Genetic Variation, Biological Pathways, and Networks

Sarah Pendergrass

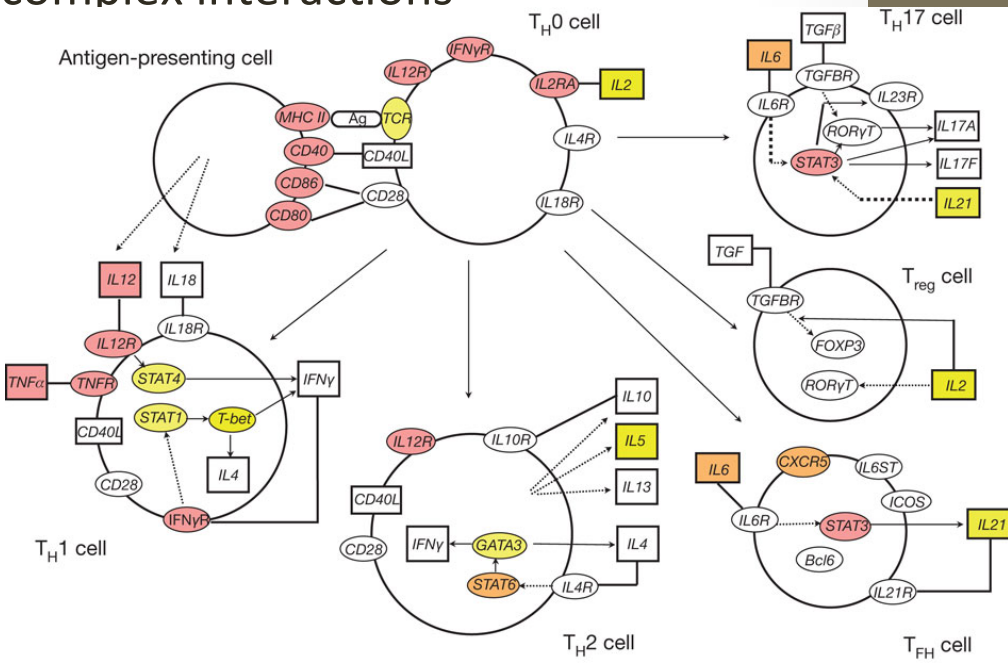
Center for Systems Genomics

Outline

- What are networks and why are they important in biology?
  - Biological Pathways
- Why do we care about genetic variants in the context of networks and pathways?
- What are tools and approaches for exploring these?

Networks

- Biological data is inherently connected
  - Networks: connecting subunits by information
  - Dynamic networks exist between genetic architecture, signaling pathways, intermediate phenotypes, and outcome traits
  - Already discussed the complicated interactions at the transcription level
  - At the translational level there are complex interactions
  - Proteins interact with each other
  - Signaling cascades
  - Temporal responses to stimuli



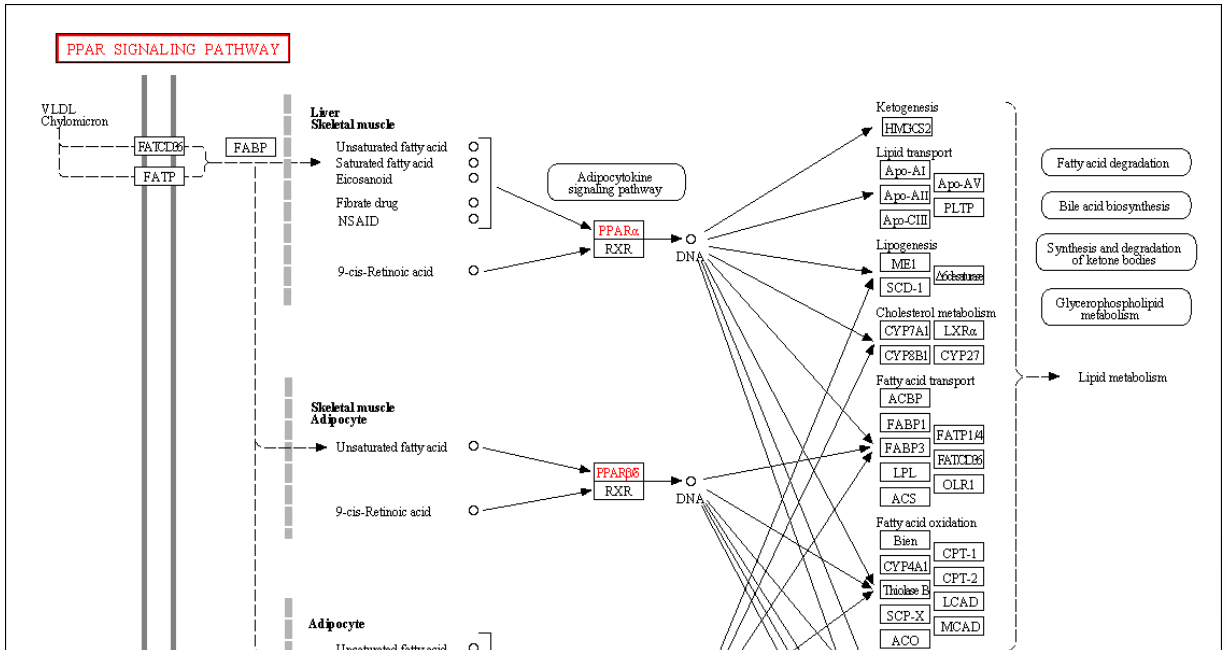
Biological Pathways

- Molecules interacting that result in changes within a cell
  - Metabolic pathways
    - Substrates being modified, usually by enzymes, to form another product
      - Breakdown of fuel into ATP
  - Gene regulation pathways
    - Turning genes on and off
  - Signal transduction pathways
    - Cell exterior signals to interior of cells
    - Binding of growth factors to cell surface receptors
      - Cascades of behaviors that follow
    - Inflammatory response
      - Cellular responses
    - Hormones and the endocrine signaling system

- Important to remember feedback here
  - Most pathways don't just end at some point but are connected to something else
- What are some ways to explore these networks/pathways if I have a gene or genes of interest?

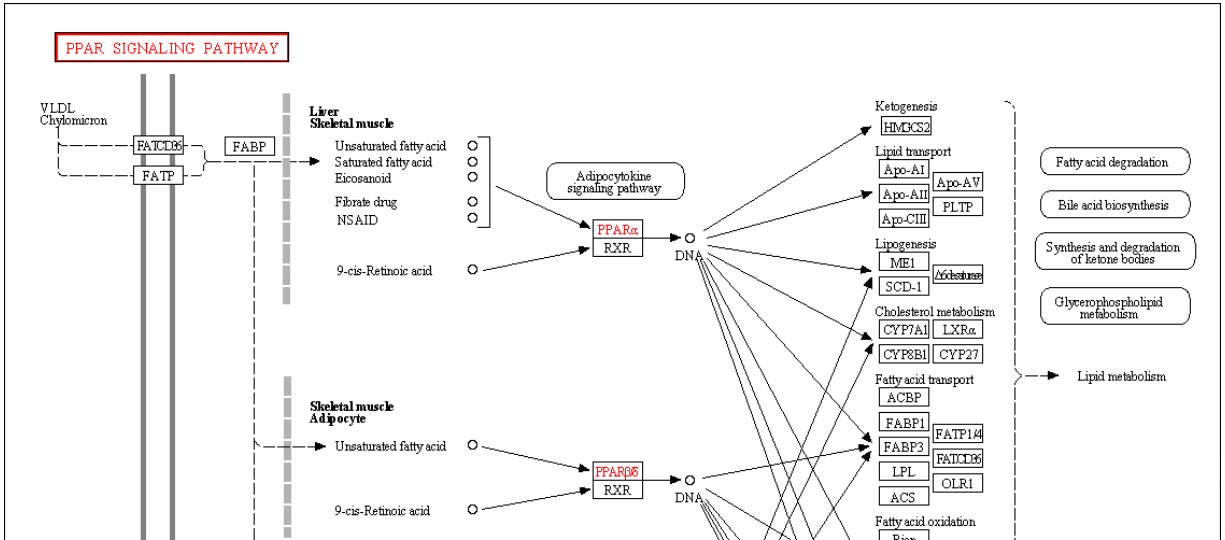
Biological Pathways

- Example sources of pathway information
  - KEGG: Kyoto Encyclopedia of Genes and Gene Interactions
    - http://www.genome.jp/kegg/
  - KEGG consists of the seventeen main databases, broadly categorized into
    - systems information
    - genomic information
    - chemical information
    - health information



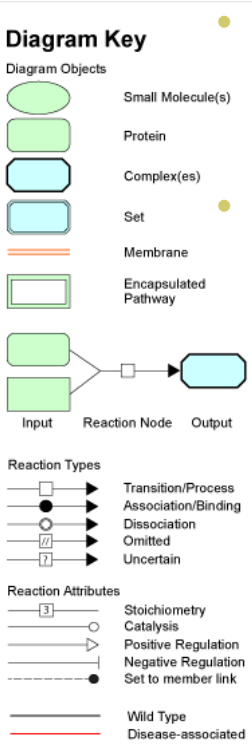
Biological Pathways

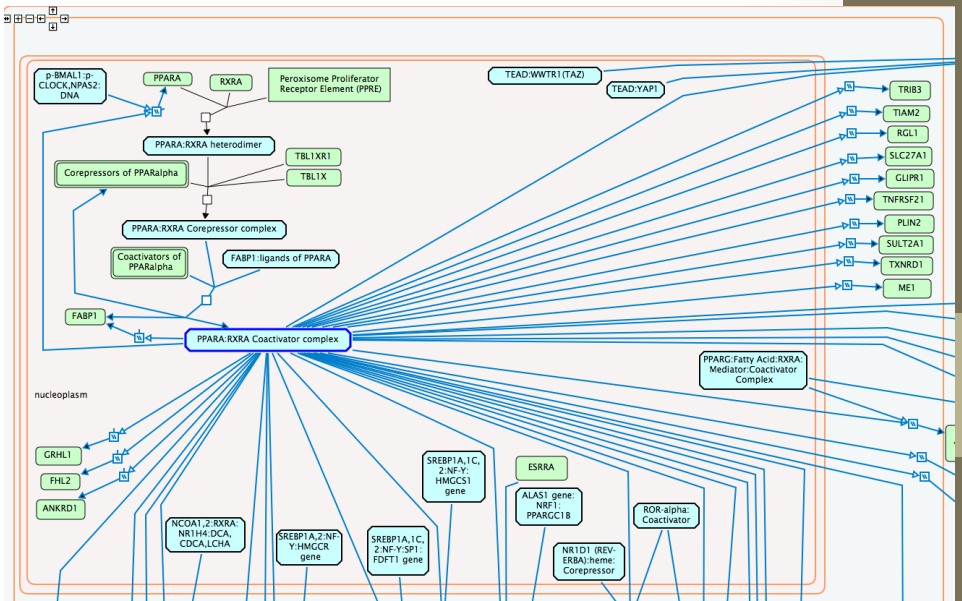
- Example sources of pathway information
  - KEGG: Kyoto Encyclopedia of Genes and Gene Interactions
    - http://www.genome.jp/kegg/
  - Visualization of connections between data
    - The global metabolic gene network covers about 1100 genes involved in approximately 15 000 unique pairwise interactions
    - Biological pathway molecular interactions and reactions, but also other other biological relationships
    - One at a time gene search
    - The pathways in KEGG are manually drawn and derived from textbooks, literature and expert knowledge



