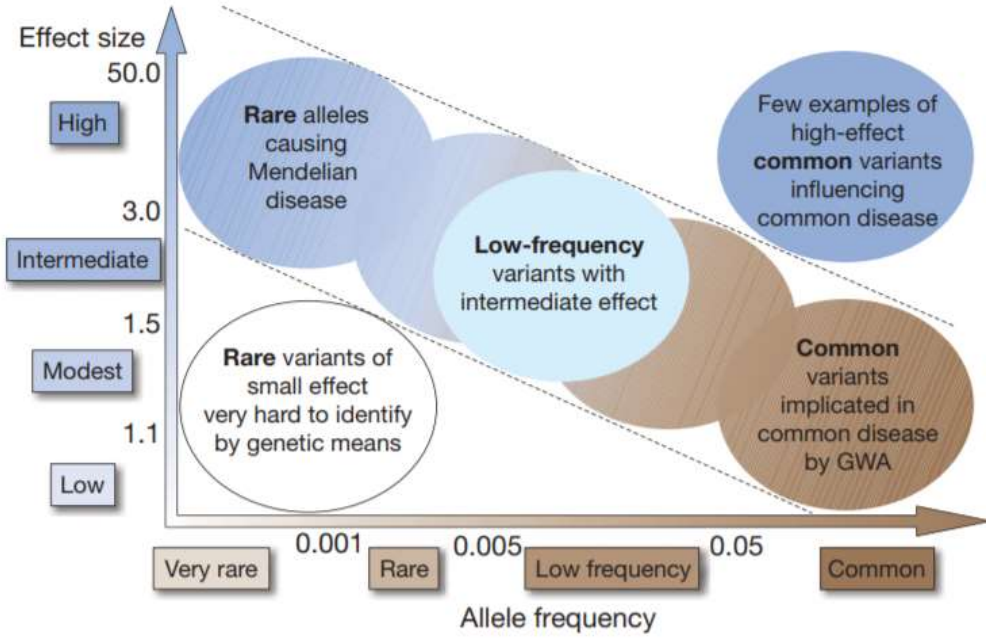
Heritability

$$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2}$$

- Proportion of variability explained by genetics
- Total Variance Trait = Genetic Variance + Environmental Variance
  - Always between 0 and 1, normally reported as a percent
- If have DNA can estimate the heritability of trait (gene expression)
- Dependent on the population you are studying
  - Ranges of heritability reported
- Easy to calculate in animal models where different inbred strains used:
  - R^2 from a 1-way ANOVA: expression = strain

Missing Heritability

- Twin and familial-based linkage studies estimate heritability
- GWAS can account for only for a small proportion
- Example: human height
  - Complex trait (numerous genetic loci involved) estimated heritability at LEAST 80% from familial studies
  - GWAS found 40 loci which IN TOTAL account for only 5% variation
- For complex traits, each loci involved has a small effect size and hard to identify
- Rare variants don't have as much power due to sample size



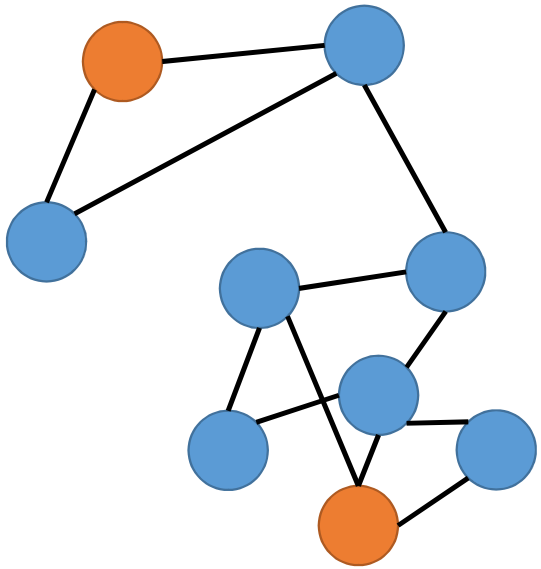
Source: Manolio et. al, 2009

Epigenetic Effects & mRNA Expression

- Easy to link because of location
- Have ChIP peak in a range of gene's TSS
- DNA methylation and gene expression
  - Same samples: correlation (expect negative correlation)
  - Different samples: take candidate list of say differential methylated positions and see if there are differential expression in corresponding gene
- miRNA and mRNA
  - Find the targets for miRNA
    - multiMiR (Dr. Katerina Kechris) http://multimir.ucdenver.edu/
  - Look at correlation or candidates (yes/no)

