Pathway Analysis in other data types

Alison Motsinger-Reif, PhD
Branch Chief, Senior Investigator
Biostatistics and Computational Biology Branch
National Institute of Environmental Health Sciences

alison.motsinger-reif@niehs.nih.gove

New “-Omes”

- Genome
- Transcriptome
- Metabolome
- Epigenome
- Proteome
- Phenome, exposome, lipidome, glycome, interactome, spliceome, mechanome, etc...

Goals

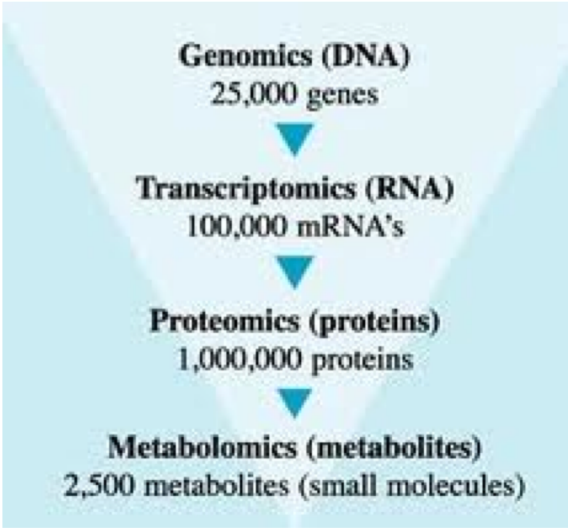
- Pathway analysis in metabolomics
- Pathway analysis in proteomics
- Issues, concerns in other data types
  - Methylation data
  - aCGH
  - Next generation sequencing technologies

- Many approaches generalize, but there are always specific challenges in different data types

- XGR – a new tool for pathway and network analysis

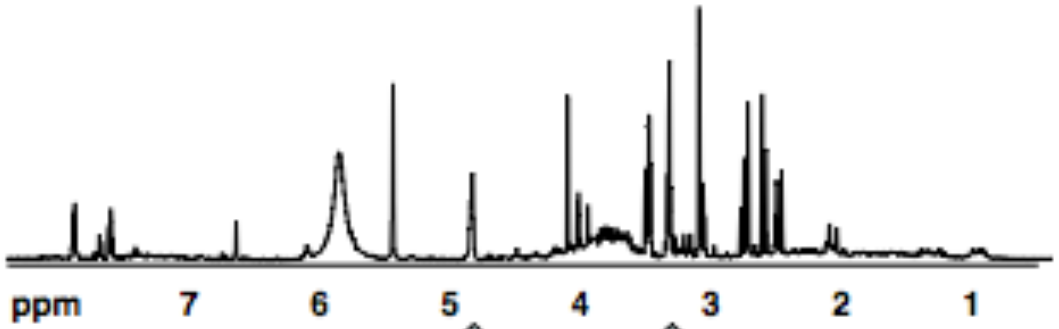
Metabolomics

- While many proteins interact with each other and the nucleic acids, the real metabolic function of the cell relies on the enzymatic interconversion of the various small, low molecular weight compounds (metabolites)
- Technology is rapidly advancing
- The frequent final product of the metabolomics pipeline is the generation of a list of metabolites who's concentrations have been (significantly) altered which must be interpreted in order to derive biological meaning



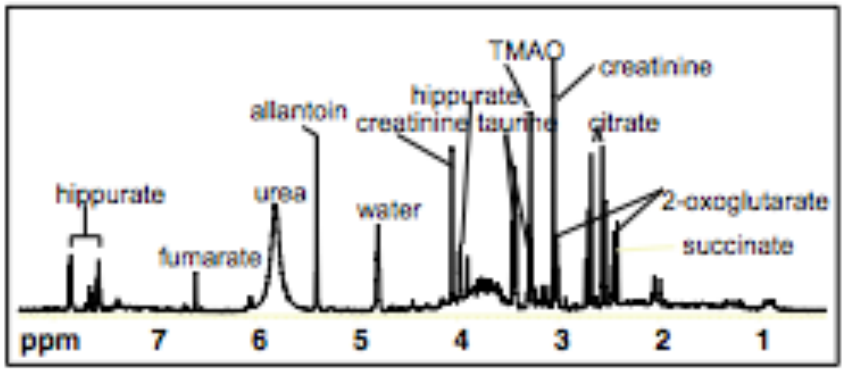
→ Perfect for pathway analysis

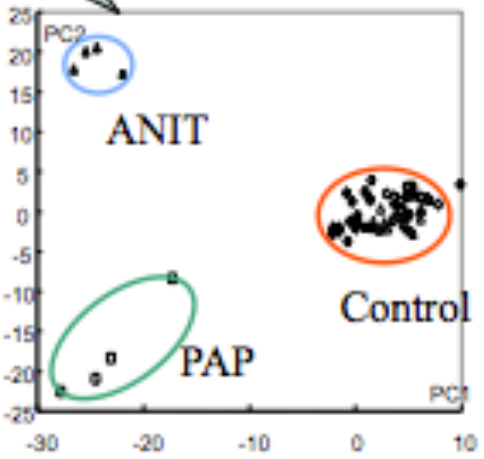
2 routes to Metabolomics



Quantitative (Targeted) Methods

Chemometric (Profiling) Methods





Data processing and annotation

- Preprocessing and the level of annotation is VERY different than in genomic and transcriptomic data

- Many steps in overall experimental design that greatly influence interpretation

- Will briefly cover some of the main issues

Analytical Platform

- Likely GC/LC-MS or NMR as they are the most common
- Choice is normally based more on available equipment, etc. more than experimental design
- GC-MS is an extremely common metabolomics platform, resulting in a high frequency of tools which allow for the direct input of GC-MS spectra.
  - Popularity is due to its relatively high sensitivity, broad range of detectable metabolites, existence of well-established identification libraries and ease of automation
  - separation-coupled MS data requires much processing and careful handling to ensure the information it contains is not artifactual

