Large-scale functional annotation

Kris Gunsalus

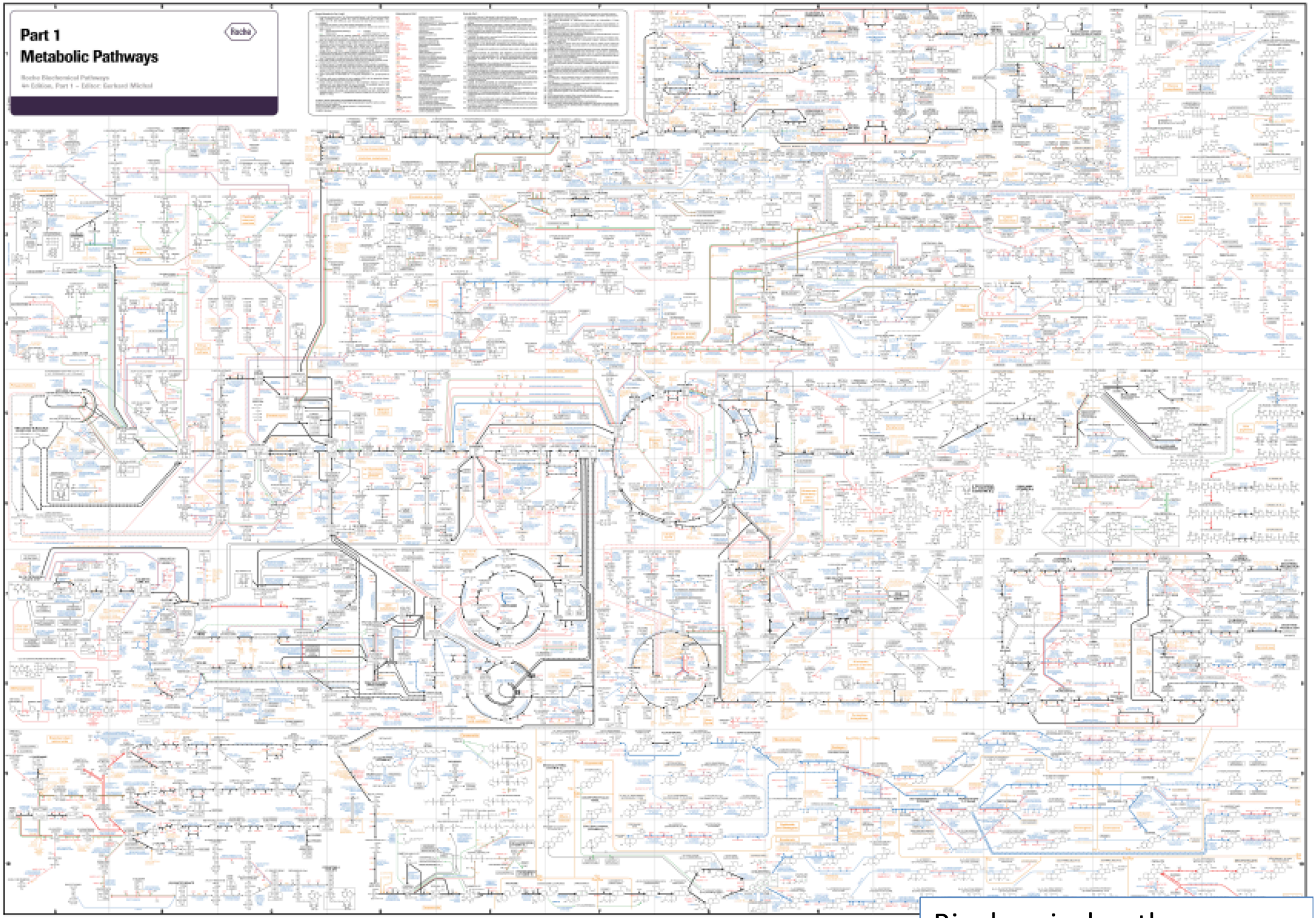
XDAS Fall 2019

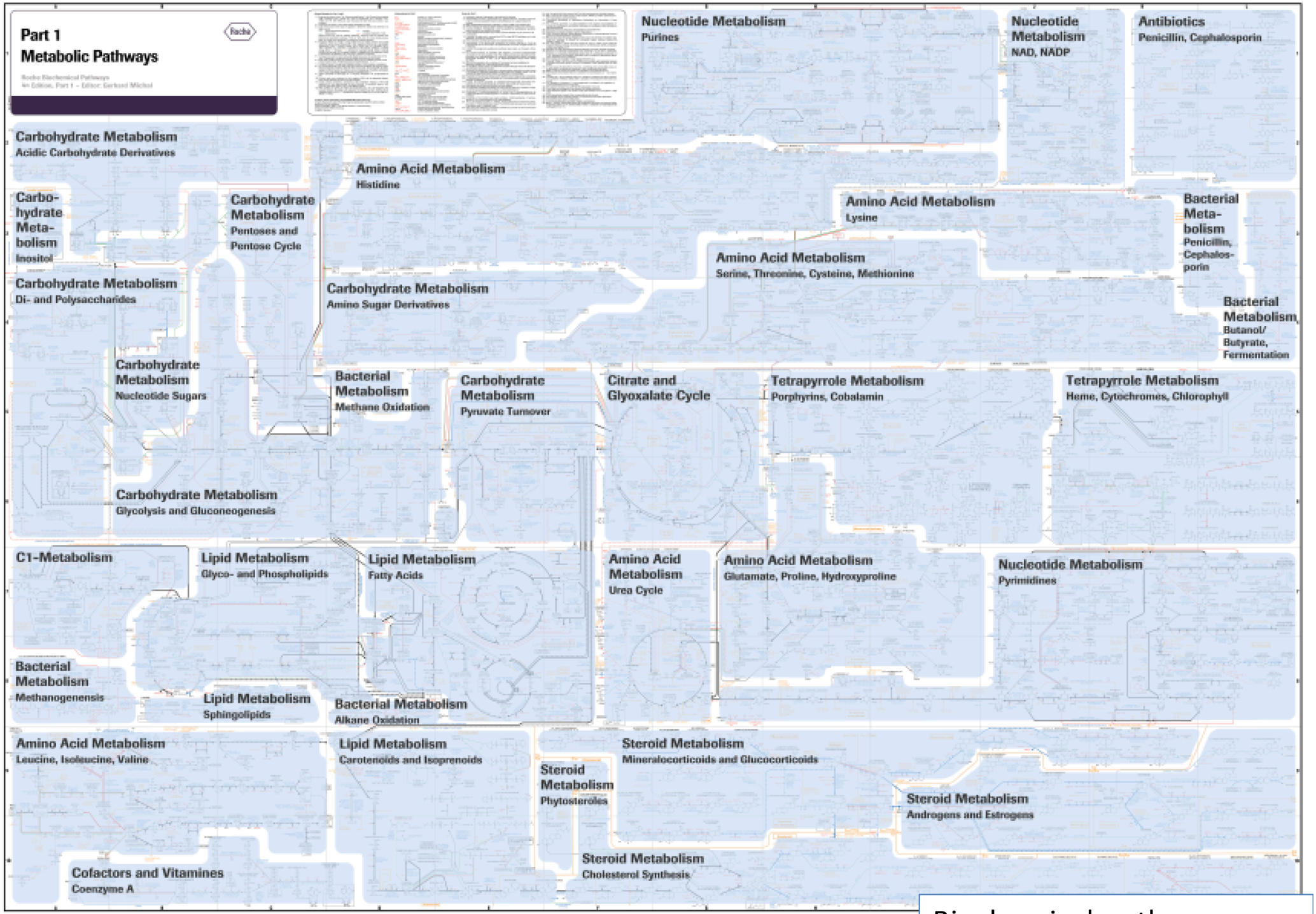
Pathway and network analysis

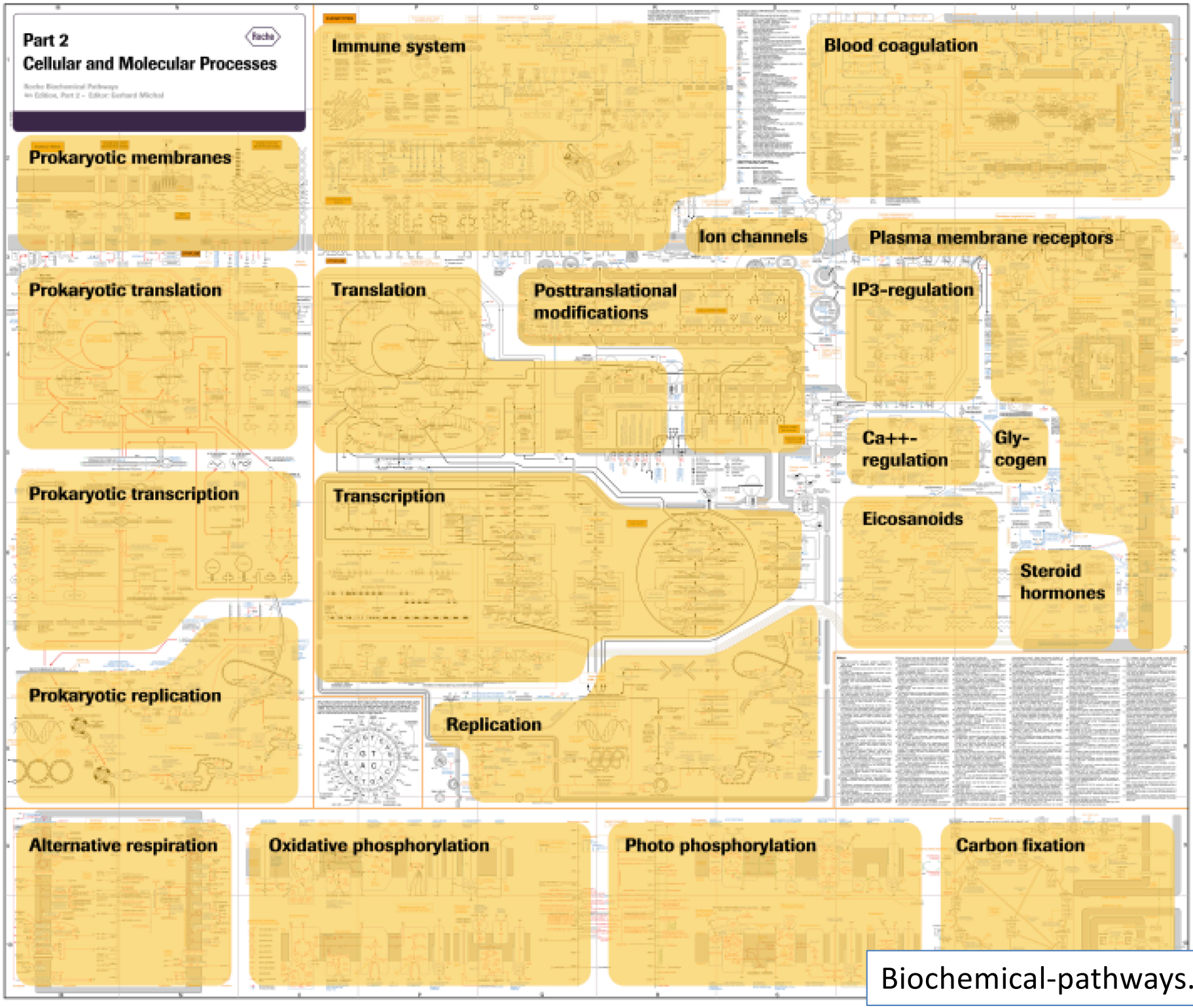
- Approaches to analyze functional enrichment
- Network visualization
- Network inference

