Genetic Variation, Biological Pathways, and Networks

Sarah Pendergrass

Center for Systems Genomics

Outline

- What about functional annotation?
  - Proteins
- 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?

Why Important

- Protein function prediction
- These methods are important for identifying more information about protein coding regions
- This information can be leveraged for understanding more about the effect of individual SNVs on proteins
  - Databases and tools that combine protein information with the potential impact of genetic variation on proteins

Homology Based Tools

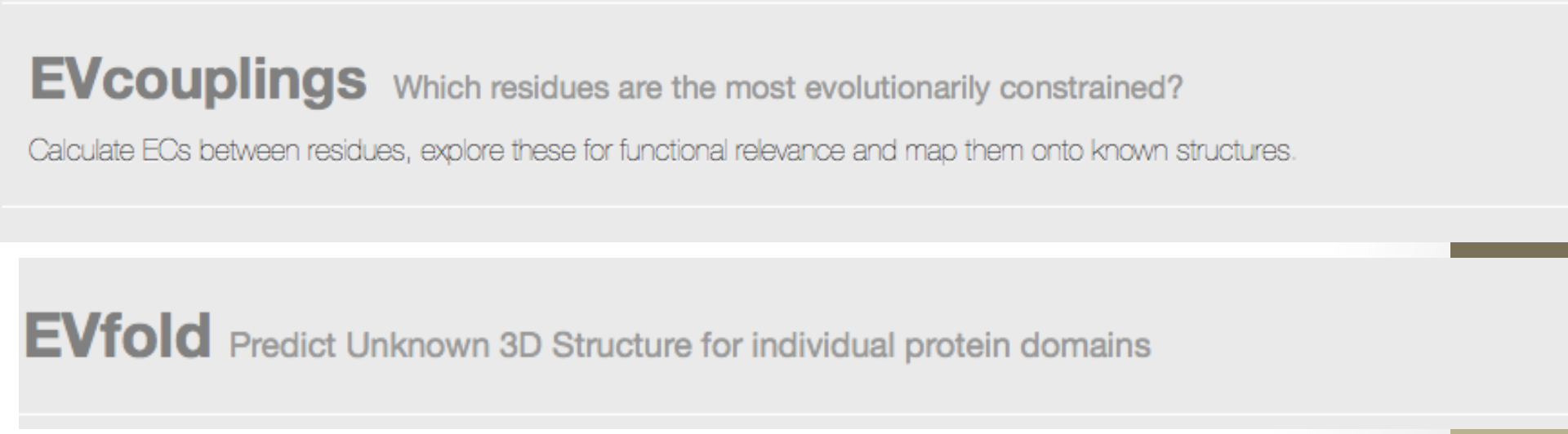
- Identify what a protein will do from a potential protein sequence
- Sequence similarity commonly done with software like BLAST
  - Aligning a pair of sequences
- P-fam can be used
  - Database of conserved protein domain families to annotate and classify proteins
- Multiple sequence alignments and protein 3-D structures can be combined for analyses

Non-Homology Based Tools

- We know many more protein sequences than protein three-dimensional structures
  - Sequencing data identifying coding regions at a fast rate
  - 40% of known human genes don't have functional classification by sequence similarity
- CHALLENGE
  - The homology based approach is not always possible
- Many proteins contain enough information in their amino acid sequence to determine their three-dimensional structure
  - Non-homology based tools

Non-Homology Based Tools

- Evolutionary Couplings
- Evfold.org
- Hundreds of sequences are needed to derive plausible causative evolutionary couplings
  - Primary limitation



Protein structure prediction from sequence variation
Nature Biotechnology 30, 1072–1080 (2012)

Novel Missense Variants

- I have SNPs or SNVS
  - Perhaps sequencing data or whole-exome, or genotype array data
- Non-synonymous protein coding SNVs may have the greatest impact on phenotypic variation
  - Excess of rare alleles among those predicted to be functional for proteins
- So for my SNPs or SNVs
  - How can I identify if they will have an impact on a protein?

Novel Missense Variants

- Why would I do try these?
- Protein information can be used to prioritize study of specific SNPs/ SNVs
- Can help identify key SNPs or SNVs for further study after identifying them in association testing
  - How to sort through hundreds of results
- Identification of genes that can then be migrated to gene based testing
- Easier translation to molecular/cellular assay

There are many tools….

Single nucleotide variations: Biological impact and theoretical interpretation

Panagiotis Katsonis,[1] Amanda Koire,[2] Stephen Joseph Wilson,[3] Teng-Kuei Hsu,[3] Rhonald C Lua,[1] Angela Dawn Wilkins,[1,4] and Olivier Lichtarge[1,2,3,4,5,*]

SNP Impact Predictors

Server	Year	Input	URL	Pubmed ID
Structural				
SDM	1997	PDB ID	http://www-cryst.bioc.cam.ac.uk/~sdm/sdm.php	9051729
Dmutant	2002	PDB ID	http://sparks.informatics.iupui.edu/hzhou/mutation.html(Unavailable)	12381853
PoPMuSiC	2009	PDB ID	http://dezyme.com/	19654118
SDS	2014	-	Cannot automate	24795746
Homology				
SIFT	2001	Protein identifier, SNP IDs, or alignment	http://sift.jcvi.org/	11337480
Panther	2003	Sequence	http://www.pantherdb.org/tools/csnpScoreForm.jsp	12952881
MAPP	2005	Alignment and phylogenetic tree	http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html	15965030
A-GVGD	2006	Alignment	http://agvgd.iarc.fr/agvgd_input.php	16014699
mutationassessor (xvar)	2011	Protein identifier or chrom. location	http://mutationassessor.org/	21727090
Provean	2012	Sequence or chrom. location	http://provean.jcvi.org/index.php	23056405
Evolutionary action	2014	Protein identifier	http://mammoth.bcm.tmc.edu/EvolutionaryAction/	
Hybrid				
PolyPhen	2002	Protein identifier or sequence	http://genetics.bwh.harvard.edu/pph/	12202775
LogR.E-value	2004	Site is down for maintenance	http://lpgws.nci.nih.gov/cgi-bin/GeneViewer.cg	14751981
nsSNPAnalyzer	2005	Sequence (requires available PDB structure)	http://snpanalyzer.uthsc.edu/	15980516
SNPeffect	2005	Sequence, PDB ID, UniProt ID	http://snpeffect.switchlab.org/menu	15608254
LS-SNP	2005	SNP, protein or pathway identifier	http://modbase.compbio.ucsf.edu/LS-SNP/	15827081
MUpro	2005	Protein sequence, structure (optional)	mupro.proteomics.ics.uci.edu	16372356
prost	2005	Sequence (on demand version) or PDB ID (precalculated version)	http://mmb2.pcb.ub.es:8080/PMut/	15879453
PhD-SNP	2006	Protein identifier or sequence	http://snps.biofold.org/phd-snp/phd-snp.html	16895930
SNPs3D	2006	SNP identifier	http://www.snps3d.org/	16551372
Parepro	2007	Alignment	http://www.mobioinfor.cn/parepro/index.htm	18005451
SAPRED	2007	Sequence and PDB files	http://sapred.cbi.pku.edu.cn/ (Login required)	17384424
Imutant 3.0	2007	Sequence or PDB ID	http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi	18387208
SNAP	2007	Sequence	http://rostlab.org/services/snap/submit	17526529
AUTO-MUTE	2010	PDB ID	http://proteins.gmu.edu/automute/AUTO-MUTE_nsSNPs.html	20573719
Mutation Taster	2010	Transcript, gene, or ORF	http://www.mutationtaster.org	20676075
PolyPhen2	2010	Protein or SNP identifier or sequence	http://genetics.bwh.harvard.edu/pph2/	20354512
Condel	2011	Protein identifier, mutation, homology tree	No server, but can get PERL pipeline scripts and then download each tool	21457909
CADD	2014	VCF file	http://cadd.gs.washington.edu/score	24487276
VarMod	2014	Sequence	http://www.wasslab.org/varmod/	24906884
SuSPect	2014	Sequence or VCF	http://www.sbg.bio.ic.ac.uk/suspect/index.html	24810707

PolyPhen

- Polymorphism Phenotyping (PolyPhen)
- Predicts possible impact of an amino acid substitution on the structure and function of a human protein
- http://genetics.bwh.harvard.edu/pph2/

PolyPhen

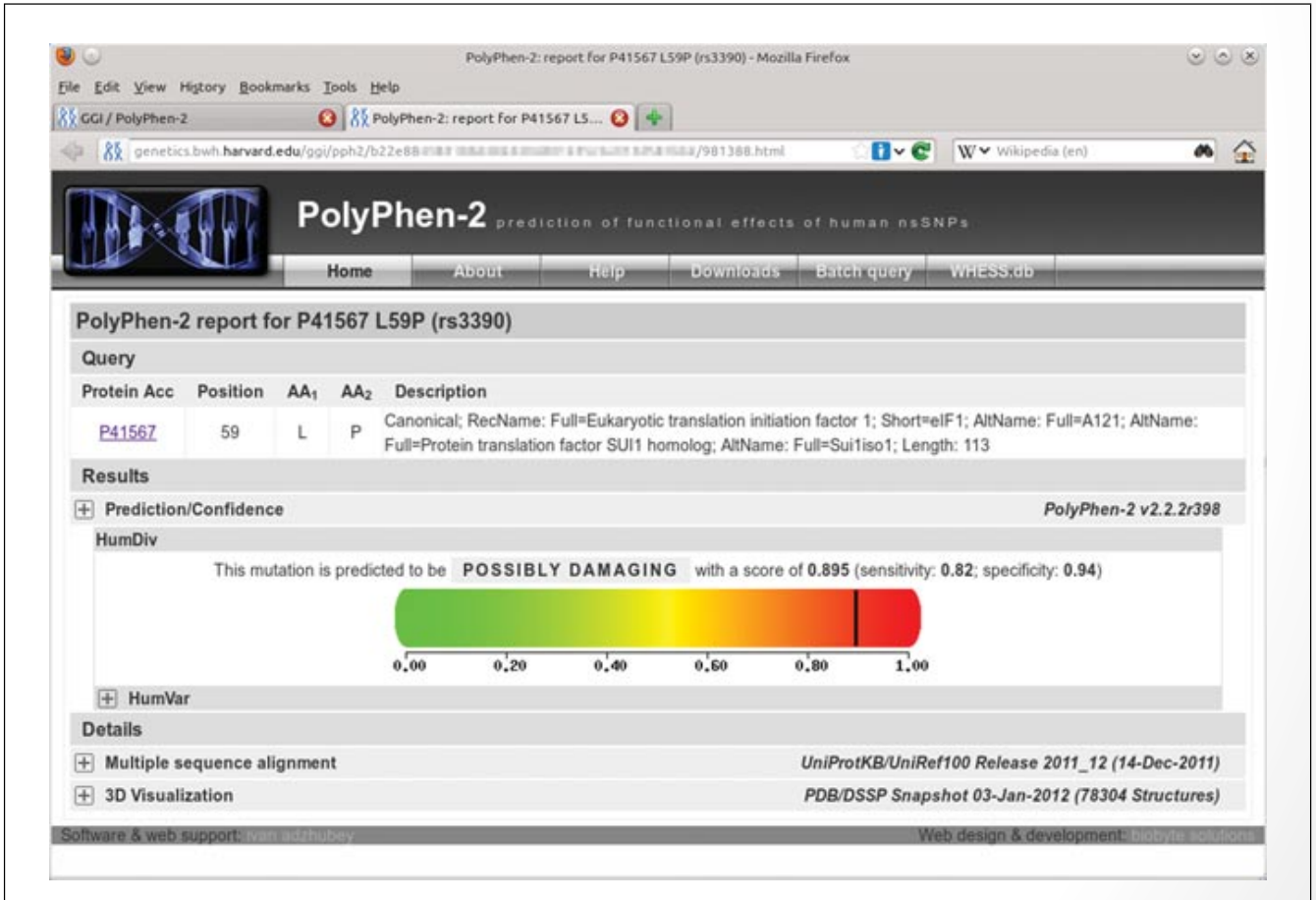
- Probabalistic classifier
- Prediction is based on a number of sequence, phylogenetic, and structural features characterizing the substitution
  - Maps coding SNPs to gene transcripts
  - Extracts protein sequence annotations and structural attributes
  - Builds conservation profiles
  - Then estimates the probability of the missense mutation being damaging based on a combination of all these properties

PolyPhen

- Can use the web interface
  - Provide protein identifier (UniProtKB assession number of entry name)
  - Enter position of the substitution in the protein sequence into the position text box
  - Indicate reference amino acid residue, and the substitution residue
  - Can also analyze large datasets of single nucleotide changes using batch mode
    - Get a web page with download options after the batch runs
- Also command line capability