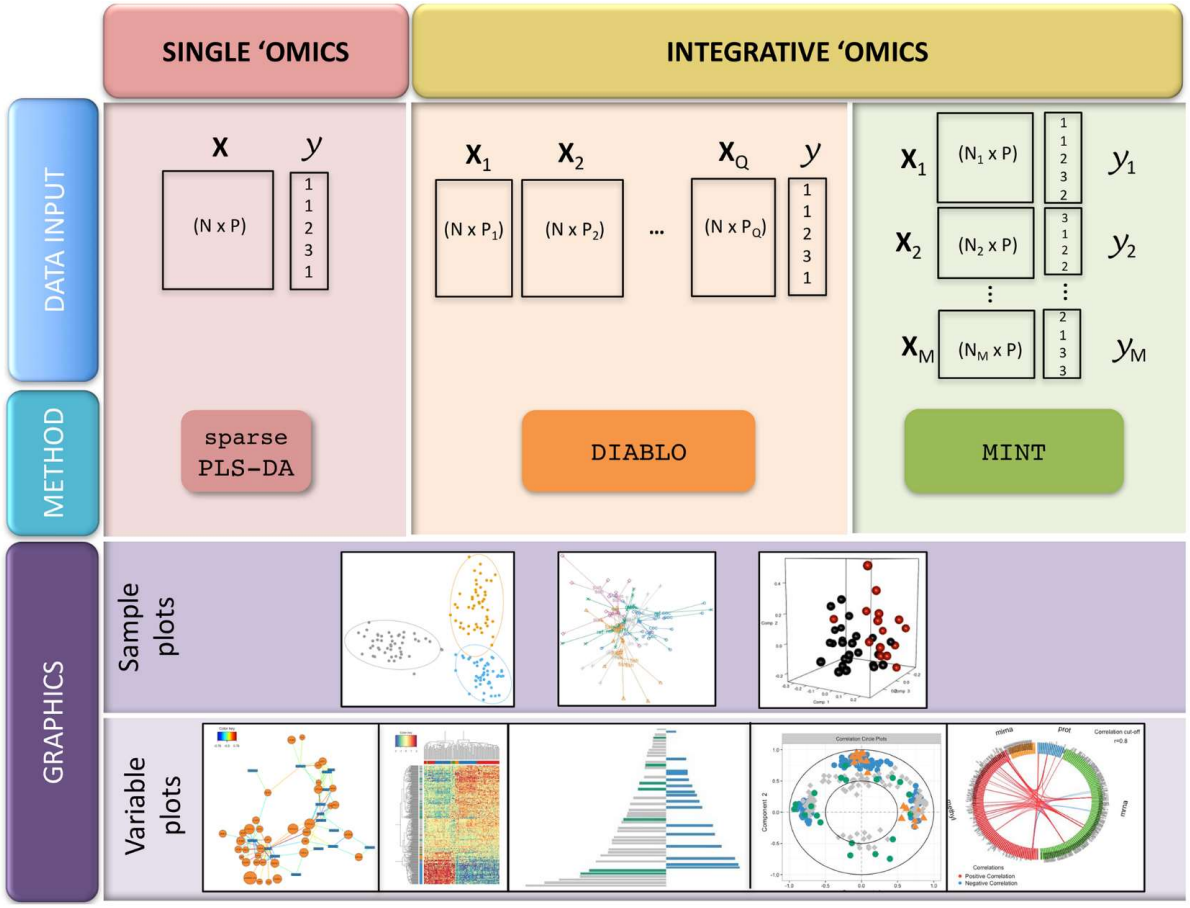
LASSO to get miRNA-mRNA network

- Datasets available:
  - miRNA dataset (~2,000)
  - mRNA dataset (~20,000)
- miRNA can target multiple genes
- Perform WGCNA on the mRNA dataset
- Identify miRNA(s) that regulate module by performing LASSO using eigengene as outcome and miRNAs as predictors
- Need to have same samples in both datasets