Targeted vs. Untargeted

- Scientists have been quantifying metabolite levels for over 50 years through targeted analysis…
- With new technologies, the focus can be on untargeted metabolomics
  - Really hard to annotate and interpret
  - Integrated –omics analysis being used to help annotate and understand untargeted metabolites
  - Analogous to candidate gene vs. genome wide testing

Key Issues in Metabolomics

- All of the metabolites within a system cannot be identified with any one analytical method due to chemical heterogeneity, which will cause downstream issues as all metabolites in a pathway have not been quantified
- Not all metabolites have been identified and characterized and so do not exist in the standards libraries, leading to large number of unannotated and/or unknown metabolites of interest
- Organism specific metabolic databases/networks only exist for the highest use model organisms making contextual interpretations difficult for many researchers
- Interpreting the huge datasets of metabolite concentrations under various conditions with biological context is an inherently complex problem requiring extremely in depth knowledge of metabolism.
- The issue of determining which metabolites are actually important in the experimental system in question.

Metabolomic Databases

- Two types of data-bases:
  - top-down (gene to protein to metabolite)
  - bottom-up (chemical entity to biological function) approaches
  - www.metabolomicssociety.org/database

- Most commonly used in biomedical applications:
  - MetaCyc
  - KEGG
    - Subdatabases LIGAND, REACTION PAIR and PATHWAY

Metabolomic Databases

- KEGG and MetaCyc are largest (in terms of number of organisms and most in depth comprehensive (i.e. contains linked information from metabolite to gene)

- Others that are rapidly growing:
  - Reactome (human)
  - KNApSAcK (plants)
  - Model SEED (diverse)
  - BiG [40] (6 model organisms)

  - can be more useful than the large databases if a specific organism is desired

Metabolomic Databases

- KEGG and MetaCyc databases each contain a generalized 'conserved' set of pathways based on metabolic pathways that are more or less the same throughout life in general
  - For KEGG, organism specific annotations are available to query
  - For MetaCyc, individual 'Cyc' databases have been generated for a number of organisms,
    - some just computationally
    - others extensively manually curated such as AraCyc for Arabidopsis

- More recent development are the cheminformatic databases like PubChem
  - provide a chemically ontological approach to cataloguing the ill-defined category of 'small molecules' active in biological systems
  - can provide additional non-biology specific information as well alternative formatting options for datasets *(watch for errors!)*

Enrichment analysis

- These databases are used to create "metabolite sets" for enrichment analysis
- Majority of available tools do early generation over-representation analysis
  - With all the advantages and caveats!
  - For more up to date analysis, will need to work to merge databases, etc. to correctly use more up-to-date approaches

Metabolomics Analysis Tools

- Comprehensive platforms
  - Provide a suite of utilities allowing comprehensive analysis from raw spectral data to pathway analysis
    - MetaboAnalyst
    - MeltDB
- Enrichment Analysis
  - Only works with processed data
    - PAPi
    - MBRole
    - MPEA
    - TICL
    - IMPaLA
- Metabolite Mapping
  - Connects metabolites to genetic/proteomic, etc. resources
    - MetaMapp
    - Masstrix
    - Paintomics
    - VANTED
    - Pathos

Metaboanalyst

- A number of utilities:
  - Data quality checking (useful for batch effects)
  - metabolite ID converter among others are also included.
  - If beginning from raw GC or LC-MS data MetaboAnalyst uses XCMS for peak fitting, identification etc.
  - Once at the peak list (NMR or MS) stage, various preprocessing options such as data-filtering and missing value estimation can be used.
  - A number of normalization, transformation and scaling operations can be performed.
  - Suite of statistical analyses including metabolomics standards like PCA, PLS-DA and hierarchically clustered heatmaps, among many other options.
  - *All these things can be done in other programs, but this is a great tool to get started if you're new to metabolomics!*

Metaboanalyst

- Enrichment Analysis tool of MetaboAnalyst was one of the earliest implementations of GSEA for metabolomics datasets (MSEA)
  - quite biased towards human metabolism unless you make custom background pathways/sets
- Three options for input
  - a single column list of compounds (Over Representation Analysis, ORA)
  - a two column list of compounds AND abundances (Single Sample Profiling, SSP)
  - a multi-column table of compound abundances in classed samples (Quantitative Enrichment Analysis, QEA).

Metaboanalyst

- ORA will calculate whether a particular set of metabolites is statistically significantly higher in the input list than a random list, which can be used to examine ranked or threshold cut-off lists

- SSP is aimed at determining whether any metabolites are above the normal range for common human biofluids

- QEA is the most canonical and will determine which metabolite sets are enriched within the provided class labels, while providing a correlation value and p-value

PAPi

- Pathway Activity Profiling is an R-based tool
- As input it takes a list with abundances (normalized and scaled)
- Works on the assumptions that the detection (i.e. presence in the list) of more metabolites in a pathway and that lower abundances of those metabolites indicates higher flux and therefore higher pathway activity
  - Assumption may not always be true
  - Ex. TCA cycle intermediates can have high abundance even when flux through the reactions in this pathway is also high

