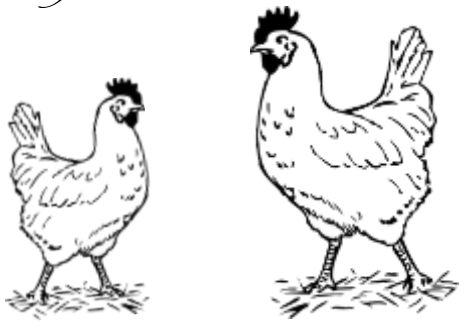
Network-guided Data Integration and Gene Prioritization

Close friends act alike



Are we close?

Lieven Verbeke
Kathleen Marchal
IBCN - Vakgroep Informatietechnologie
Faculteit Ingenieurswetenschappen en Architectuur

About us

- Situated at the Faculty of Sciences / Faculty of Engineering and Architecture, Ghent University
- PI Kathleen Marchal – Jan Fostier
- Department of Information Technology
- Main interest: method development
  - Network analysis in the broad sense / Systems biology
  - Machine learning / data mining
  - High performance computing
  - Study of clonal systems: bacteria = tumour cells
  - Increasing emphasis on medical applications
    - Tumour subtyping
    - Uncovering mechanisms of actions underpinning subtypes / phenotypes
    - Drug repurposing / synergy prediction
  - Drylab in constant search of wetlab partners

Outline

- Networks for the unitiated
- eQTL prioritization
- Linking genes to traits
- A unified tumour analysis framework
- Extra: non-coding somatic variants in cancer

