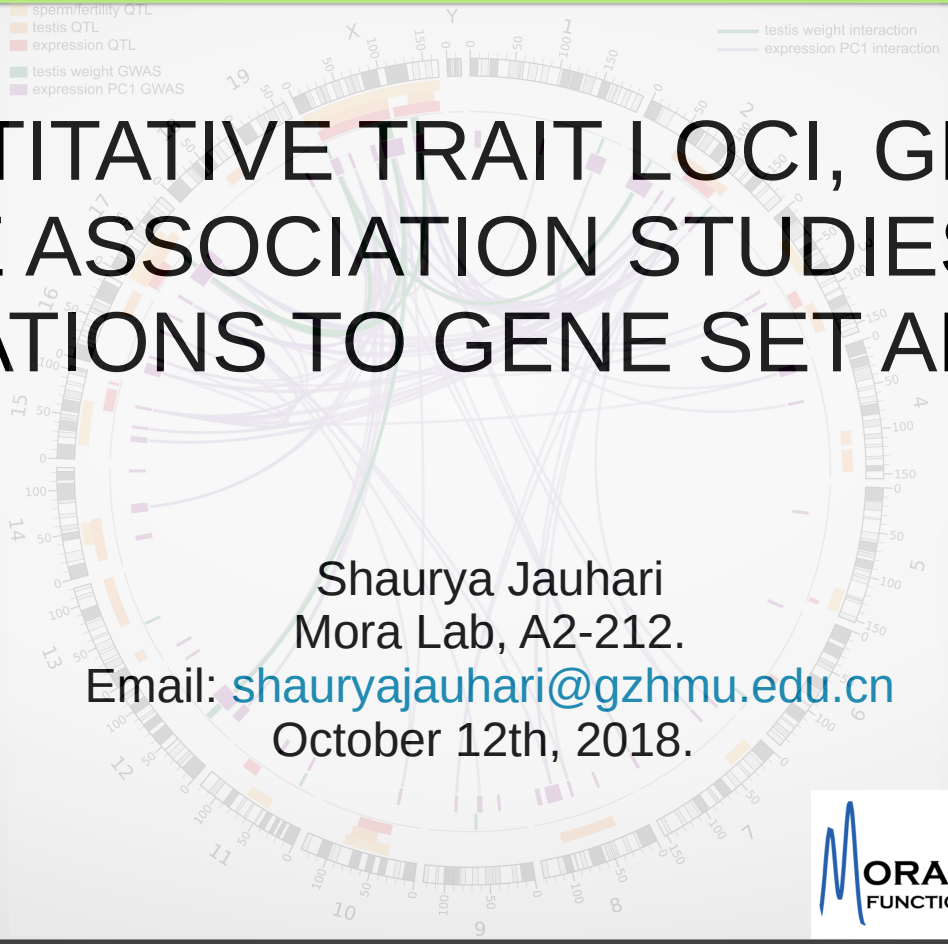


QUANTITATIVE TRAIT LOCI, GENOME WIDE ASSOCIATION STUDIES AND APPLICATIONS TO GENE SET ANALYSES

Shaurya Jauhari
Mora Lab, A2-212.
Email: shauryajauhari@gzhmu.edu.cn
October 12th, 2018.