PAPi

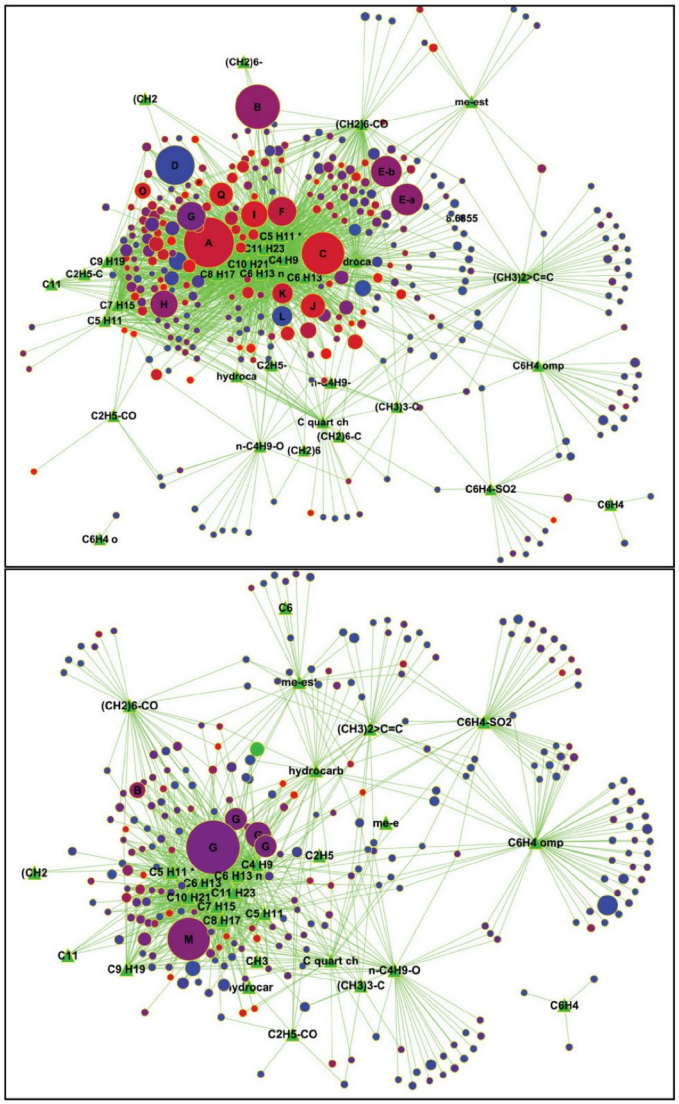
- PAPi calculates an activity score (AS) for each pathway
- The metabolic pathways are taken from the general KEGG database
- The AS indicates the probability of this pathway being active in the cell
- These scores can then be used to compare experimental and control conditions by performing ANOVA or a t-test to compare two sample types.

MetaMapp

- Performs metabolic mapping for unknown and unannotated metabolites
- Since biochemistry is the interconversion of chemically similar entities, compounds can be clustered solely by their chemical similarity
  - Highly beneficial for metabolites without reaction annotation
- Also uses KEGG reactant pair information
  - chemical similarity misclustered some obviously biologically-related metabolites

MetaMapp

- Can also map metabolites based on their mass spectral similarity (for unknowns)
- Can be used to make custom/novel sets for pathway analysis



Summary on Metabolomics Pathway Analysis

- Metabolomics is a maturing area
- “Easy” implementations of tools often behind best practices in pathway approaches
- Issues with time dependencies, tissue dependencies, etc. are more exaggerated in metabolomics
- As the technology is maturing, we are just getting to understand the biases, sources of variation, etc.
  - Data quality control best practices are evolving
  - Will have major impact on the pathway analysis

Specific Issues for other -omics

- Will consider some issues that are both specific to the "-ome" and to particular technologies

- Proteomics
- Epigenomics
- Array CGH data
- RNA seq
- Next generation sequencing
- …….

Proteomics

- After genomics and transcriptomics, proteomics is the next step in central dogma
- Genome is more or less constant, but the proteome differs from cell to cell and from time to time
- Distinct genes are expressed in different cell types, which means that even the basic set of proteins that are produced in a cell needs to be identified

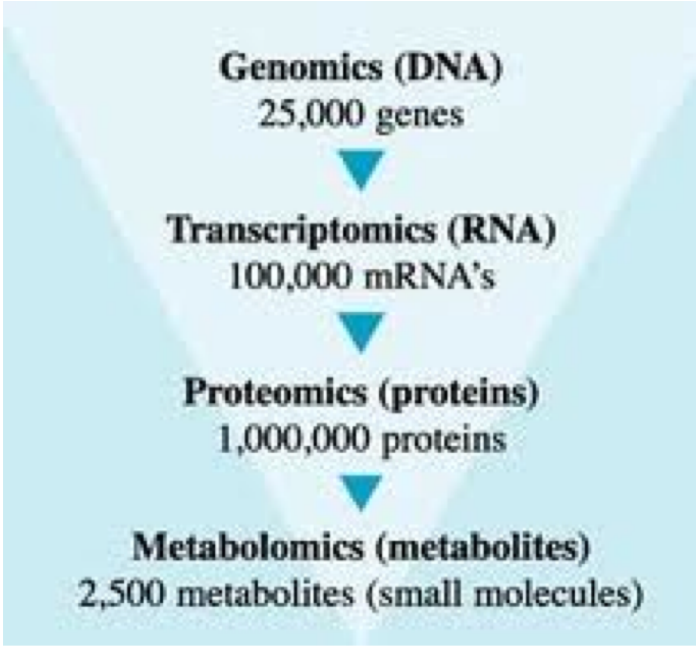
- It was assumed for a long time that microarrays would capture much of this information → NO!