A mystery finally solved

HOXB8

The hipster gene

PLOS | GENETICS

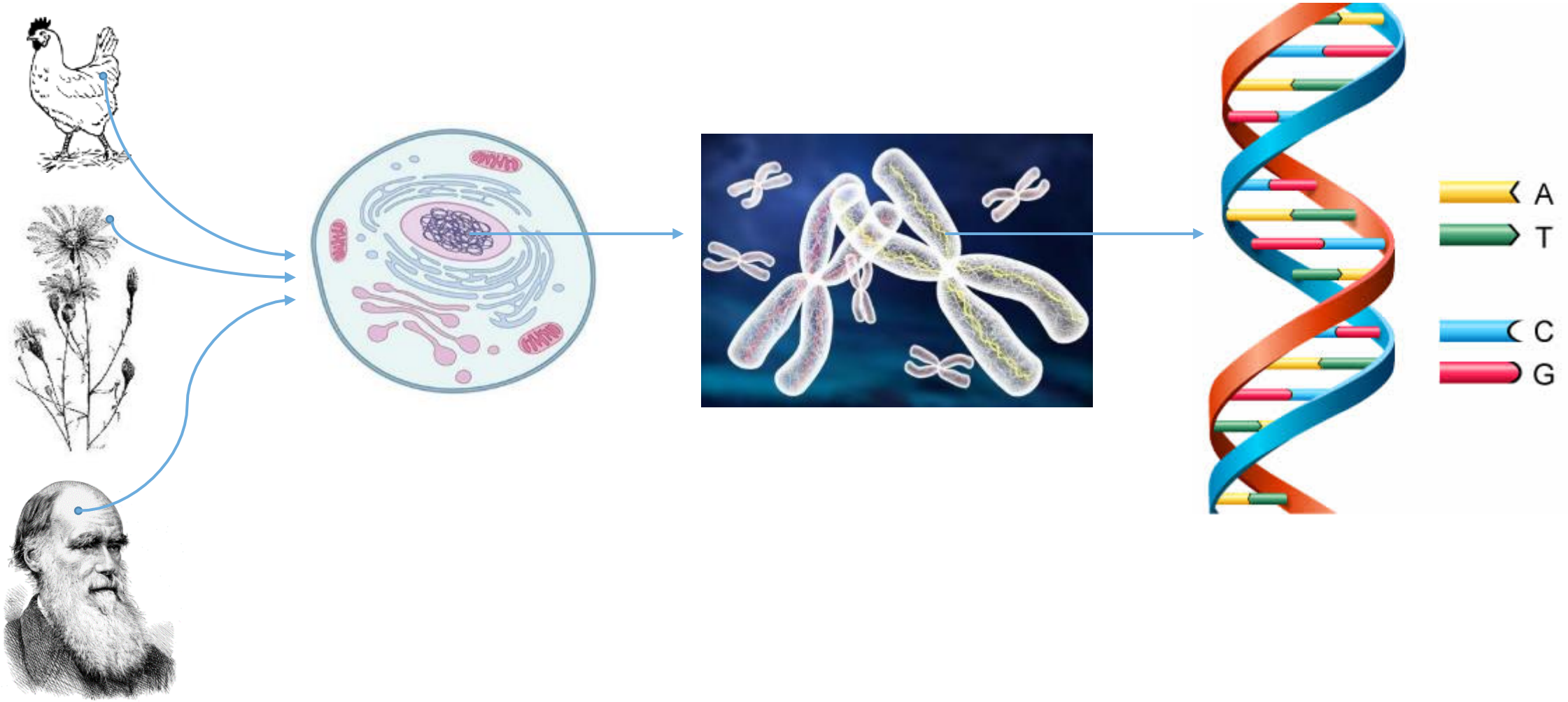
RESEARCH ARTICLE

A Complex Structural Variation on Chromosome 27 Leads to the Ectopic Expression of *HOXB8* and the Muffs and Beard Phenotype in Chickens

Ying Guo[1,2☯], Xiaorong Gu[1,2☯], Zheya Sheng[1,2,3☯¤], Yanqiang Wang[1,2], Chenglong Luo[4], Ranran Liu[5], Hao Qu[4], Dingming Shu[4], Jie Wen[5], Richard P. M. A. Crooijmans[6], Örjan Carlborg[3], Yiqiang Zhao[1,2], Xiaoxiang Hu[1,2]*, Ning Li[1,2]

1 State Key Laboratory for Agro-Biotechnology, China Agricultural University, Beijing, China, **2** National Engineering Laboratory for Animal Breeding, China Agricultural University, Beijing, China, **3** Division of Computational Genetics, Department of Clinical Sciences, Swedish University of Agricultural Sciences, Uppsala, Sweden, **4** Institute of Animal Science, Guangdong Academy of Agricultural Sciences, Guangzhou, Guangdong, China, **5** Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing, China, **6** Animal Breeding and Genomics Centre, Wageningen University, Wageningen, the Netherlands

An extremely short introduction to molecular genetics



An extremely short introduction to molecular genetics

Double helix, four letter sequence {ACTG}

…

CTGAAACCGTCCCCCAAGCGTTCAGGGTGGGGTTTGCTACGACTTCCGAGTCCAAAGTGTCCGTGTT
TTTGATATATACGCTCAAGGGCGAGAATTGGACCTGGCTTACGTCTTAGTACGTAGCATGGTGACAC
AAGCACAGTAGATCCTGCCCGCGTTTCCTATATATTAAGTTAAATCTTATGGAATATAATAACATGTG
GATGGCCAGTGGTCGGTTGTTACACGCCTACCGCAATGCTGAAAGACCCGGACTAGAGTGGCGAGA
TCTATGGCGTGTGACCCGTTATGCTCCATTTCGGTCAGTGGGTCACAGCTAGTTGTGGATTGGATTG
CCATTCTCCGAGTGTTTTAGCGTGACAGCCGCAGGGATCCCATAAAATGCAATCGTAGTCCACCTGA
TCGTACTTAGAAATGAGGGTCCGCTTTTGCCCACGCACCTGATCGCTCCTCGTTTGCTTTTAAGAACC
GGACGAACCACAGAGCATAAGGAGAACCTCTAGCTGCTTTACAAAGTACTGGTTCCCTTTCCAGCGG
GATGCTTTATCTAAACGCAATGAGAGAGGTATTCCTCAGGCCACATCGCTTCCTAGTTCCGCTGGGA
TCCATCGTTGGCGGCCGAAGCCGCCATTCCATAGTGAGTTCTTCGTCTGTGTCATTCTGTGCCAGATC
GTCTGGCAAATAGCCGATCCAGTTTATCTCTCGAAACTATAGTCGTACAGATCGAAATCTTAAGTCAA
ATCACGCGACTAGACTCAGCTCTATTTTAGTGGTCATGGGTTTTGGTCCCCCCGAGCGGTGCAACCG
ATTAGGACCATGTAGAACATTAGTTATAAGTCTTCTTTTAAACACAATCTTCCTGCTCAGTGGTACAT
GGTTATCGTTATTGCTAGCCAGCCTGATAAGTAACACCACCACTGCGACCCTAATGCGCCCTTTCCAC
GAACACAGGGCTGTCCGATCCTATATTACGACTCCGGGAAGGGGTTCGCAAGTCGCA…