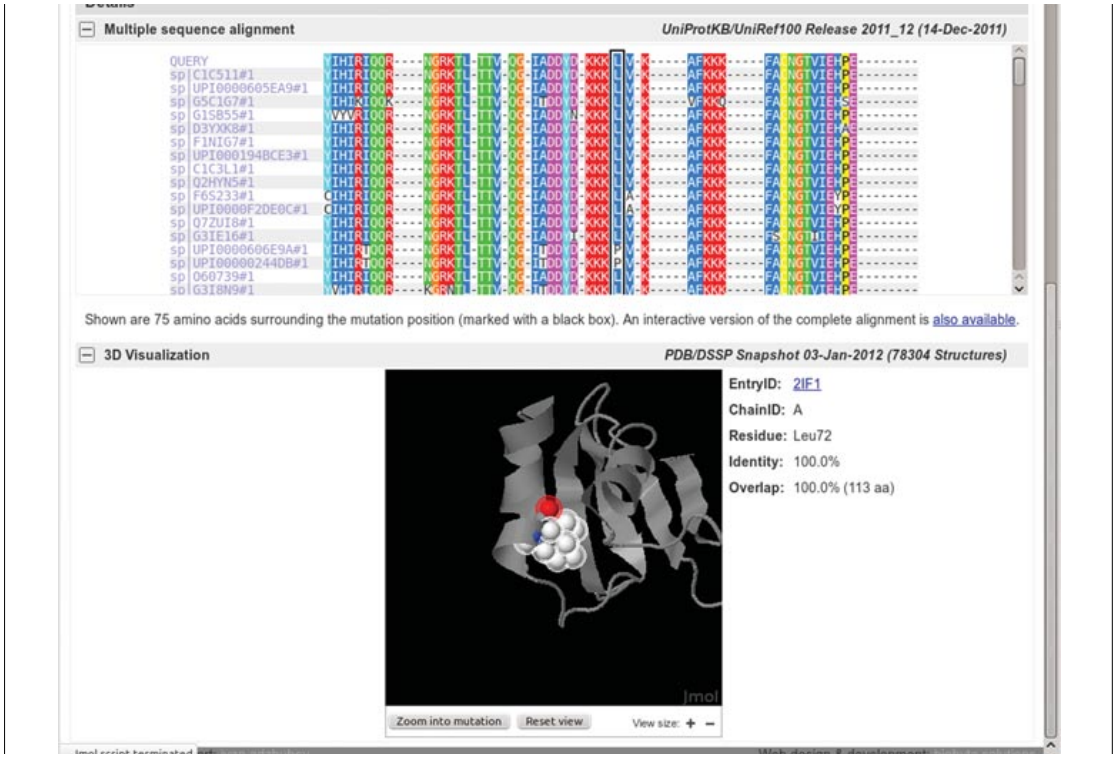
PolyPhen

- Web interface
  - Heatmap color bar with the black indicator illustrating the strength of the putative damaging effect for the variant



PolyPhen

- Web interface
  - Multiple sequence alignment and black box around variant
  - 3D protein structure with variant location marked in red
    - Interactive



VEP

- Variant Effect Predictor (VEP)
- Determines effect of SNPs, insertions, deletions, CNVs, or structural variants on
  - Genes, transcripts, protein sequences, and regulatory regions
- Input coordinates of variants and nucleotide changes to find
  - Genes affected by the variants
  - Location of the variants
    - Upstream of the transcript
    - In a coding sequence
    - In non-coding RNA
    - In regulatory regions
    - Exons, introns
    - PolyPhen predictions
- Consequence of your variants
  - Stop gained, missense, stop lost, frame shift

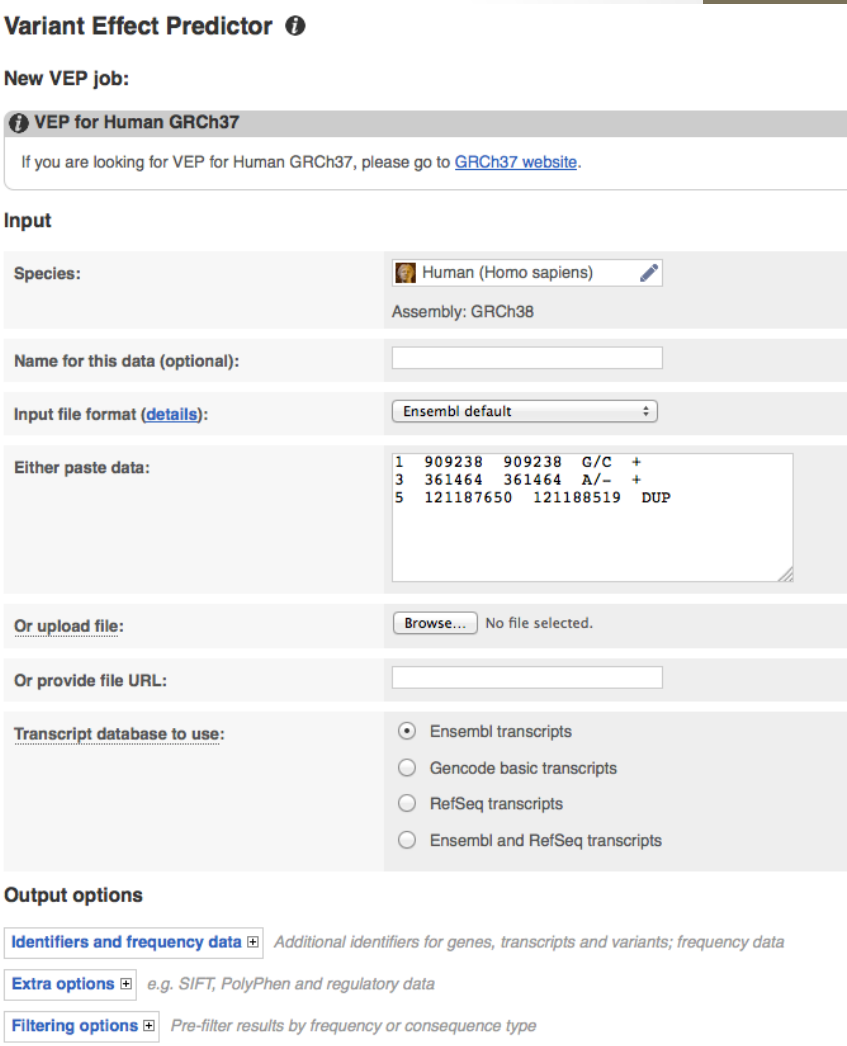
Sometimes Analysis Sources and Databases are Like This:

VEP

- Find co-located known variants
  - Report known variants from the Ensembl Variation database
- Report of minor allele frequency data from the 1000 Genomes data and NHLBI-ESP

VEP

- Web interface that suits small volumes of data
- User-friendly command line for Unix, Linux



SIFT

- Sorting Intolerant from Tolerant (SIFT)
- http://sift.jcvi.org/
- SIFT prediction is based on the degree of conservation of amino acid residues in sequence alignments derived from closely related sequences, collected through PSI-BLAST

SnpEff

- SNPEff: Genetic variant annotation and effect prediction toolbox
  - Annotates and predicts the effects of variants on genes (such as amino acid changes)
  - SnpEff is really fast: calculated predictions for all the SNPs in the 1000 Genomes project in less than 15 minutes
- http://snpeff.sourceforge.net/
- Also has SnpSift
  - Once you annotated your files using SnpEff, you can use SnpSift to help you filter large genomic datasets in various ways

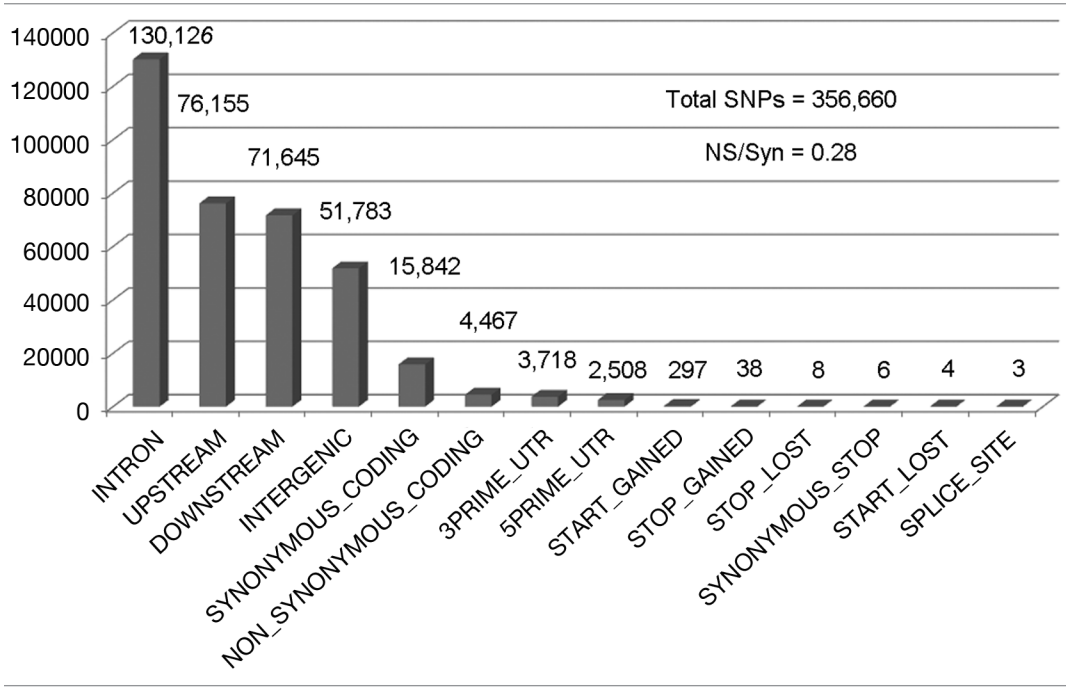
SnpEff

- SNPEff: Genetic variant annotation and effect prediction toolbox
  - Speed—the ability to make thousands of predictions per second
  - Flexibility—the ability to add custom genomes and annotation
  - Can integrate with Galaxy, an open access and web-based platform for computational bioinformatic research
  - Compatibility with multiple species and multiple codon usage tables (including mitochondrial genomes)
  - Integration with the Genome Analysis Toolkit (GATK)
  - Ability to perform non-coding annotations

Sub-field	Notes
Effect	Effect of this variant. See details below
Codon_Change	Codon change: old_codon/new_codon
Amino_Acid_change	Amino acid change: old_AA/new_AA
Warnings	Any warnings or errors
Gene_name	Gene name
Gene_BioType	BioType, as reported by ENSEMBL
Coding	[CODING | NON_CODING]. If information reported by ENSEMBL (e.g., has 'protein_id' information in GTF file)
Trancript	Transcript ID (usually ENSEMBL)
Exon	Exon ID (usually ENSEMBL)
Warnings	Any warnings or errors (not shown if empty)

SnpEff

- SNPEff: Genetic variant annotation and effect prediction toolbox
  - More genome versions
  - Open source for any user
  - Supports VCF files



SnpEff

- SNPEff: Genetic variant annotation and effect prediction toolbox
  - Similar in many ways to ANNOVAR and VAAST
    - http://www.openbioinformatics.org/annovar/
    - http://www.yandell-lab.org/software/vaast.html

VAT

- http://varianttools.sourceforge.net/Association/HomePage
- Variant Association Tools
- Large collection of utilities devoted to data exploration, quality control and association analysis of rare/common single nucleotide variants and indels

STRING

- http://string-db.org/
- A database of known and predicted protein interactions
- The interactions include direct (physical) and indirect (functional) associations
- Can search by protein name, or protein sequence

MAKER

- MAKER
  - http://www.yandell-lab.org/software/maker.html
- Purpose is to allow smaller eukaryotic and prokaryotic genome projects to independently annotate their genomes and to create genome databases
- Identifies repeats, aligns contigs and proteins to a genome, produces gene predictions and automatically synthesizes these data into gene annotations having evidence-based quality values

Which One Do I Choose?

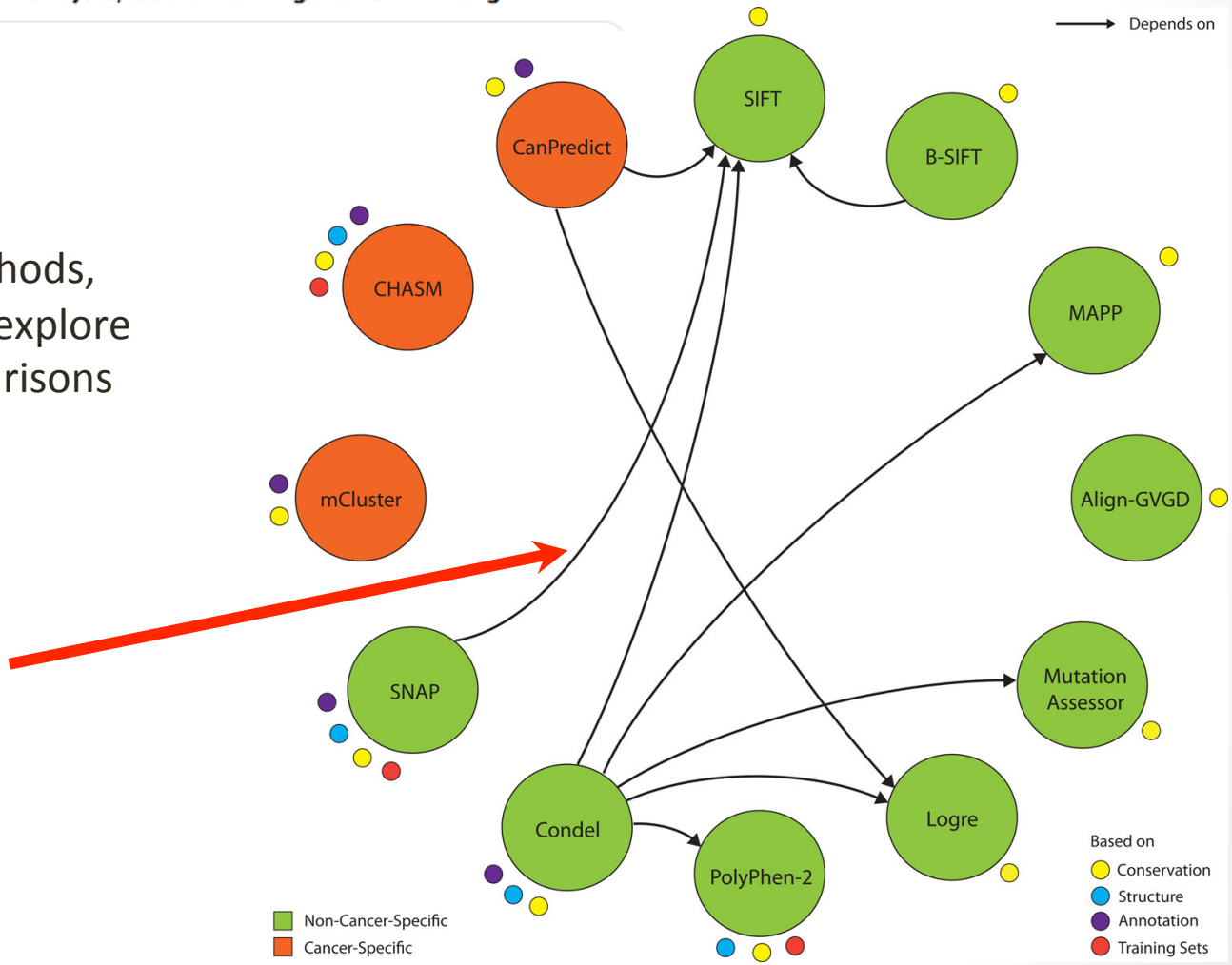
- It is a challenge
  - Variety of similar and also different information
  - Different approaches for estimating effect on proteins
  - It does become a matter of opinion and specific needs of a given project
  - The good news is that work is in progress to unify multiple sources so you can gather information from multiple sources
  - *Don't forget growing repositories of other kinds of functional data:*
    - ENCODE/GENCODE
    - Haploreg
    - RegulomeDB
    - SCAN database

Which One Do I Choose?