Proteomics vs. Transcriptomics

- mRNA levels do not correlate with protein content
- mRNA is not always translated into protein
- The amount of protein produced for a given amount of mRNA depends on the gene it is transcribed from and on the current physiological state of the cell
- Many proteins are also subjected to a wide variety of chemical modifications after translation
  - Affect function
  - Ex: phosphorylation, ubiquitination
- Many transcripts give rise to more than one protein, through alternative splicing or alternative post-translational modifications

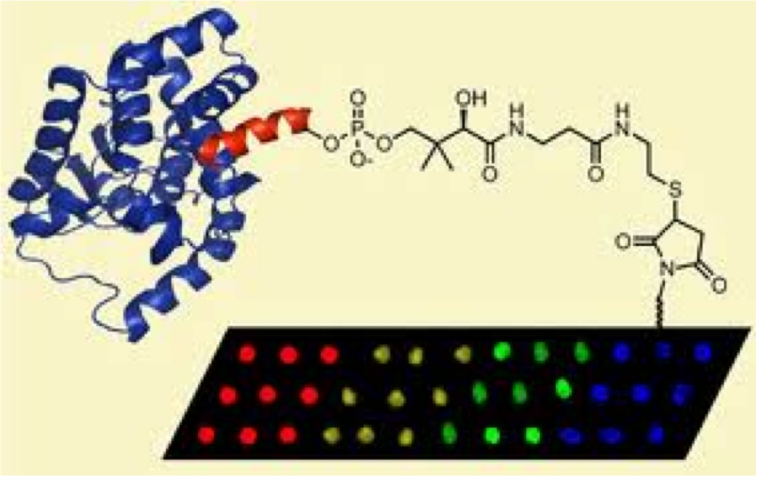
Proteomics

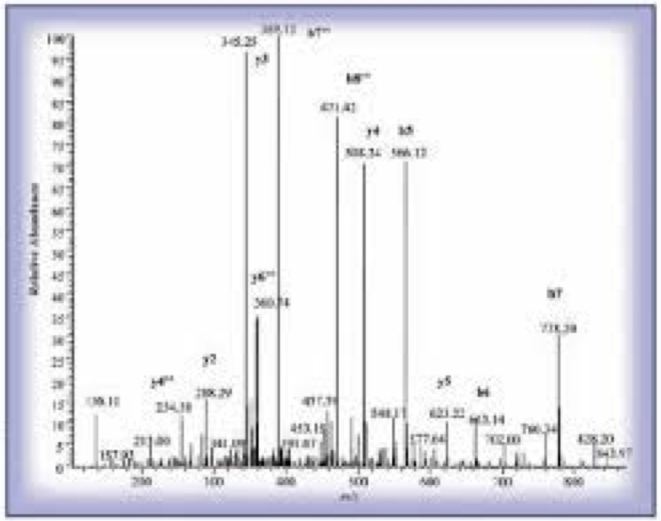
- Technological advances for proteomics has slowed
  - Like metabolomics, the lack of any PCR-like amplification is limited
  - Unlike metabolomics that has a reasonable search space, there estimated to be more than a million transcripts



Proteomics

- Available technologies have different challenges
  - Protein microarrays vs. mass spec based methods
  - General concerns with reproducibility dampened initial excitement





Proteomics

- The high complexity and technical instability mean that the level of annotation is often quite low
- Same challenges as with metabolomics, but more exaggerated given the large annotation space
- Many of the same issues …..

Epigenomics

- “Complete” set of epigenetic modifications on the genetic material of a cell
  - epigenetic modifications are reversible modifications on a cell’s DNA or histones that affect gene expression without altering the DNA sequence
  - DNA methylation and histone modification most commonly assayed

- Rapidly advancing technologies
  - Histone modification assays
  - CHIP-CHIP and CHIP-Seq
  - Methylation arrays

Epigenomics

- Recent studies have focused on issues related to differential numbers of probes in genes
  - Most microarrays were designed with the same number
  - For methylation data, this is not the case, and extreme bias can be seen
  - Bias results in a large number of false positives
- Can be corrected by applying methods that models the relationship between the number of features associated with a gene and its probability of appearing in the foreground list
  - CpG probes in the case of microarrays
  - CpG sites in the case of high-throughput sequencing
  - Chip annotation
- Can also be corrected with careful application of permutation approaches

Next Generation Sequencing

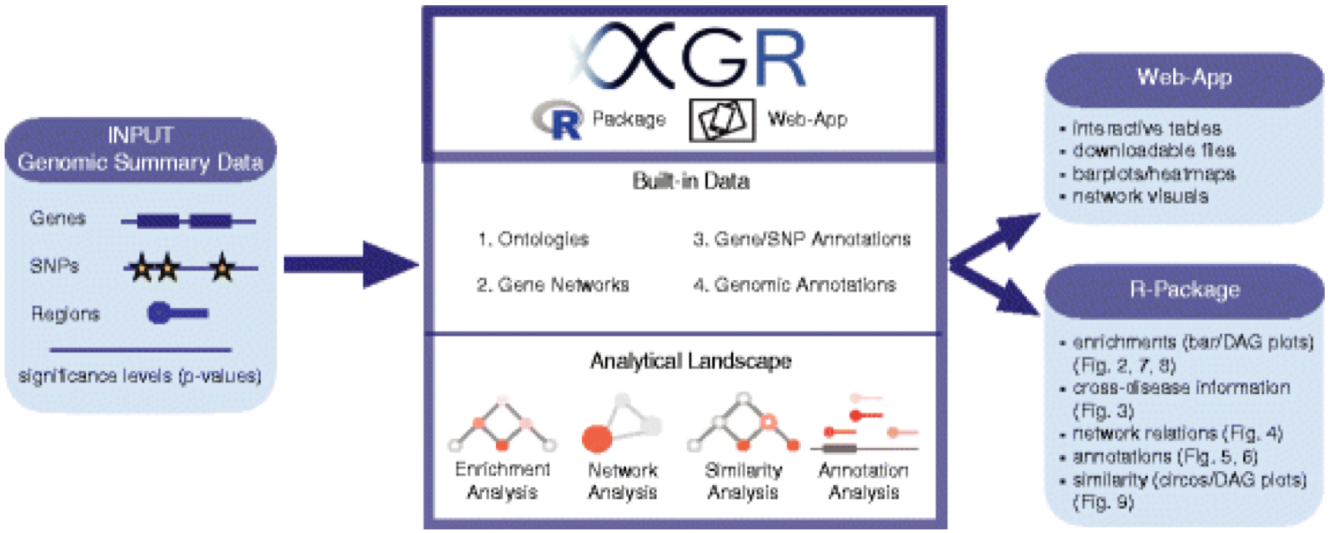
- Variant calling in NGS can detect single nucleotide variants (SNVs) and SNPs
- For SNPs, the exact same pathway methods can be used as designed for GWAS studies (assuming genotyping in genome wide)
- For rare variants, standard approaches are a challenge
  - highly inflated false-positive rates and low power in pathway-based tests of association of rare variants
  - due to their lack of ability to account for gametic phase disequilibrium
  - New area of methods development