R/mixOmics

- Feature Selection
- Data integration
- Supervised analysis
  - Classify or discriminate sample groups
- Sparse Partial Least Squares Discriminant analysis (sPLS-DA)
  - Original 1 dataset approach
- DAIBLO
  - Integration of same biological samples (N) measured on different platforms
  - N-integration
- MINT (Meta Analysis)
  - Integration of independent datasets on measured on same predictors
  - P-integration



Methods Available in mixOmics

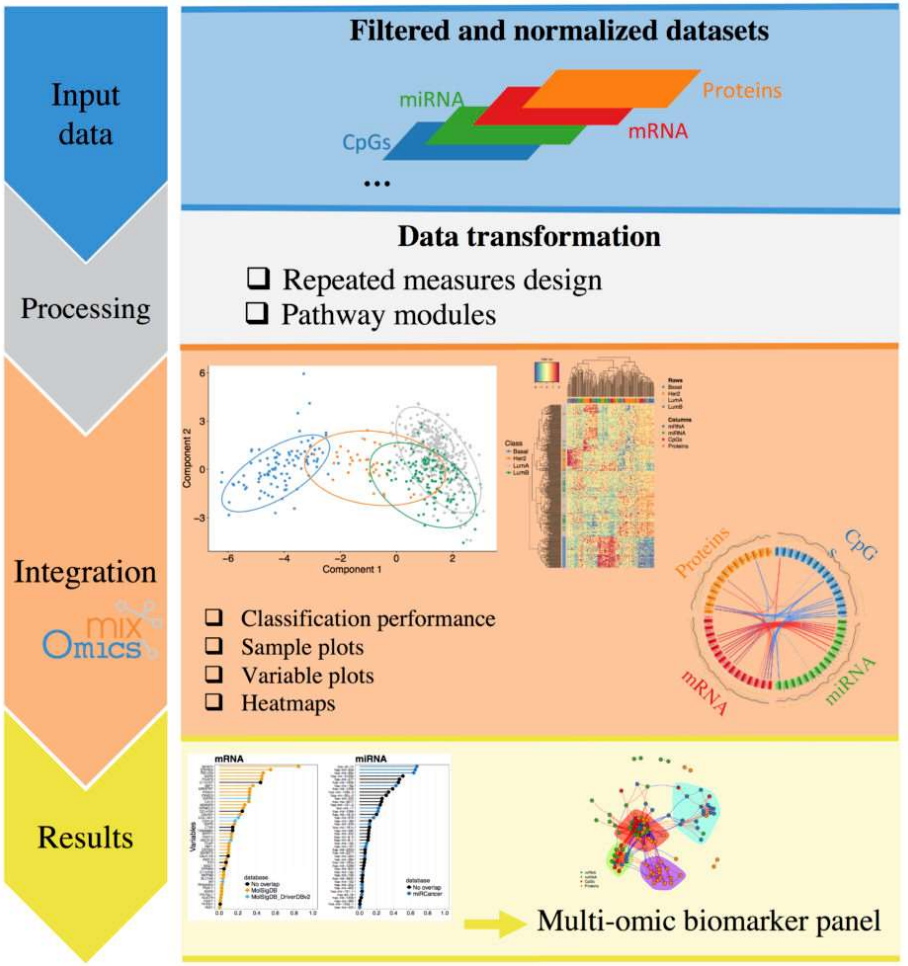
Framework		Sparse	Function name	Predictive model
Single 'omics	unsupervised	-	pca	-
		-	ipca	-
		✓	spca	-
	supervised	-	plsda	✓
		✓	splsda	✓
Two 'omics	unsupervised	-	rcca	-
		-	pls	✓
		✓	spls	✓
N-integration	unsupervised	-	wrapper.rgcca	-
		✓	wrapper.sgcca	-
		-	block.pls	✓
		✓	block.spls	✓
	supervised	-	block.plsda	✓
		✓	block.splsda (DIABLO)	✓
P-integration	unsupervised	-	mint.pls	✓
		✓	mint.spls	✓
	supervised	-	mint.plsda	✓
		✓	mint.splsda	✓

https://doi.org/10.1371/journal.pcbi.1005752.t001

DIABLO

- Data Integration Analysis for Biomarker discovery using Latent variable approaches for 'Omics studies
- Builds on