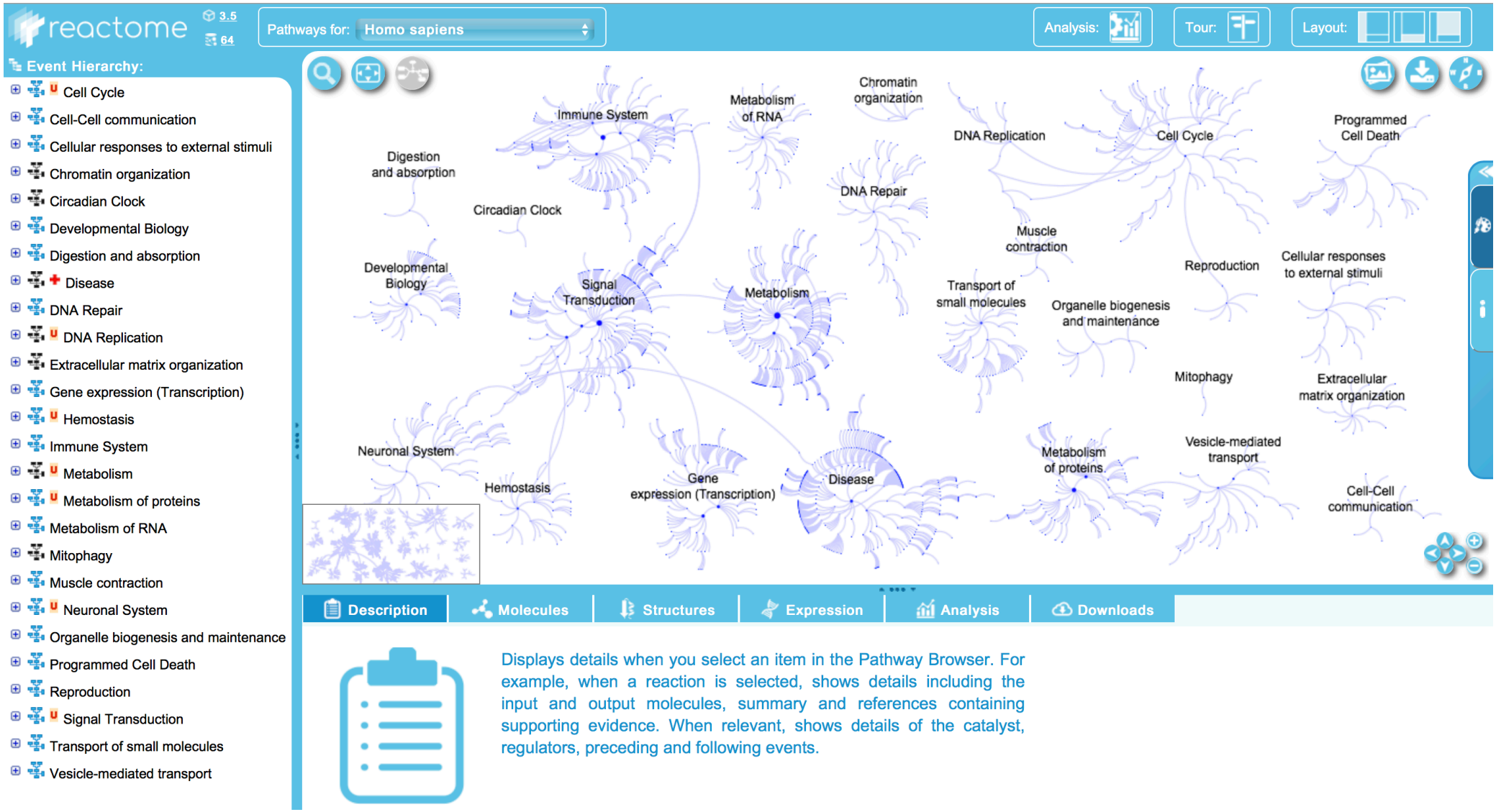
How do we define "function"?

- **KEGG: Kyoto Encyclopedia of Genes and Genomes**
- **Reactome: expert-curated pathways**
- **GO: Gene Ontology**
  - Directed acyclic graph
  - Three ontologies:
    - Biological Process (BP)
    - Cellular Component (CC)
    - Molecular Function (MF)
  - Multiple sources of annotation
    - Biochemistry, genetics, electronically inferred (e.g. from homologs)

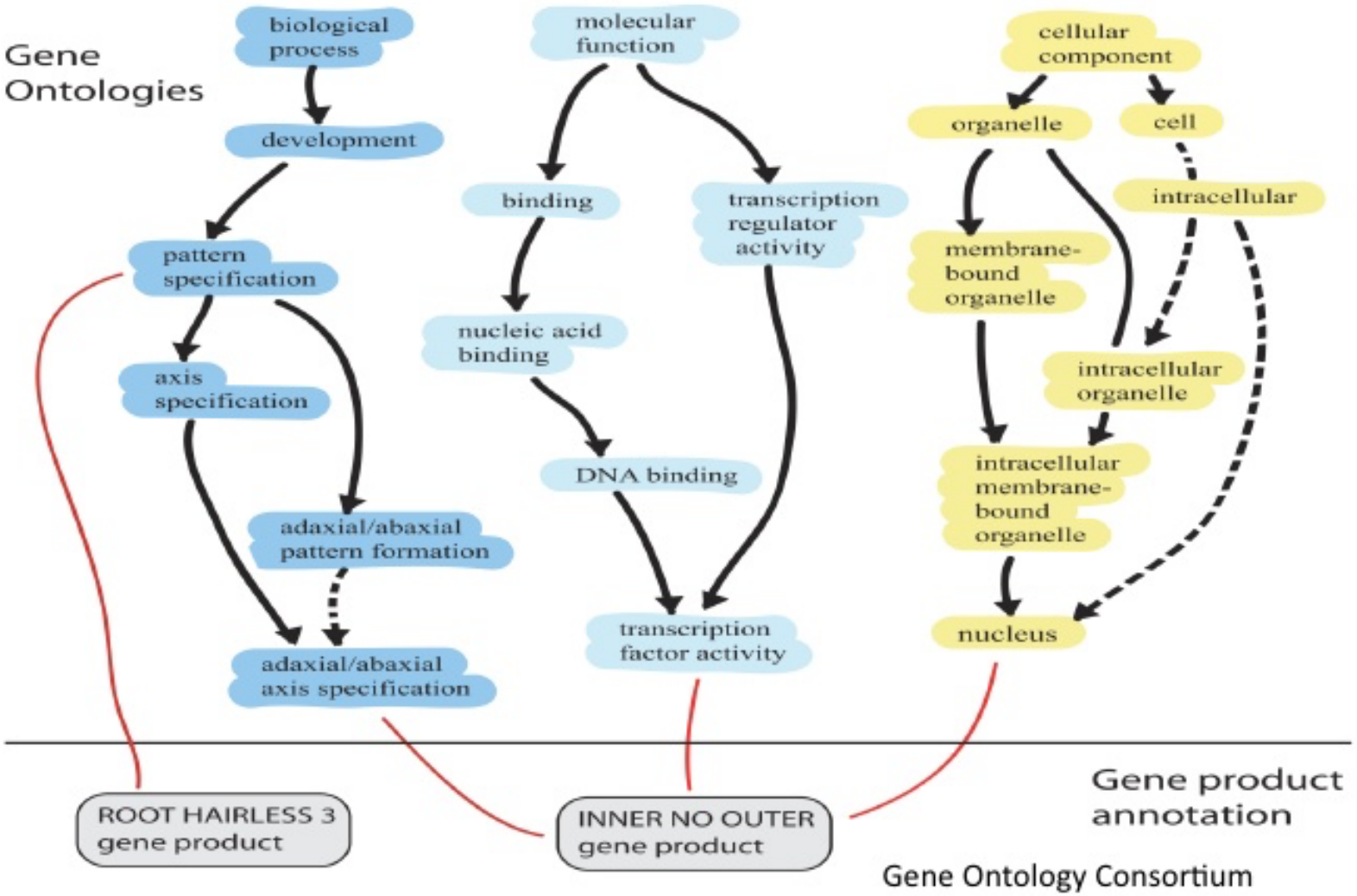

Part 1
Metabolic Pathways
Roche Biochemical Pathways
4th Edition, Part 1 – Editor: Gerhard Michal
Roche
Biochemical-pathways.com



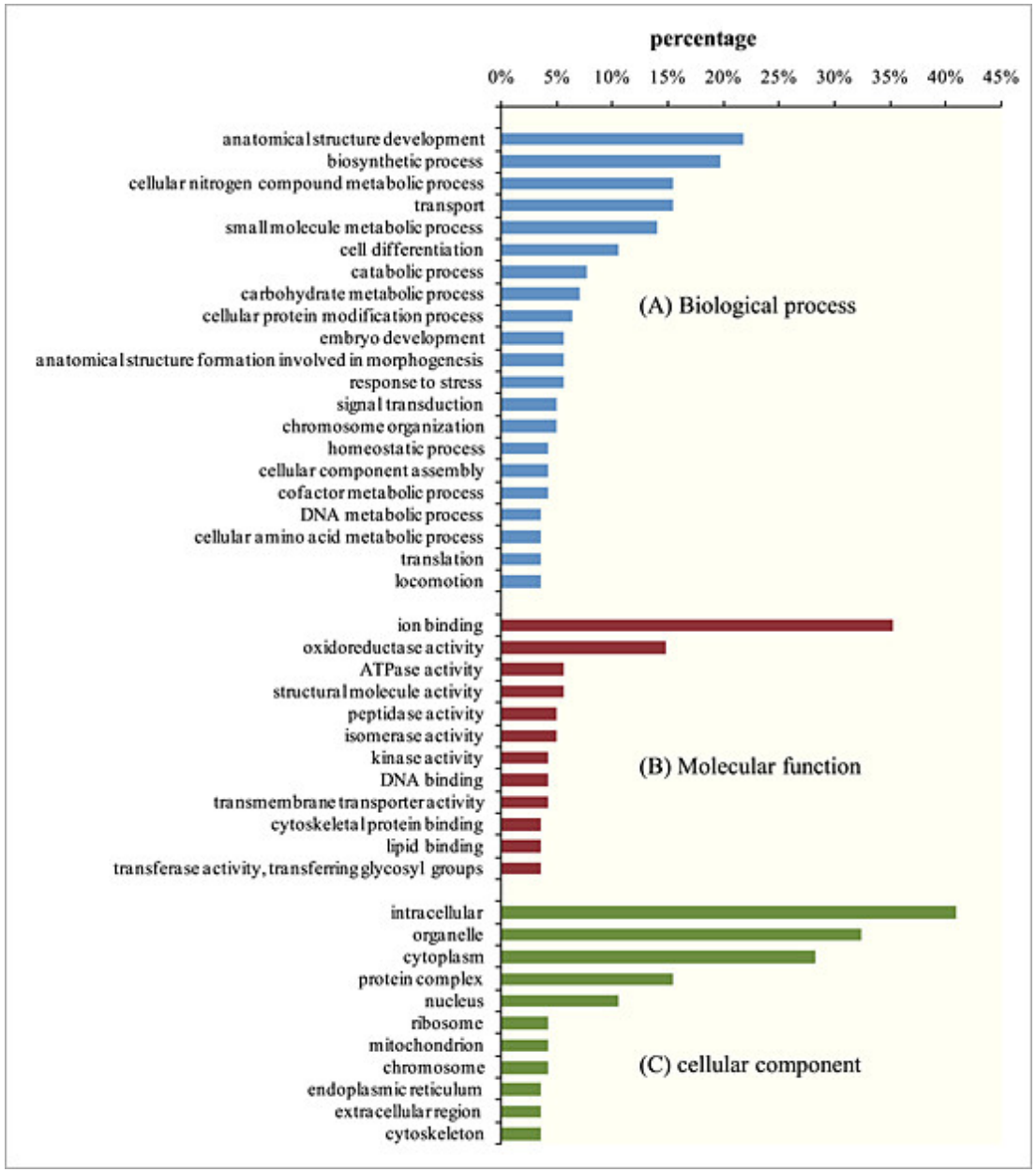

Part 2
Cellular and Molecular Processes
Roche Biochemical Pathways
4th Edition, Part 2 – Editor: Gerhard Michal
Prokaryotic membranes
Immune system
Blood coagulation
Ion channels
Plasma membrane receptors
Prokaryotic translation
Translation
Posttranslational modifications
IP3-regulation
Ca++-regulation
Gly-cogen
Prokaryotic transcription
Transcription
Eicosanoids
Steroid hormones
Prokaryotic replication
Replication
Alternative respiration
Oxidative phosphorylation
Photo phosphorylation
Carbon fixation
Biochemical-pathways.com


reactome
3.5
64
Pathways for: Homo sapiens
Analysis:
Tour:
Layout:
Event Hierarchy:
Cell Cycle
Cell-Cell communication
Cellular responses to external stimuli
Chromatin organization
Circadian Clock
Developmental Biology
Digestion and absorption
Disease
DNA Repair
DNA Replication
Extracellular matrix organization
Gene expression (Transcription)
Hemostasis
Immune System
Metabolism
Metabolism of proteins
Metabolism of RNA
Mitophagy
Muscle contraction
Neuronal System
Organelle biogenesis and maintenance
Programmed Cell Death
Reproduction
Signal Transduction
Transport of small molecules
Vesicle-mediated transport
Immune System
Metabolism of RNA
Chromatin organization
Digestion and absorption
DNA Replication
Cell Cycle
Programmed Cell Death
Circadian Clock
DNA Repair
Muscle contraction
Developmental Biology
Signal Transduction
Metabolism
Transport of small molecules
Reproduction
Cellular responses to external stimuli
Organelle biogenesis and maintenance
Mitophagy
Extracellular matrix organization
Neuronal System
Hemostasis
Gene expression (Transcription)
Disease
Metabolism of proteins
Vesicle-mediated transport
Cell-Cell communication
Description
Molecules
Structures
Expression
Analysis
Downloads
Displays details when you select an item in the Pathway Browser. For example, when a reaction is selected, shows details including the input and output molecules, summary and references containing supporting evidence. When relevant, shows details of the catalyst, regulators, preceding and following events.