Next Generation Sequencing

- RNA-seq data
  - Not truly quantitative
  - With experience, know that there are very different variance distributions at different levels of expression
  - Will matter for methods that test for differences in variance as well as mean
    - Two sided K-S tests….

XGR

- Fang H, Knezevic B, Burnham KL, Knight JC. XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits. Genome Med. 2016 Dec 13;8(1):129.



- Schematic workflow of XGR: achieving enhanced interpretation of genomic summary data. This flowchart illustrates the basic concepts behind XGR. The user provides an input list of either genes, SNPs, or genomic regions, along with their significance levels (collectively referred to as genomic summary data). XGR, available as both an R package and a web-app, is then able to run enrichment, network, similarity, and annotation analyses based on this input. The analyses themselves are run using a combination of ontologies, gene networks, gene/SNP annotations, and genomic annotation data (built-in data). The output comes in various forms, including bar plots, directed acyclic graphs (DAG), circos plots, and network relationships. Furthermore, the web-app version provides interactive tables, downloadable files, and other visuals (e.g. heatmaps)

XGR Functions

Functions	Tasks achieved	Runtime[a]
Enrichment analysis		
xEnricher	A template for enrichment analysis	~40
xEnricherGenes	Gene-based enrichment analysis using a wide variety of ontologies[b]	~40
xEnricherSNPs	SNP-based enrichment analysis using Experimental Factor Ontology on GWAS traits	~70
xEnricherYours	Custom-based enrichment analysis using user-defined ontologies	~5
xEnrichConciser	Removing redundant ones from enrichment outputs	~15
xEnrichBarplot	Barplot of enrichment outputs	<1
xEnrichCompare	Side-by-side barplots of comparative enrichment outputs	<1
xEnrichDAGplot	DAG plot of enrichment outputs	<1
xEnrichDAGplotAdv	DAG plot of comparative enrichment outputs	<1
Annotation analysis		
xGRviaGeneAnno	Annotation analysis using nearby gene annotations by a wide variety of ontologies[b]	~60
xGRviaGenomicAnno	Annotation analysis using a wide variety of genomic annotations[c]	~30
Similarity analysis		
xSocialiser	A template for similarity analysis	~60
xSocialiserGenes	Gene-based similarity analysis using structured ontologies on functions, diseases, and phenotypes	~70
xSocialiserSNPs	SNP-based similarity analysis using Experimental Factor Ontology on GWAS traits	~60
xCircos	Circos plot of similarity outputs	~10
xSocialiserDAGplot	DAG plot of one set of terms used for similarity analysis	<1
xSocialiserDAGplotAdv	DAG plot of two sets of terms used for similarity analysis	<1
Network analysis		
xSubneterGenes	Gene-based network analysis	~60
xSubneterSNPs	SNP-based network analysis	~60
xVisNet	Network visualisation	<1