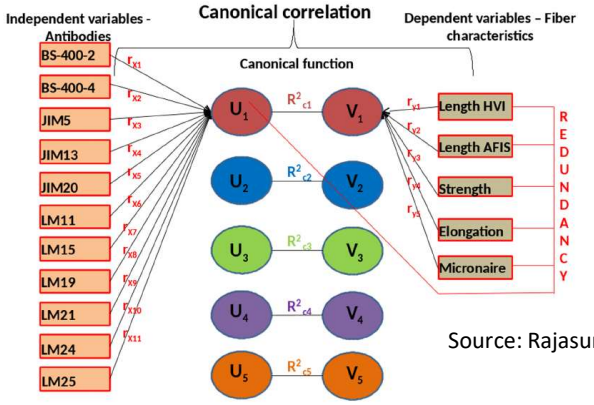
1. generalized canonical correlation analysis (CCA)
2. Sparse sGCCA method



Canonical Correlation Analysis (CCA)

- Multivariate correlation
- Accounts for multi-collinearity
  - Features in a dataset not independent
  - Know not all genes are independent from each other
- Compare sets of variables to sets of variables
- Canonical loadings
  - Variables relationship to own set
  - E.g. gene A expression to gene B expression
- Canonical weight
  - Variables relationship to other set
  - E.g. gene A expression to miRNA X expression
- Canonical Cross-loadings
  - Variables relationship to other set
  - E.g. gene A expression to whole miRNA set CONSIDERING all of gene A's other interaction with other genes in it's own set

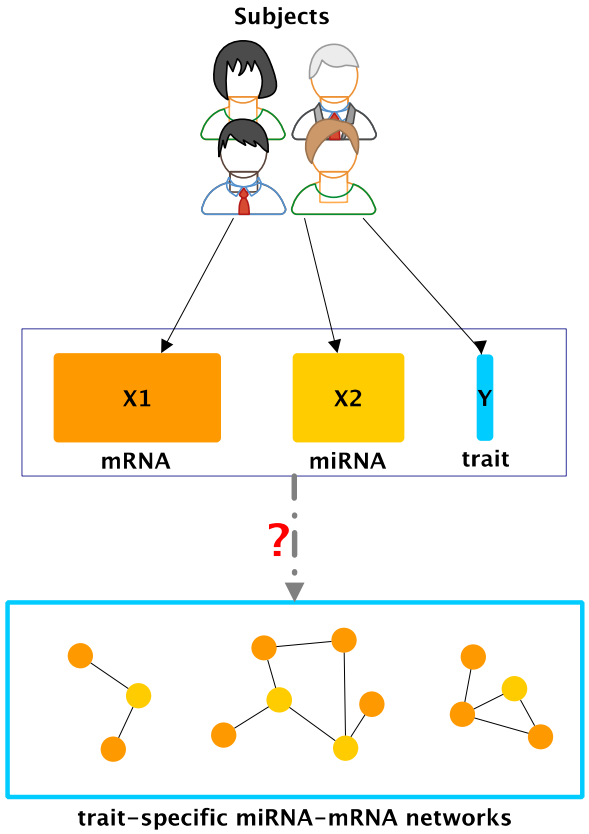
- Canonical loadings and cross-loadings are called structure coefficients
- Canonical weight is a function coefficient
- Canonical correlation is the correlation between sets
- Redundancy coefficients
  - Shared loading variance (variance explained within set)
  - Shared cross-loading variance (variance explained between sets)



Source: Rajasundaram et. Al, 2014

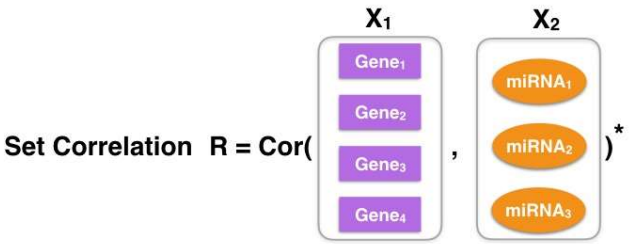
Sparse Multiple Canonical Correlation Network Analysis (SmCCNet)