Biological Pathways

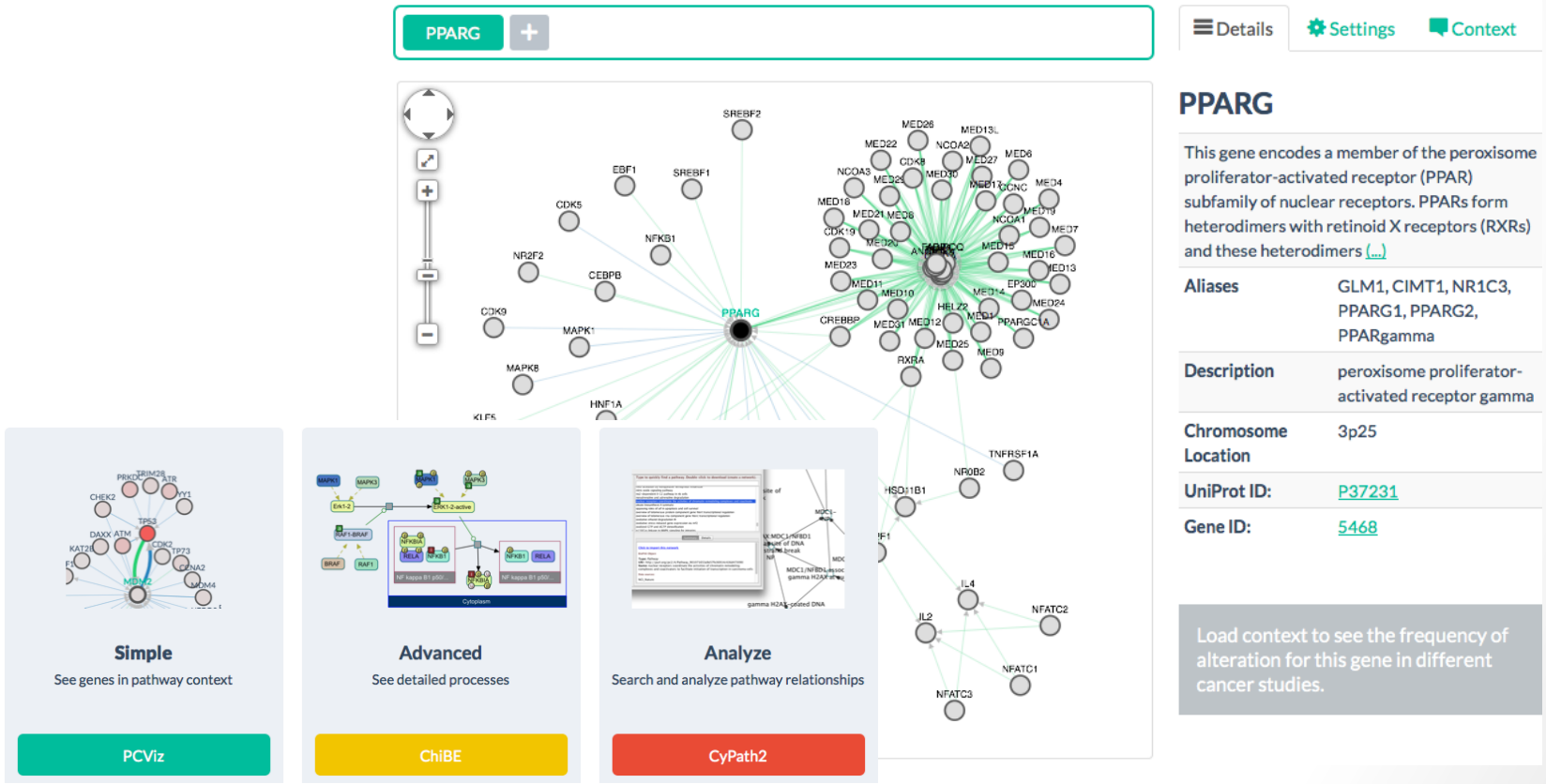
- Example sources of pathway information
  - Reactome
    - http://www.reactome.org/
  - Little more user friendly than KEGG (can search multiple genes)
  - Free, open-source, curated and peer reviewed pathway database
  - Pathways and reactions (pathway steps) in human biology
  - 3700 proteins (including proteins from non-human species that interact with human proteins) involved in approximately 83 000 unique pairwise interactions
  - Easy to search on list of genes





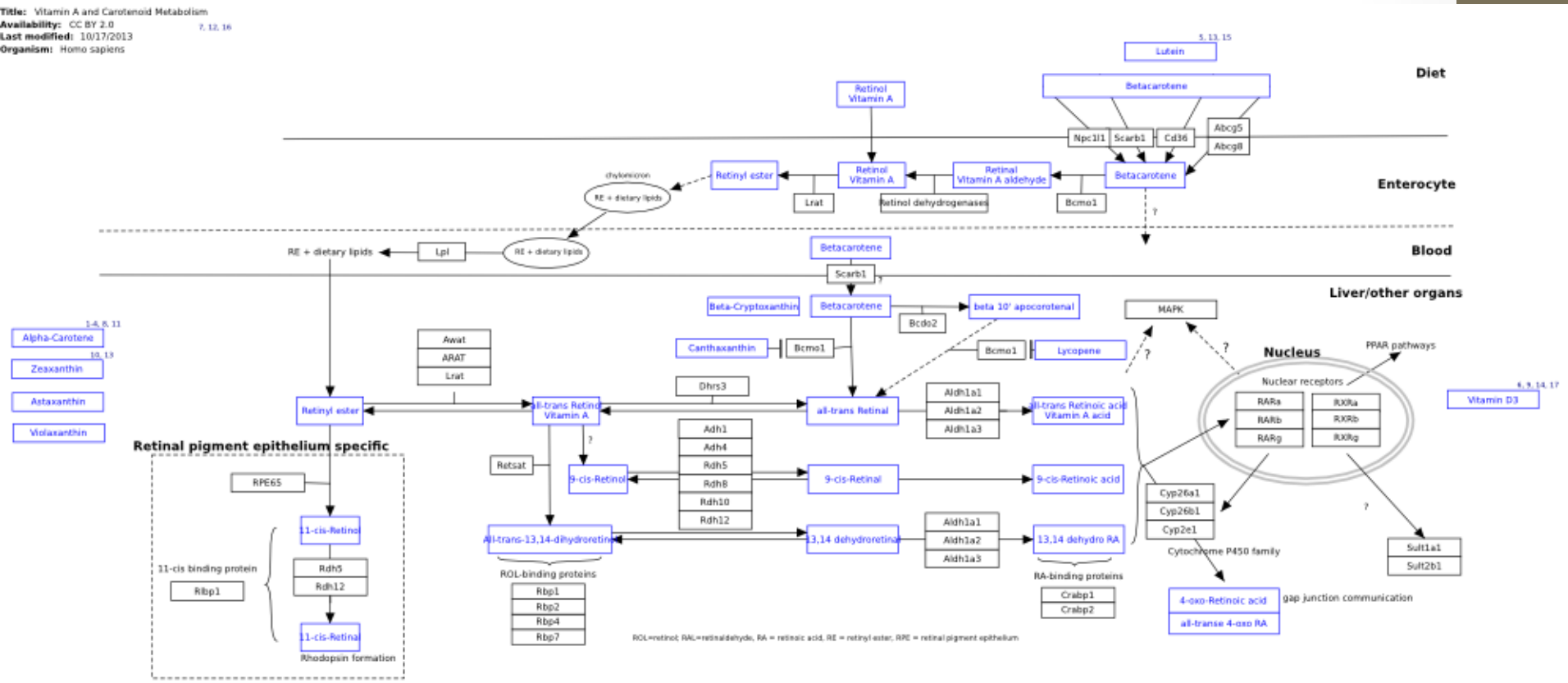
Biological Pathways

- Example sources of pathway information
  - Pathway Commons
  - http://www.pathwaycommons.org/about/
  - Can provide gene list to see connections



Biological Pathways

- Example sources of pathway information
  - Wikipathways
  - http://en.wikipedia.org/wiki/WikiPathways
  - Articles each devoted to a different biological pathway
    - Peer review the responsibility of the user committee

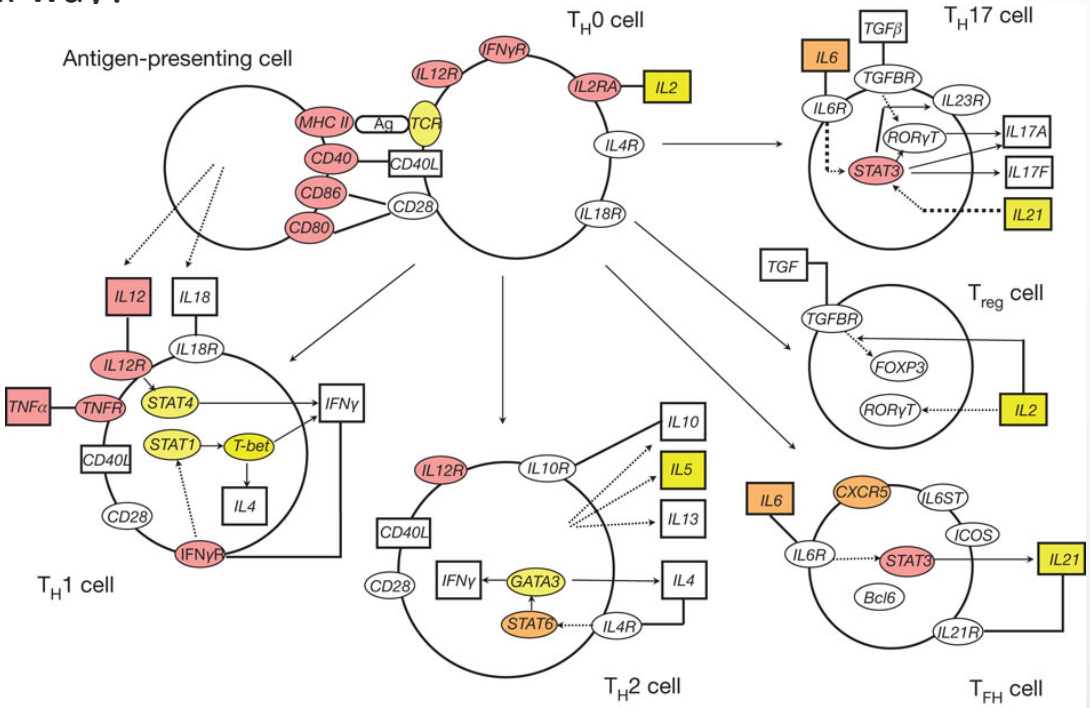


Biological Pathways

- Example sources of pathway information
  - Gene Ontology
  - http://geneontology.org
  - A collaborative effort to address the need for consistent descriptions of gene products across databases
    - Three structured, controlled vocabularies (ontologies) describing gene products in terms of
      - Associated biological processes
      - Cellular components
      - Molecular functions
      - All in a species-independent manner
  - These concepts have been used to "annotate" gene functions based on experiments reported in over 100,000 peer-reviewed scientific papers
  - Another way to think about how your gene of interest relates to other genes via biological processes

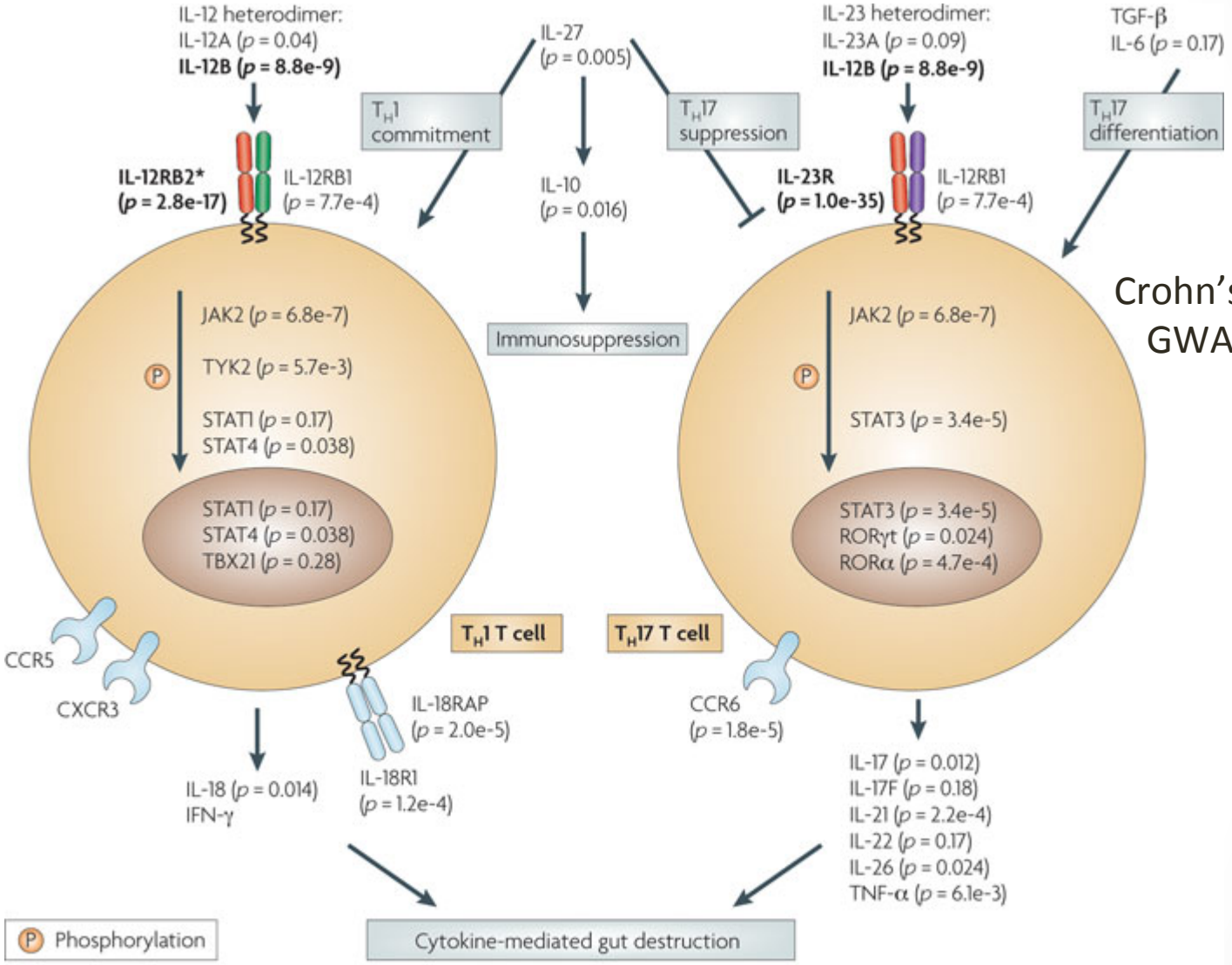
Networks/Pathways

- Why do I care about network/pathway context for my SNPs or SNVs?
  - If you have found a series of SNPs associated with an outcome, such as multiple sclerosis
    - You can identify genes those SNPs are in, near, or have a relationship to
    - Do you see all of these genes have protein products that interact in a biologically functional way?