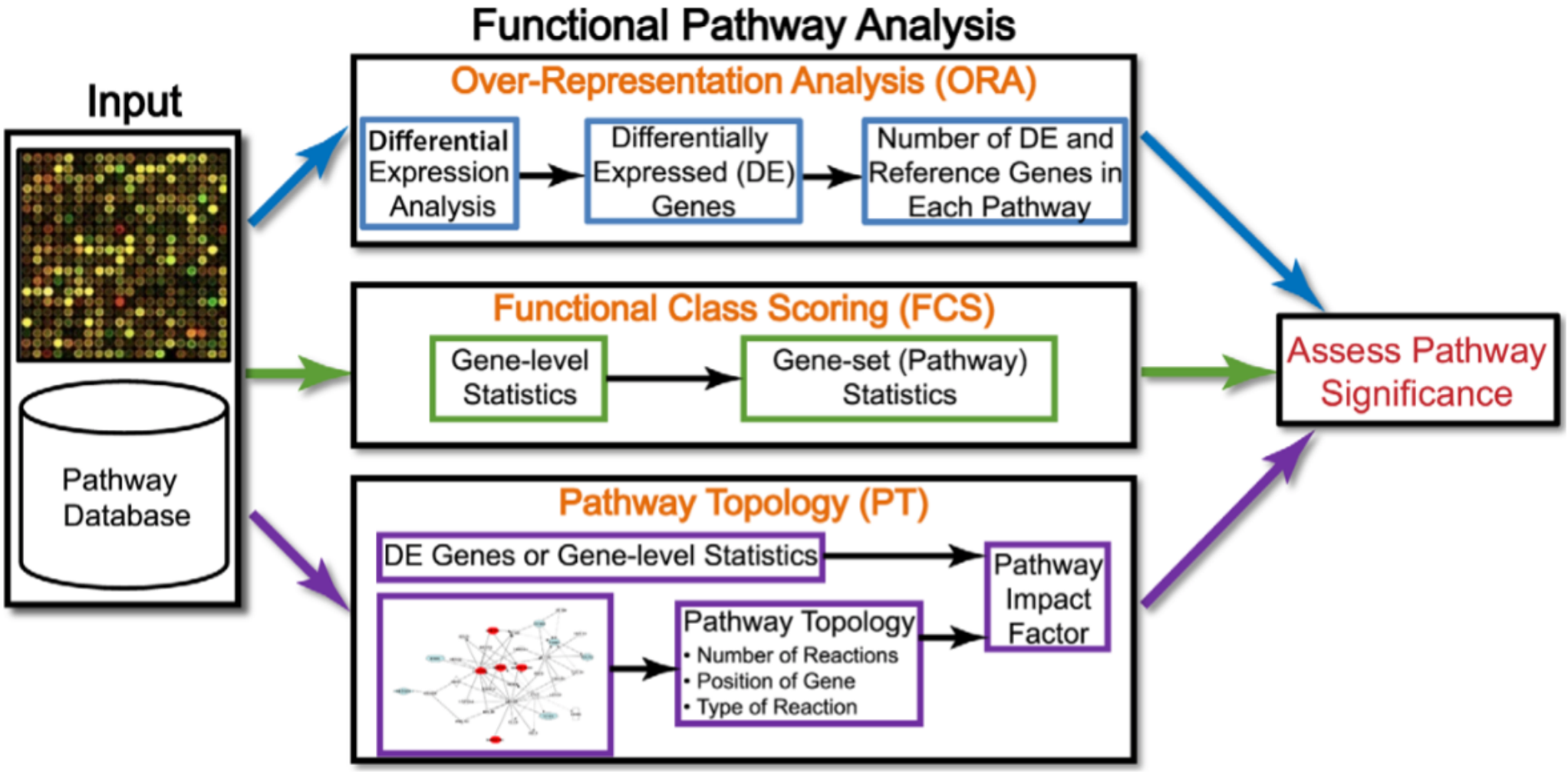
Gene Ontology: example



Gene Ontology: example



Three generations of pathway analysis



⇒ *Large-scale studies often produce* ***lists of genes*** *that are perturbed in some way*

⇒ *Pathway analysis* ***reduces complexity*** *and* ***increases explanatory power***

Khatri et al. (2012) *PLoS Comp Biol* 8:e1002375

Over-representation Analysis (ORA)

- Approach
  - Input list
    - up-or down-regulated genes, RNAi hits, etc.
  - Assess proportion of genes with a particular annotation
    - GO term, pathway membership
  - Compare against universe of all possible genes
  - Assess over- or under-representation
  - "2x2" method: Tabular Statistics
    - hypergeometric, chi-squared, binomial distributions
- Limitations
  - Statistical tests independent of measured changes
  - Uses only most significant genes
  - Assumes genes (and annotations) are independent of each other

Functional Class Scoring (FCS)

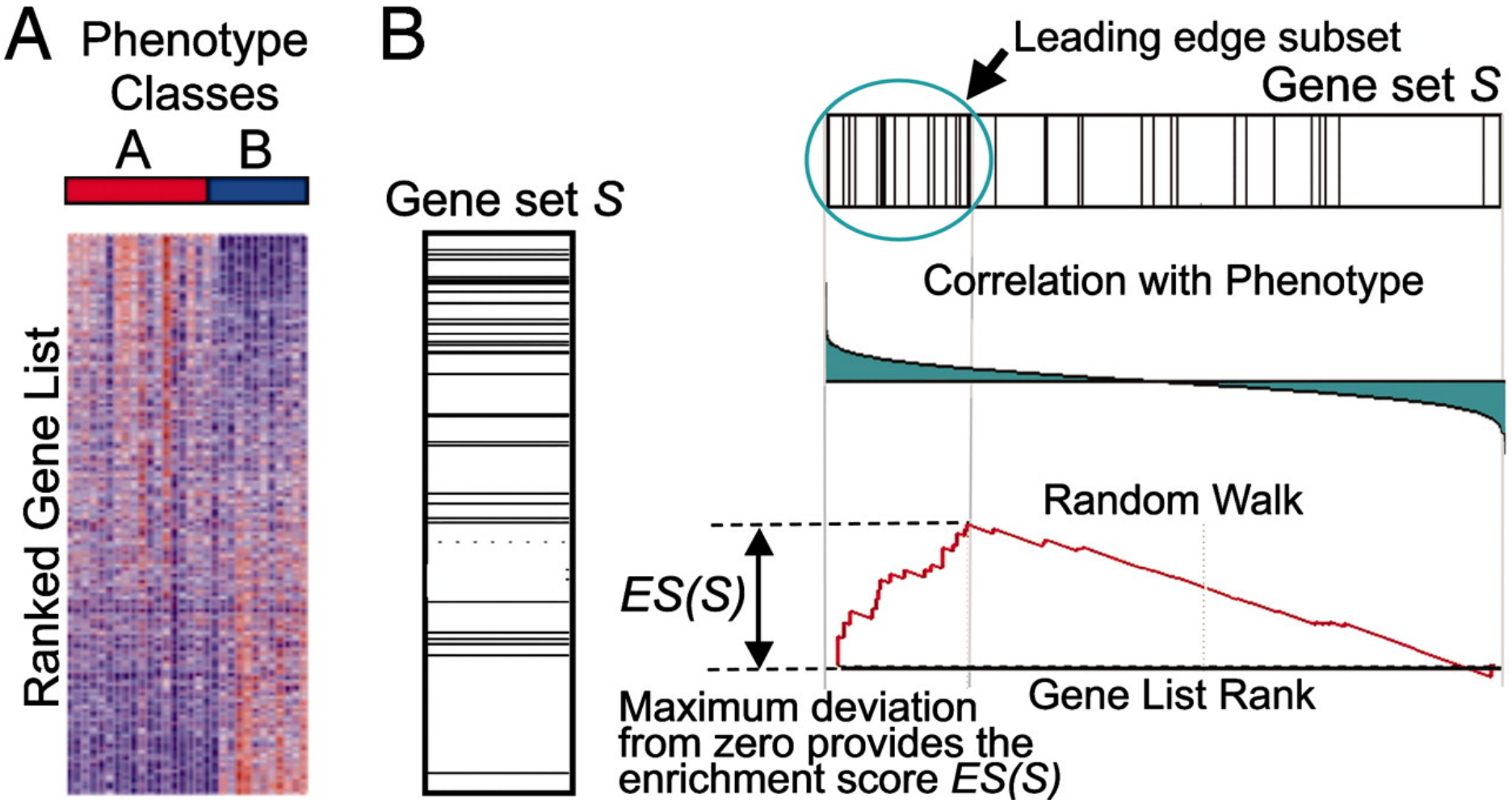
- Approach
  - Try to capture weaker **coordinated changes** that can cumulatively be informative
  - Compute **gene-level statistics**: significance of deviation from background distribution
    - t-test, Z-score, Q-statistic, ANOVA, signal-to-noise, correlation with phenotype
  - Aggregate gene-level statistics into a single **pathway-level statistic**
    - Univariate or multivariate (account for dependencies b/w genes)
    - E.g. KS test, Wilcoxon rank sum (Mann-Whitney), ANCOVA
  - Assess pathway-level **significance**
    - **Competitive** null hypothesis (permute pathway labels)
      vs. **self-contained** null hypothesis (permute class labels)

Functional Class Scoring (FCS)

- Advantages
  - Do not require arbitrary threshold for significance of groups
  - Use all available molecular measurements for pathway analysis
  - Detect coordinated changes in pathway components
  - Account for dependence between genes in pathway
- Limitations
  - Analyzes each pathway independently
  - May use changes to rank genes in a pathway only

(Exception: some methods use sum or mean of gene-level scores to compute pathway-level score)

GSEA: Gene Set Enrichment Analysis



Subramanian et al., *PNAS* 2005

Controlling errors

Table of error types		Null hypothesis (H_0) is	
		True	**False**
Decision About Null Hypothesis (H_0)	**Reject**	Type I error (False Positive)	Correct inference (True Positive)
	Fail to reject	Correct inference (True Negative)	Type II error (False Negative)