MORA LAB
FUNCTIONAL BIOINFORMATICS

GUANGZHOU MEDICAL UNIVERSITY
1958

OUTLINE

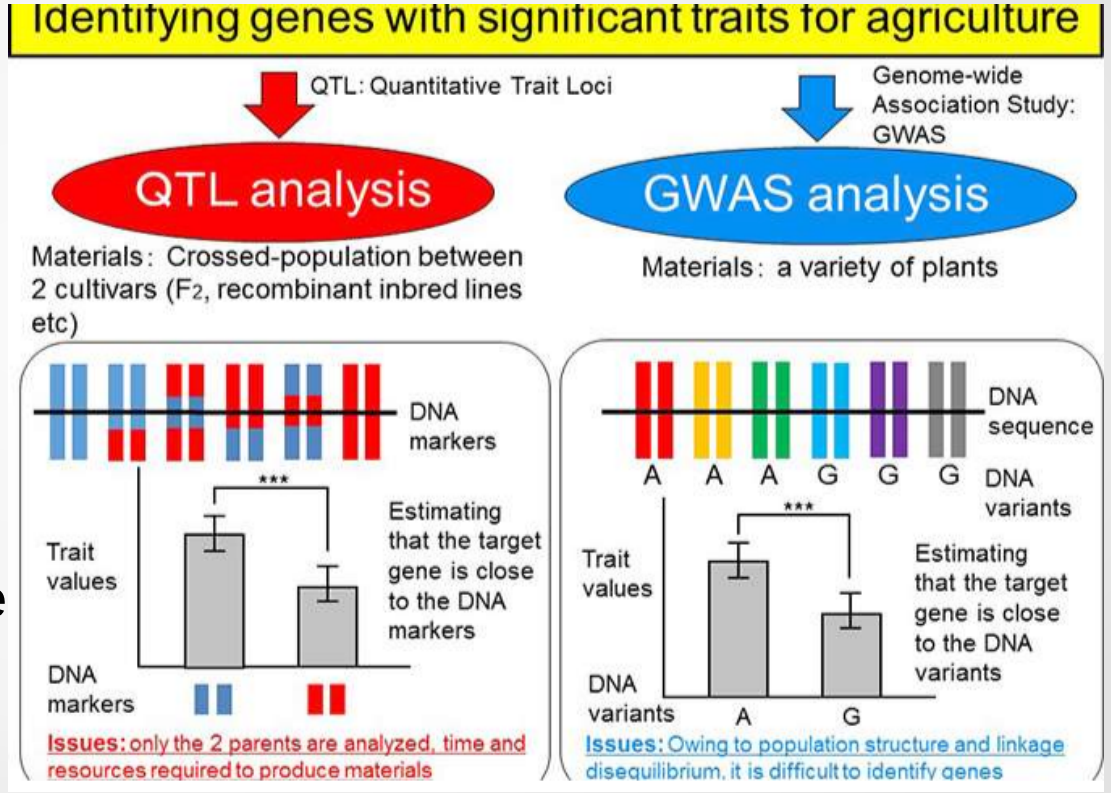
- Single Nucleotide Polymorphism (SNP)
- Copy Number Variations (CNV)
- Genome Wide Association Studies (GWAS)
- Quantitative Trait Loci (QTL)
- Data Sources

GENOME WIDE ASSOCIATION STUDIES

- *Genetics, more than genomics !*
  - heritability of complex traits
- Comparing differential loci in correlation to functional traits
- SNPs, CNVs comprehend largely to the GWAS doctrine
- Almost drawing parallels with the QTL

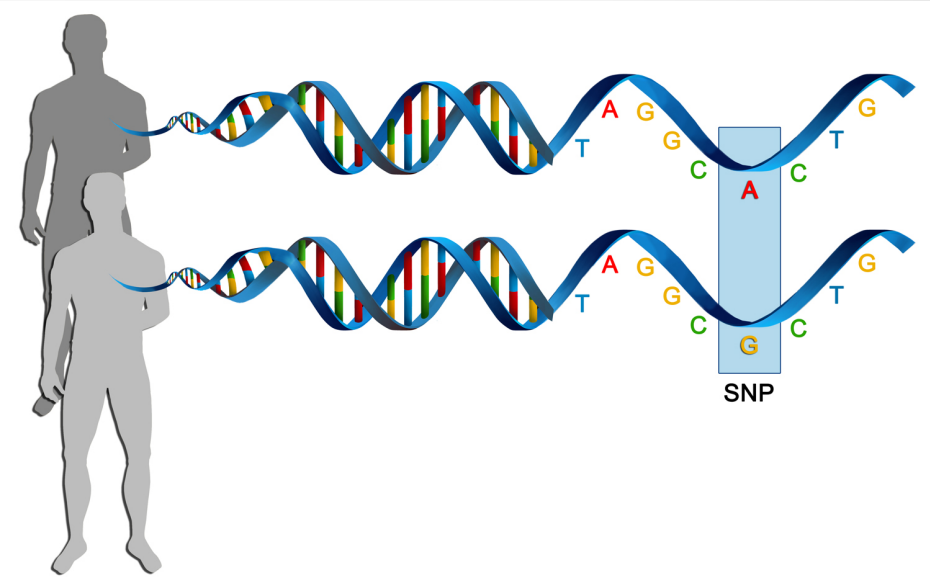
HAND IN HAND

- Credit: Kobe University
- Kenji Yano et al. *Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice*, Nature Genetics (2016). DOI: 10.1038/ng.3596
- GWAS: discrete classification/ qualitative trait; usually single gene
- QTL: continuous data reflecting polygenic enactment