Red/Yellow
Significantly Associated

Networks/Pathways



Crohn's Disease GWAS SNPs

Analysing biological pathways in genome-wide association studies
Nature Reviews Genetics 11, 843-854 (December 2010)

Networks/Pathways

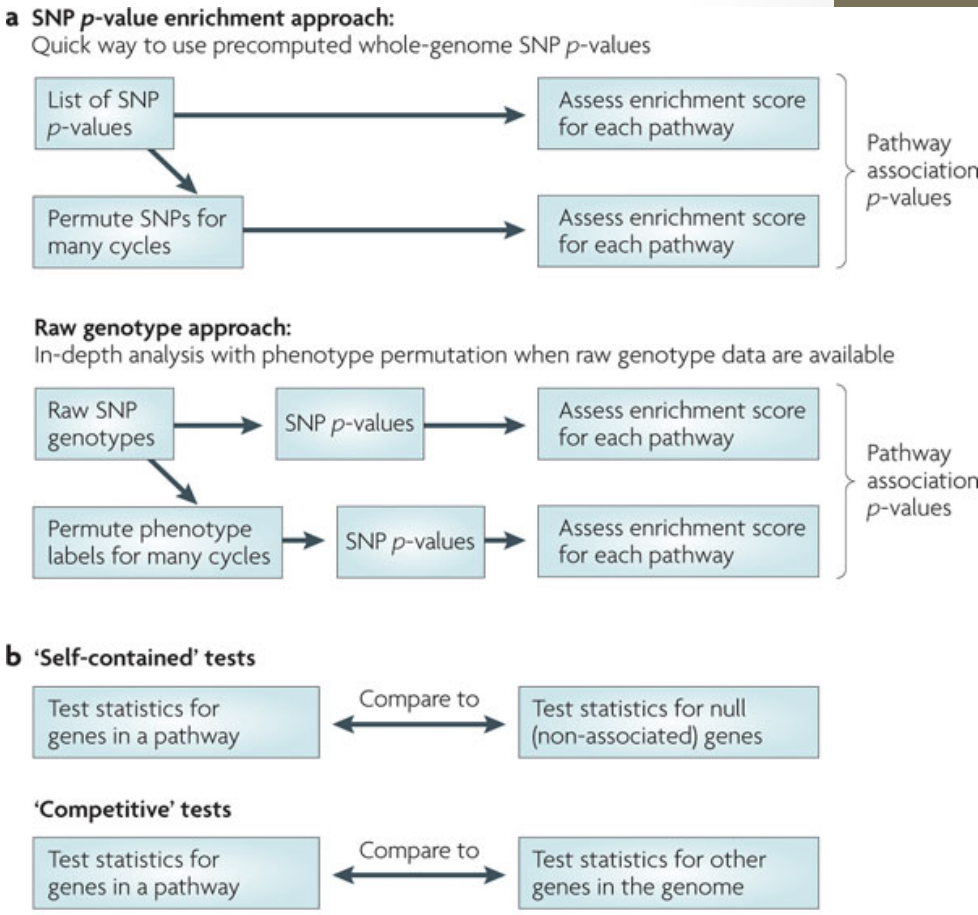
- Why do I care about network/pathway context for my SNPs or SNVs?
  - What if I have a series of genetic variants and
    - They are so low frequency I can't evaluate them one at a time?
    - Across cases/controls individually there is not a consistent pattern
      - No single genetic variant seems to be a "smoking gun"
    - Similarly: what if this disease seems to have different mutational contributions across multiple people?
      - Different mutations leading to the same type of cancer

Networks/Pathways

- Why do I care about network/pathway context for my SNPs or SNVs?
  - Evaluating genetic variation across a network/pathway can provide a clear pattern
    - Significant association for binned genetic variants at low frequency
    - Signals that are consistent for cases or controls when evaluating genetic variation across a specific pathway
      - Reduction of heterogeneity
    - Different mutations but they all affect small number of specific pathways
      - Drug targets (“block the door to the building”)

Pathway Based Approaches

- For GWAS data
- Basically divided into two methods
  - Providing SNP p-values
  - Providing SNP genotypes
  - There are many methods
    - So going over a few characteristics and rules of thumb
    - Have to carefully evaluate each method individually before use

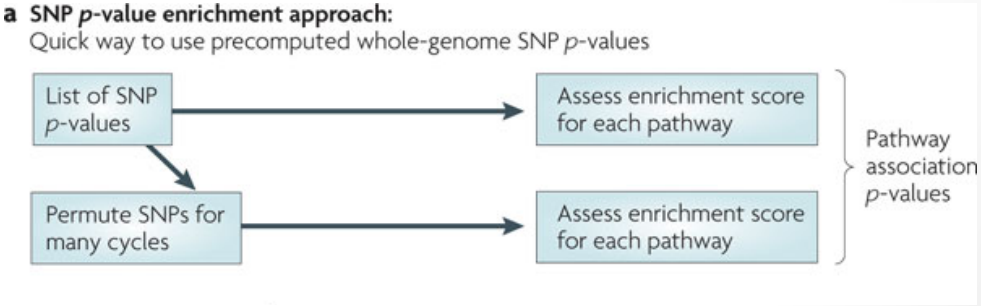


Analysing biological pathways in genome-wide association studies
Nature Reviews Genetics 11, 843-854 (December 2010)

Nature Reviews | Genetics

Pathway Based Approaches

- For GWAS data
- Basically divided into two methods
  - Providing SNP p-values
  - Determine whether a group of p-values for SNPs or genes is enriched for association signals
  - Choose a p-value cutoff for identifying significant SNPs for further analysis
  - Look out for gene-size, pathway-size biases
    - Lots of base pairs, more chances for your SNPs to be both significantly associated in GWAS and co-located
  - Look out for LD



Analysing biological pathways in genome-wide association studies
Nature Reviews Genetics 11, 843-854 (December 2010)

PARIS

- Pathway Analysis by Randomization Incorporating Structure (PARIS)
- For GWAS data
- Independent of study design and don't have to have the original dataset
- Provide p-values and choose your source of pathway information

Genetic analysis of biological pathway data through genomic randomization
Human Genetics
May 2011, Volume 129, Issue 5, pp 563-571

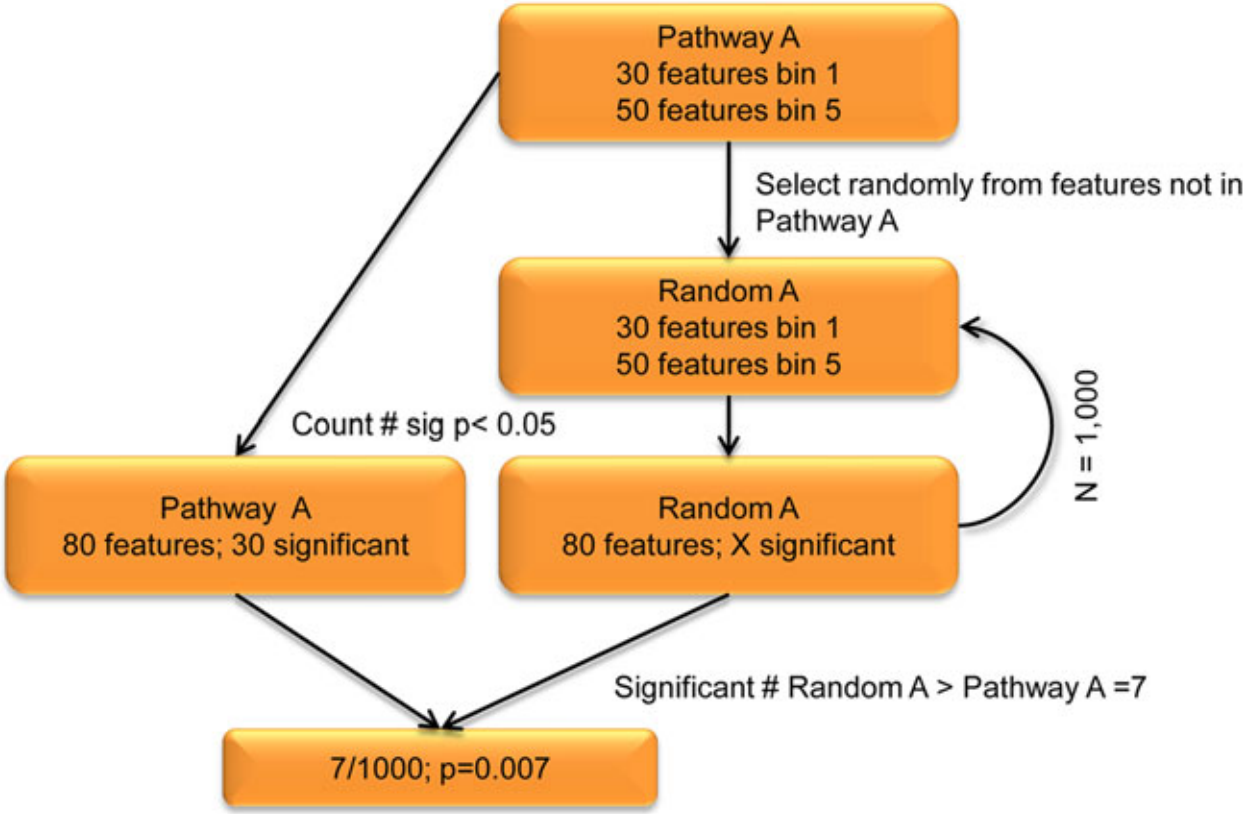
PARIS

- Pathway Analysis by Randomization Incorporating Structure (PARIS)
- First determines structure of the pathway being tested by taking LD into account
  - LD features overlapping pathway members
- Counting the number of significant p-value associations in a pathway
- PARIS creates randomized feature collections from the remainder of the genome that mimic the size and number of features of the actual pathway being tested

Genetic analysis of biological pathway data through genomic randomization
Human Genetics
May 2011, Volume 129, Issue 5, pp 563-571

PARIS

- Pathway Analysis by Randomization Incorporating Structure (PARIS)



Genetic analysis of biological pathway data through genomic randomization
Human Genetics
May 2011, Volume 129, Issue 5, pp 563-571

PARIS

- Pathway Analysis by Randomization Incorporating Structure (PARIS)
- For pathways of interest
  - Is it significant because of one gene with many significant features?
  - Or many genes contributing to the signal?
  - Assessment of contribution of each gene to the overall pathway signal
    - Permutation test based on features present in the single gene to also assign a p-value to each gene in the pathway
    - Identical to a PARIS pathway in which the pathway contains one gene

Genetic analysis of biological pathway data through genomic randomization
Human Genetics
May 2011, Volume 129, Issue 5, pp 563-571

PARIS

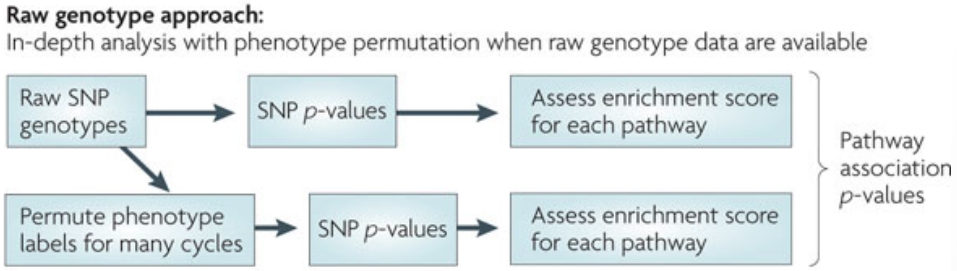
- Pathway Analysis by Randomization Incorporating Structure (PARIS)
- First example use
  - Used KEGG pathways
  - Autism GWAS dataset
- Revealed pathways with a significant enrichment of positive association results