Controlling errors

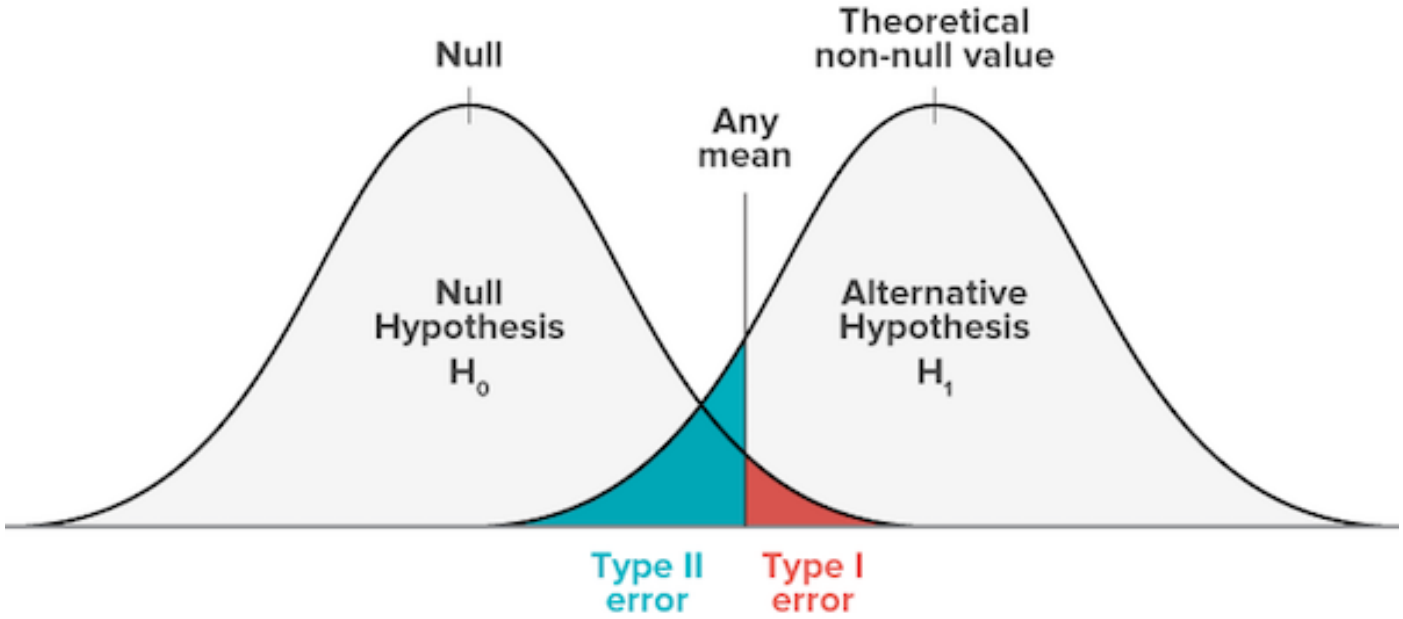
		predicted condition			
	total population	prediction positive	prediction negative	Prevalence = $\frac{\Sigma \text{ condition positive}}{\Sigma \text{ total population}}$	
true condition	condition positive	**True Positive (TP)**	**False Negative (FN)** (type II error)	True Positive Rate (TPR), Sensitivity, Recall, Probability of Detection $= \frac{\Sigma \text{ TP}}{\Sigma \text{ condition positive}}$	False Negative Rate (FNR), Miss Rate $= \frac{\Sigma \text{ FN}}{\Sigma \text{ condition positive}}$
	condition negative	**False Positive (FP)** (Type I error)	**True Negative (TN)**	False Positive Rate (FPR), Fall-out, Probability of False Alarm $= \frac{\Sigma \text{ FP}}{\Sigma \text{ condition negative}}$	True Negative Rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ TN}}{\Sigma \text{ condition negative}}$
	Accuracy $= \frac{\Sigma \text{ TP} + \Sigma \text{ TN}}{\Sigma \text{ total population}}$	Positive Predictive Value (PPV), Precision $= \frac{\Sigma \text{ TP}}{\Sigma \text{ prediction positive}}$	False Omission Rate (FOR) $= \frac{\Sigma \text{ FN}}{\Sigma \text{ prediction negative}}$	Positive Likelihood Ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic Odds Ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$
		False Discovery Rate (FDR) $= \frac{\Sigma \text{ FP}}{\Sigma \text{ prediction positive}}$	Negative Predictive Value (NPV) $= \frac{\Sigma \text{ TN}}{\Sigma \text{ prediction negative}}$	Negative Likelihood Ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	

$\Rightarrow$ *Precision-recall (AUC)*

$\Rightarrow$ *Sensitivity-specificity*

$\Rightarrow$ *Log odds ratio*

Effect size, error rates, power



		Reality	
		Positive	Negative
Study Finding	Positive	**True Positive** (Power) (1-β)	False Positive **Type I Error** (α)
	Negative	False Negative **Type II Error** (β)	**True Negative**

GSEA: Gene Set Enrichment Analysis

- **Step 1: Compute enrichment score (ES)**
  - The degree to which a set *S* is overrepresented at the extremes (top or bottom) of the entire ranked list *L*
  - Procedure: Walk down *L*, increase (decrease) running-sum statistic if gene is (not) in *S*
    - The magnitude of the increment depends on the correlation of the gene with the phenotype
  - ES = maximum deviation from 0
- **Step 2: Estimate significance of ES (P value)**
  - Permutation test: generate null distribution by permuting class (phenotype) labels
    - Permuting class labels preserves gene-gene correlation structure
  - Empirical P-value is observed ES relative to null distribution
- **Step 3: Calculate false discovery rate (FDR)**
  - Normalize ES for each gene set to account for size of set (NES)
  - Compare tails of observed and null distributions

Subramanian et al., *PNAS* 2005

FWER vs FDR

Null hypothesis testing methods adjusted for multiple comparisons (multiple hypothesis testing)

- **Family-wise error rate (e.g. Bonferroni correction):**
  - Control the probability of *at least one* Type I error
  - Most conservative

- **False Discovery rate (Benjamini Hochberg):**
  - Control the expected proportion of rejected null hypotheses that are false (incorrect rejections)
  - Less conservative, more Type I errors
  - Greater power (probability that the test correctly rejects the null hypothesis when the alternative hypothesis is true)

Pathway Topology (PT)

- Approach
  - Use annotation databases that catalog **interactions**
    - Activation, inhibition, physical, metabolic
    - Cellular compartment
  - Same steps as FCS (functional class scoring), BUT use **pathway topology** to compute gene-level statistics
    - Similarity between genes based on **proximity** in network (metabolic networks)
    - E.g. correlation, covariance weighted by graph distance
  - IF: impact factor (signaling pathways)
    - Model signaling pathways as graphs (nodes, edges)
    - Gene-level statistic is PF: perturbation factor
      - Expression AND linear function of PF for all genes in pathway
    - IF is sum of all PFs for pathway

Pathway Topology (PT)

- Limitations
  - Networks are not necessarily the same in each cell
  - Condition-specific changes not known or annotated
  - Do not model dynamic states
  - Do not model interdependence between pathways

⇒ *All methods will improve as resolution, accuracy, and completeness of annotations increase*

PathCORE-T: identifying and visualizing globally co-occurring pathways in large transcriptomic compendia

Kathleen M. Chen[1], Jie Tan[2], Gregory P. Way[1], Georgia Doing[3], Deborah A. Hogan[3] and Casey S. Greene[1*]

* Correspondence: csgreene@upenn.edu
[1]Department of Systems Pharmacology and Translational Therapeutics. Perelman School of Medicine, University of Pennsylvania, 3400 Civic Center Blvd., Philadelphia, PA 19104, USA
Full list of author information is available at the end of the article

Abstract

Background: Investigators often interpret genome-wide data by analyzing the expression levels of genes within pathways. While this within-pathway analysis is routine, the products of any one pathway can affect the activity of other pathways. Past efforts to identify relationships between biological processes have evaluated overlap in knowledge bases or evaluated changes that occur after specific treatments. Individual experiments can highlight condition-specific pathway-pathway relationships; however, constructing a complete network of such relationships across many conditions requires analyzing results from many studies.

Results: We developed PathCORE-T framework by implementing existing methods to identify pathway-pathway transcriptional relationships evident across a broad data compendium. PathCORE-T is applied to the output of feature construction algorithms; it identifies pairs of pathways observed in features more than expected by chance as *functionally co-occurring*. We demonstrate PathCORE-T by analyzing an existing eADAGE model of a microbial compendium and building and analyzing NMF features from the TCGA dataset of 33 cancer types. The PathCORE-T framework includes a demonstration web interface, with source code, that users can launch to (1) visualize the network and (2) review the expression levels of associated genes in the original data. PathCORE-T creates and displays the network of globally co-occurring pathways based on features observed in a machine learning analysis of gene expression data.

Conclusions: The PathCORE-T framework identifies transcriptionally co-occurring pathways from the results of unsupervised analysis of gene expression data and visualizes the relationships between pathways as a network. PathCORE-T recapitulated previously described pathway-pathway relationships and suggested experimentally testable additional hypotheses that remain to be explored.

Keywords: Gene expression, Unsupervised feature construction, Crosstalk, Pathway interactions

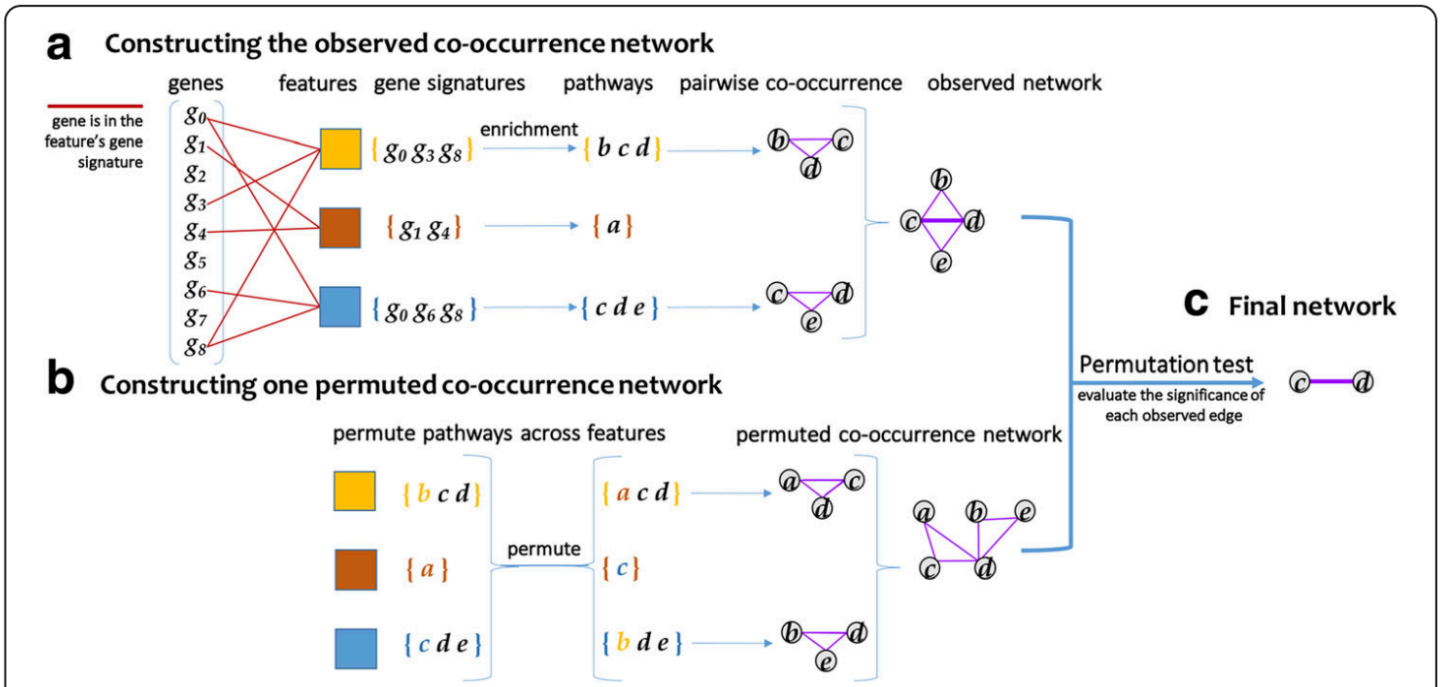
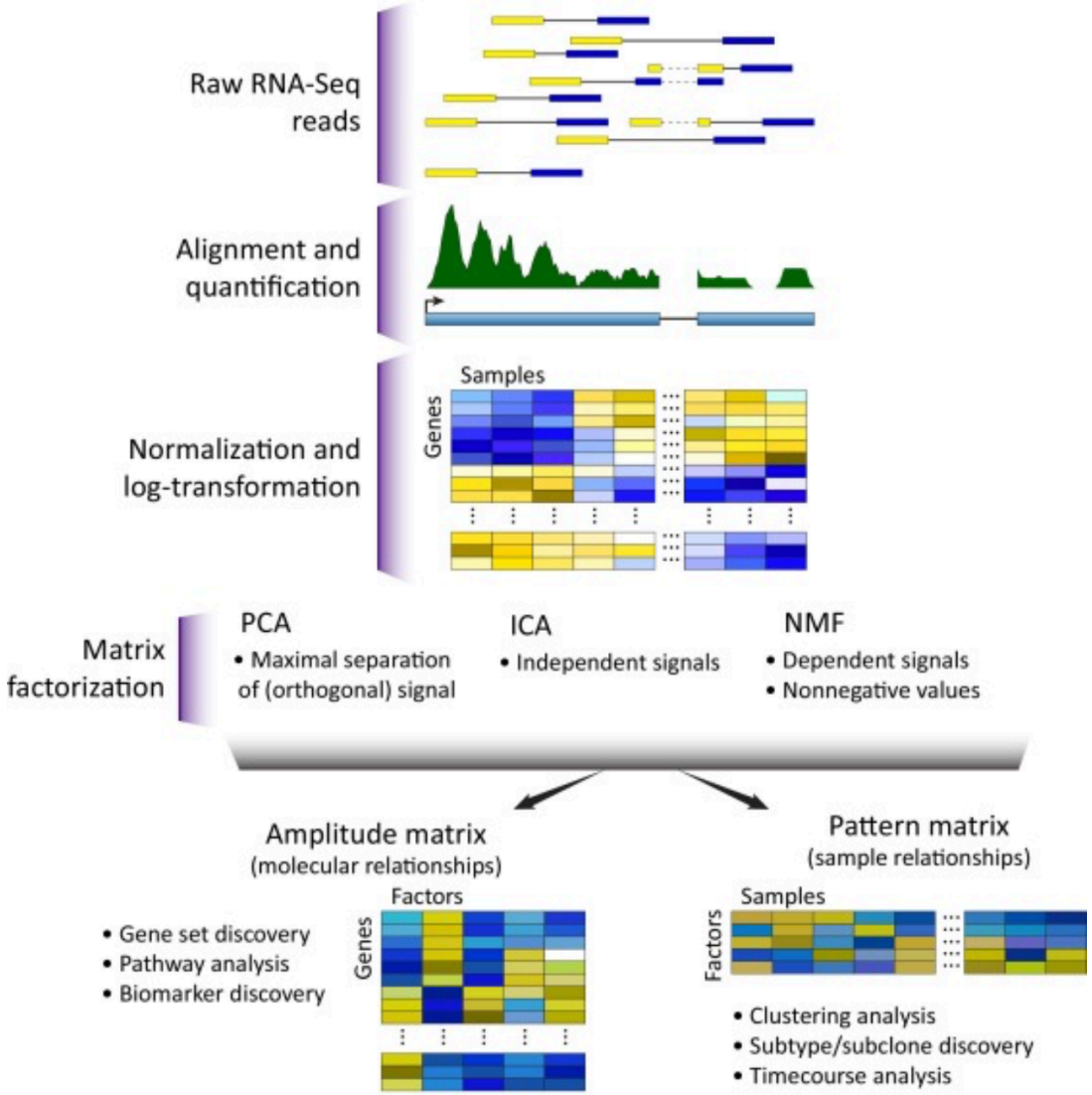