The central dogma of molecular biolog

DNA

ACGCCTACCGCAATGCTGAAA

Does stuff

Gene expression *the activity of a gene measured as the amount of mRNA*

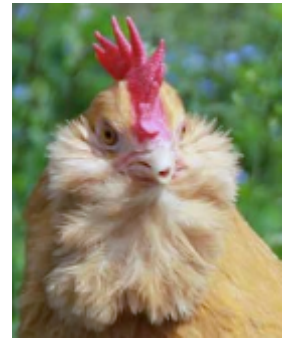
Genetic variability can cause different phenotypes

Individual 1

ACGCCTACC**TCT**ATGCTGAAA

ACGCCTACC**CCT**ATGCTGAAA

Individual 2



Phenotype

the set of observable characteristics of an individual resulting from the interaction of its genotype with the environment

Sources and types of genetic variability

Single nucleotide variations / mutations

ACGC**C**TACCG → ACGC**G**TACCG

Structural variations – copy number

A**CGCCTA**CCG → A**CGCCTA CCGCCTA CCGCCTA**CCG

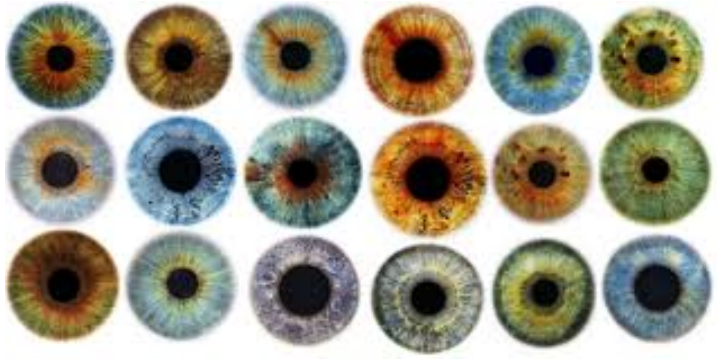
Epigenetic modifications

ACGCCTACCG → ACGCCTACCG

Where does this genetic variation come from?

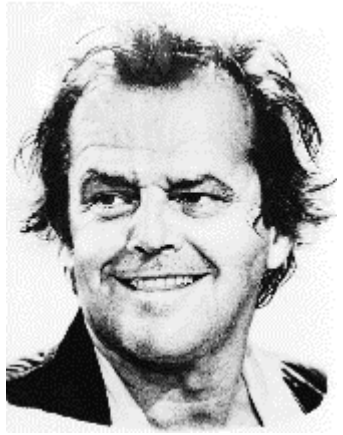
- Natural variation
- New mutational variants

Problem: phenotype is rarely determined by genetic variation in only a single gene