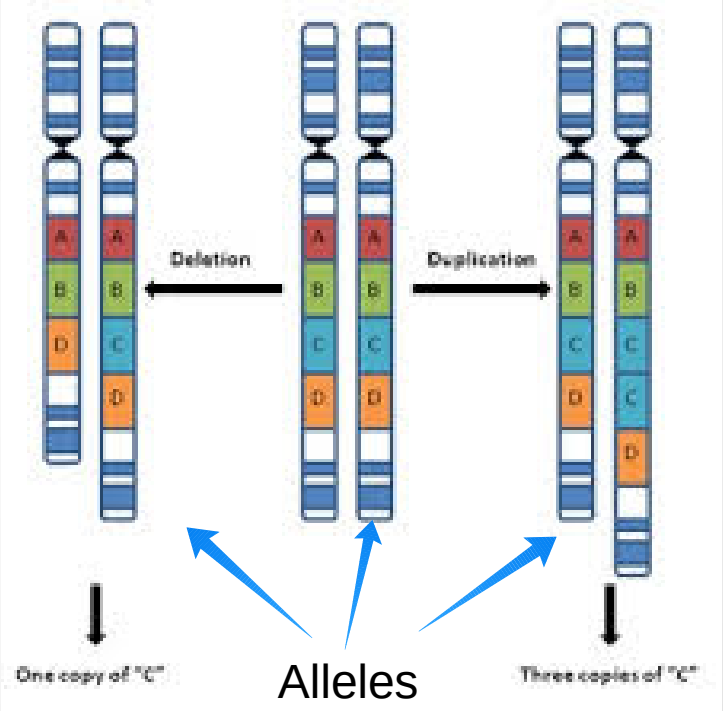
SINGLE NUCLEOTIDE POLYMORPHISM (SNP)

- Most common genetic variation/ genomic polymorphism
  - Occur (on average) in every 1000 bp window
  - ~ 3 million SNPs across two individuals (humans)
- Suffuse coding as well as non-coding regions
- Manifest disease causing abilities, response to therapy, and other genomic stimuli.
- Data sources:
  - NCBI- dbSNP (rsxxxxx)



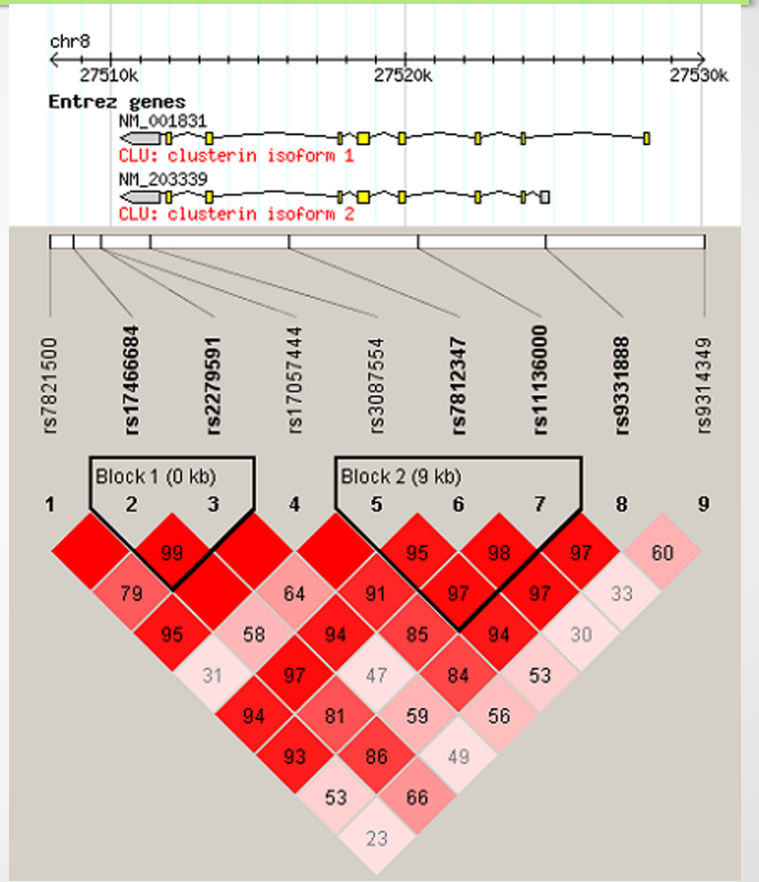
COPY NUMBER VARIATIONS

- Structural Variation
- Discrepancy in occurrence of the sections in the genome
  - Deletion (loss) and Duplication (gain)
- “Manifest disease causing abilities, response to therapy, and other genomic stimuli.“