- R/SmCCNet
  - Congrat to Drs. Katerina Kechris & Jenny Shi!
- CCA: Relationship between 2 multivariate datasets measured on same samples
  - E.g. Gene A to Gene B within mRNA dataset
- Multiple: multiple 'omics datasets
  - miRNA and mRNA
- Sparse: not expecting many connections



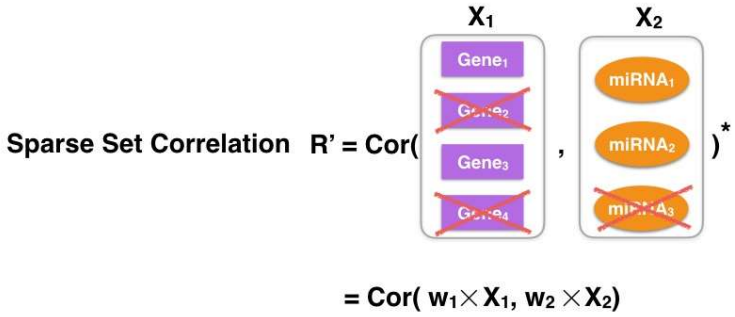
Source: Katerina Kechris

CCA vs Sparse CCA



where w_1 & w_2 are 4×1 and 3×1 unit vectors respectively.

Sample canonical weights:
$w_1 = (0.10, -0.48, 0.83, 0.22)^t$,
$w_2 = (0.67, 0.14, 0.73)^t$.