Assessment of computational methods for predicting the effects of missense mutations in human cancers

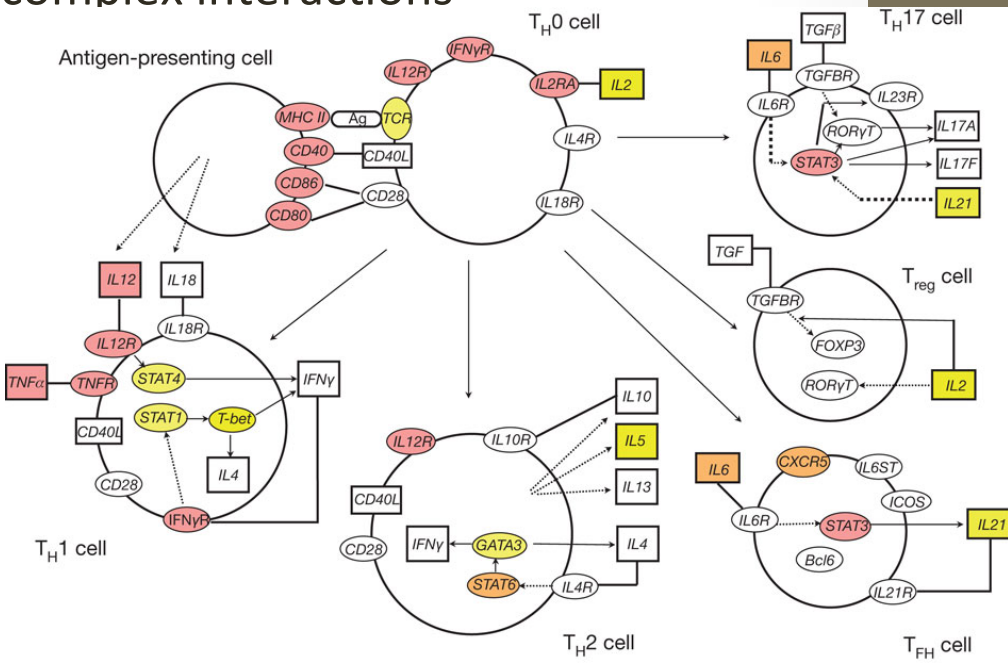
Florian Gnad, Albion Baucom, Kiran Mukhyala, Gerard Manning and **Zemin Zhang***

Growing sources of comparisons of methods, before using these, explore what current comparisons have been made



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



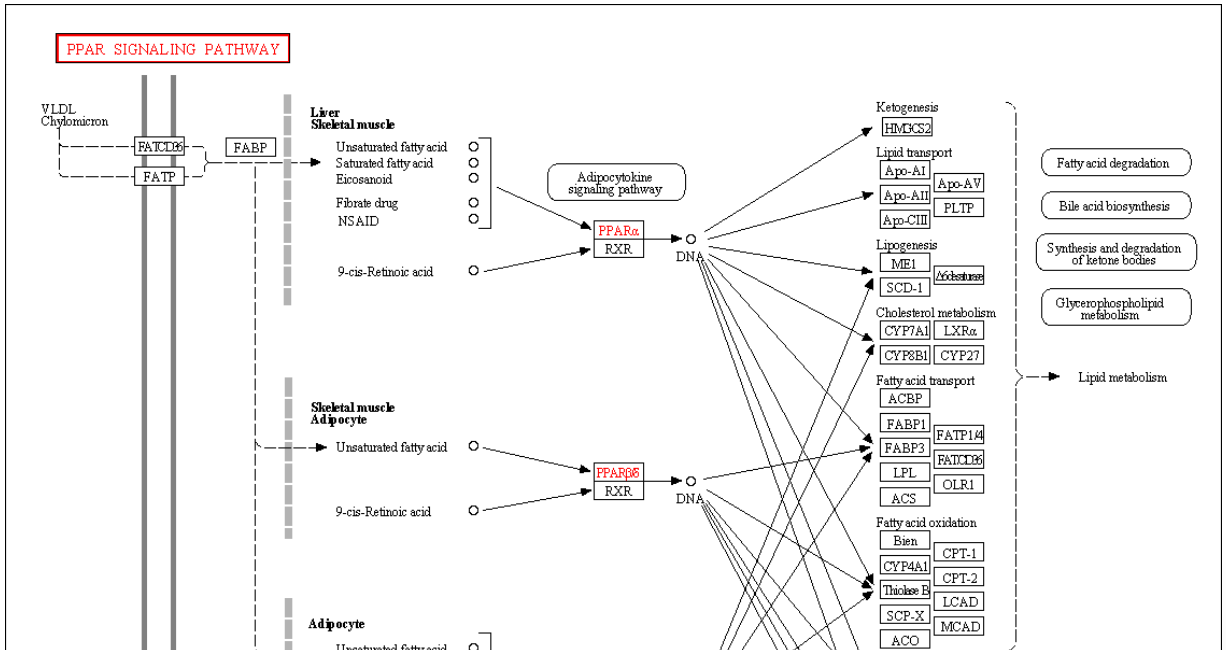
Biological Pathways

- 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?

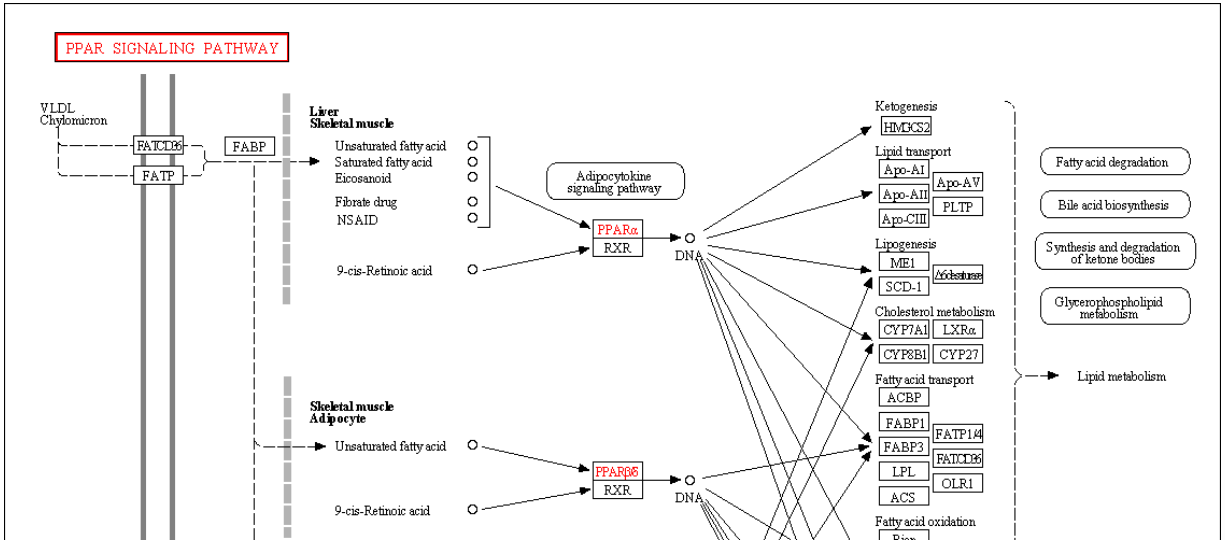
Biological Pathways

- 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



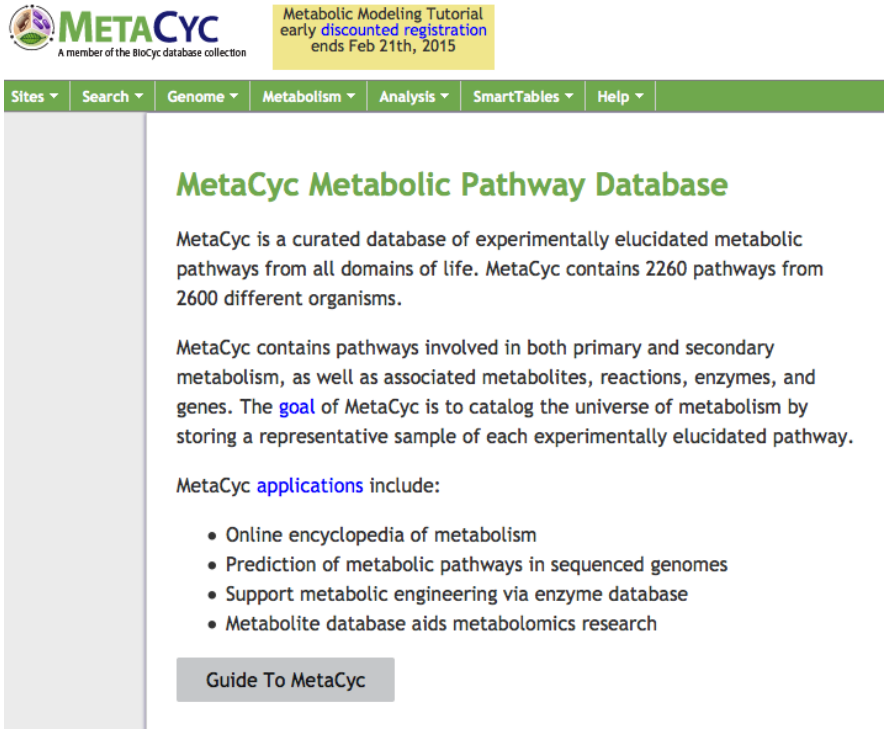
Biological Pathways

- 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



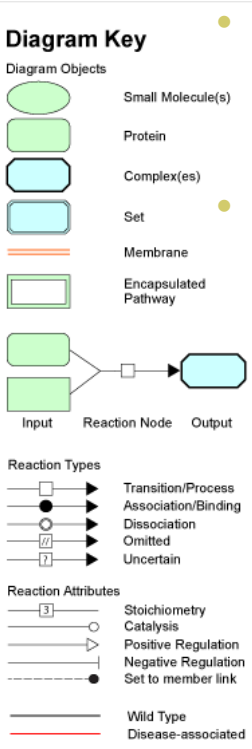
Biological Pathways

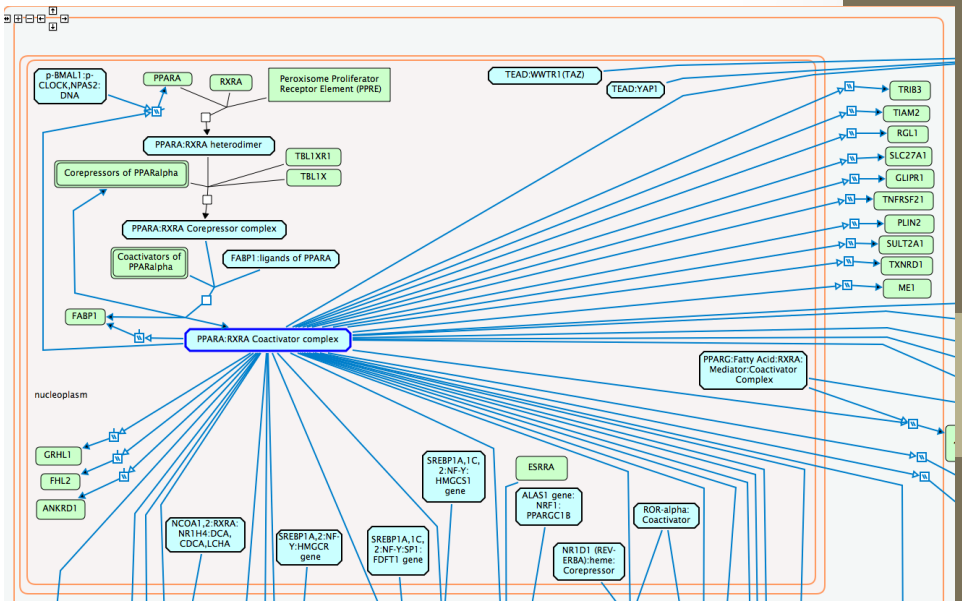
- MetaCyc: http://metacyc.org/
  - MetaCyc is a curated database of experimentally elucidated metabolic pathways from all domains of life
    - MetaCyc contains 2260 pathways from 2600 different organisms
    - Superpathways of combined information