Pathway name	Total SNP count	Description	p value	Gene count	Gene count $p < 0.05$
path:hsa00040	370	Pentose and glucuronate interconversions	<0.001	16	10
path:hsa04120	3,803	Ubiquitin mediated proteolysis	<0.001	133	62
path:hsa00072	219	Synthesis and degradation of ketone bodies	0.001	9	6
path:hsa00740	476	Riboflavin metabolism	0.001	16	6
path:hsa00053	458	Ascorbate and aldarate metabolism	0.008	16	10
path:hsa04060	5,911	Cytokine-cytokine receptor interaction	0.011	262	105
path:hsa04710	414	Circadian rhythm—mammal	0.011	13	8
path:hsa05211	2,323	Renal cell carcinoma	0.011	70	43
path:hsa05221	1,792	Acute myeloid leukemia	0.012	56	25
path:hsa00534	1,242	Heparan sulfate biosynthesis	0.014	26	18
path:hsa05220	2,558	Chronic myeloid leukemia	0.018	75	42
path:hsa04330	1,521	Notch signaling pathway	0.019	46	26
path:hsa00980	1,180	Metabolism of xenobiotics by cytochrome P450	0.02	58	21
path:hsa00480	1,021	Glutathione metabolism	0.03	47	24
path:hsa00860	808	Porphyrin and chlorophyll metabolism	0.037	32	19
path:hsa00760	763	Nicotinate and nicotinamide metabolism	0.039	24	14
path:hsa00730	307	Thiamine metabolism	0.048	8	3

Genetic analysis of biological pathway data through genomic randomization
Human Genetics
May 2011, Volume 129, Issue 5, pp 563-571

Pathway Based Approaches

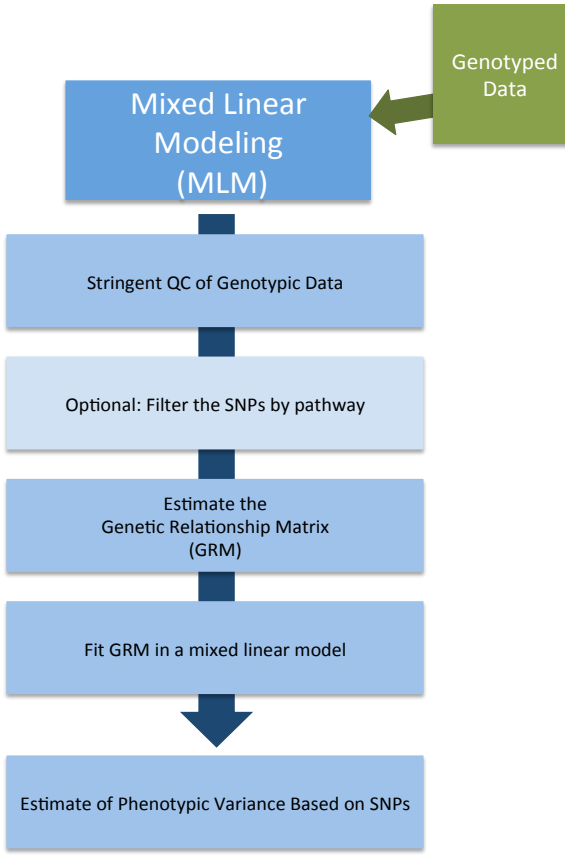
- For GWAS data
- Basically divided into two methods
  - Providing SNP genotypes
  - Gene-level and pathway-level test statistics
  - Can require just multi-markers
  - Some require single marker p-values and genotypic data



Analysing biological pathways in genome-wide association studies
Nature Reviews Genetics 11, 843-854 (December 2010)

Pathway Based Approaches

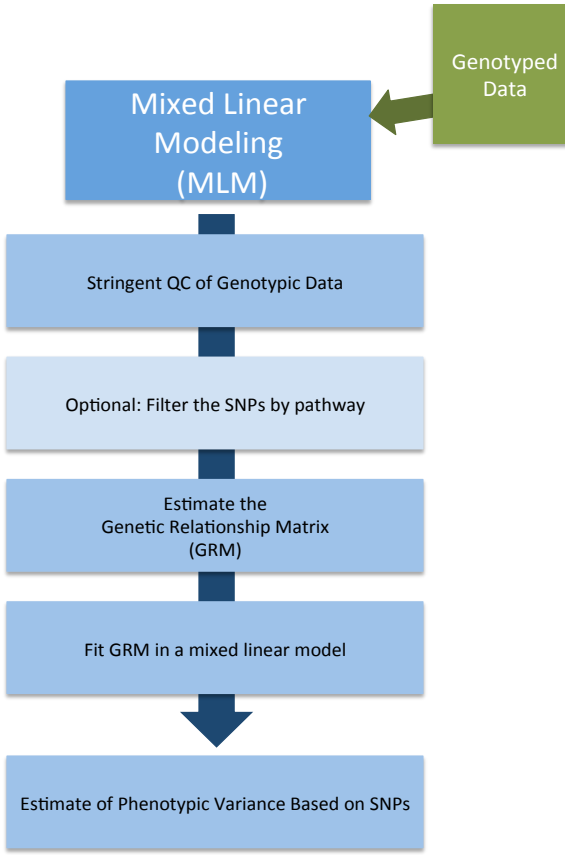
- Example of using genotypic data
- Polygenic etiology of paclitaxel-induced neuropathy
- Estimated the variance explained by common SNPs (MAF > 1%) for two outcomes
  - Maximum grade of sensory peripheral neuropathy
  - Dose at first instance of peripheral neuropathy



Polygenic inheritance of paclitaxel-induced sensory peripheral neuropathy driven by axon outgrowth gene sets in CALGB 40101 (Alliance)
Pharmacogenomics J. 2014 Aug;14(4):336-42. doi: 10.1038/tpj.2014.2. Epub 2014 Feb 11.

Pathway Based Approaches

- Example of using genotypic data
- Polygenic etiology of paclitaxel-induced neuropathy
  - Used the GCTA software tool
    - http://www.complextraitgenomics.com/software/gcta/
    - Mixed Linear Modeling
  - Axonogenesis GO Term set (GO: 0007409) had significant estimates of heritability close to 20%
    - Suggesting portion of the heritability of paclitaxel-induced neuropathy is driven by genes involved in the regulation of axon extension
    - Disruption of axon outgrowth may be one of the mechanisms by which paclitaxel treatment results in sensory peripheral neuropathy in susceptible patients



Polygenic inheritance of paclitaxel-induced sensory peripheral neuropathy driven by axon outgrowth gene sets in CALGB 40101 (Alliance)
Pharmacogenomics J. 2014 Aug;14(4):336-42. doi: 10.1038/tpj.2014.2. Epub 2014 Feb 11.

Pathway Based Approaches

- For GWAS data
- Important to remember
  - If there is an interplay with multiple genes in a pathway or across multiple pathways, these approaches could highlight this
  - Different sources have differences in data presented
  - However: one strongly associated gene in a pathway, or in multiple pathways, may make it seem like that pathway or pathways is VERY significant
    - Removing a very strong susceptibility gene or SNP can help in this case
    - Also take into account if there is extensive LD for a SNP with a strong relationship to an outcome trait

Analysing biological pathways in genome-wide association studies
Nature Reviews Genetics 11, 843-854 (December 2010)

Pathway Based Approaches

- Can get different answers via different methods and sources
- Worthwhile exploring different methods and contrasting/ comparing
  - Different sources used will have different data
  - Different properties of statistical tests
- The importance of replication in another dataset

Analysing biological pathways in genome-wide association studies
Nature Reviews Genetics 11, 843-854 (December 2010)

Always Remember

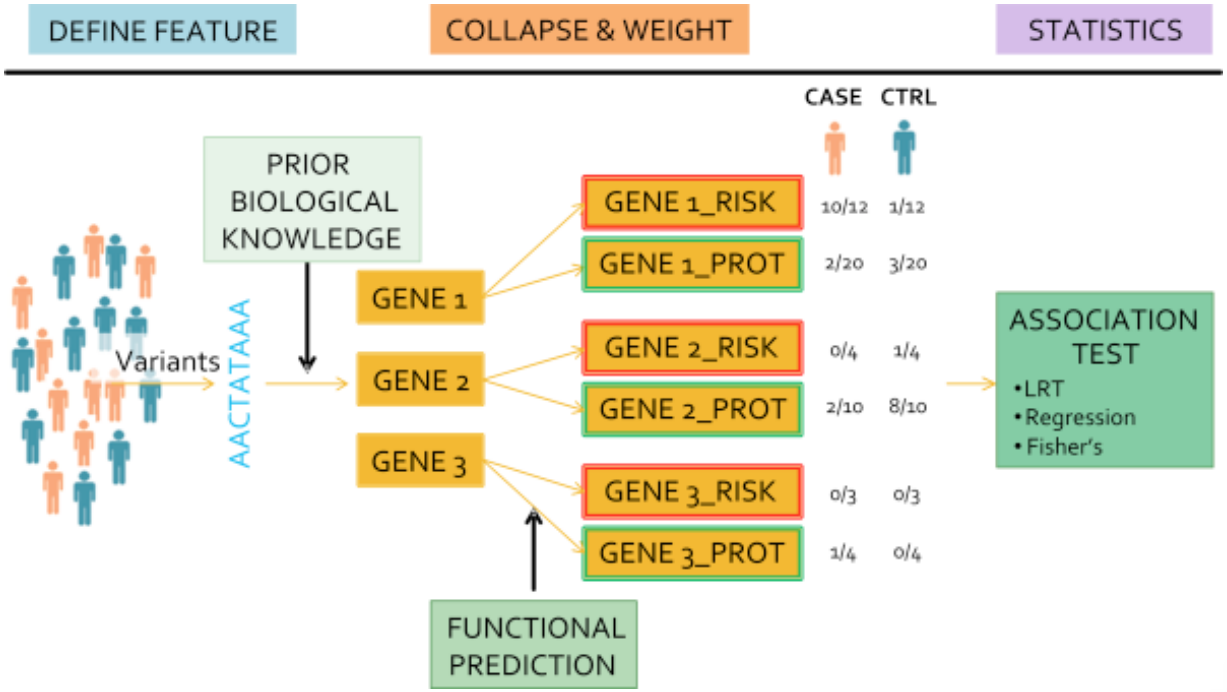
- We can't know what we don't know
  - When moving to these analyses we can only investigate known pathways or connections
    - We are still elucidating many biological networks, protein functions, and protein interactions
    - As a result, over time the results of your analyses can change
      - Always record WHEN you did an analysis
    - Your search space is defined by what is known and it affects your ability to do unbiased discovery
      - But on the other hand – we have to start somewhere right???
- There are also "de-novo" pathway based approaches
  - Direction for microarray analysis of genes…

Pathways: Rare Variant Approaches

- Low frequency variants
  - Regression not a good choice with so few individuals with these genetic variants
  - So how about binning variants?
    - Binning by gene
    - Binning by pathway

BioBin

- Low frequency variants
- BioBin is a novel method to collapse sequence data and detect disease associations using prior biological knowledge
- Enrichment for low frequency variants in your controls?



https://ritchielab.psu.edu/software/biobin-download

BioBin

- For analysis of whole-exome or whole-genome sequence data
- Does not rely on the selection of candidate genes
- Utilizes collapsing strategy as a means of reducing the search space
  - Enriches association signals
  - Reduces penalty of multiple testing
- Can be applied to case-control data
- Can prioritize bins using biological information
- Results can be used in a regression framework to test for association

BioBin: a bioinformatics tool for automating the binning of rare variants using publicly available biological knowledge.
BMC Med Genomics. 2013;6 Suppl 2:S6. doi: 10.1186/1755-8794-6-S2-S6. Epub 2013 May 7.