LINKAGE DISEQUILIBRIUM

- Non-random association of alleles at two-or more loci.
  - If random, then linkage equilibrium (lower correlation).
- LD coefficient: r^2
- Visualization available in R:
  - *LD.plot()*, *LDheatmap()*, *snp.plotter()*
- The cells represent correlation coefficient (in percentage). We are interested to know if two SNPs were INHERITED TOGETHER.
- Blocks represent haplotypes.



GWAS Impediments

- SNPs to Genes
  - Definitions of gene boundaries
  - Overlapping gene signatures
  - LD problem; arbitrary boundary selection
  - *Biasness* via gene set size and gene length
  - Annotation bias; certain genes are better studied

HEAD ON

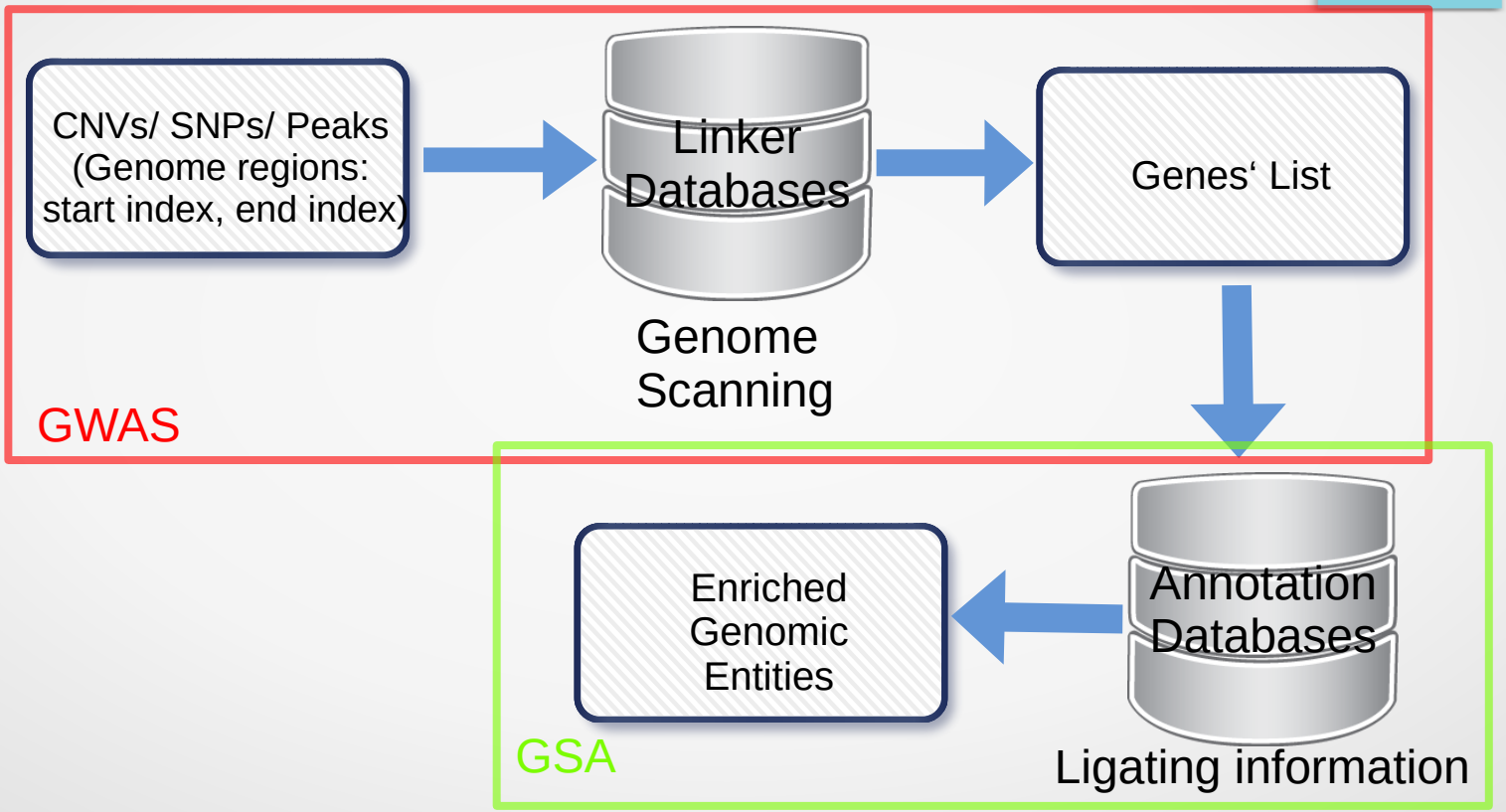
GWAS

- SNP/Gene level association to diseases
- Correcting for multiple testing may overlook weak associations
- Overwhelming sensitivity for GSA methods
- Linkage Disequilibrium (LD): SNP-Gene affiliations if other SNPs lie in the same genic region; induces count bias.

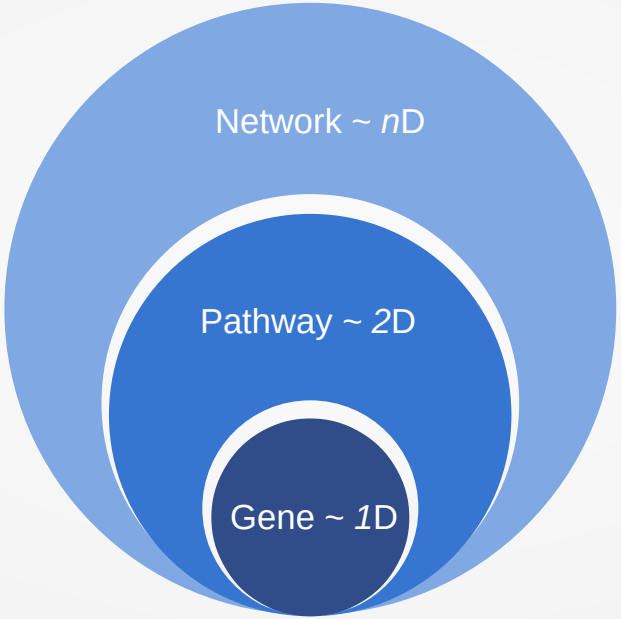
GSA

- Pathway level association to diseases
- Pathways (generally) outnumber genes
- GSA methods inappropriate for GWAS data
- Normalization warranted on the grounds of gene-set/ pathway size and multiple associations of gene-pathway linkages.

A Generic Schema



Contd...



WHY?

- High throughput sequencing techniques spit out *DATA*, **not** *INFORMATION*.

- Listing isn‘t enough!