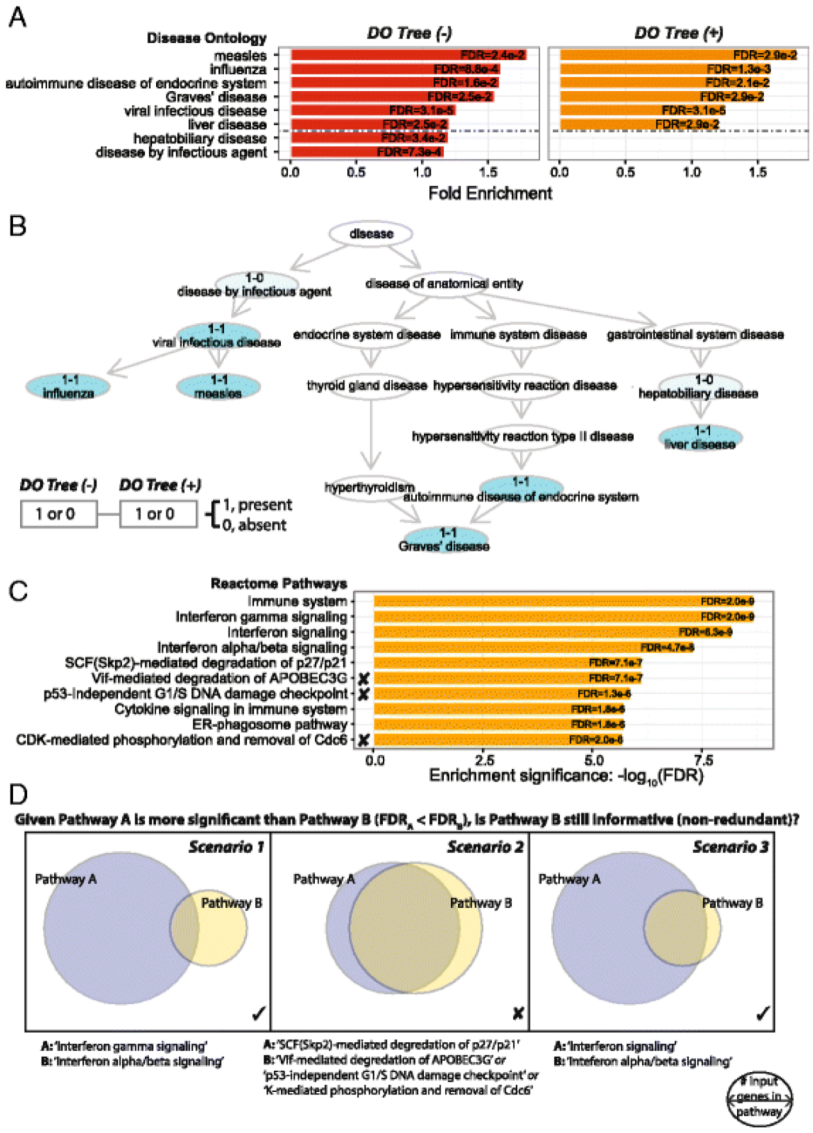


Fig. 2

Necessity of respecting ontology tree-like structure and of removing redundant non-structured pathways in enrichment analysis. This is demonstrated by analysing differentially expressed genes induced by 24-h interferon gamma in monocytes. The effect of taking ontology tree-like structure into account is demonstrated using Disease Ontology (DO) and the removal of redundant non-structured ontologies using Reactome pathways. **a** Side-by-side bar plots comparing the significant DO terms between the analysis without considering the tree structure (*DO Tree(-)*) versus the analysis considering the tree structure (*DO Tree(+)*). The horizontal dotted line separates commonly identified terms (*top section*) and redundant terms in the *DO Tree(-)* analysis. **b** DAG plot comparing commonly identified terms (coloured in *cyan*) and redundant terms from the *DO Tree(-)* analysis (coloured in *light cyan*). The term name (if significant) is prefixed in the form 'x1-x2'. x1 represents '*DO Tree (-)*' and x2 '*DO Tree (+)*'. The value of x1 (or x2) can be '1' or '0', denoting whether this term is identified (present) or not (absent). **c** The top pathway enrichments, with the redundant pathways to be removed indicated (*X*). **d** Illustrations of whether a less significant pathway B is redundant considering a more significant pathway A. Pathway B is counted redundant if it meets both criteria. Criterion 1: more than 90% of input genes annotated with pathway B are also covered by pathway A. Criterion 2: more than 50% of input genes annotated with pathway A are also covered by pathway B. Scenario 1 does not meet either criteria, scenario 2 meets both, and scenario 3 meets criterion 1 but not criterion 2. Notably, criterion 2 ensures the resulting pathways (as shown in scenario 3) are informative in capturing knowledge spheres of different granularities; otherwise, pathway B would be considered redundant in scenario 3, leading to loss of information. *FDR*: false discovery rate