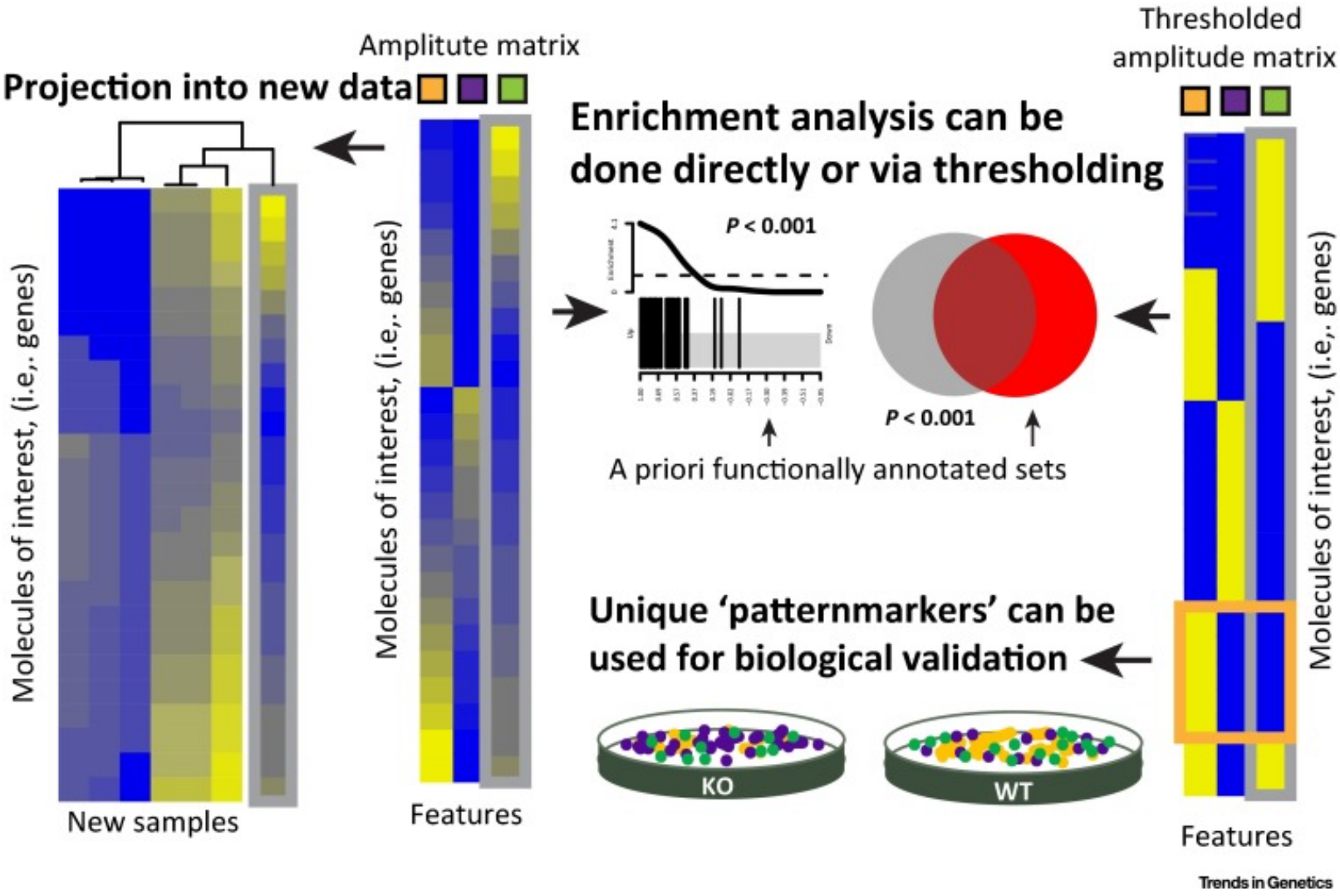
Fig. 2 The approach implemented in PathCORE-T to construct a pathway co-occurrence network from an expression compendium. **a** A user-selected feature extraction method is applied to expression data. Such methods assign each gene a weight, according to some distribution, that represents the gene's contribution to the feature. The set of genes that are considered highly representative of a feature's function is referred to as a feature's gene signature. The gene signature is user-defined and should be based on the weight distribution produced by the unsupervised method of choice. In the event that the weight distribution contains both positive and negative values, a user can specify criteria for both a positive and negative gene signature. A test of pathway enrichment is applied to identify corresponding sets of pathways from the gene signature(s) in a feature. We consider pathways significantly overrepresented in the same feature to co-occur. Pairwise co-occurrence relationships are used to build a network, where each edge is weighted by the number of features containing both pathways. **b** *N* permuted networks are generated to assess the statistical significance of a co-occurrence relation in the graph. Here, we show the construction of one such permuted network. Two invariants are maintained during a permutation: (1) pathway side-specificity (if applicable, e.g. positive and negative gene signatures) and (2) the number of distinct pathways in a feature's gene signature. **c** For each edge observed in the co-occurrence network, we compare its weight against the weight distribution generated from *N* (default: 10,000) permutations of the network to determine each edge's *p*-value. After correcting the *p*-value by the number of edges observed in the graph using the Benjamini—Hochberg procedure, only an edge with an adjusted *p*-value below alpha (default: 0.05) is kept in the final co-occurrence network.

Enrichment analysis: feature extraction by matrix factorization



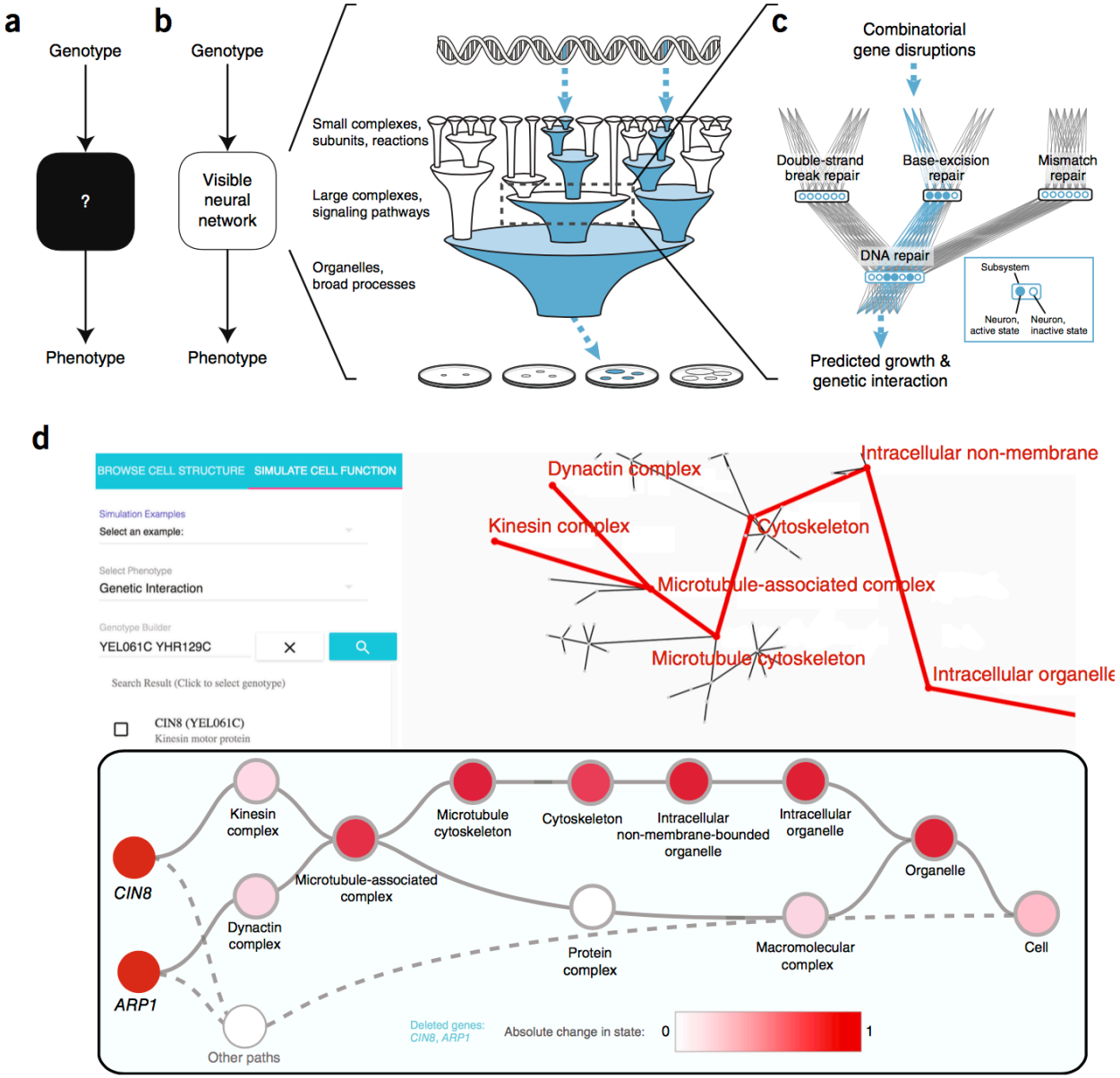
Stein-O'Brien et al., Trends Genet 2018
https://doi.org/10.1016/j.tig.2018.07.003

Enrichment analysis: feature extraction by matrix factorization



Stein-O'Brien et al., Trends Genet 2018
https://doi.org/10.1016/j.tig.2018.07.003