  - Genes do not work in isolation (polygenic associations)

- **Crucial to bring out the biological meaning.**
  - Discering etiology to illnesses
  - Macro-view of a physiology

ANNOTATION DATABASES

- Ontology databases
  - Gene Ontology
  - Disease Ontology, etc.
- Open pathway databases
  - KEGG
  - Reactome
  - WikiPathways, etc.
- Gene set databases
  - MsigDB, etc.

CAVEATS

- Centralized repositories holding clues
  - Discrepancies in nomenclature
  - Dearth of information
  - Redundancy, eg. GO stores information in hierarchical order.

- *Quantitative*, not biological.
  - Critical to definition of gene sets, statistical methods
  - Validation via 3D-FISH imaging, etc.

- There is no *BEST* method

METHODS FOR EVALUATING STATISTICAL SIGNIFICANCE | HYPOTHESIS

- *Set Permutations | Run Simulations | for empirical p-value estimation*
  - *sample gene lists of variable sizes.*
  - *Phenotype*

- *Competitive*
  - Comparing associations within the target gene set to a random gene set, of similar size.
  - *Null hypothesis*: there is no difference
  - *Cons*: Magnitude of linkage to the phenotype cannot be ascertained
  - eg. Over-representation analysis (2x2 table methods)

- *Self- Contained*
  - Comparing associations within the target gene set only.
  - *Null hypothesis*: there is no association with a phenotype.
  - *Cons*: Cannot ascertain the credibility of the target gene set with respect to others.