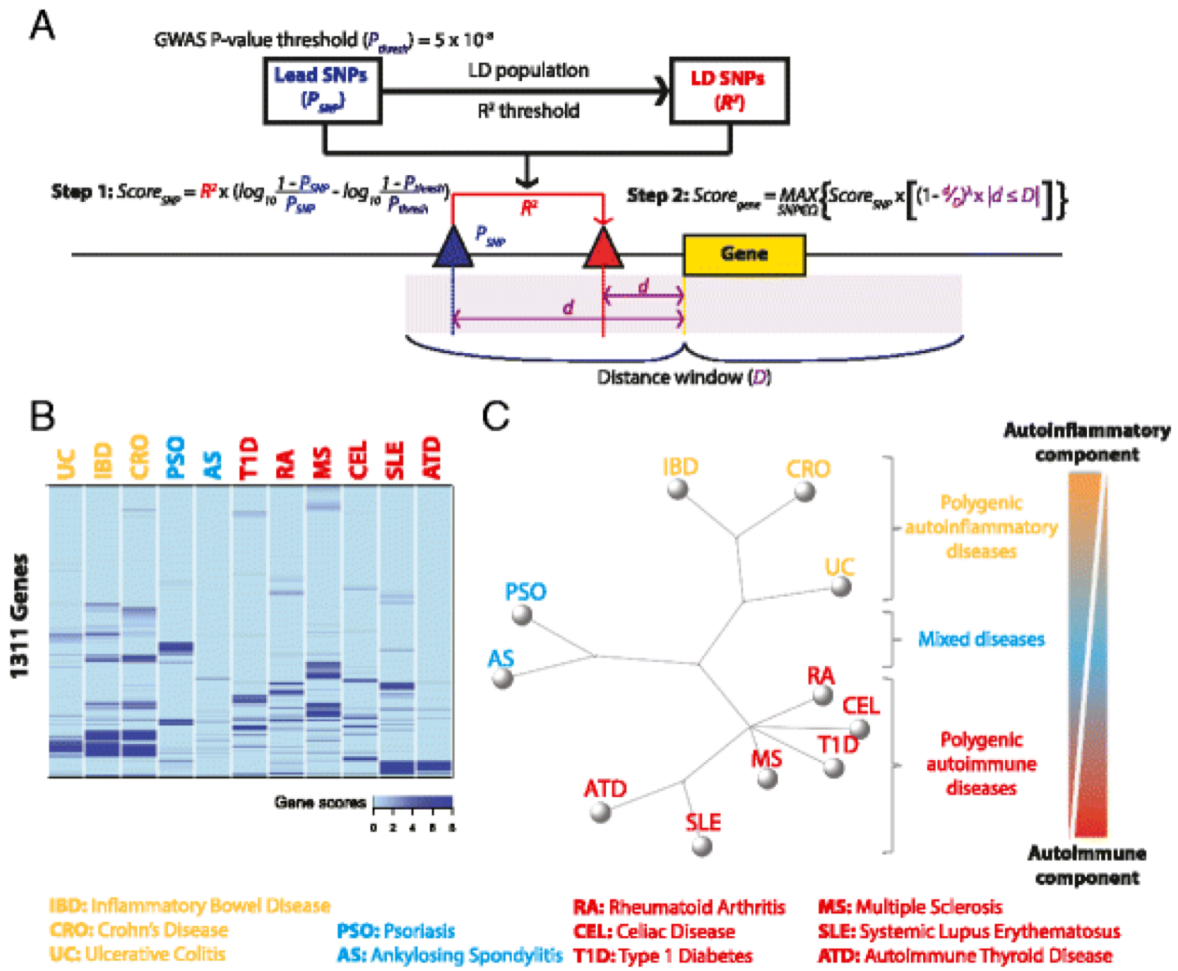


Fig. 3

Informativeness of using cross-disease GWAS summary data in characterising relationships between immunological disorders. **a** Gene scoring from GWAS SNPs prior to network analysis. **b** Heatmap of cross-disease gene scores for 11 common immunological disorders based on ImmunoBase GWAS summary data. **c** Consensus neighbour-joining tree based on the gene-scoring matrix resolves disease classification/taxonomy according to the genetic and cellular basis of autoinflammation and autoimmunity. Subdivided into 1) polygenic autoinflammatory diseases with a prominent autoinflammatory component, 2) polygenic autoimmune diseases with a prominent autoimmune component, and 3) mixed diseases having both components. Inter-disease distance is defined as the cumulative difference in gene scores

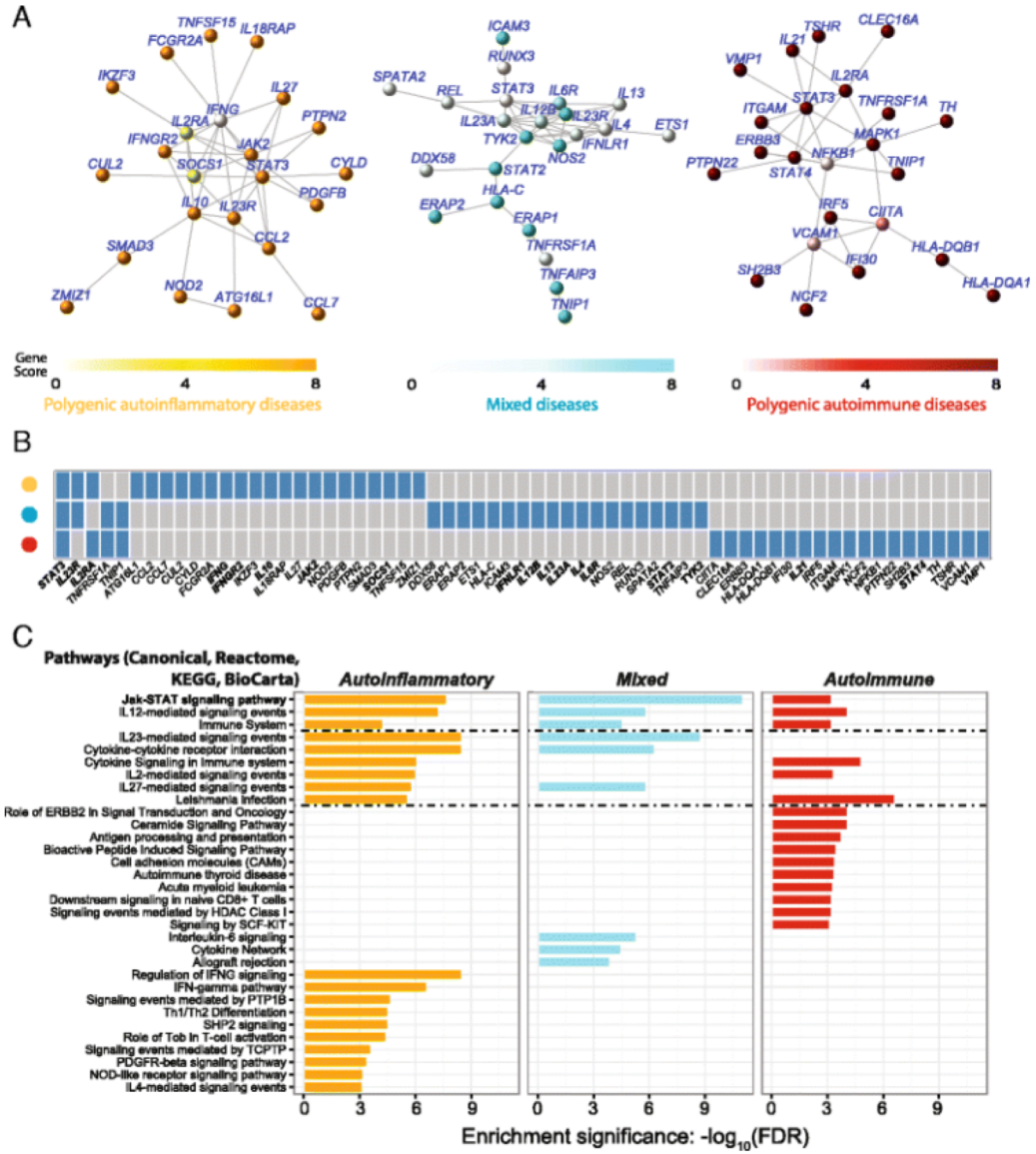


Fig. 4

SNP-modulated gene networks underlying three immunological disease categories. **a** The top-scoring gene network for the three disease categories: autoinflammatory diseases (*orange*), mixed diseases (*cyan*), and autoimmune diseases (*red*). **b** Network genes shared by and unique to disease categories. Genes involved in the Jak-STAT signalling pathway are in *bold text*. **c** Pathway enrichment analysis of network genes using all pathway ontologies and eliminating redundant pathways. The *horizontal dotted line* separates pathways common to all three disease categories (*top section*; e.g. Jak-STAT signalling pathway), those shared by any two categories (*middle*), and those only enriched in one category (*bottom*)