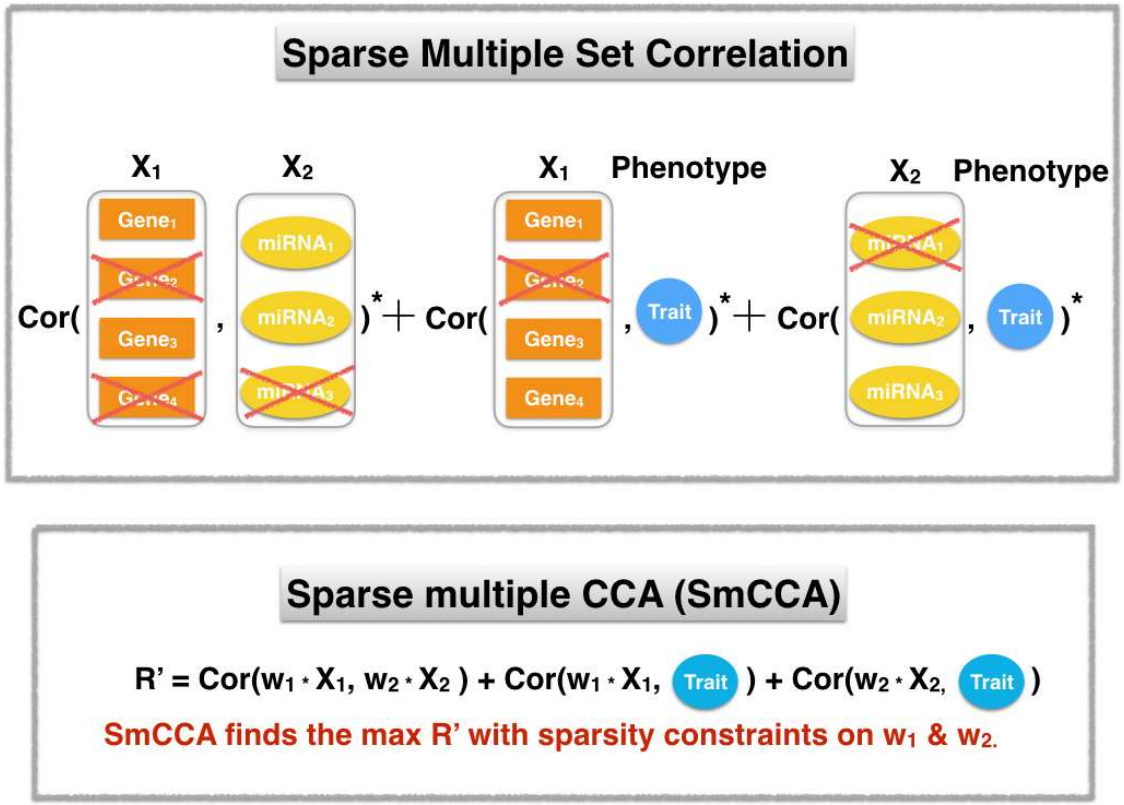


where w_1 & w_2 are 4×1 and 3×1 unit vectors, satisfying some constraints $p_1(w_1) < c_1$ and $p_2(w_2) < c_2$, respectively.

Sample canonical weights:
$w_1 = (0.17, 0, 0.37, 0)^t$,
$w_2 = (0.25, 0.28, 0)^t$.

Source: Katerina Kechris

SmCCA

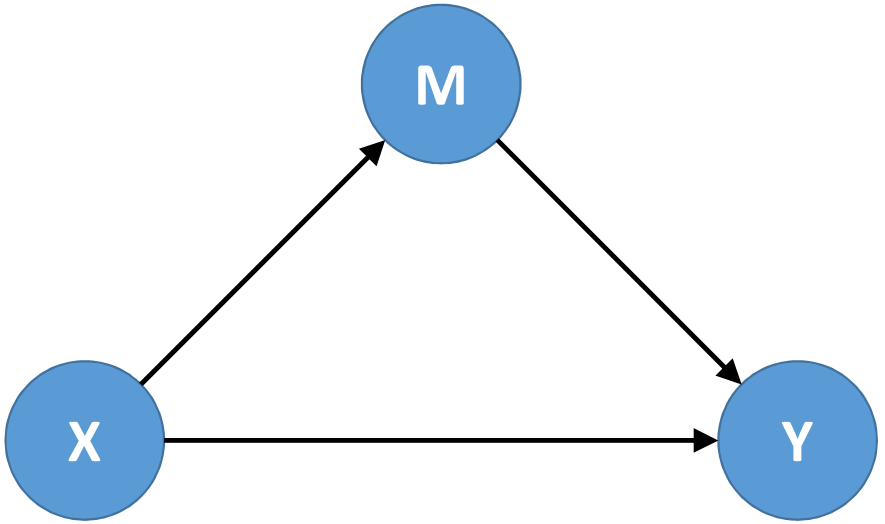


Subsampling of miRNA/mRNA & cross-validation of samples for sparse penalties on weights

Source: Katerina Kechris

Mediation Analyses

- Observe a relationship between independent (X) and dependent variables (Y)
- Mediator (M) is in the causal pathway of X -> Y
- Complete & partial mediation
- We have seen something like this!
- Build on this an get Directed Acyclic Graph (DAG)