- Null hypothesis: No genes are associated with the phenotype (recently proposed)

QTL/ GWAS data | GSA processing

QTL/ GWAS data inclusion in the standard GSA pipeline

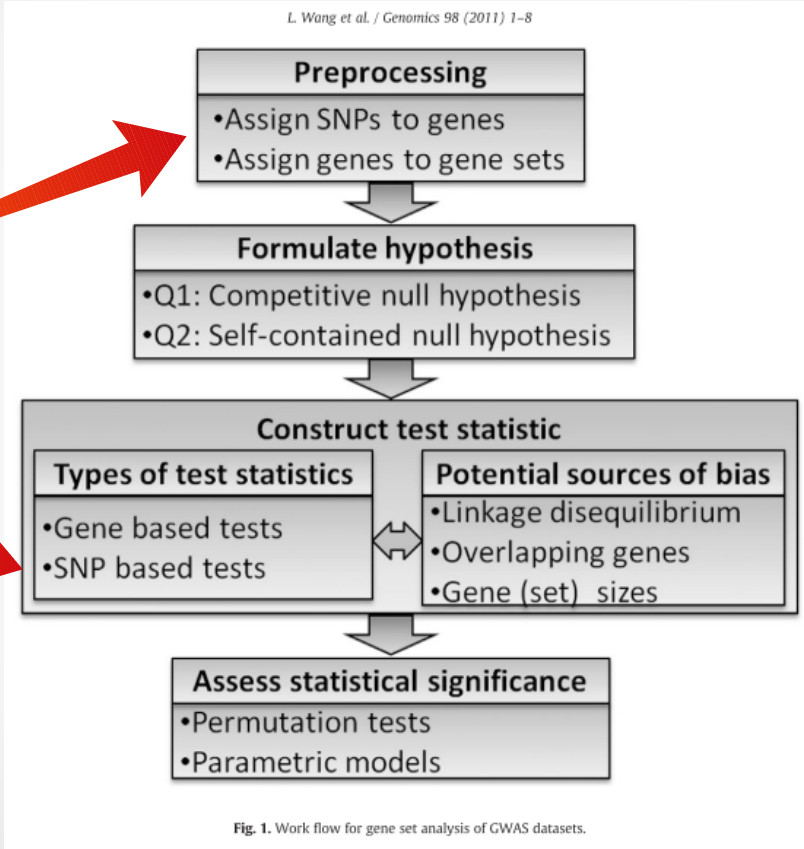


Fig. 1. Work flow for gene set analysis of GWAS datasets.

STATISTICALLY SIGNIFICANT PATHWAYS

- Sample vs Population
- Hypergeometric test on gene set
  - *phyper()* in R
- Filtering based on p-value
  - probability of observing a result at least as extreme as the one that was randomly observed.

GENE SET ANALYSES [Mooney et al. 2015]

- Step 1 : Quality control of GWAS resultant data (SNPs/ CNVs)
- Step 2 : Establish program objective
  - Which database to query?
- Step 3 : Map SNPs/ CNVs to Genes
- Step 4 : Select GSA tool/ method
- Step 5 : Evaluate statistical significance

DATA SOURCES

- QTL:
  - seeQTL (humans), QTLdb (animals)
  - Searchable by dbSNP ID, Gene Symbol and genomic coordinates
  - qtl() in R

- GWAS:
  - https://www.ebi.ac.uk/gwas/
  - Searchable by disease, dbSNP ID, Gene Symbol and genomic coordinates

BIBLIOGRAPHY

- Mooney, M. A., & Wilmot, B. (2015). Gene set analysis: A step-by-step guide. American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics, 168(7), 517–527. https://doi.org/10.1002/ajmg.b.32328
- Wang, L., Jia, P., Wolfinger, R. D., Chen, X., & Zhao, Z. (2011). Gene set analysis of genome-wide association studies: Methodological issues and perspectives. Genomics, 98(1), 1–8. https://doi.org/10.1016/j.ygeno.2011.04.006
- Fridley, B. L., & Biernacka, J. M. (2011). Gene set analysis of SNP data: Benefits, challenges, and future directions. European Journal of Human Genetics, 19(8), 837–843. https://doi.org/10.1038/ejhg.2011.57

TAKE HOME POINTS

- GWAS/ QTL analyses adds another dimension to GSA
  - bolsters findings

- Genes are ranked in the order of their expression values (ORA/ FCS), while SNPs are ranked on the degree of association with genes.
  - *SNPs → Genes → Pathways*
  - Every tier represents an exclusive magnitude of analyses

- Databases: necessary, but *not sufficient*.


Thank You!