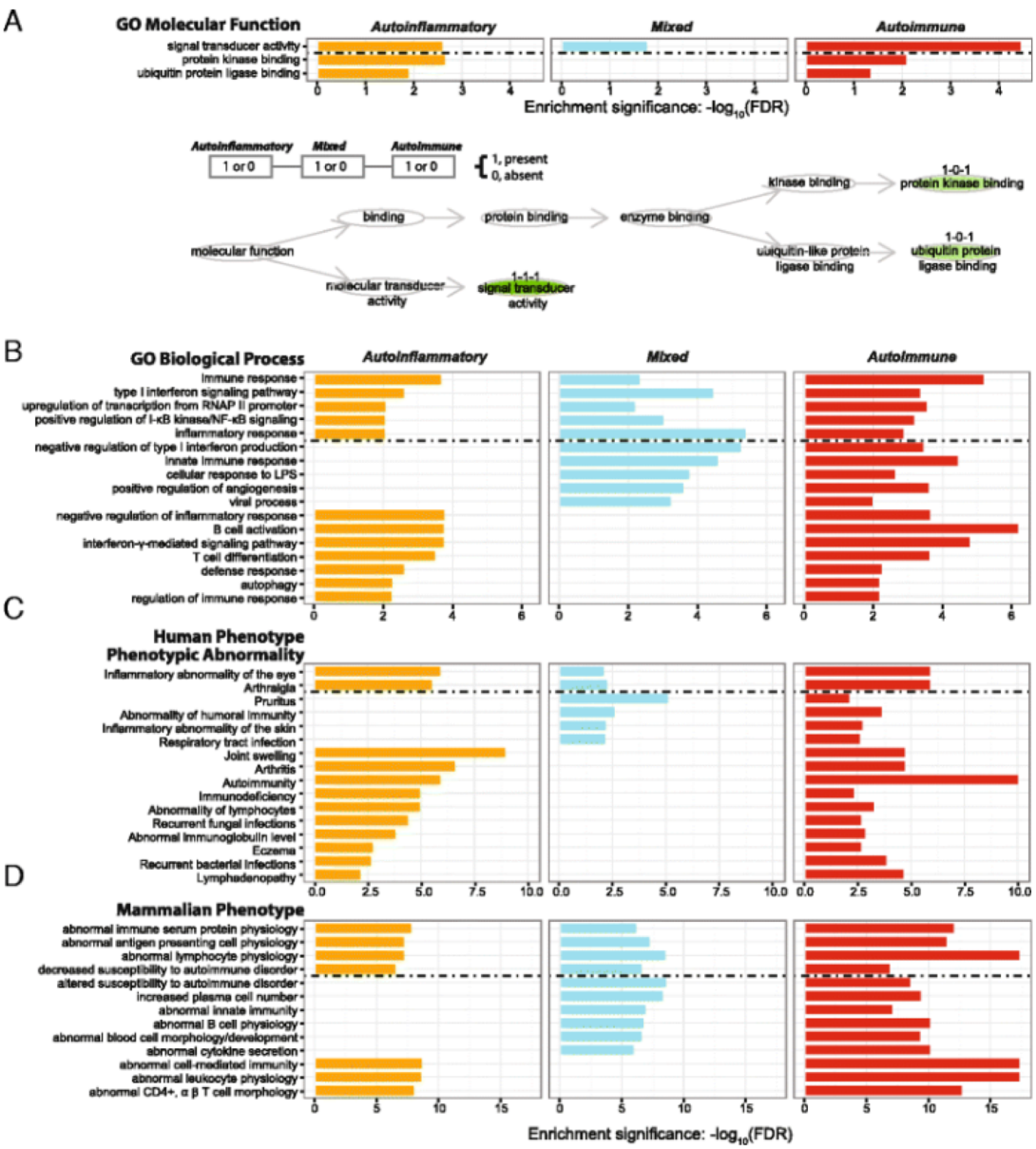


Fig. 5

Functional and phenotypic annotation analysis of genes harbouring GWAS SNPs for three immunological disease categories. Visualised in aside-by-side bar plot and/or DAG plot using functional ontologies, including **a** GO molecular function and **b** GO biological process; and using phenotype ontologies in human and mouse, including **c** human phenotype phenotypic abnormality, and **d** mammalian phenotype

Other Functionalities

- Cross-condition comparative enrichment analysis
- SNP similarity analysis based on disease trait profiles
  - eQTLs
- Epigenetic annotation/enrichment

Summary on Integrated Analysis

- Technology advances across the “omics” is an exciting opportunity for better understanding complexity
- Technologies have unique properties that need to be understood and accounted for in analysis
- Metabolomics resources are rapidly maturing

Summary on Integrated Analysis

- Database development, curation, editing, etc. always lags behind technology
- Issues with incomplete and inaccurate annotation accumulate as more “omes” are considered
- With more complex data, this complexity is not readily captured in the databases the gene set analysis relies on
  - Differences in cell types, exposure, time, etc.
  - Major needs for methods development…..

Questions?