Biological Pathways

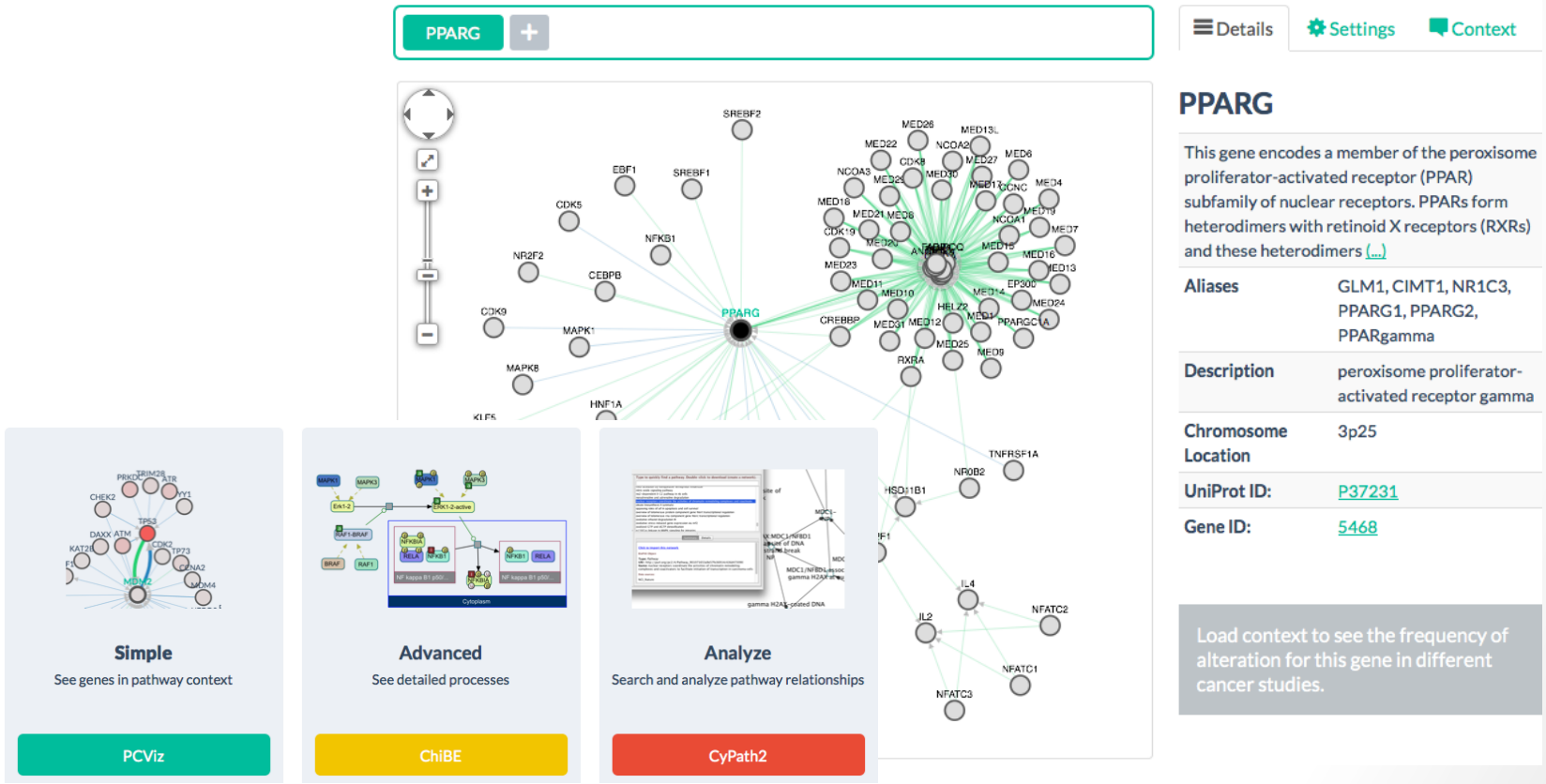
- Example sources of pathway information
  - Reactome
    - http://www.reactome.org/
  - Little more user friendly than KEGG (can search multiple genes)
  - Free, open-source, curated and peer reviewed pathway database
  - Pathways and reactions (pathway steps) in human biology
  - 3700 proteins (including 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





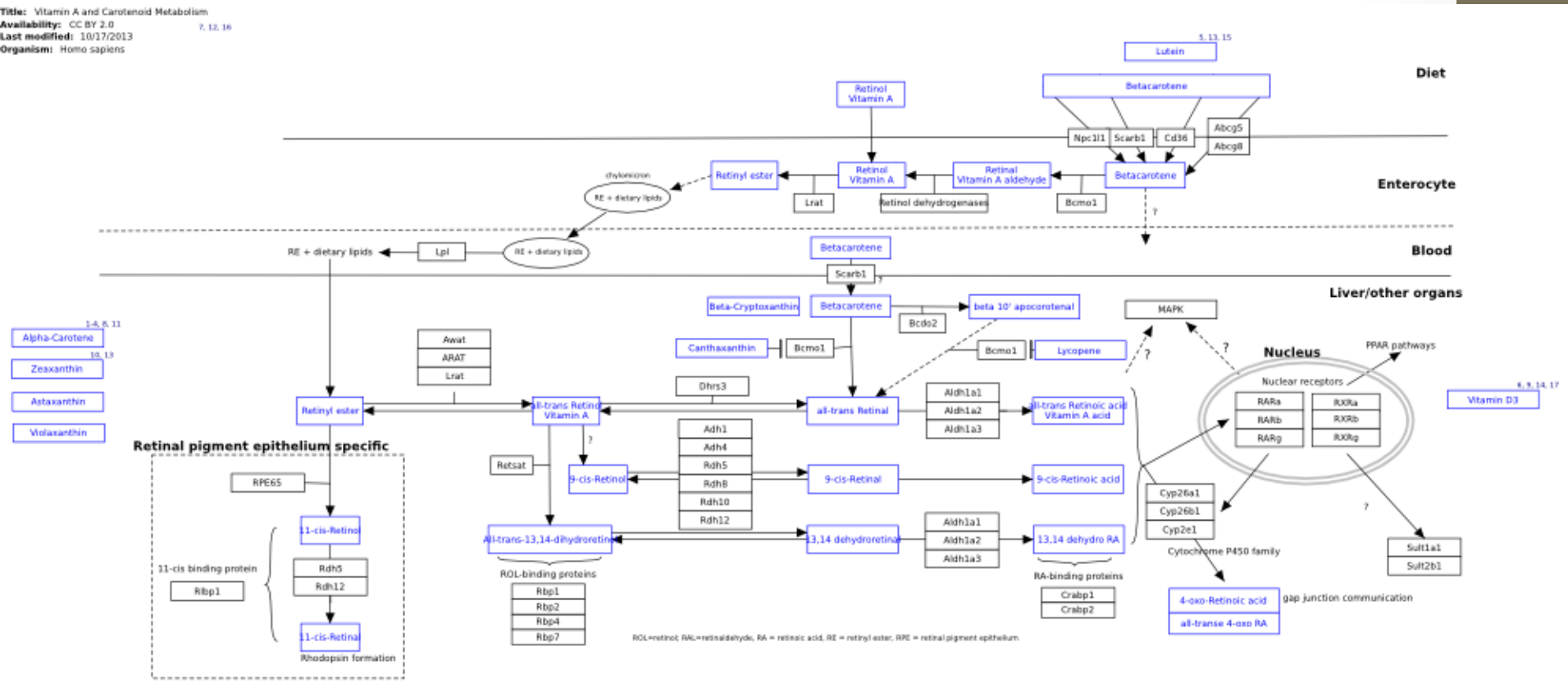
Biological Pathways

- 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

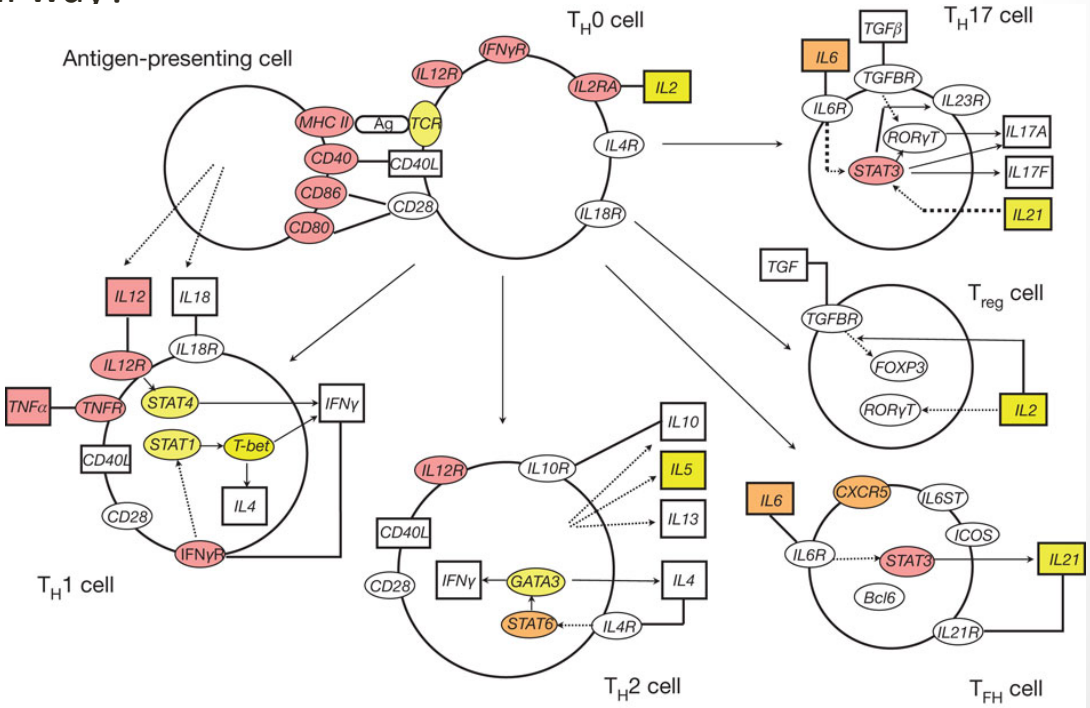


Biological Pathways

- 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

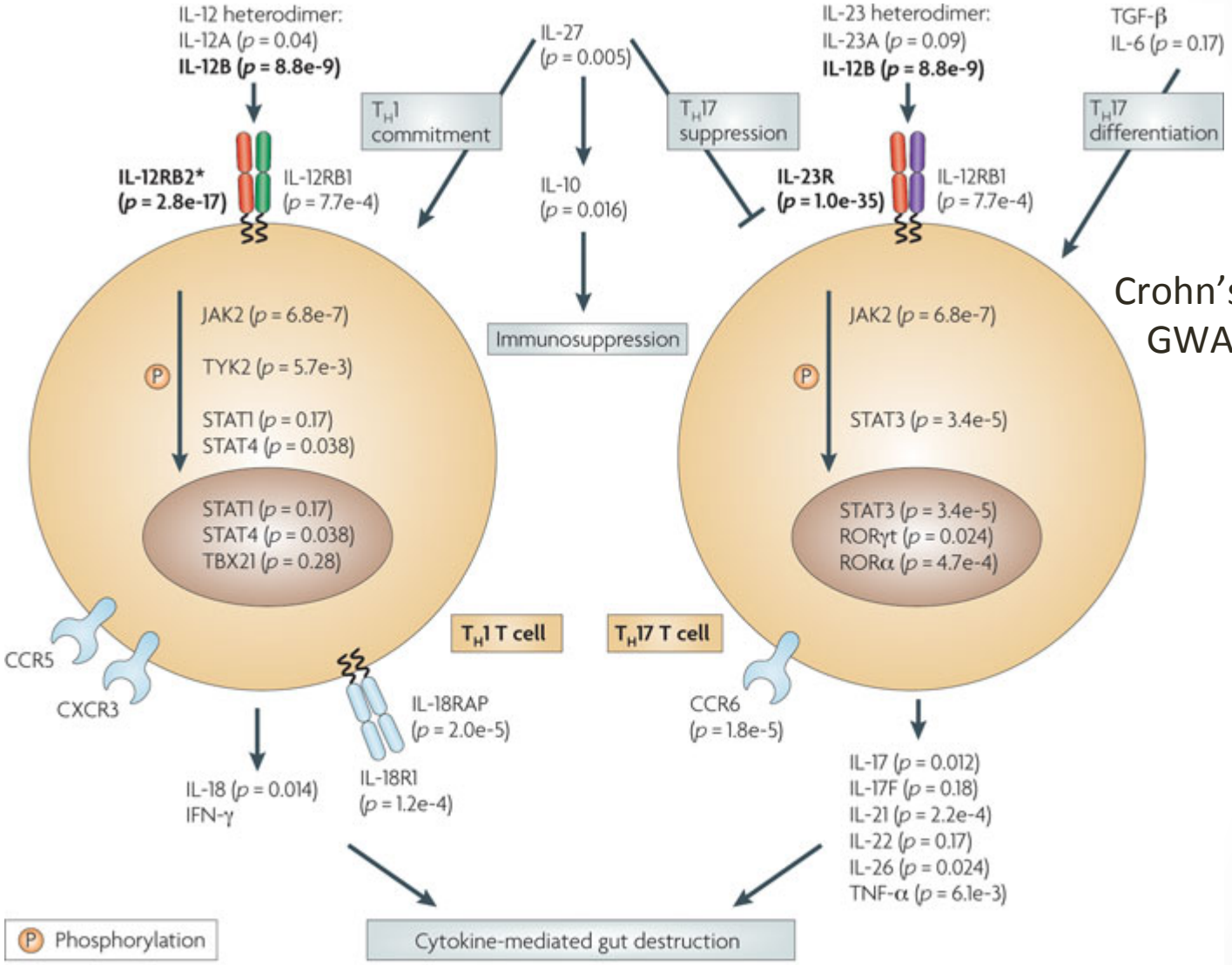
Networks/Pathways

- 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

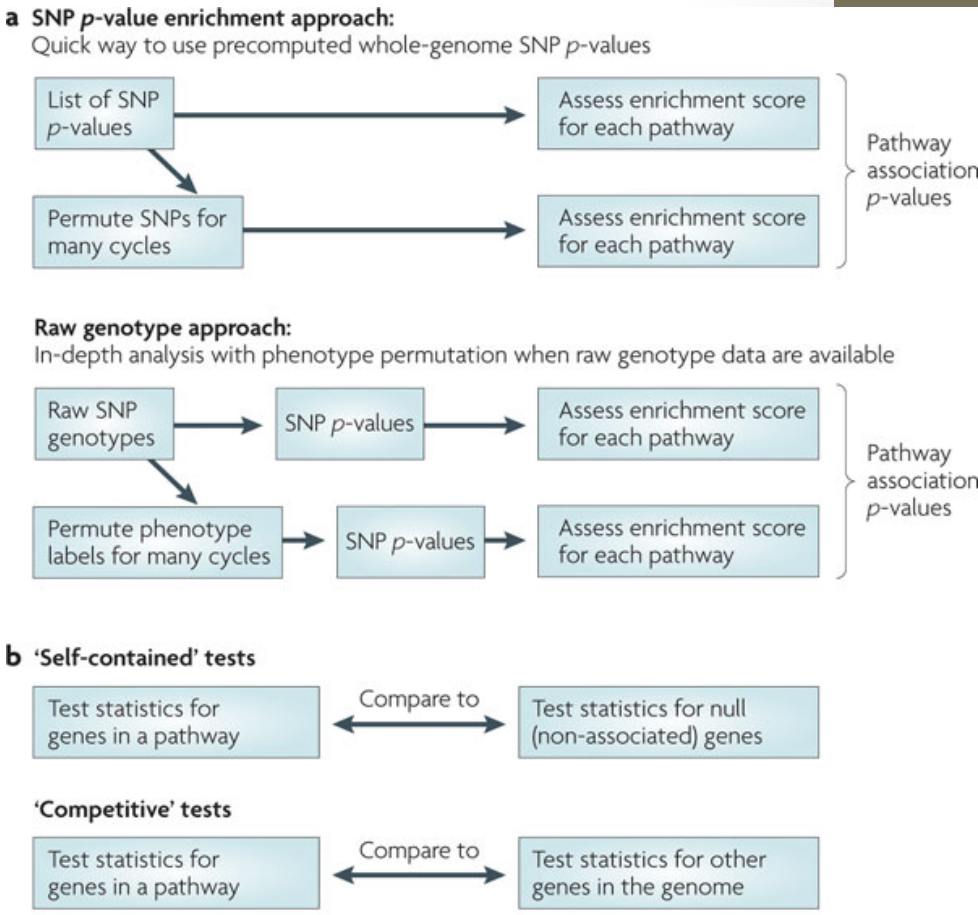
- Why do I care about network/pathway context for my SNPs or SNVs?
  - What if I have a series of genetic variants and
    - They are so low frequency I can't evaluate them one at a time?
    - Across cases/controls individually there is not a consistent pattern
      - No single genetic variant seems 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