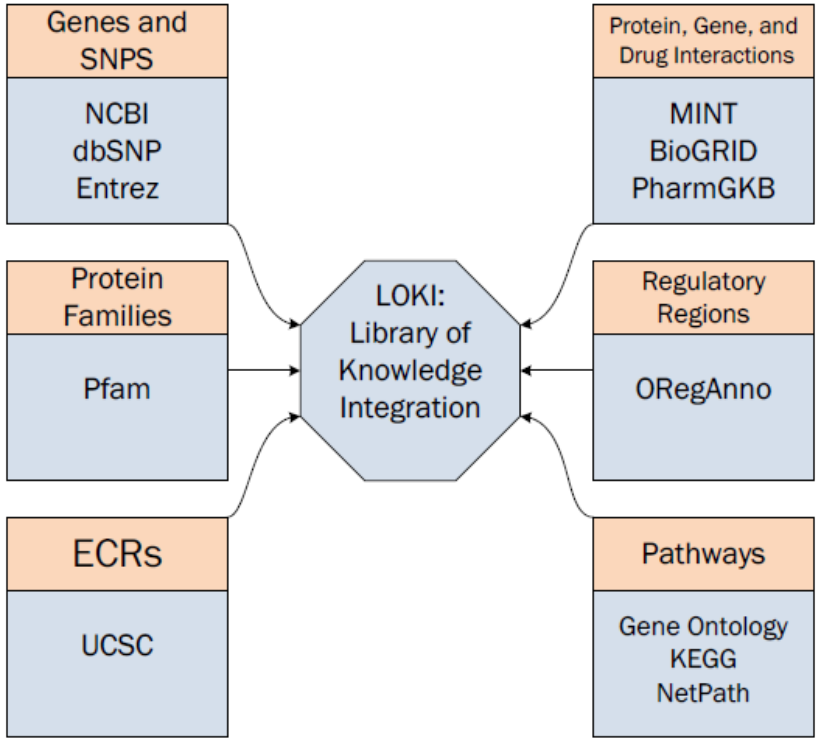
Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data.
PLoS Genet. 2013;9(12):e1003959. doi: 10.1371/journal.pgen.1003959. Epub 2013 Dec 26.

BioBin

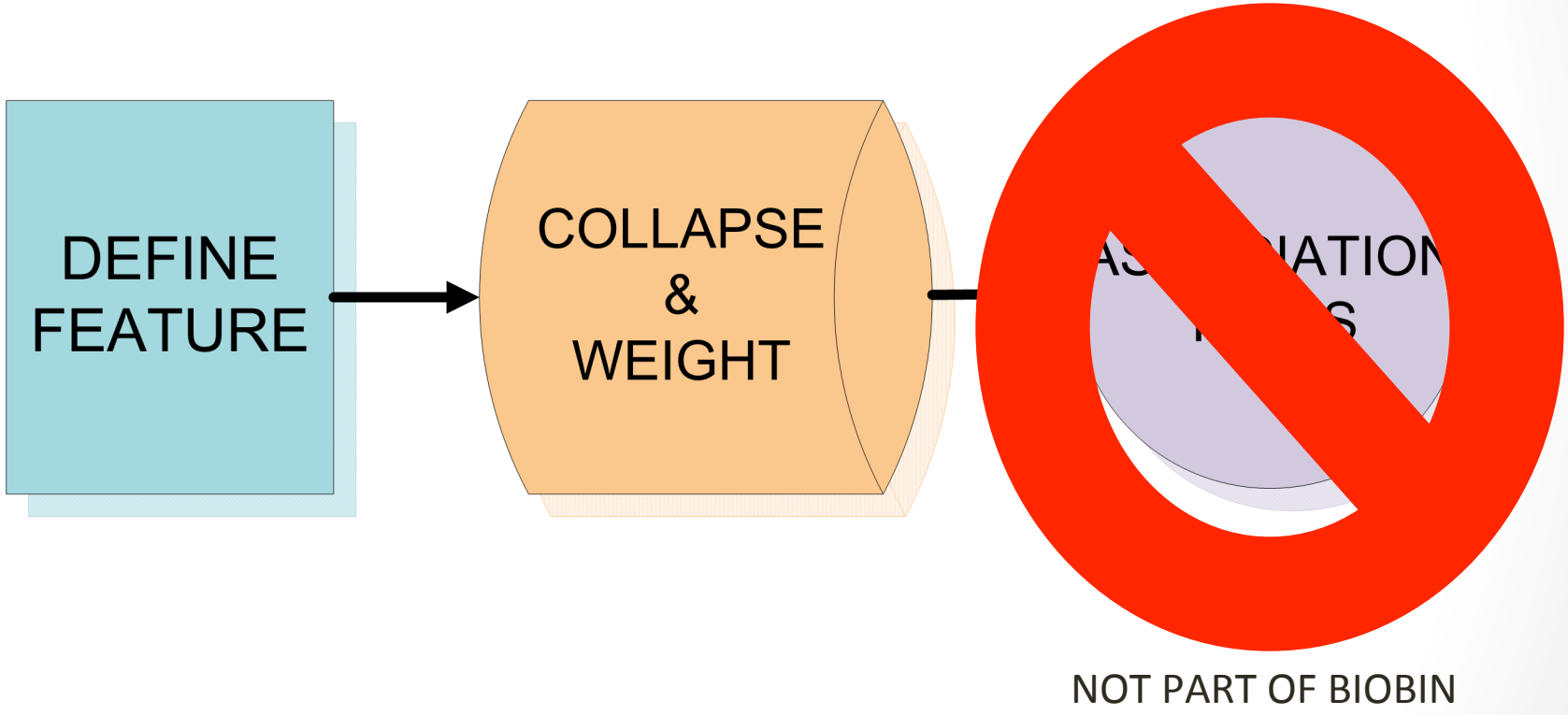
- Integrate collapsing method for rare variants using Biofilter
  - Include genetic information (pathway, gene boundaries, etc)
  - Create flexible binning structure
- Incorporate functional information
  - Bin variants by biological features
    - Gene
    - Intron
    - Exon
    - Intergenic
    - Pathway
    - Regulatory region
    - Evolutionary conserved region
    - Region of natural selection

BioBin

- Using Library of Knowledge integration (LOKI)



BioBin



BioBin

- Imagine you have a set of SNPs, specific bp locations:
  - * Denote variants with MAF below the threshold

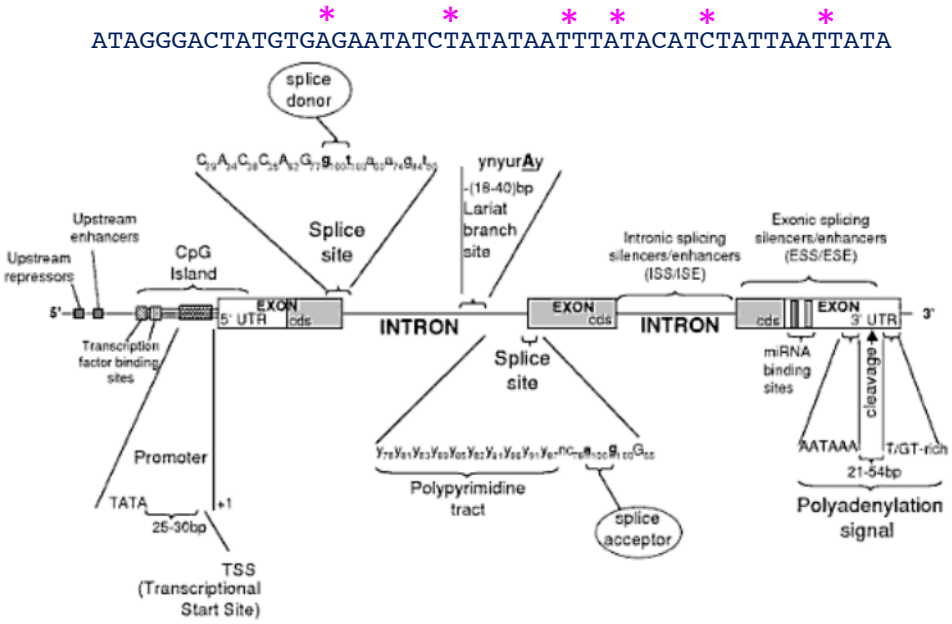


Figure 11.2 The anatomy of a gene. This figure illustrates some of the key regulatory regions that control the transcription, splicing and post-transcriptional processing of genes and transcripts. Polymorphisms in these regions should be investigated for functional effects

Plumpton and Barnes. "Predictive Functional Analysis of Polymorphisms: An Overview". Bioinformatics for Geneticists. Wiley, 2007.

Example: Bin by gene

- Bin genetic variants by genes
  * Denote variants with MAF below the threshold

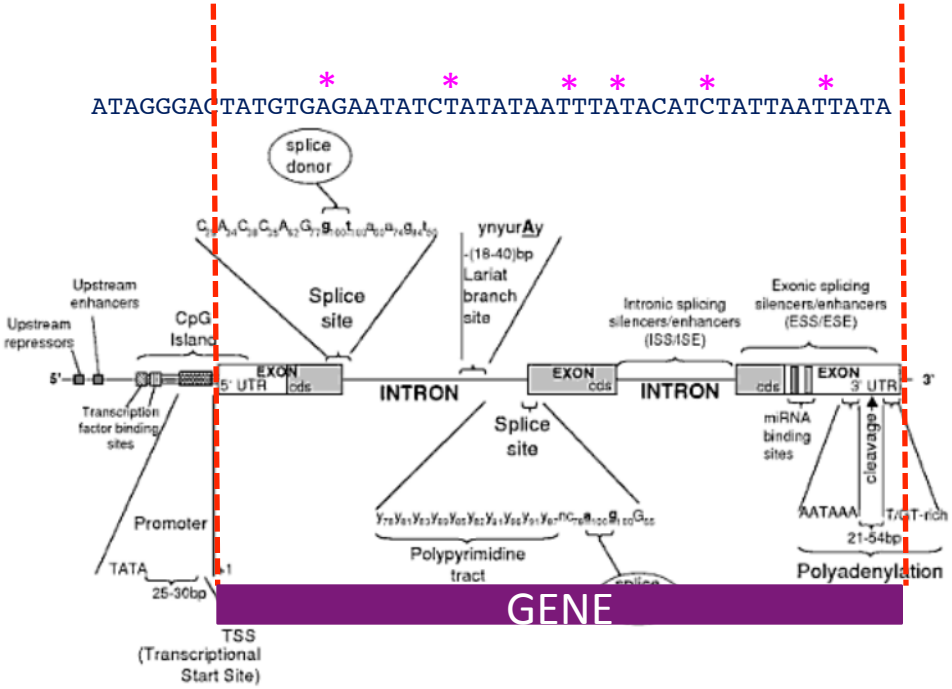


Figure 11.2 The anatomy of a gene. This figure illustrates some of the key regulatory regions that control the transcription, splicing and post-transcriptional processing of genes and transcripts. Polymorphisms in these regions should be investigated for functional effects

Plumpton and Barnes. "Predictive Functional Analysis of Polymorphisms: An Overview". Bioinformatics for Geneticists. Wiley, 2007.

Example: Bin by gene

- Bin genetic variants by genes
  - * Denote variants with MAF below the threshold

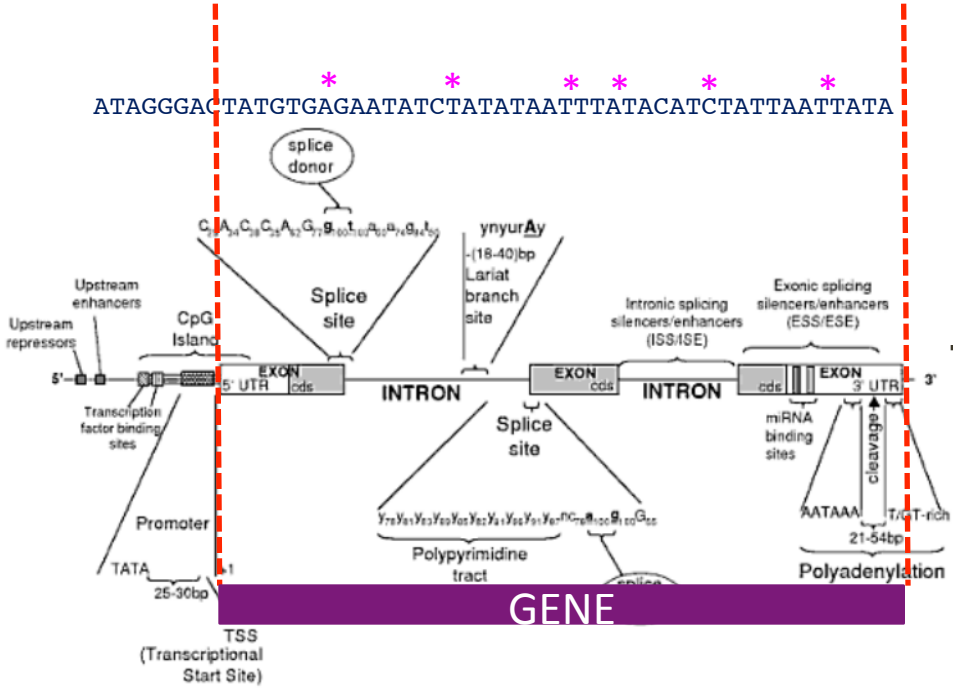


Figure 11.2 The anatomy of a gene. This figure illustrates some of the key regulatory regions that control the transcription, splicing and post-transcriptional processing of genes and transcripts. Polymorphisms in these regions should be investigated for functional effects

Plumpton and Barnes. "Predictive Functional Analysis of Polymorphisms: An Overview". Bioinformatics for Geneticists. Wiley, 2007.