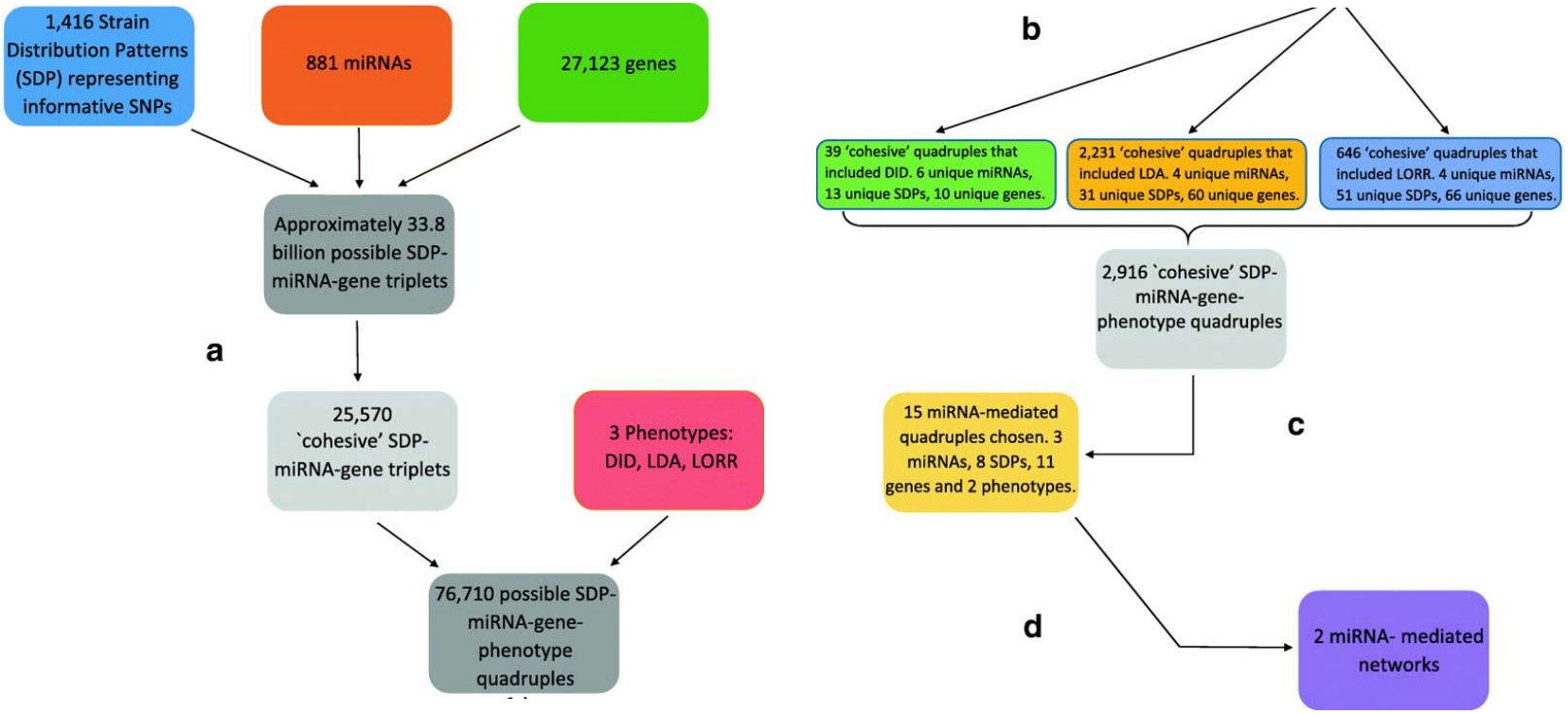


Mediation Example 1

1. Find what metabolites best represent dietary information
2. Candidates from metabolites -> (T1D)
3. Candidates from DNA methylation -> T1D
4. Candidates from metabolites -> DNA methylation
5. See resulting combinations of metabolites, DNA methylation sites left
6. Perform a traditional Baron & Kenny (et al, 1986) between 3 combos
   - Significant X -> Y (model Y = X)
   - Significant M -> Y (model Y = M)
   - Non-significant X coefficient (complete mediation) in Y = X + M