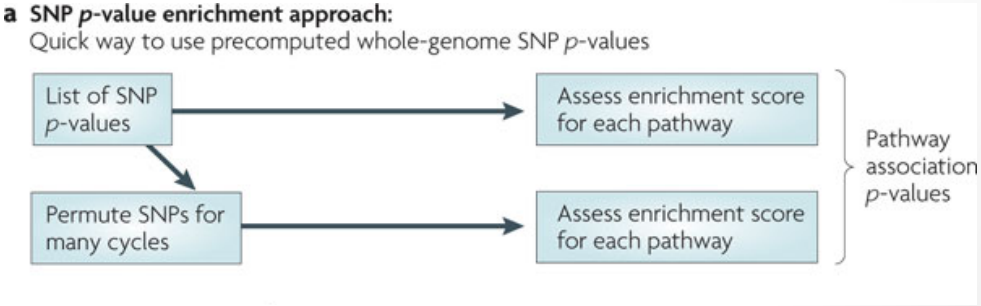


Analysing biological pathways in genome-wide association studies
Nature Reviews Genetics 11, 843-854 (December 2010)

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

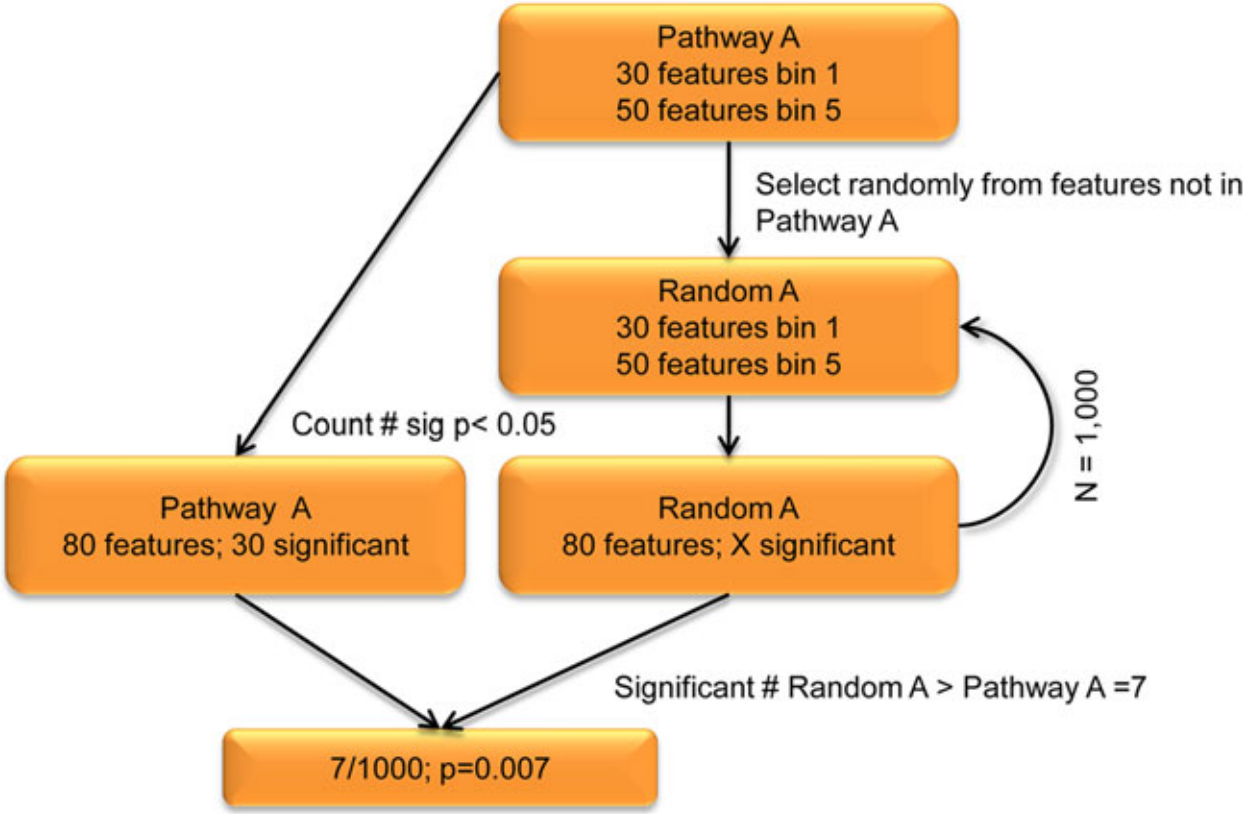
PARIS

- 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

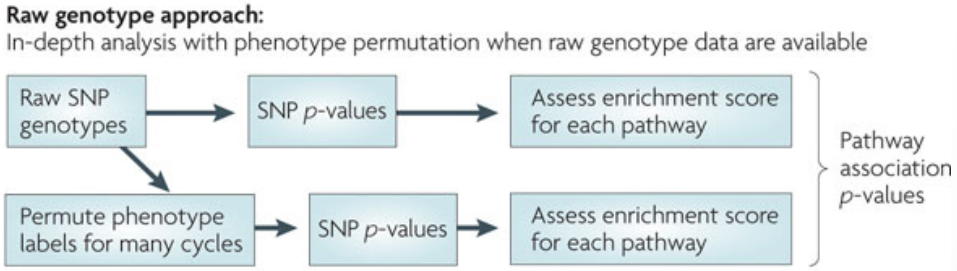
- Pathway Analysis by Randomization Incorporating Structure (PARIS)
- First example use
  - Used KEGG pathways
  - Autism GWAS dataset
- Revealed pathways with a significant enrichment of positive association results

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

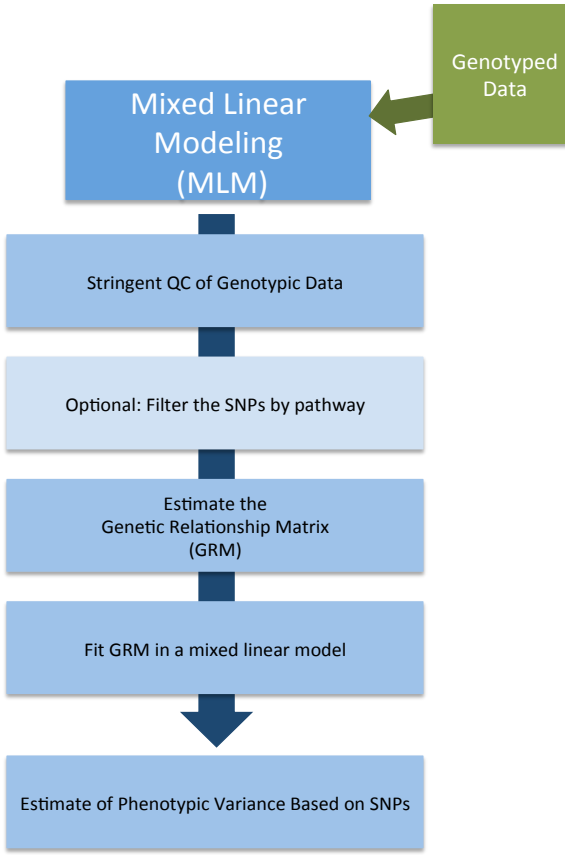
- For GWAS data
- Basically divided into two methods
  - Providing SNP genotypes
  - Gene-level and pathway-level test statistics
  - Can require just multi-markers
  - Some require single marker p-values and genotypic data



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

Pathway Based Approaches

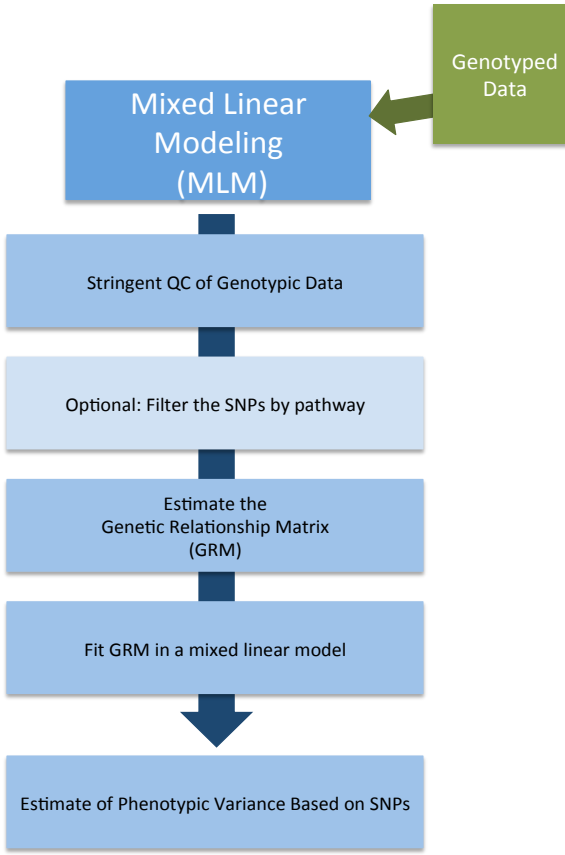
- Example of using genotypic data
- Polygenic etiology of paclitaxel-induced neuropathy
- Estimated the variance explained by common SNPs (MAF > 1%) for two outcomes
  - Maximum grade of sensory peripheral neuropathy
  - Dose at first instance of peripheral neuropathy



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

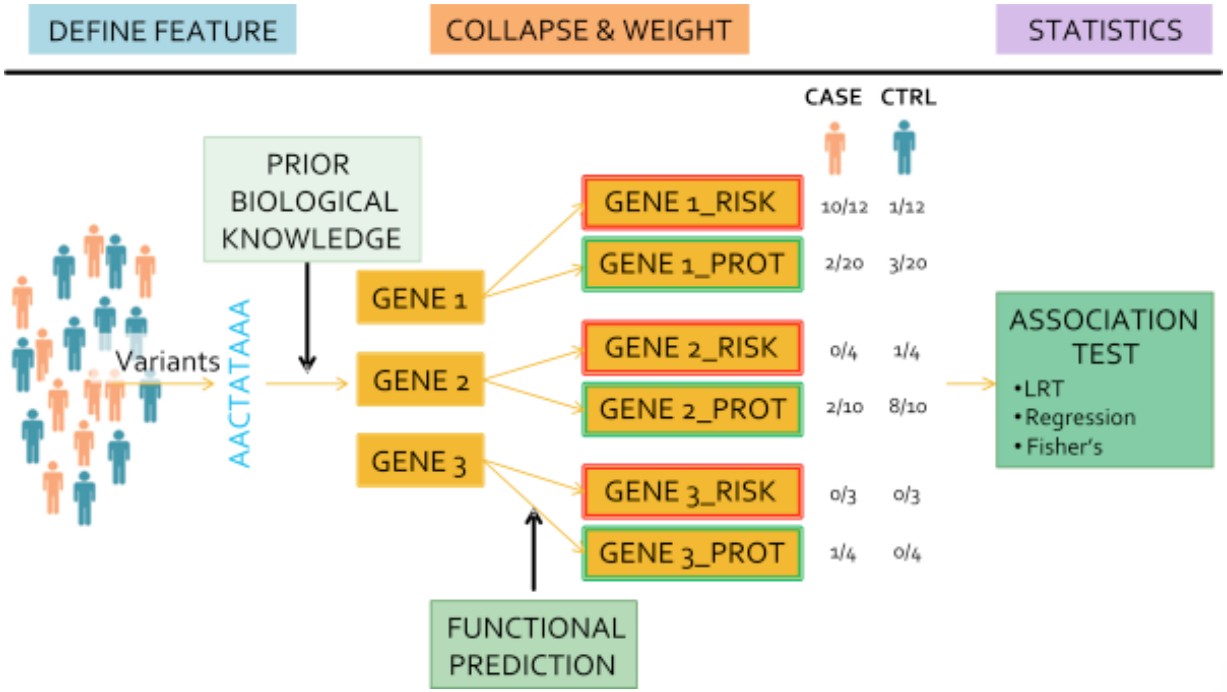
- We can't know what we don't know
  - When moving to these analyses we can only investigate known pathways or connections
    - We are still elucidating many biological networks, protein functions, and protein interactions
    - As a result, over time the results of your analyses can change
      - 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.