Example: Bin by gene

- Now choose a statistical test: more rare variants in cases than controls for that bin?
  - * Denote variants with MAF below the threshold

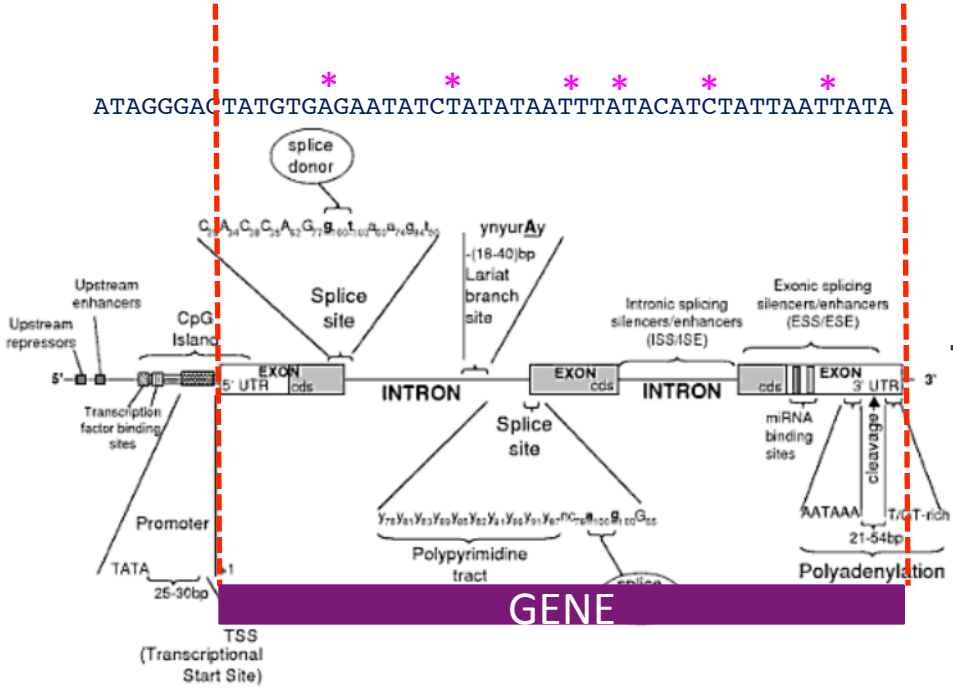
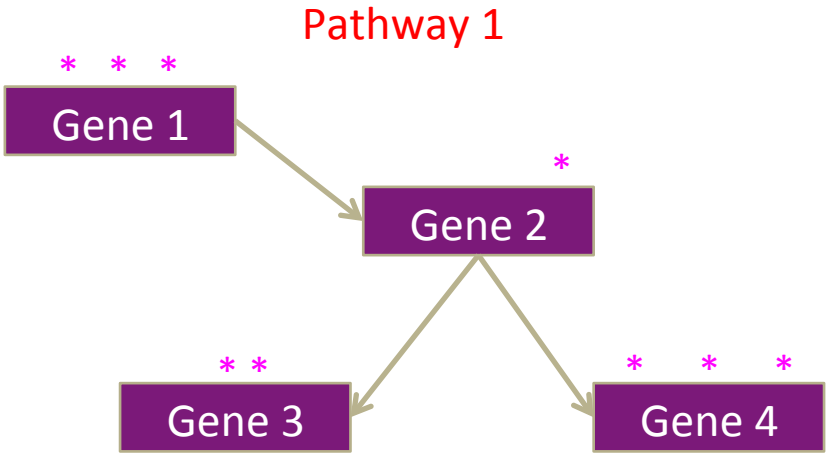


Figure 11.2 The anatomy of a gene. This figure illustrates some of the key regulatory regions that control the transcription, splicing and post-transcriptional processing of genes and transcripts. Polymorphisms in these regions should be investigated for functional effects

Plumpton and Barnes. "Predictive Functional Analysis of Polymorphisms: An Overview". Bioinformatics for Geneticists. Wiley, 2007.

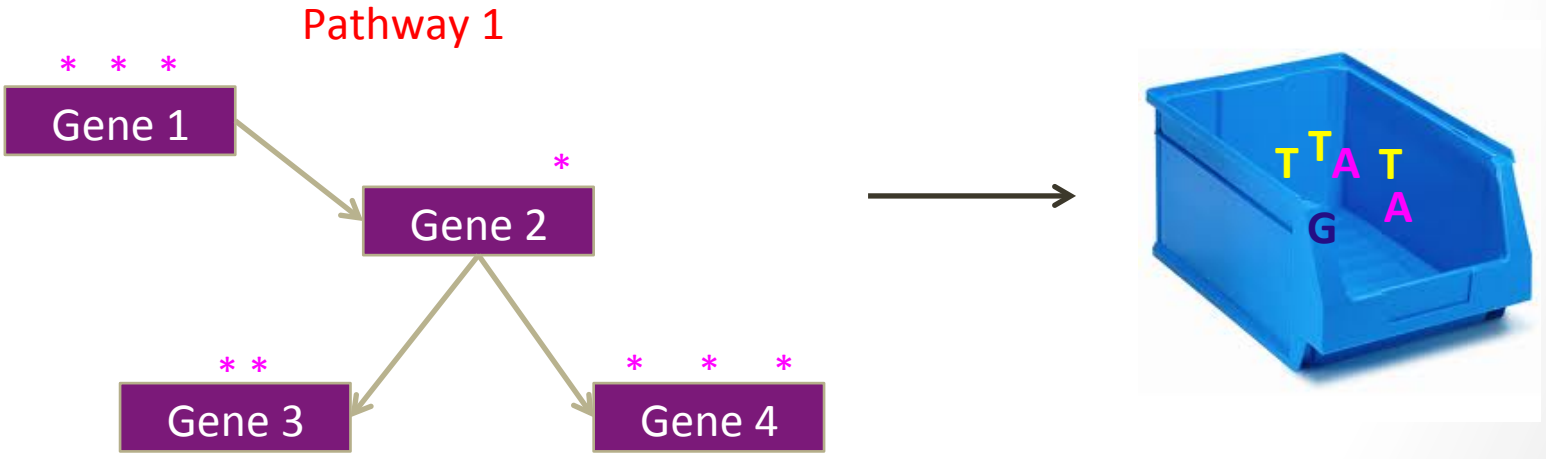
BioBin

- Ok so I can bin by gene…
  - But I thought we were talking about pathways?
- The same approach can be used to bin variants by pathways
- So you could use a pathway source in LOKI
  - So for all genes in a pathway, bin all of the variants together



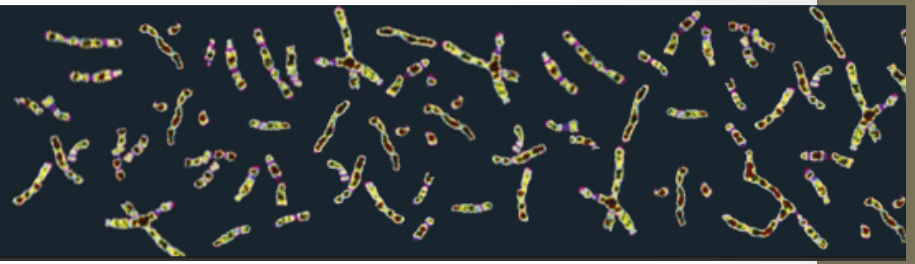
BioBin

- Ok so I can bin by gene…
  - But I thought we were talking about pathways?
- The same approach can be used to bin variants by pathways
- So you could use a pathway source in LOKI
  - So for all genes in a pathway, bin all of the variants together
  - Are there a statistically significant number of rare variants for cases vs. controls in this pathway?