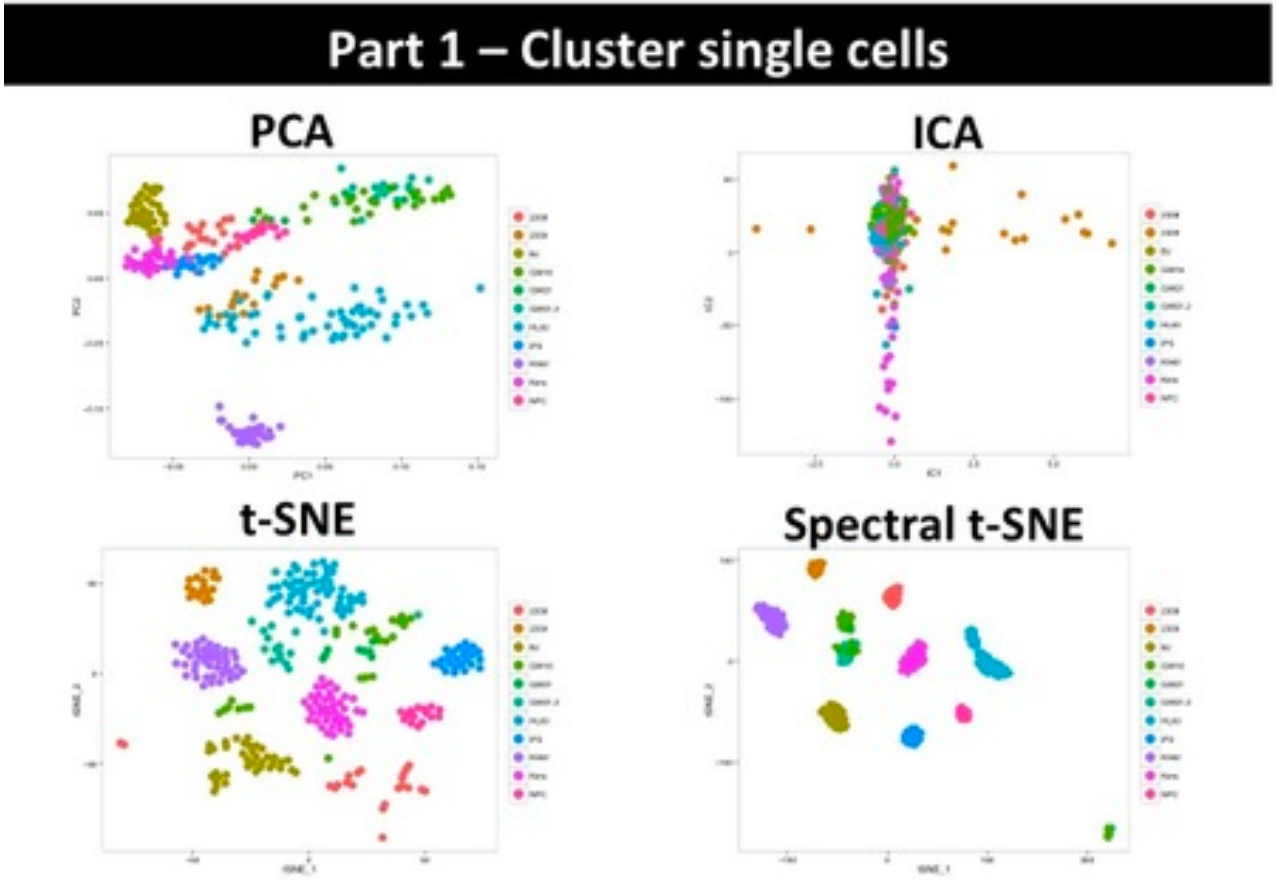
Dcell: Visual neural network for prediction of phenotypes and GIs



Ma et al, *Nature* 2018 (Ideker lab)