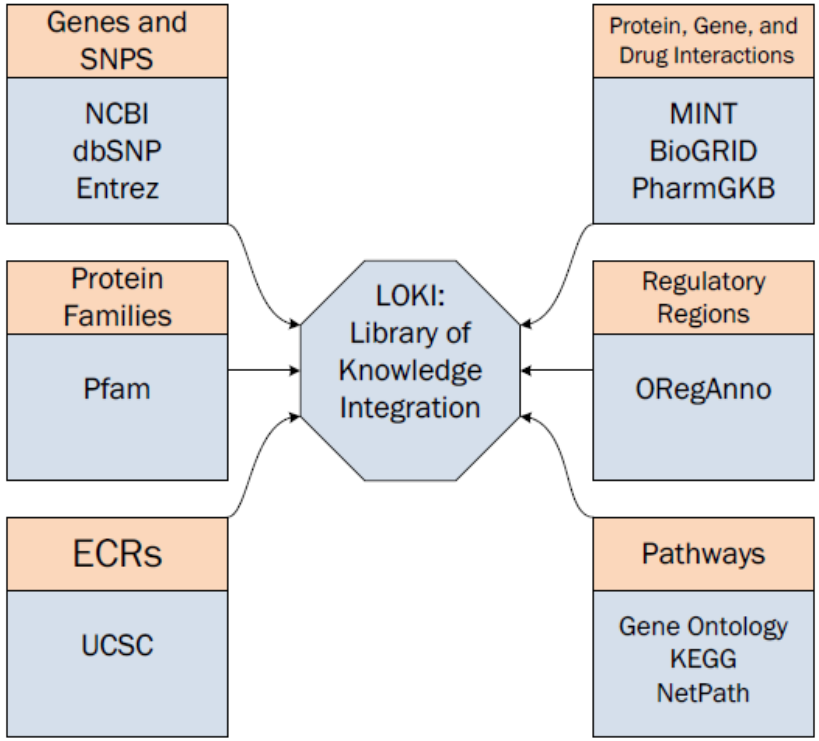
Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data.
PLoS Genet. 2013;9(12):e1003959. doi: 10.1371/journal.pgen.1003959. Epub 2013 Dec 26.

BioBin

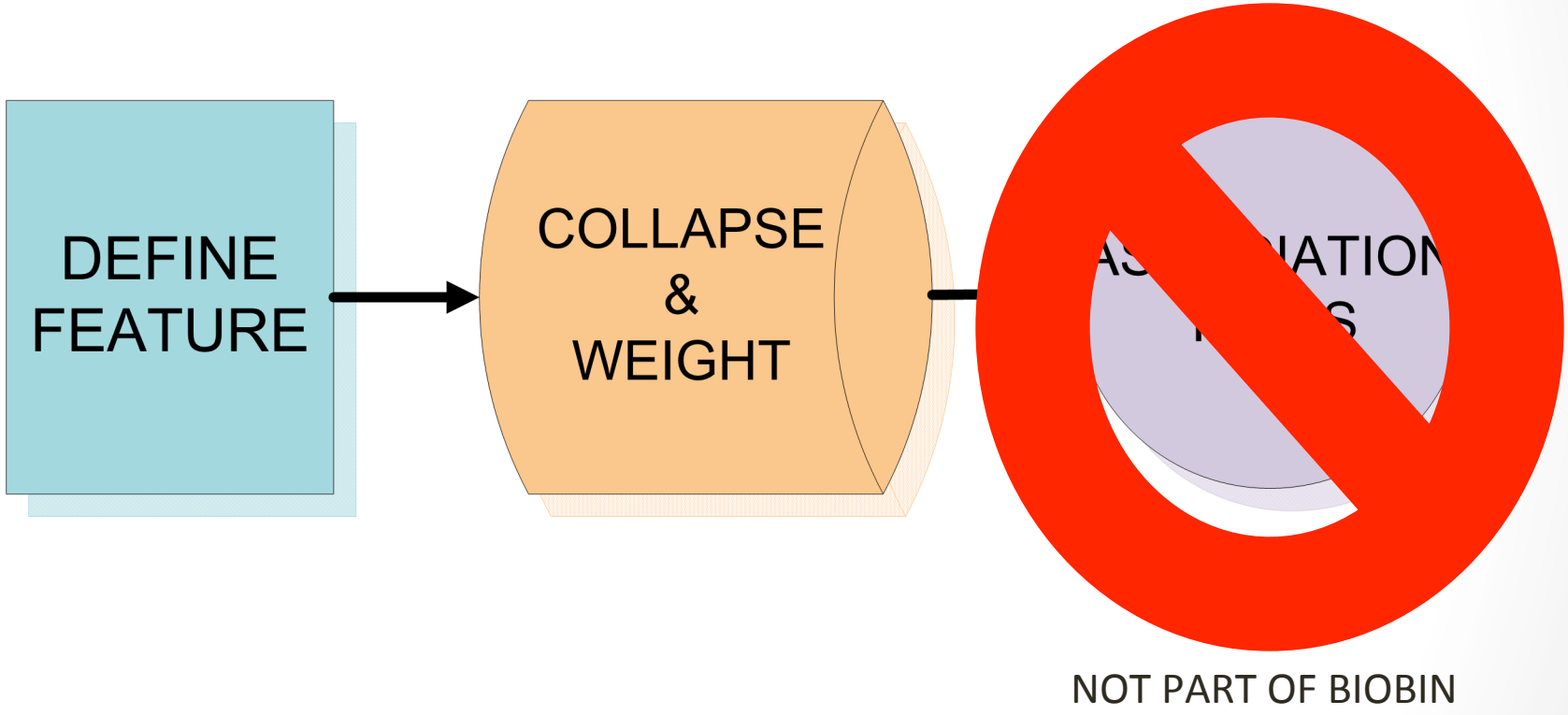
- Integrate collapsing method for rare variants using Biofilter
  - Include genetic information (pathway, gene boundaries, etc)
  - Create flexible binning structure
- Incorporate functional information
  - Bin variants by biological features
    - Gene
    - Intron
    - Exon
    - Intergenic
    - 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

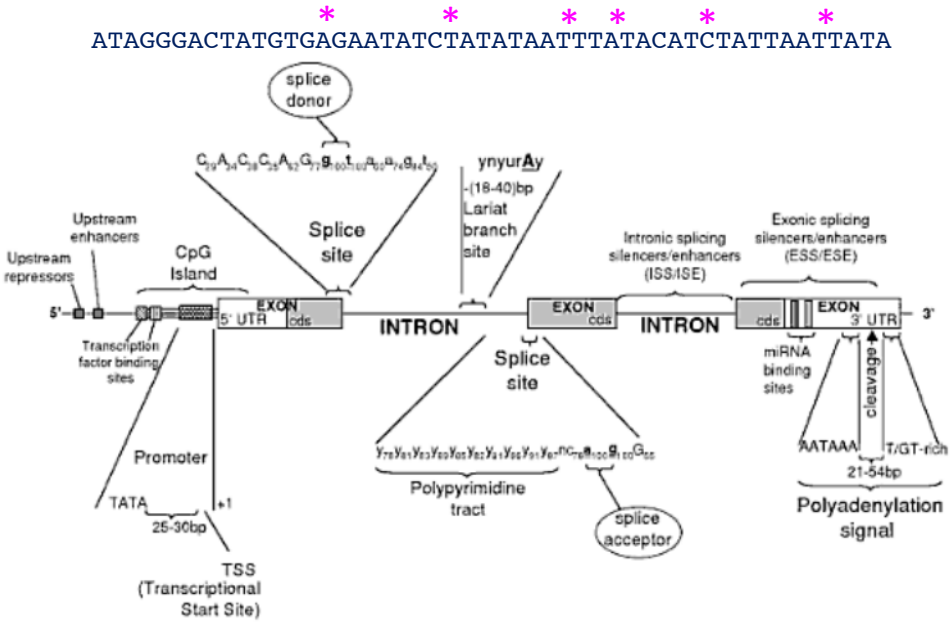


Figure 11.2 The anatomy of a gene. This figure illustrates some of the key regulatory regions that control the transcription, splicing and post-transcriptional processing of genes and transcripts. Polymorphisms in these regions should be investigated for functional effects

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

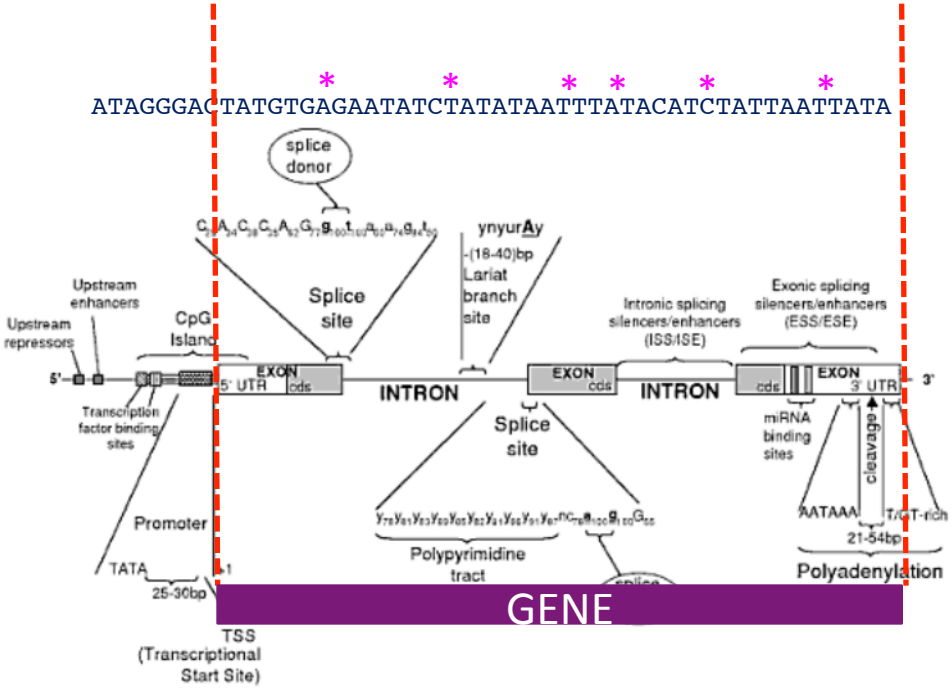


Figure 11.2 The anatomy of a gene. This figure illustrates some of the key regulatory regions that control the transcription, splicing and post-transcriptional processing of genes and transcripts. Polymorphisms in these regions should be investigated for functional effects

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

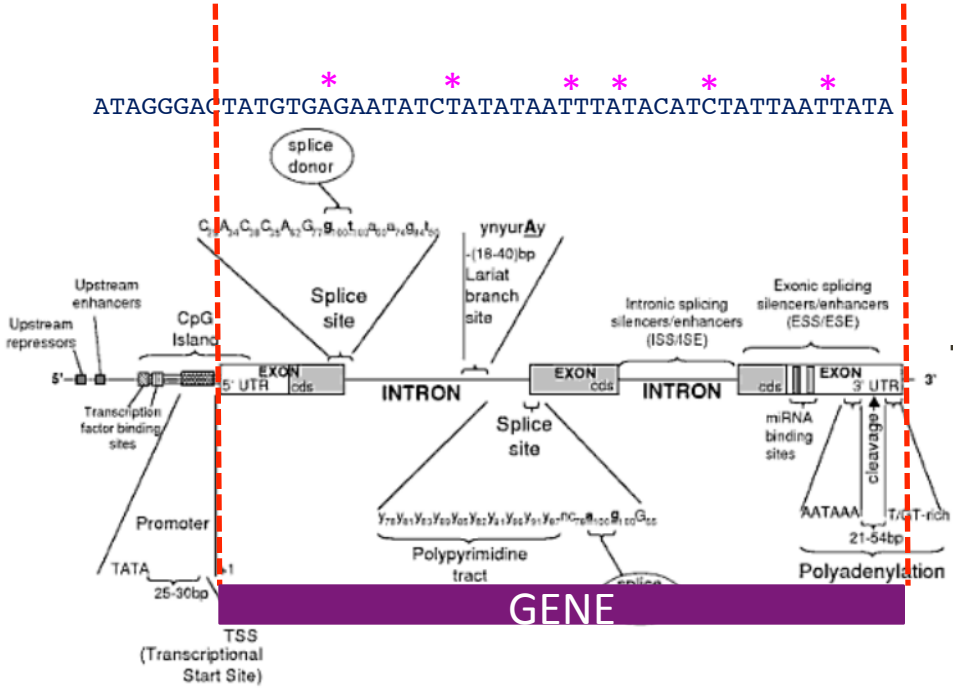


Figure 11.2 The anatomy of a gene. This figure illustrates some of the key regulatory regions that control the transcription, splicing and post-transcriptional processing of genes and transcripts. Polymorphisms in these regions should be investigated for functional effects