- Big thing is filtering down to a reasonable amount of combinations

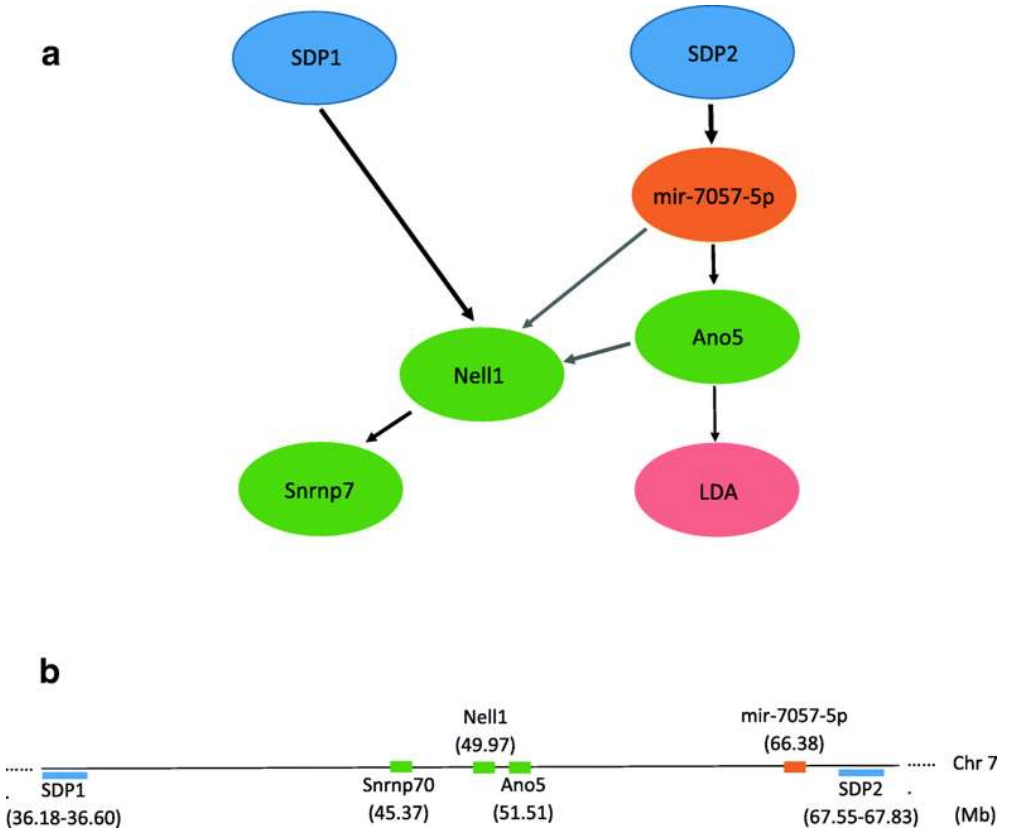
Mediation Example 2



Source: Rudra et. al, 2018

Mediation Example 2

- Bayesian networks adds the arrows between nodes of same dataset
  - Ideal for taking an event that occurred and predicting the likelihood that any one of several possible known causes was the contributing factor
  - Computationally intensive
- Once again, start small and build up



Source: Rudra et. al, 2018

Other Integration Software

- R/Omic
- R/integrOmics
- R/ STATegRasPLS
- R/OMICsPCA
- R/MultiAssayExperiment
- R/iCluster
- R/CNAmet
- R/OmicKriging
- matlab/JIVE
- JAVA/OmicsAnalyzer
- JAVA/VANTED
- JAVA/Lemon-Tree
- C++/DASS-GUI
- C++/GeneTrail2
- Perl/3Omics
- Perl and Python/PaintOmics

References

Romanoski CE, Lee S, Kim MJ, Ingram-Drake L, Plaisier CL, Yordanova R, Tilford C, Guan B, He A, Gargalovic PS, Kirchgessner TG, Berliner JA, Lusis AJ. *Systems genetics analysis of gene-by-environment interactions in human cells.* Am J Hum Genet. **2010** Mar 12;86(3):399-410.

Yao C, Joehanes R, Johnson AD, Huan T, Liu C, Freedman JE, Munson PJ, Hill DE, Vidal M, Levy D. *Dynamic Role of trans Regulation of Gene Expression in Relation to Complex Traits*. Am J Hum Genet. **2017** Apr 6;100(4):571-580.

Manolio TA. et al, *Finding the missing heritability of complex diseases.* 08 October **2009** Nature volume 461, pages 747–753.

Rajasundaram D, Runavot JL, Guo X, Willats WG, Meulewaeter F, Selbig J. *Understanding the relationship between cotton fiber properties and non-cellulosic cell wall polysaccharides.* PLoS One. **2014** Nov 10;9(11):e112168.

Baron, R. M. and Kenny, D. A. (**1986**) "The Moderator-Mediator Variable Distinction in Social Psychological Research – Conceptual, Strategic, and Statistical Considerations", Journal of Personality and Social Psychology, Vol. 51(6), pp. 1173–1182.

Rudra P, Shi WJ, Russell P, Vestal B, Tabakoff B, Hoffman P, Kechris K, Saba L. *Predictive modeling of miRNA-mediated predisposition to alcohol-related phenotypes in mouse.* BMC Genomics. **2018** Aug 29;19(1):639.