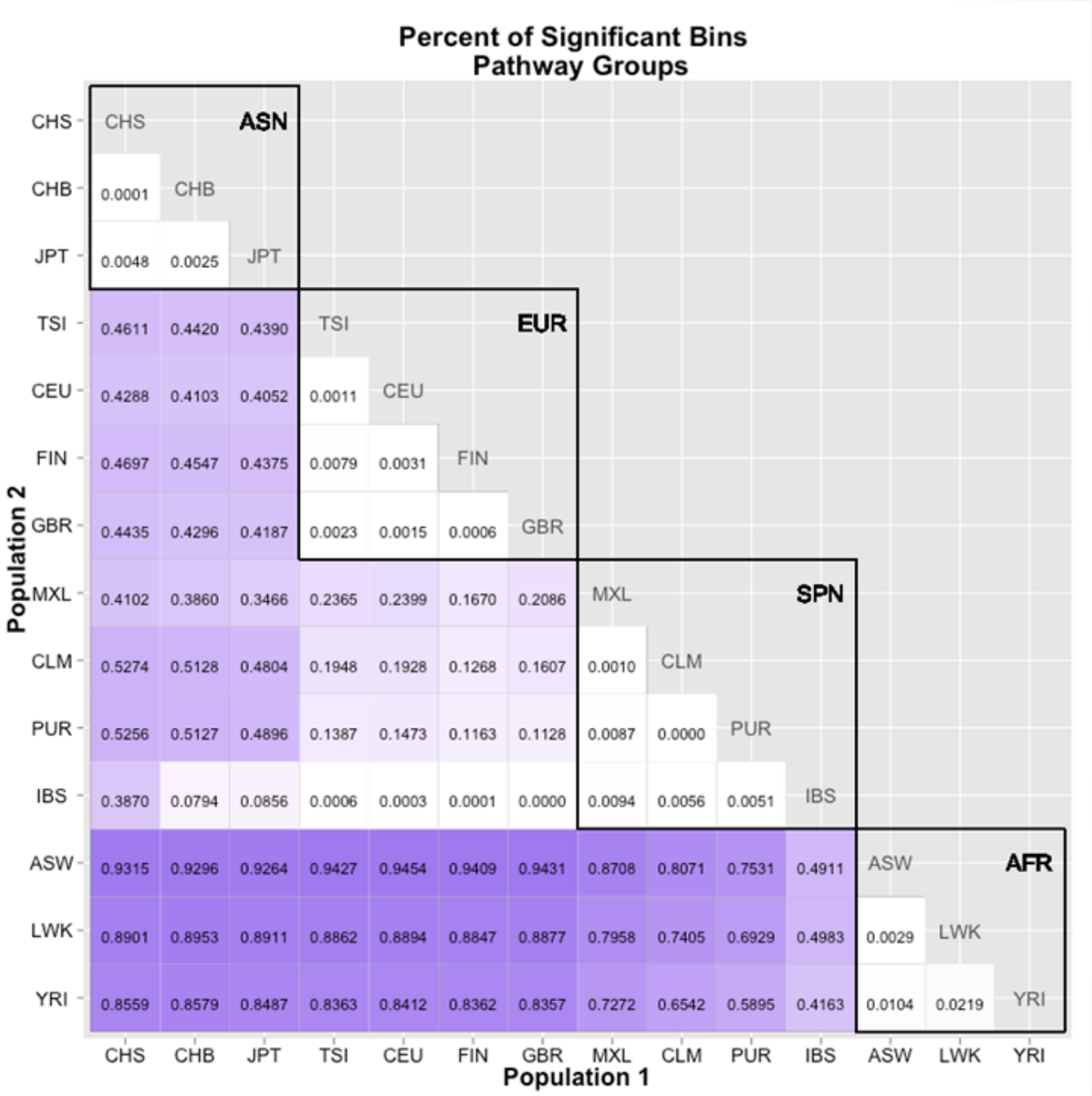
1000 Genomes

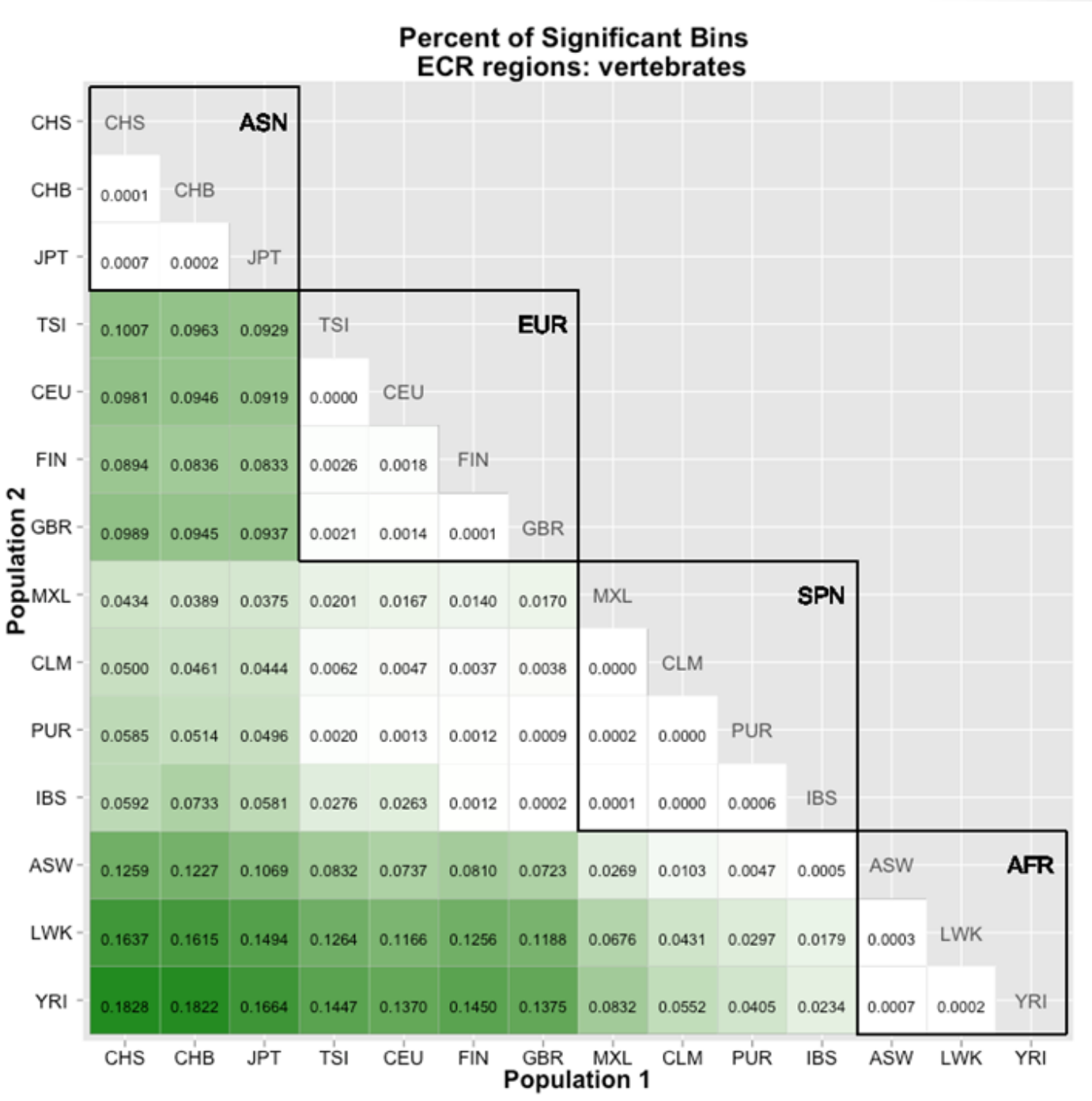
A Deep Catalog of Human Genetic Variation



- Moore *et al.* for proof of principle evaluated rare variant differences between 1000 Genomes populations

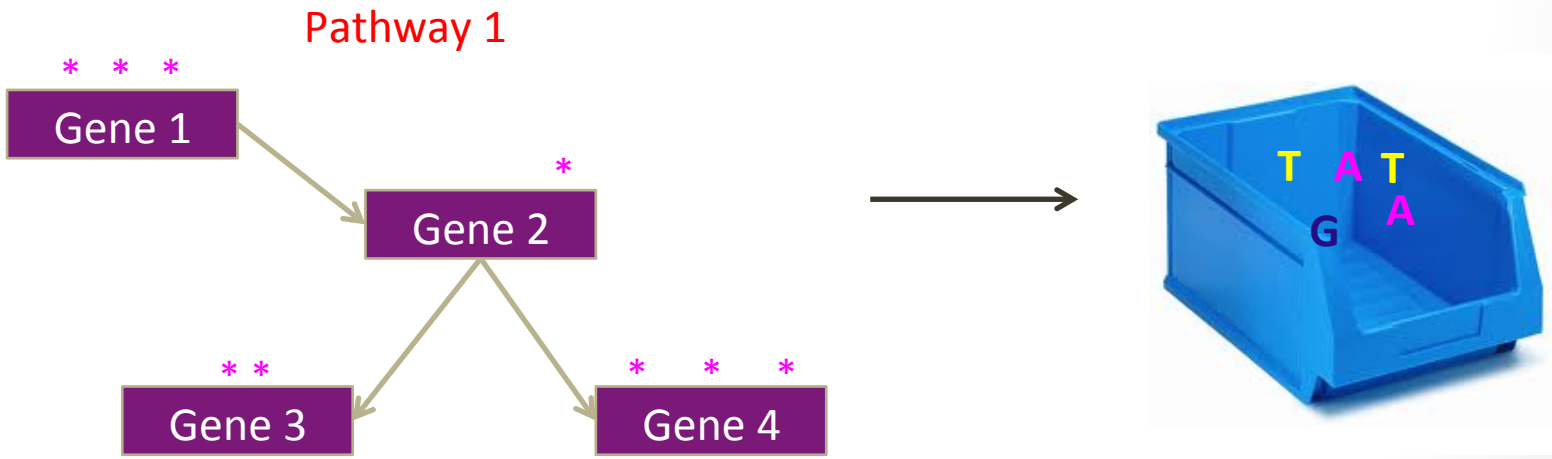
POP	N	VARIANTS	POPULATION
ASW	61	18819173	HapMap African ancestry individuals from SW US
CEU	87	11198921	CEPH individuals
CHB	97	10566371	Han Chinese in Beijing
CHS	100	10547019	Han Chinese South
CLM	60	13869201	Colombian in Medellin, Colombia
FIN	93	11005104	HapMap Finnish individuals from Finland
GBR	88	11388832	British individuals from England and Scotland
IBS	14	8424366	Iberian populations in Spain
JPT	89	10368186	Japanese individuals
LWK	97	19936728	Luhya individuals
MXL	66	12929352	HapMap Mexican individuals from LA California
PUR	55	14066653	Puerto Rican in Puerto Rico
TSI	98	11858607	Toscan individuals
YRI	88	18022152	Yoruba individuals


Percent of Significant Bins
Pathway Groups
ASN
EUR
SPN
AFR
CHS
CHB
JPT
TSI
CEU
FIN
GBR
MXL
CLM
PUR
IBS
ASW
LWK
YRI
0.0001
0.0048 0.0025
0.4611 0.4420 0.4390
0.4288 0.4103 0.4052 0.0011
0.4697 0.4547 0.4375 0.0079 0.0031
0.4435 0.4296 0.4187 0.0023 0.0015 0.0006
0.4102 0.3860 0.3466 0.2365 0.2399 0.1670 0.2086
0.5274 0.5128 0.4804 0.1948 0.1928 0.1268 0.1607 0.0010
0.5256 0.5127 0.4896 0.1387 0.1473 0.1163 0.1128 0.0087 0.0000
0.3870 0.0794 0.0856 0.0006 0.0003 0.0001 0.0000 0.0094 0.0056 0.0051
0.9315 0.9296 0.9264 0.9427 0.9454 0.9409 0.9431 0.8708 0.8071 0.7531 0.4911
0.8901 0.8953 0.8911 0.8862 0.8894 0.8847 0.8877 0.7958 0.7405 0.6929 0.4983 0.0029
0.8559 0.8579 0.8487 0.8363 0.8412 0.8362 0.8357 0.7272 0.6542 0.5895 0.4163 0.0104 0.0219
Population 2
Population 1


Percent of Significant Bins
ECR regions: vertebrates
ASN
EUR
SPN
AFR
Population 2
Population 1
CHS CHB JPT TSI CEU FIN GBR MXL CLM PUR IBS ASW LWK YRI
0.0001
0.0007 0.0002
0.1007 0.0963 0.0929
0.0981 0.0946 0.0919 0.0000
0.0894 0.0836 0.0833 0.0026 0.0018
0.0989 0.0945 0.0937 0.0021 0.0014 0.0001
0.0434 0.0389 0.0375 0.0201 0.0167 0.0140 0.0170
0.0500 0.0461 0.0444 0.0062 0.0047 0.0037 0.0038 0.0000
0.0585 0.0514 0.0496 0.0020 0.0013 0.0012 0.0009 0.0002 0.0000 0.0006
0.0592 0.0733 0.0581 0.0276 0.0263 0.0012 0.0002 0.0001 0.0000
0.1259 0.1227 0.1069 0.0832 0.0737 0.0810 0.0723 0.0269 0.0103 0.0047 0.0005
0.1637 0.1615 0.1494 0.1264 0.1166 0.1256 0.1188 0.0676 0.0431 0.0297 0.0179 0.0003
0.1828 0.1822 0.1664 0.1447 0.1370 0.1450 0.1375 0.0832 0.0552 0.0405 0.0234 0.0007 0.0002

Rare Variants

- Direction of effect
  - What if the rare variant has a protective effect or a risk effect?
  - You could have cases that have many rare variants that contribute to protection, and vice versa
    - Do you loose signal because you are just counting variants?
  - Dispersion methods
    - No assumption of burden tests of the same direction of effect of all rare variants on the trait within the same functional unit or genomic region



Rare Variants

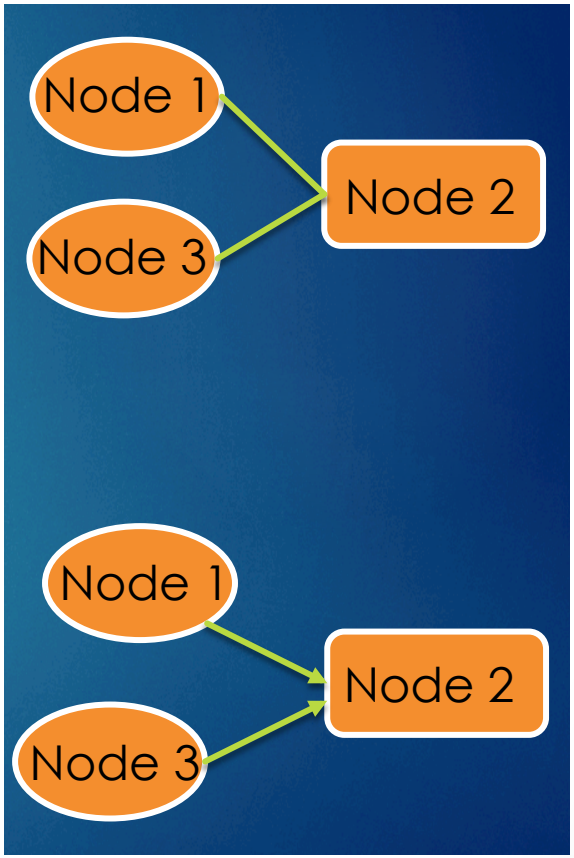
- Direction of effect
  - Dispersion methods
    - No assumption of burden tests of the same direction of effect of all rare variants on the trait within the same functional unit or genomic region
  - SKAT Method
    - SNP-set (Sequence) Kernel Association Test (SKAT)
    - http://www.hsph.harvard.edu/skat/
    - Gene or a region level test for association between a set of rare (or common) variants and dichotomous or quantitative phenotypes
    - SKAT aggregates individual score test statistics of SNPs in a SNP set
      - Computing SNP-set level p-values (gene or a region level p-value)
      - Adjustments can be made for covariates, such as principal components to account for population stratification