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

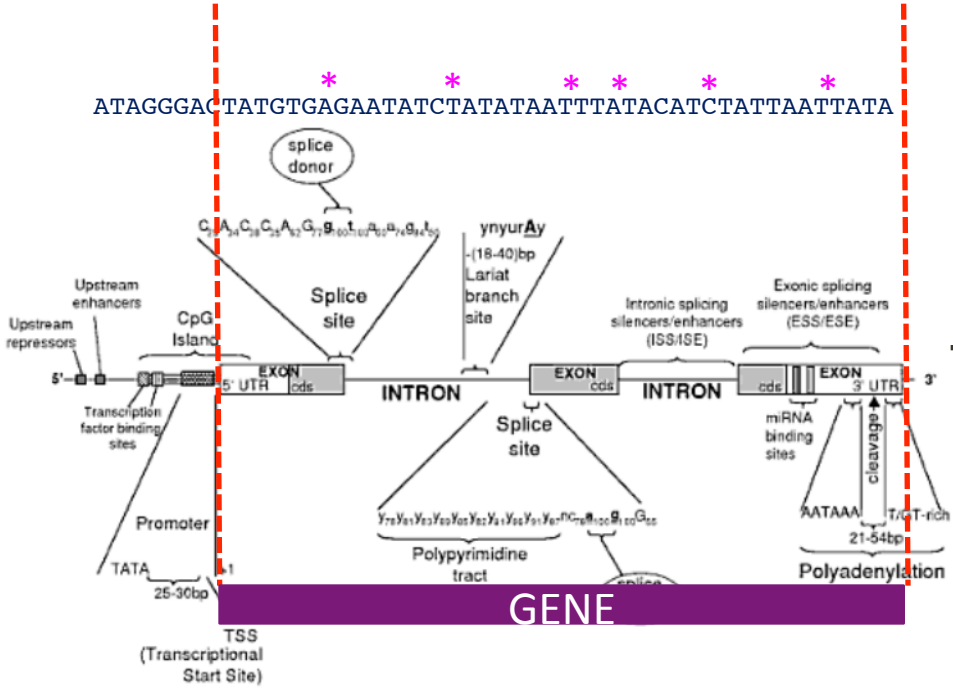
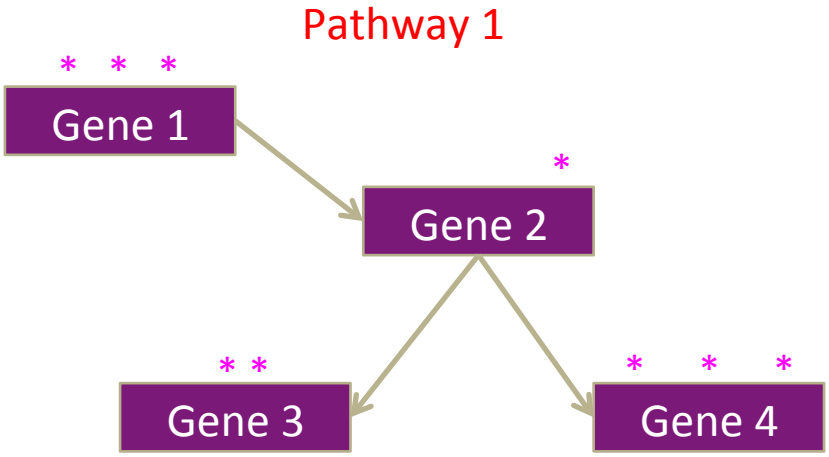


Figure 11.2 The anatomy of a gene. This figure illustrates some of the key regulatory regions that control the transcription, splicing and post-transcriptional processing of genes and transcripts. Polymorphisms in these regions should be investigated for functional effects

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

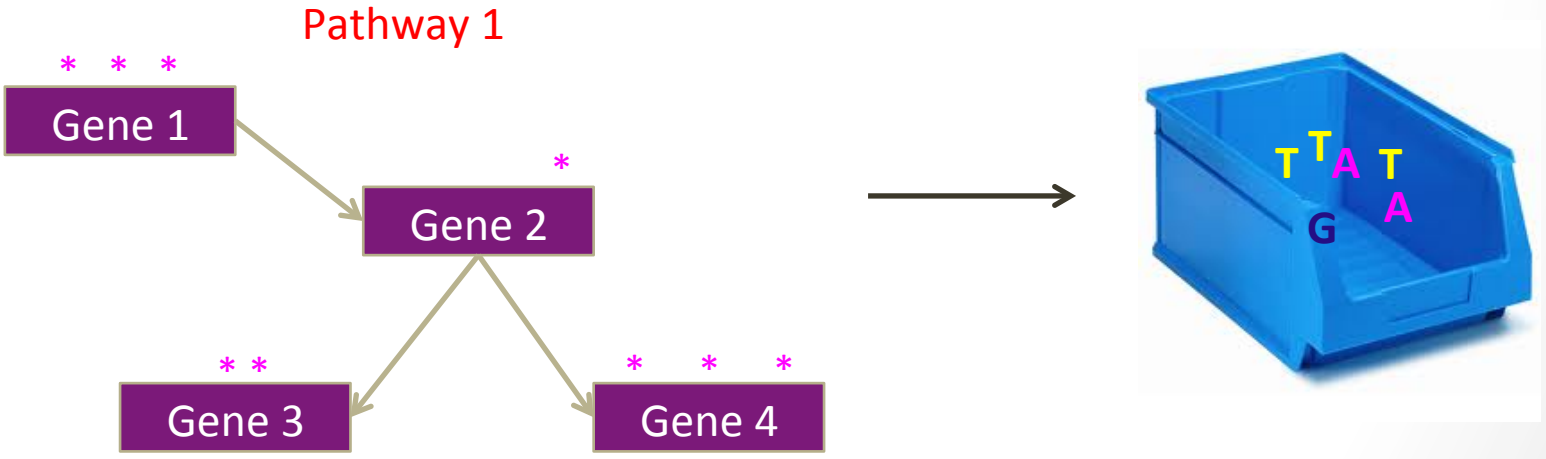
BioBin

- 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



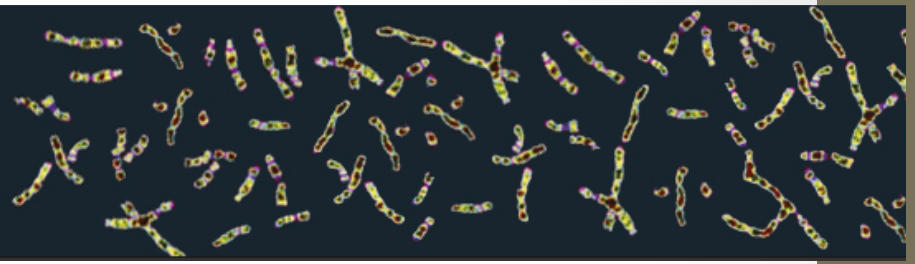
BioBin

- 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?



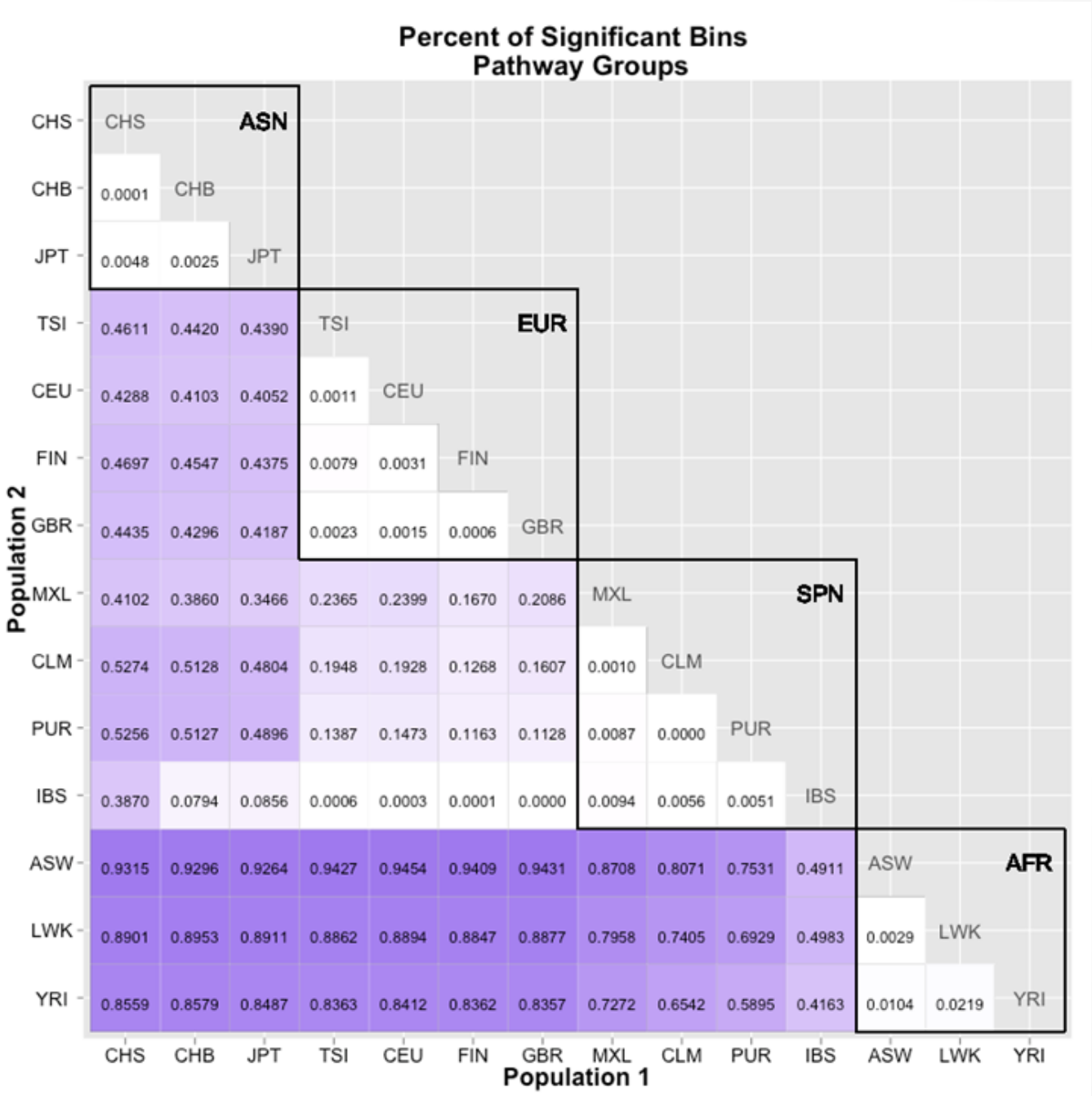
1000 Genomes

A Deep Catalog of Human Genetic Variation



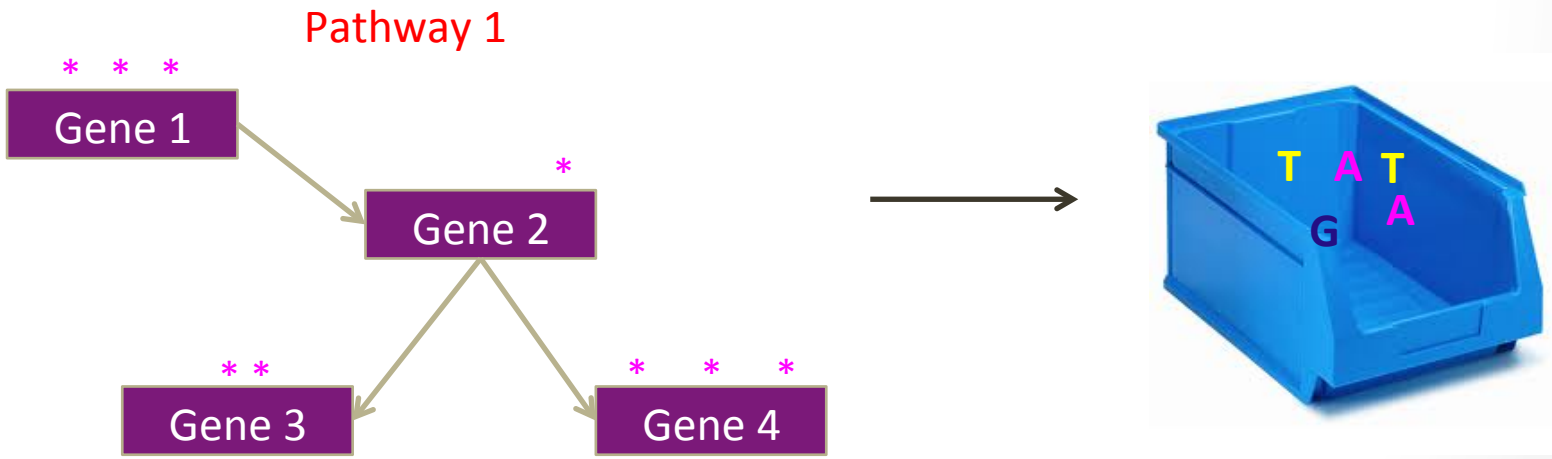
- Moore *et al.* for proof of principle evaluated rare variant differences between 1000 Genomes populations

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

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

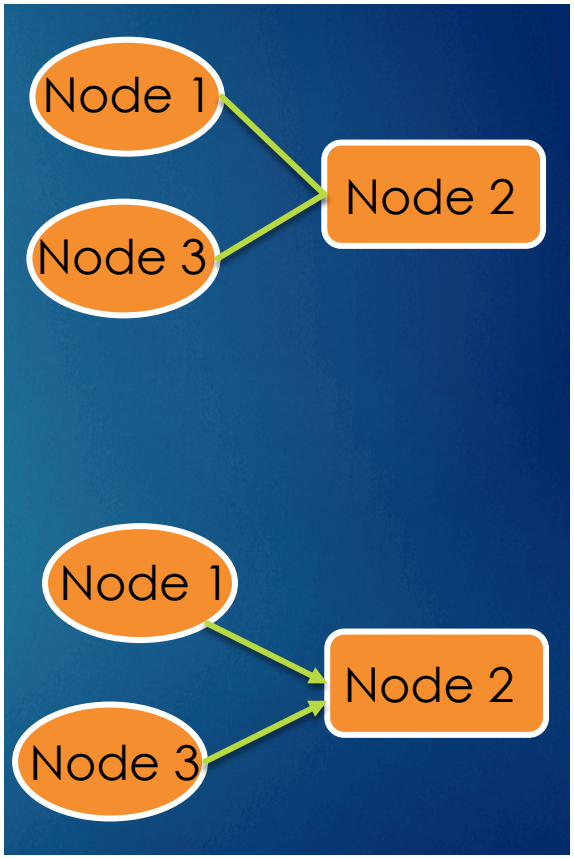
- Direction of effect
  - Dispersion methods
    - No assumption of burden tests of the same direction of effect of all rare variants on the trait within the same functional unit or genomic region
  - SKAT Method
    - SNP-set (Sequence) Kernel Association Test (SKAT)
    - http://www.hsph.harvard.edu/skat/
    - 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