A really nice introduction to neural networks in 4 short videos

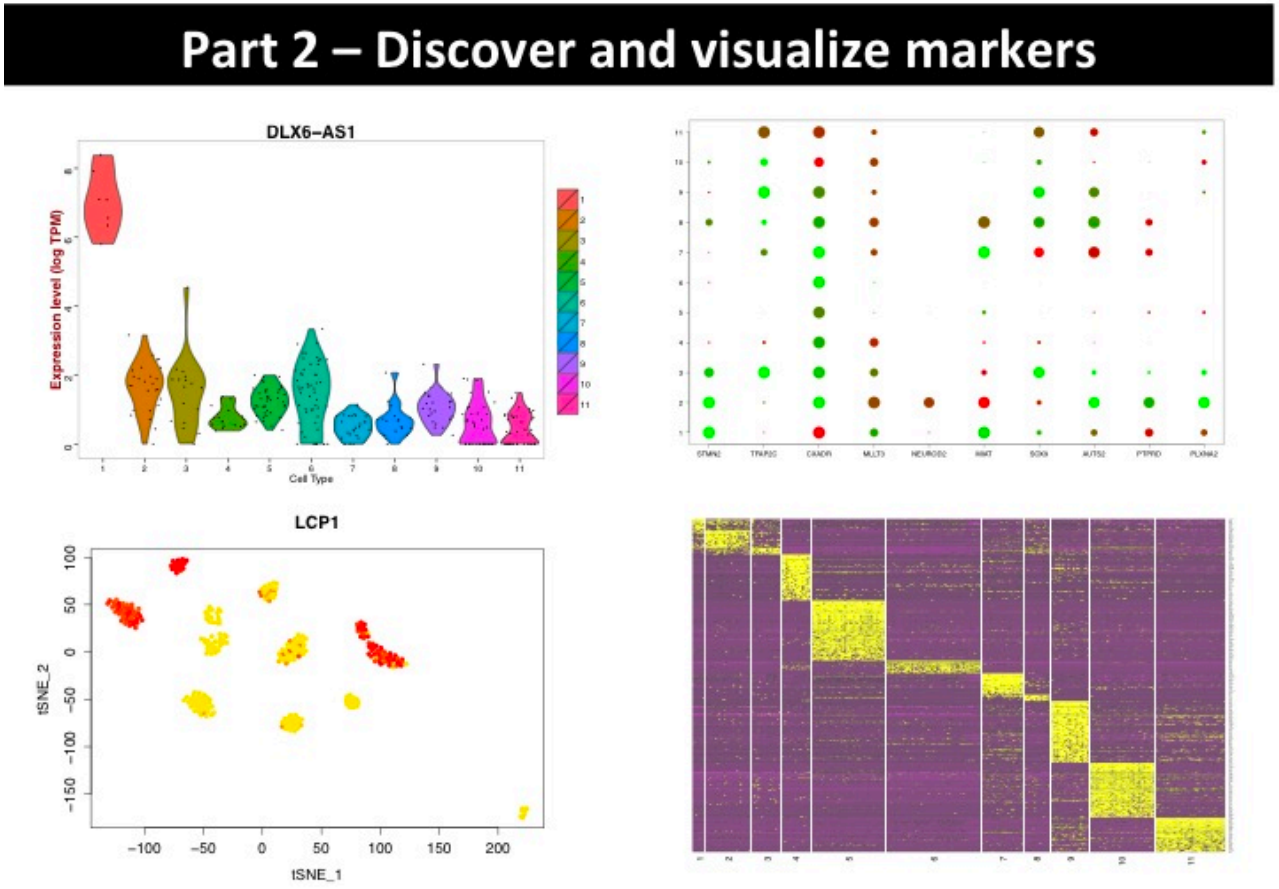
Enrichment Analysis: Seurat



⇒*Now use UMAP*

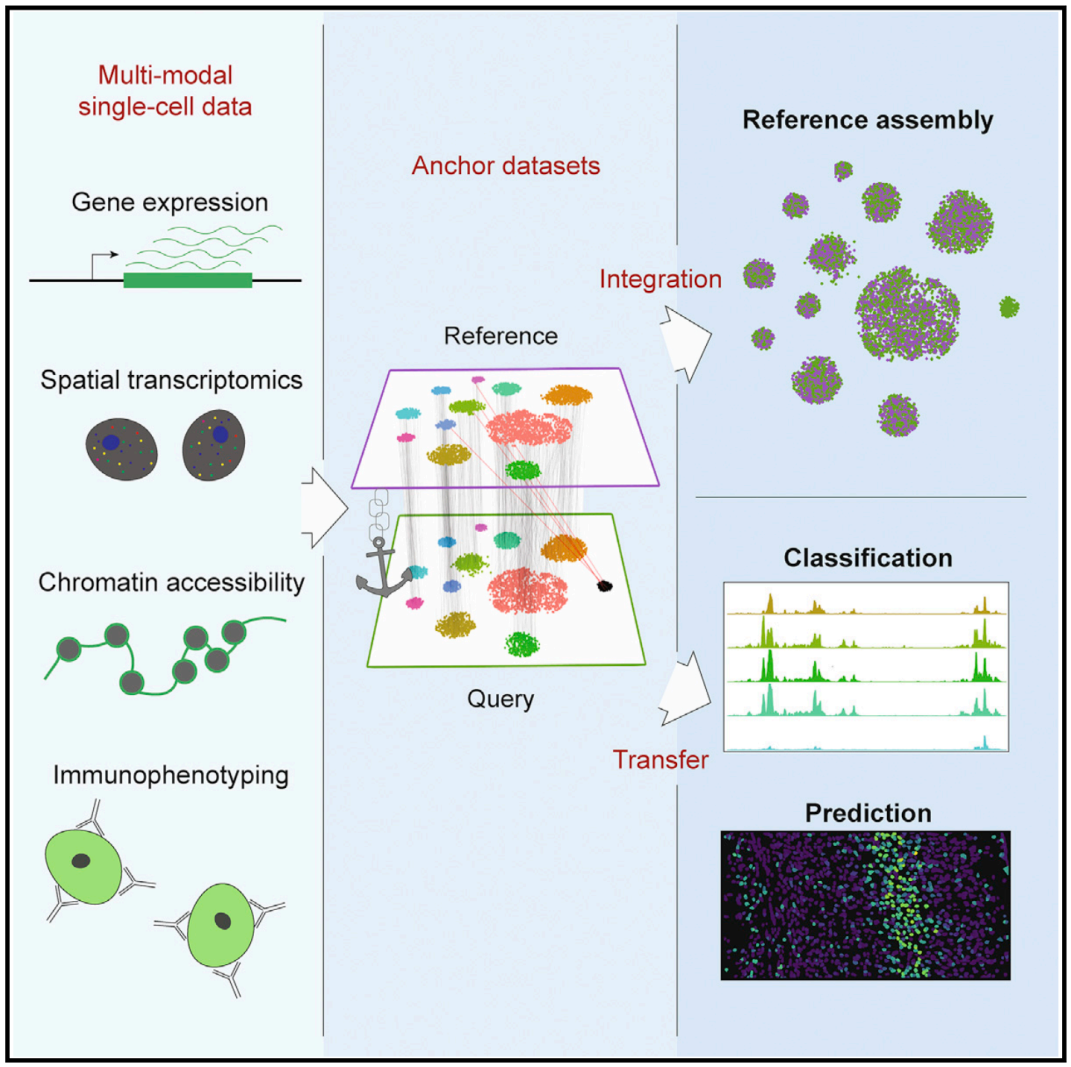
Stuart, Butler et al. *Cell* 2019

Enrichment Analysis: Seurat



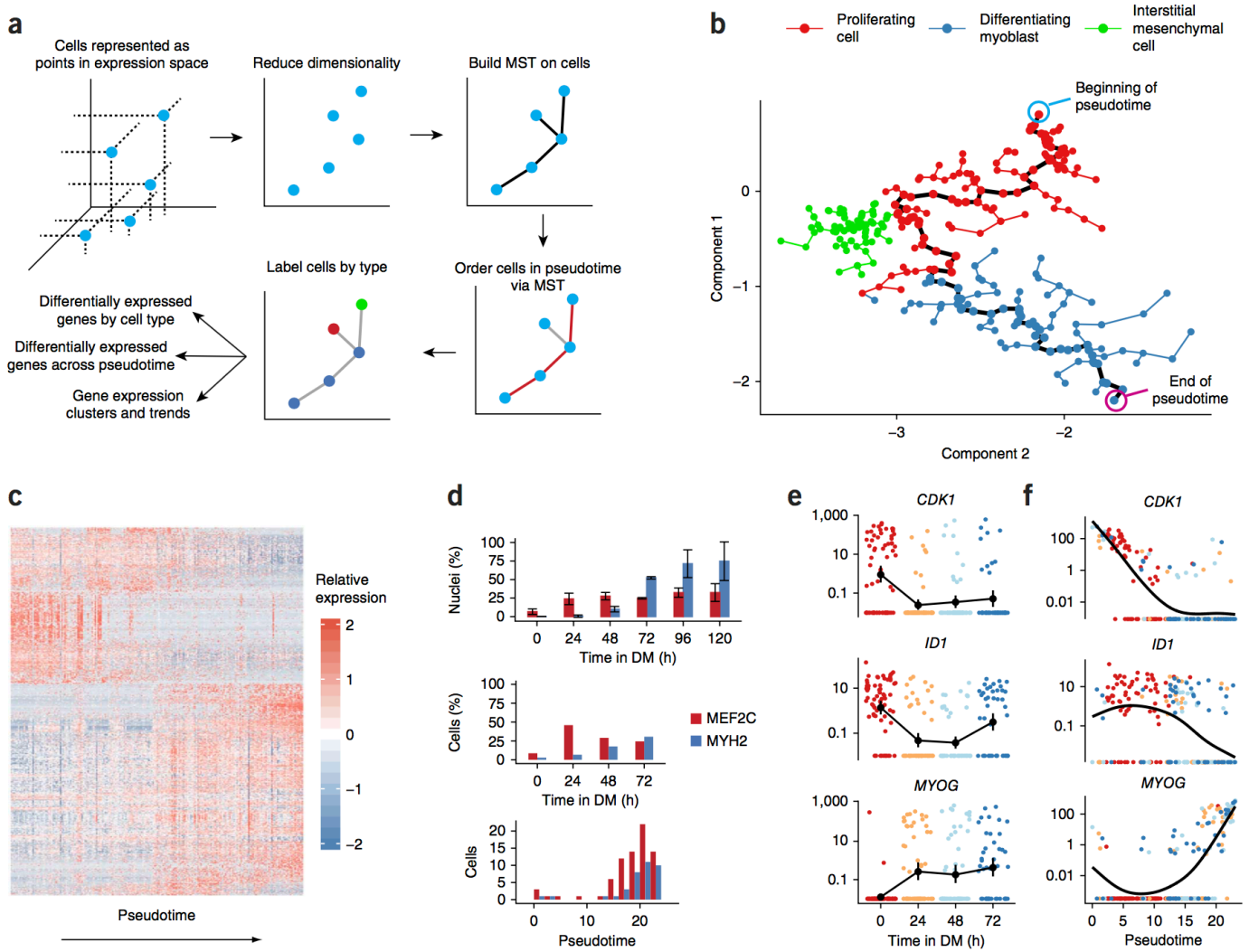
Stuart, Butler et al. *Cell* 2019

Enrichment Analysis: Seurat



Stuart, Butler et al. *Cell* 2019

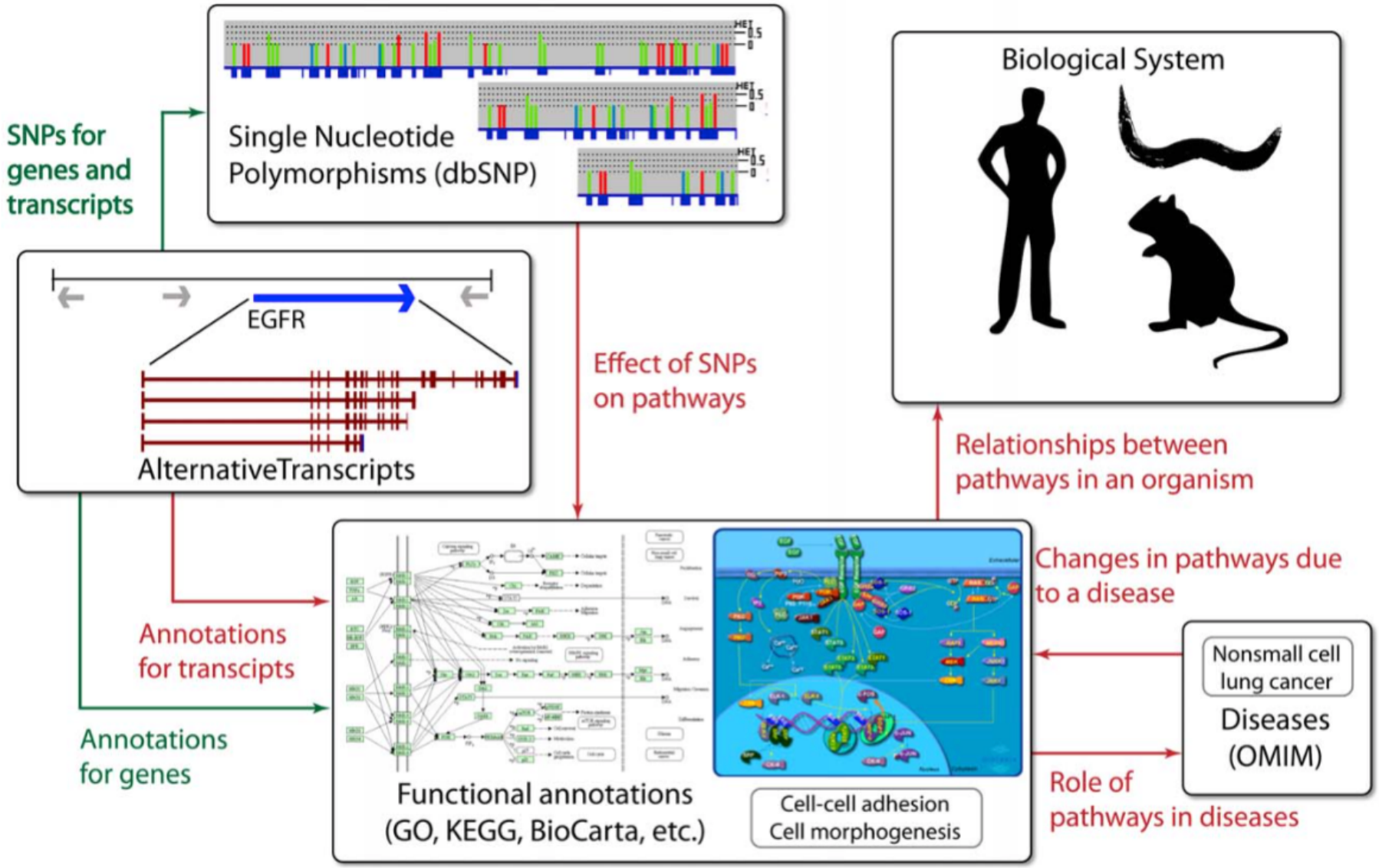
Enrichment Analysis: Pseudotime



Trapnell et al. *Nat Biotech* 2014

http://bioconductor.jp/packages/3.2/bioc/vignettes/monocle/inst/doc/monocle-vignette.pdf

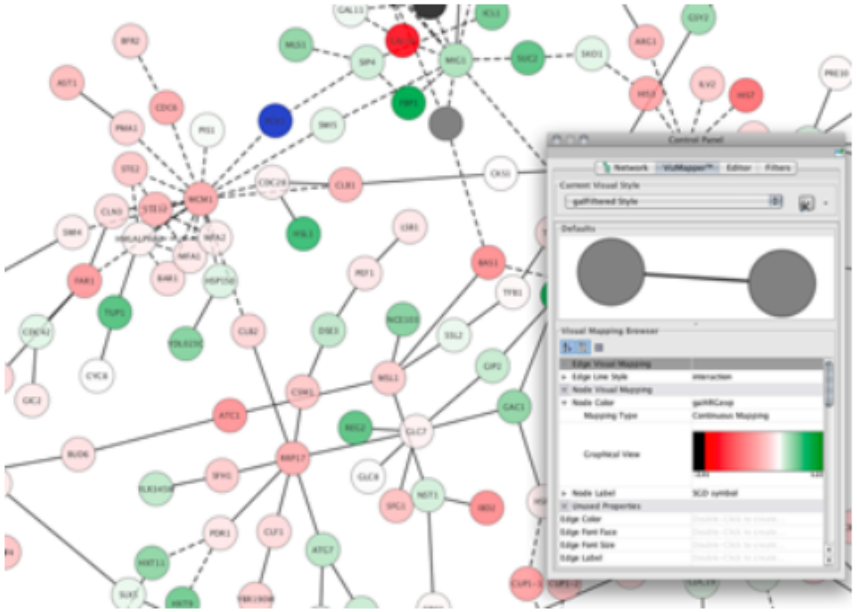
Outstanding challenges



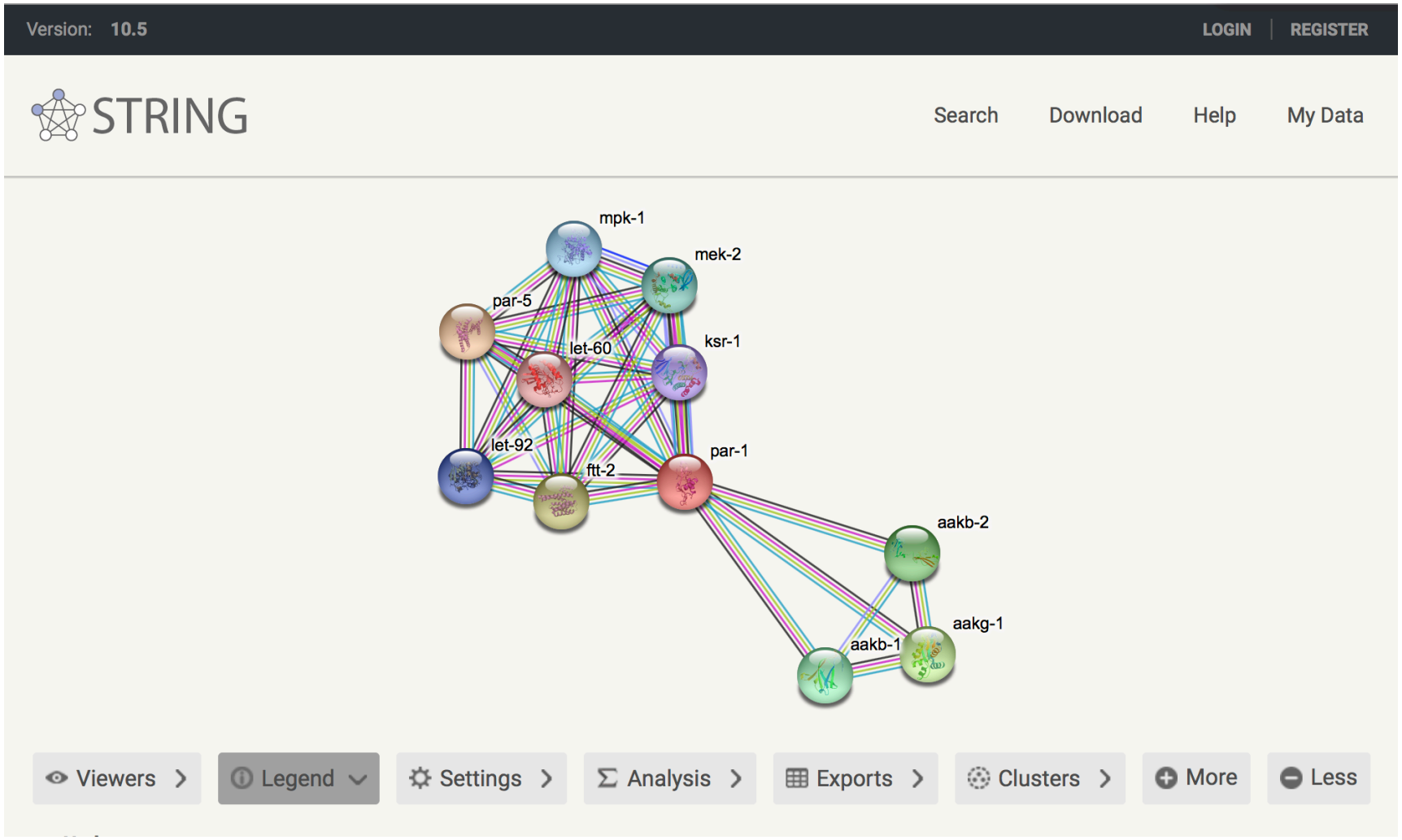
Khatri et al. (2012) PLoS Comp Biol 8:e1002375

Cytoscape supports many use cases in molecular and systems biology, genomics, and proteomics:

- Load molecular and genetic interaction data sets in many standards formats
- Project and integrate global datasets and functional annotations
- Establish powerful visual mappings across these data
- Perform advanced analysis and modeling using Cytoscape Apps
- Visualize and analyze human-curated pathway datasets such as WikiPathways, Reactome, and KEGG.