Networks

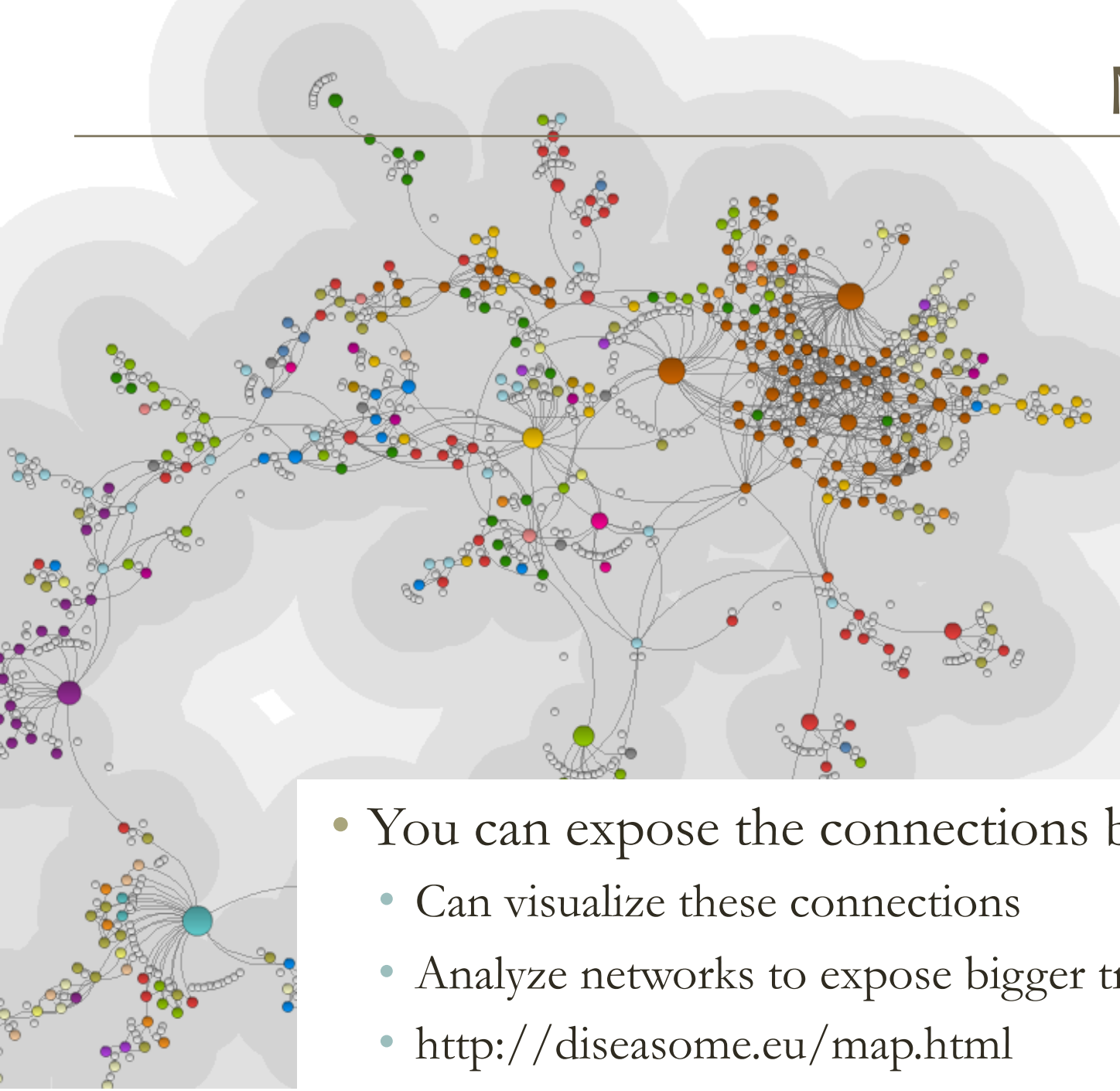
- Remember how biological data is inherently connected?
- I found a series of SNPs in different genes
- I identified that these genes are in shared KEGG pathways
  - Can I visualize this information?
- Cytoscape and Gephi are two free software packages that allow for network visualization

Networks

- Vocabulary
  - Nodes
  - Edges
- Directionality is when you have nodes that depend on other nodes
  - Target depends on source
  - Frequent in pathways that require one step then another

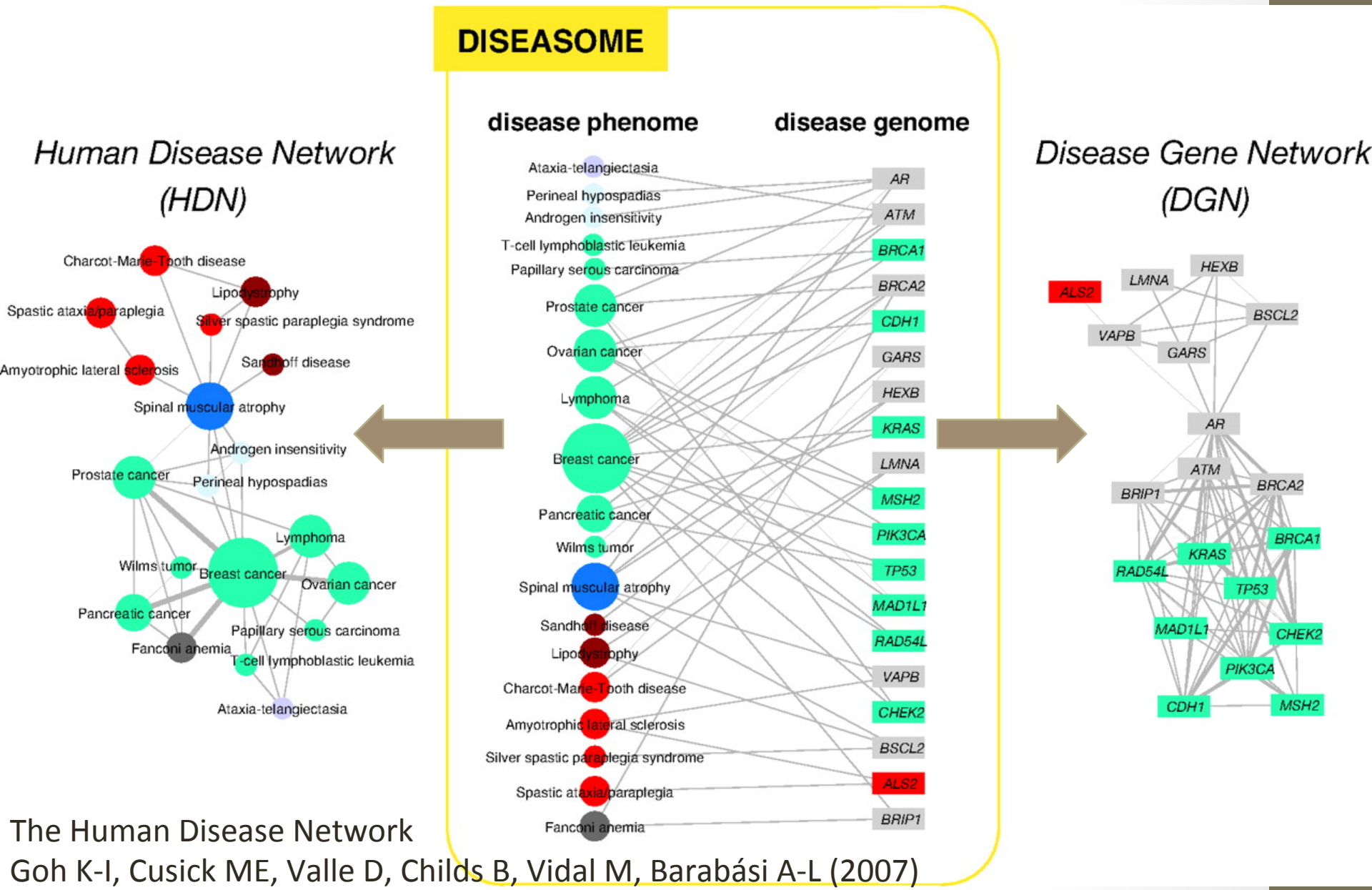


Networks



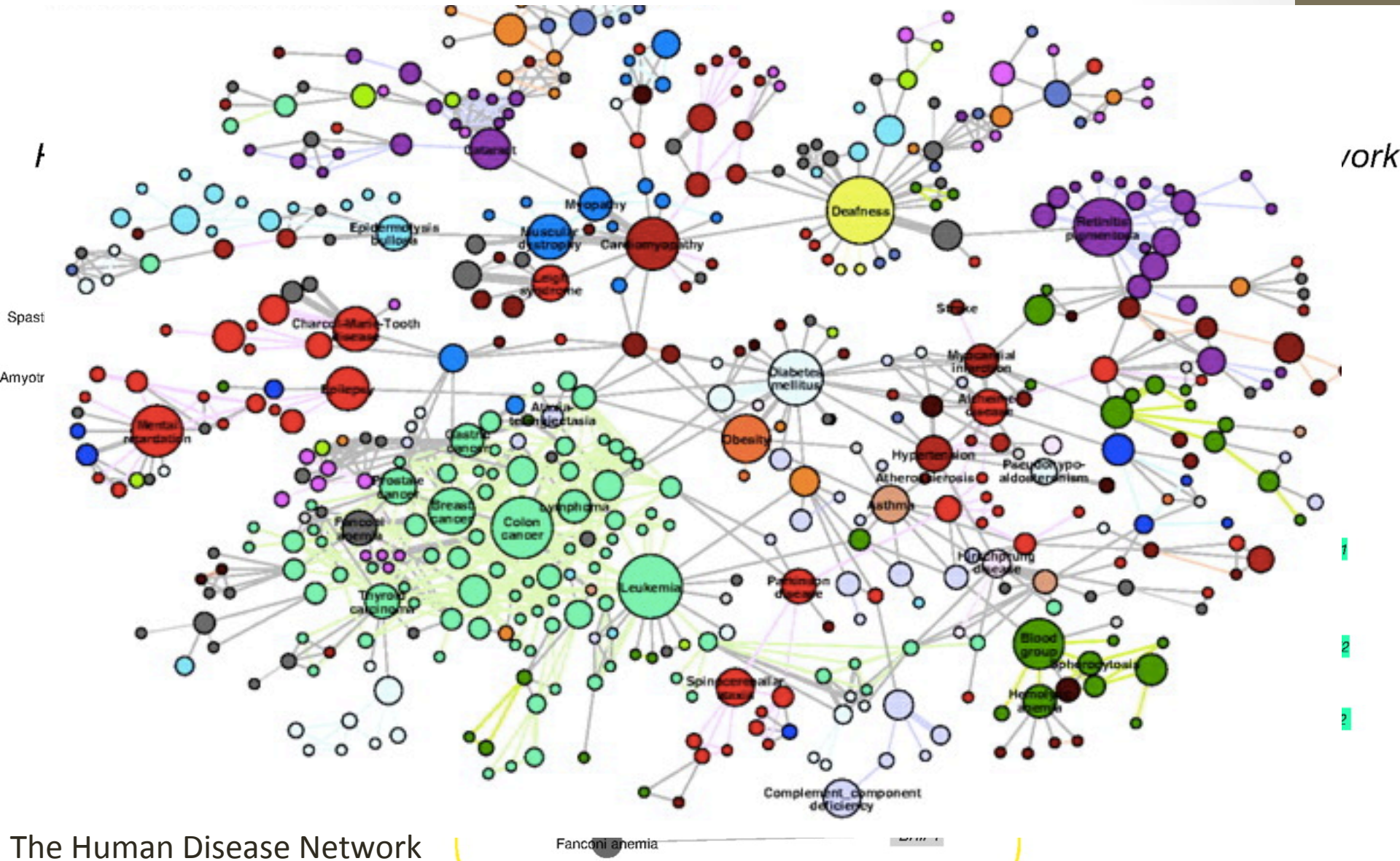
- You can expose the connections between data
  - Can visualize these connections
  - Analyze networks to expose bigger trends
  - http://diseasome.eu/map.html

Networks



The Human Disease Network
Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L (2007)

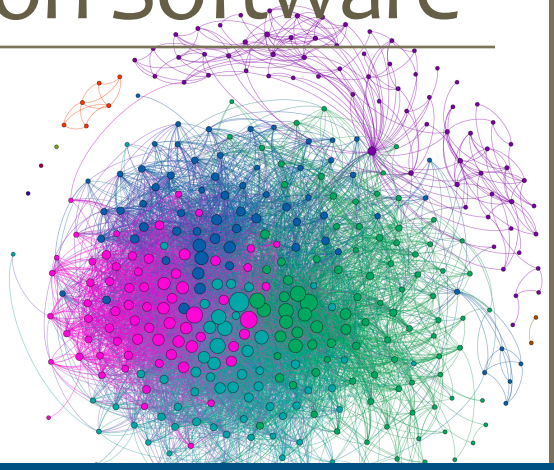
Networks



The Human Disease Network
Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L (2007)

Data Visualization Software

- Networks
  - Cytoscape and Gephi
  - http://www.cytoscape.org/
  - https://gephi.github.io/



Cytoscape

Network Data Integration, Analysis, and Visualization in a Box

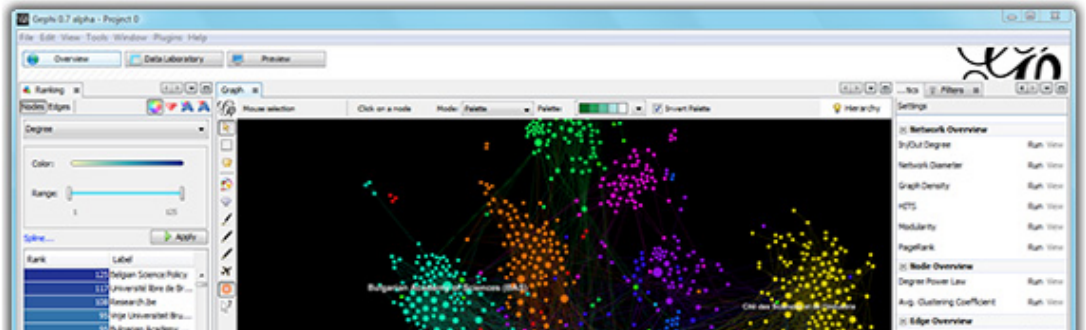
Gephi
makes graphs handy

Home Features Learn Develop Plugins Services Consortium

The Open Graph Viz Platform

Gephi is an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs.

Runs on Windows, Linux and Mac OS X. Gephi is open-source and free.



Questions?