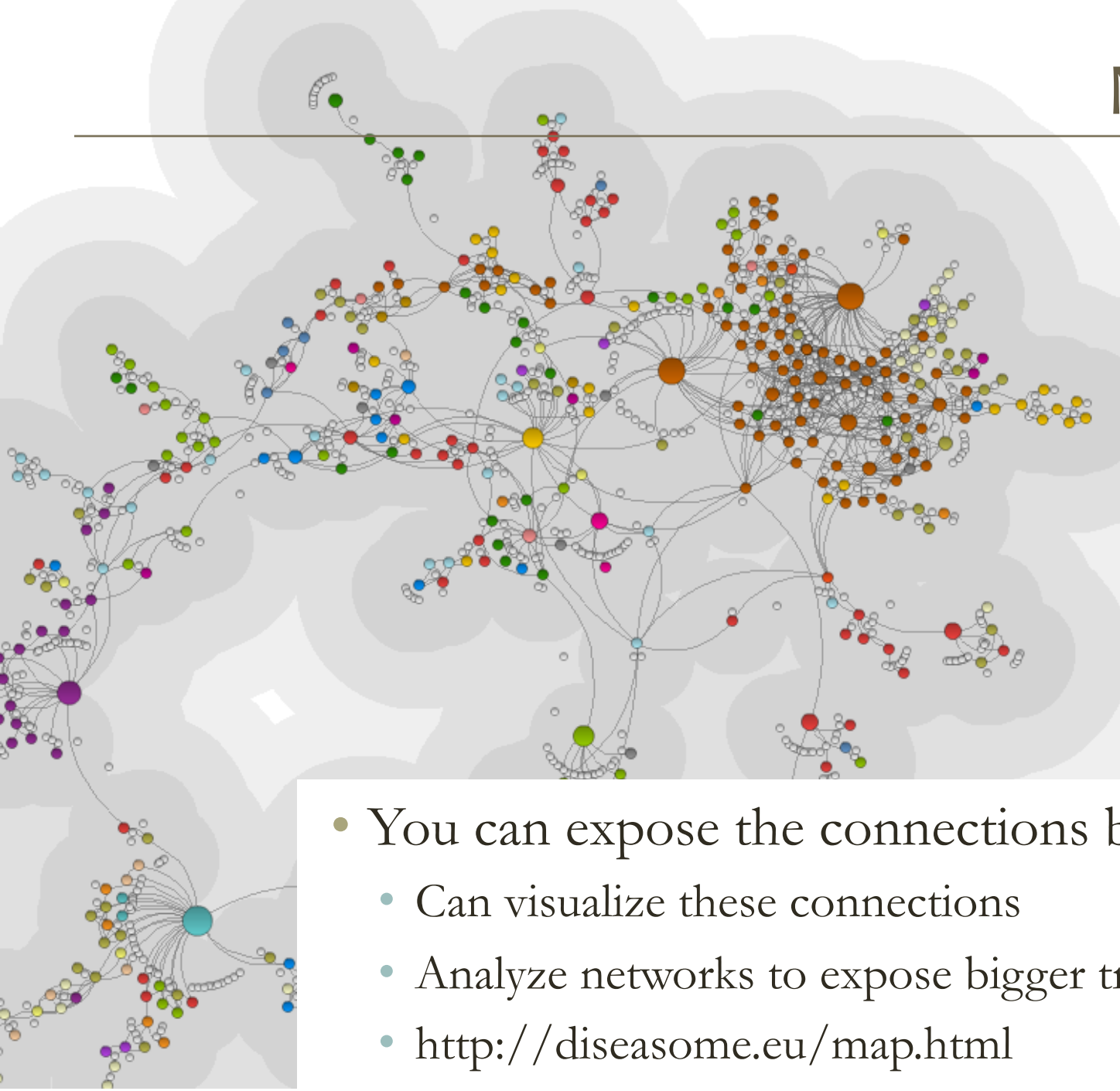
- Remember how biological data is inherently connected?
- I found a series of SNPs in different genes
- I identified that these genes are in shared KEGG pathways
  - Can I visualize this information?
- Cytoscape and Gephi are two free software packages that allow for network visualization

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

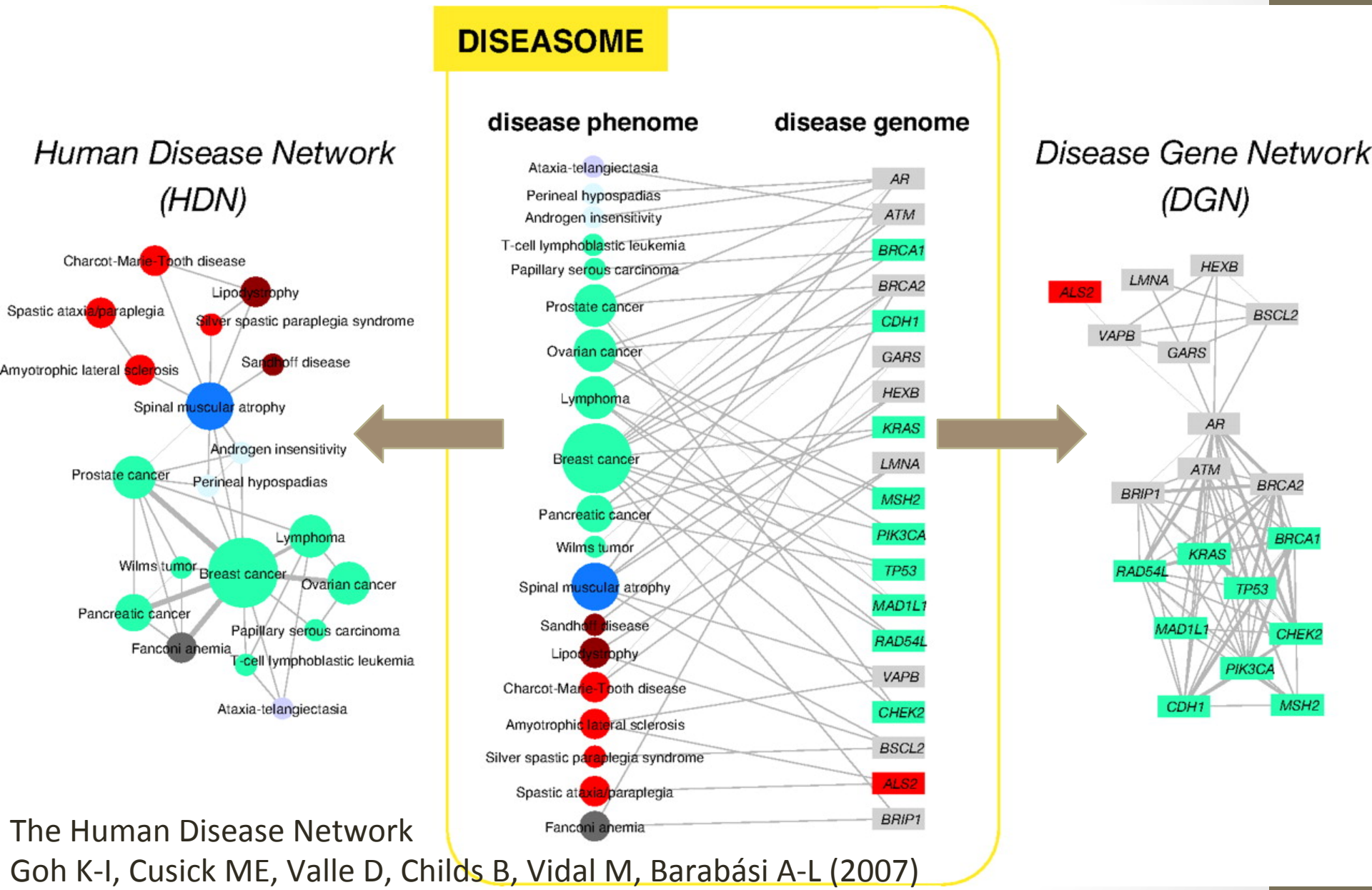


Networks



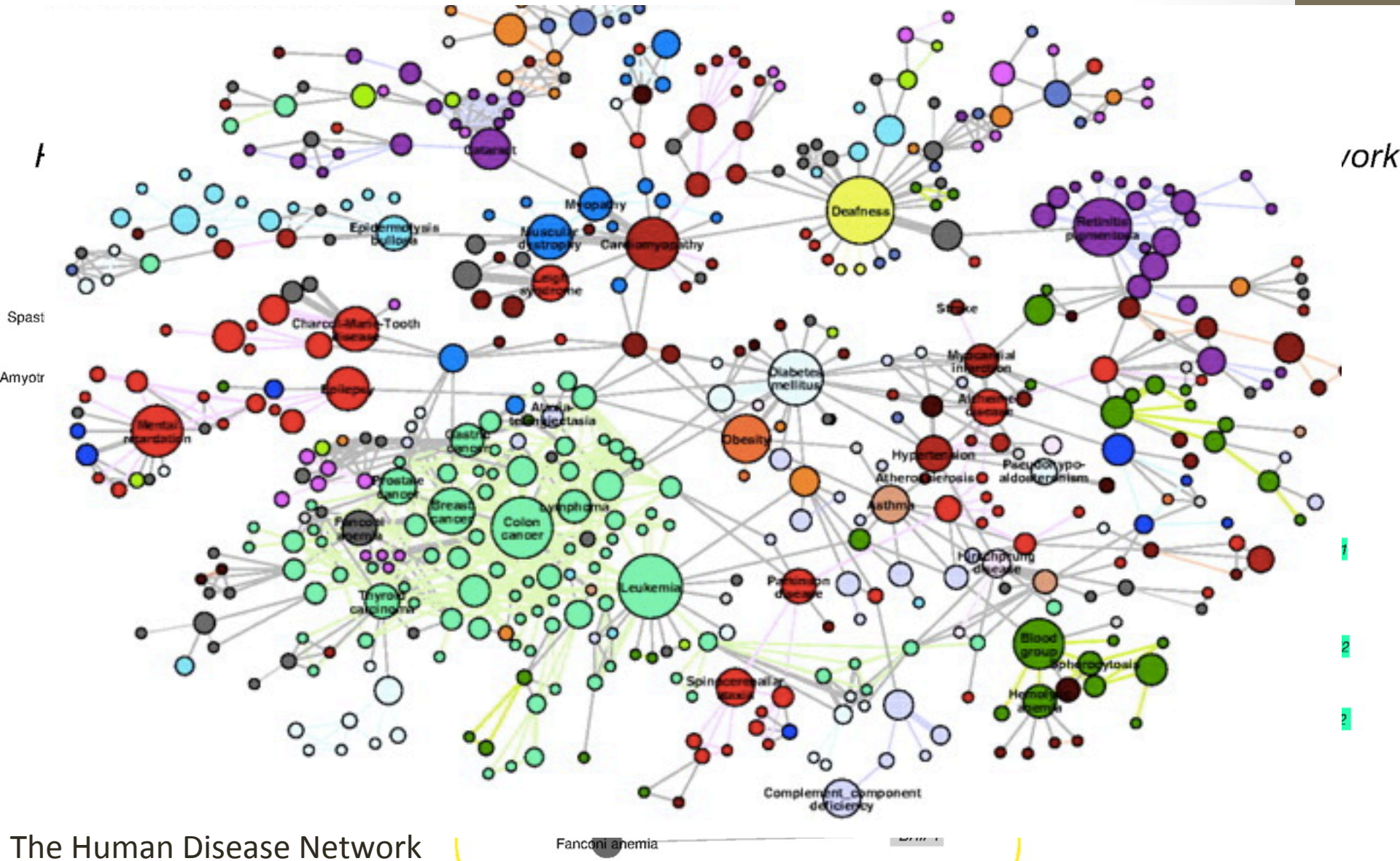
- You can expose the connections between data
  - Can visualize these connections
  - Analyze networks to expose bigger trends
  - http://diseasome.eu/map.html

Networks



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

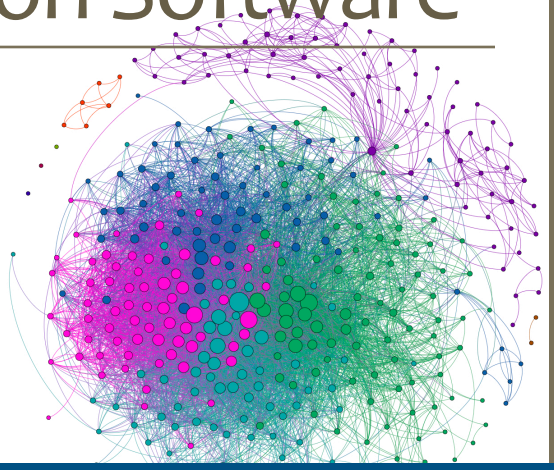
Networks



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

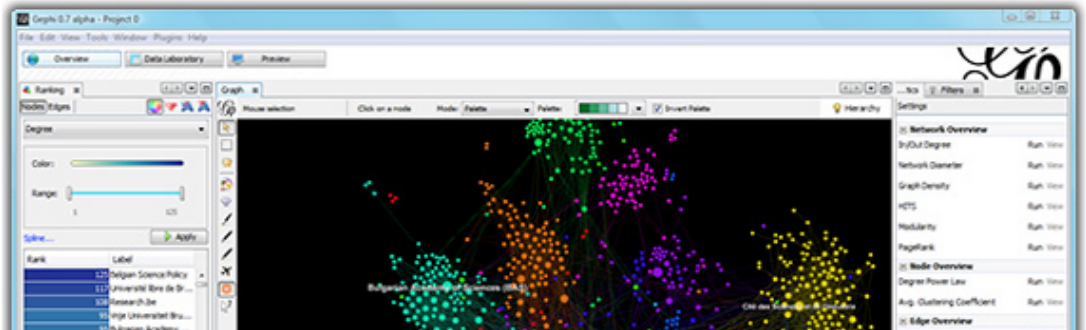
Gephi
makes graphs handy

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?