STRING database



Nodes:

Network nodes represent proteins

splice isoforms or post-translational modifications are collapsed, i.e. each node represents all the proteins produced by a single, protein-coding gene locus.

Node Color

colored nodes:
query proteins and first shell of interactors

white nodes:
second shell of interactors

Node Content

empty nodes:
proteins of unknown 3D structure

filled nodes:
some 3D structure is known or predicted

Edges:

Edges represent protein-protein associations

associations are meant to be specific and meaningful, i.e. proteins jointly contribute to a shared function; this does not necessarily mean they are physically binding each other.

Known Interactions

from curated databases

experimentally determined

Predicted Interactions

gene neighborhood

gene fusions

gene co-occurrence

Others

textmining

co-expression

protein homology

Your Input:

		Neighborhood	Gene Fusion	Cooccurence	Coexpression	Experiments	Databases	Textmining	[Homology]	Score
par-1	*Serine/threonine-protein kinase par-1 ; Required for cytoplasmic partitioning and asymmetric cell division in early embryogenesis (PubMed-7758115). Phosphorylates and restricts the asymmetry effector mex-5 (and possibly also mex-6) to the anterior cytoplasm of the zygote (PubMed-18842813). Regulates mes-1 expression during early embryogenesis (PubMed-11003841). Critical role in postembryonic vulval morphogenesis (PubMed-12490197). Involved in the establishment of neuronal polarity (PubMed-20023164) (1216 aa)*									

Predicted Functional Partners:

par-5	*14-3-3-like protein 1 (248 aa)*					●	●	●		0.981
ftt-2	*14-3-3-like protein 2 ; Required for extension of life-span by sir-2.1 (PubMed-16777605). Promotes nuclear export of yap-1 ...*					●	●			0.973
aakg-1	*Protein AAKG-1 (582 aa)*					●		●		0.967
aakb-2	*AMP-Activated Kinase Beta subunit family member (aakb-2) (274 aa)*					●		●		0.955
aakb-1	*AMP-Activated Kinase Beta subunit family member (aakb-1) (269 aa)*					●		●		0.951
mek-2	*Dual specificity mitogen-activated protein kinase kinase mek-2 ; Functions in the let-60 Ras signaling pathway; acts downs...*						●	●	●	0.931
mpk-1	*Mitogen-activated protein kinase mpk-1 ; Function in let-60 Ras signaling pathway; acts downstream of lin-45 raf kinase, b...*						●		●	0.924
let-92	*LEThal family member (let-92) (318 aa)*						●			0.922
ksr-1	*Kinase Suppressor of activated Ras family member (ksr-1) ; Serine/threonine-protein kinase which positively regulates Ra...*						●	●	●	0.919
let-60	*Ras protein let-60 ; The level of let-60 controls the switch between vulval and hypodermal cell fates during C.elegans vulva...*						●			0.915

Your Current Organism:

Caenorhabditis elegans

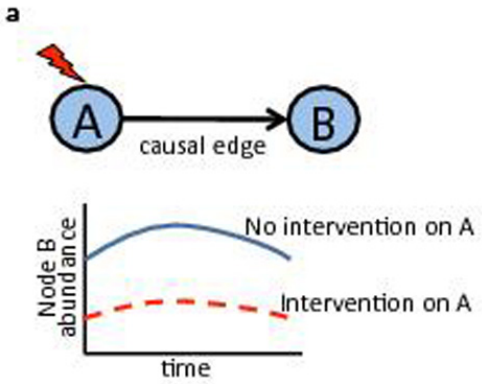
NCBI taxonomy Id: 6239
Other names: C. elegans, Caenorhabditis elegans, Rhabditis elegans, nematode

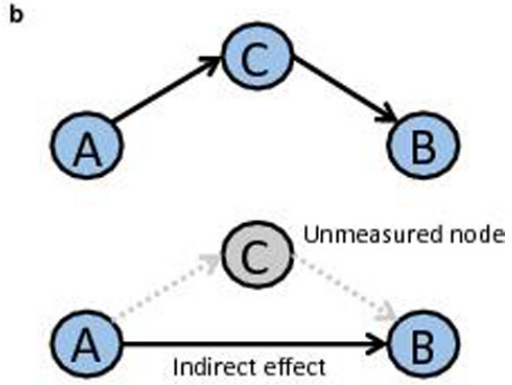
Gene regulatory network (GRN) inference

- Integrate many types of molecular and perturbation studies to generate predictive models
- Gene expression, physical interactions, chromatin accessibility, histone marks, TF-gene interactions…
- Combine network priors and measured features from many large-scale datasets
- Mathematical models (large linear systems)

DREAM challenge: inferring causal networks

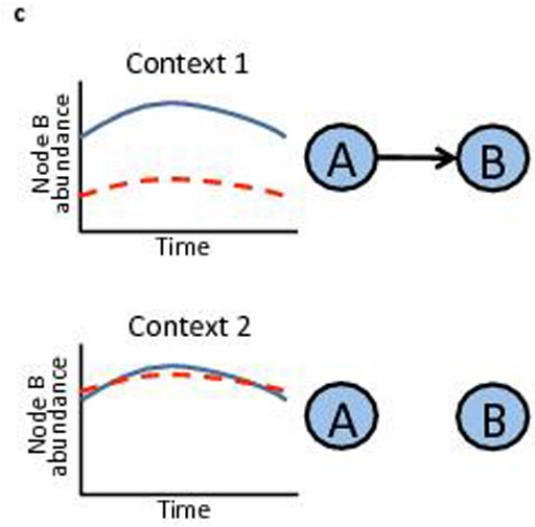
Direct, causal (A influences B)

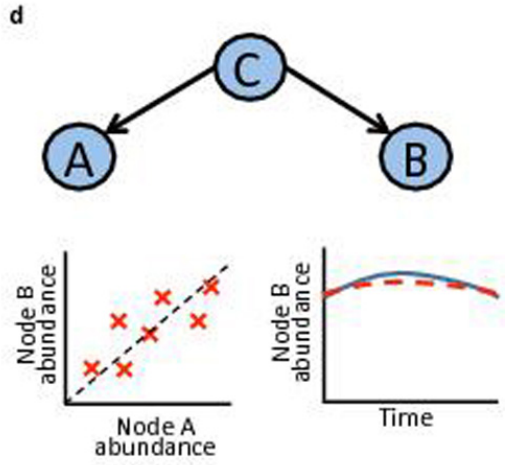




Indirect, causal (A influences B)

Context-dependent interaction

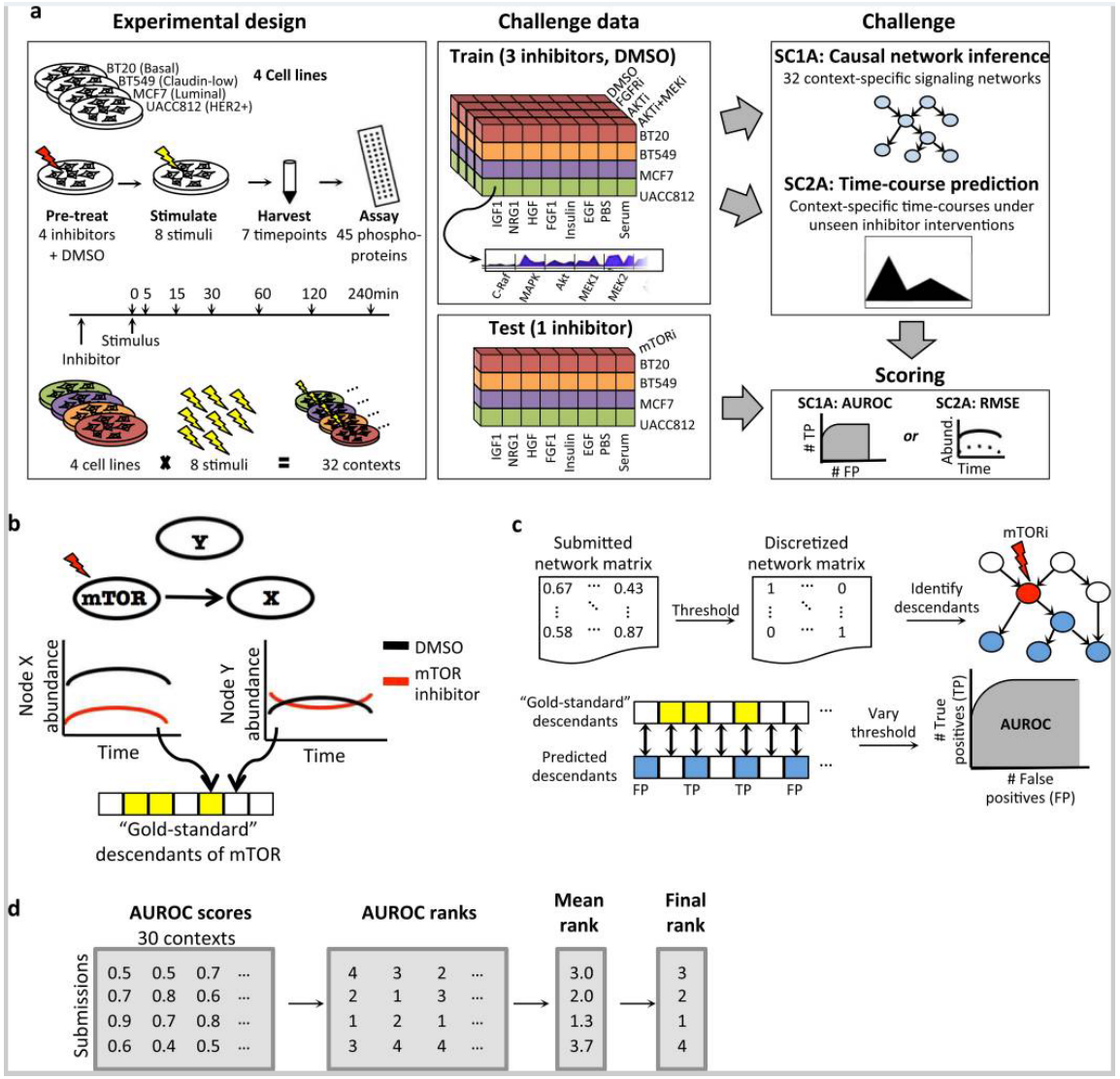




Indirect, correlated but not causal (A does NOT influence B)

Nat Methods (2016) 13:310-8. doi: 10.1038/nmeth.3773

DREAM challenge: inferring causal networks



Nat Methods (2016) 13:310-8. doi: 10.1038/nmeth.3773

Gene regulatory network inference software tools |...

Submit tool

Elucidating gene regulatory network (GRN) from large scale experimental data remains a central challenge in systems biology. The advent of high-throughput data generation technologies has allowed researchers to fit theoretical models to experimental data on gene-expression profiles.

CMGRN
Web

CMGRN / Constructing Multilevel Gene Regulatory Networks
★★★★★ (2) 0 discussions
An integrative web server to unravel hierarchical interactive networks at different regulatory levels. The developed method used the Bayesian network modeling to infer causal interrelationships among…

MIPRIP
Desktop

MIPRIP / Mixed Integer linear Programming based Regulatory Interaction Pre
☆☆☆☆☆ (0) 0 discussions
Predicts regulators of a gene of interest from gene expression profiles of the samples under study and known regulator binding information (from e.g. ChIP-seq/ChIP-chip databases). MIPRIP is…

AMIGO2
Desktop

AMIGO2
☆☆☆☆☆ (0) 0 discussions
A MATLAB environment to solve mathematical optimization problems which in dynamic modeling and control of biological systems. AMIGO2 is organized in four main modules: the pre-processor, the…

SpidermiR
Desktop

SpidermiR
☆☆☆☆☆ (0) 0 discussions
Offers an easy access to both Gene Regulatory Networks (GRNs) and miRNAs to the end user. SpidermiR integrates co-expression, physical interaction, co-localization, genetic influence, pathways, and…

RELATED WEBSITES

Gene regulatory networks
Wikipedia

INFORMATION

Share the latest tools and make your expertise visible

CONTRIBUTE