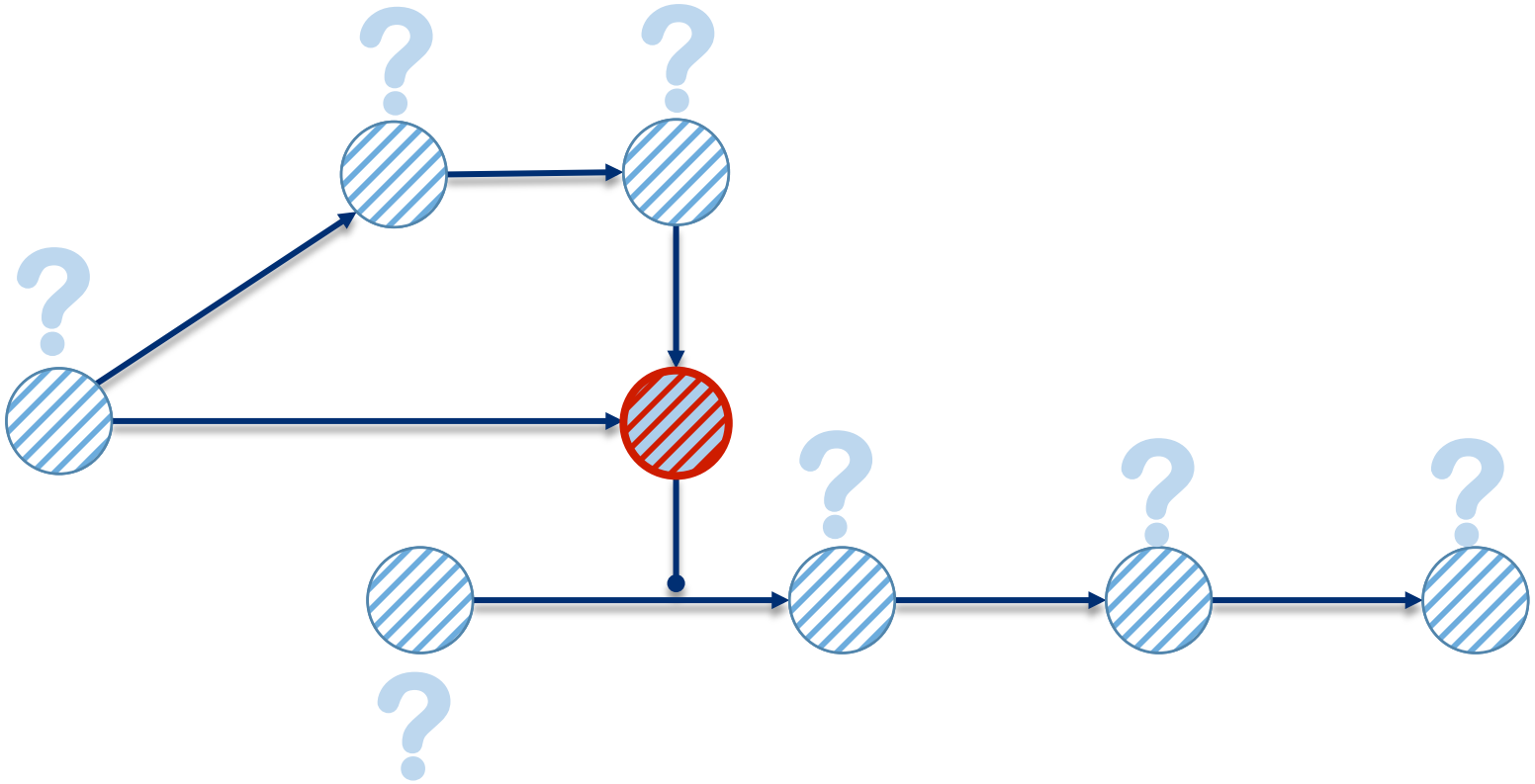
10 genes for eye color
50 genes for iris structure

400 genes for body weight



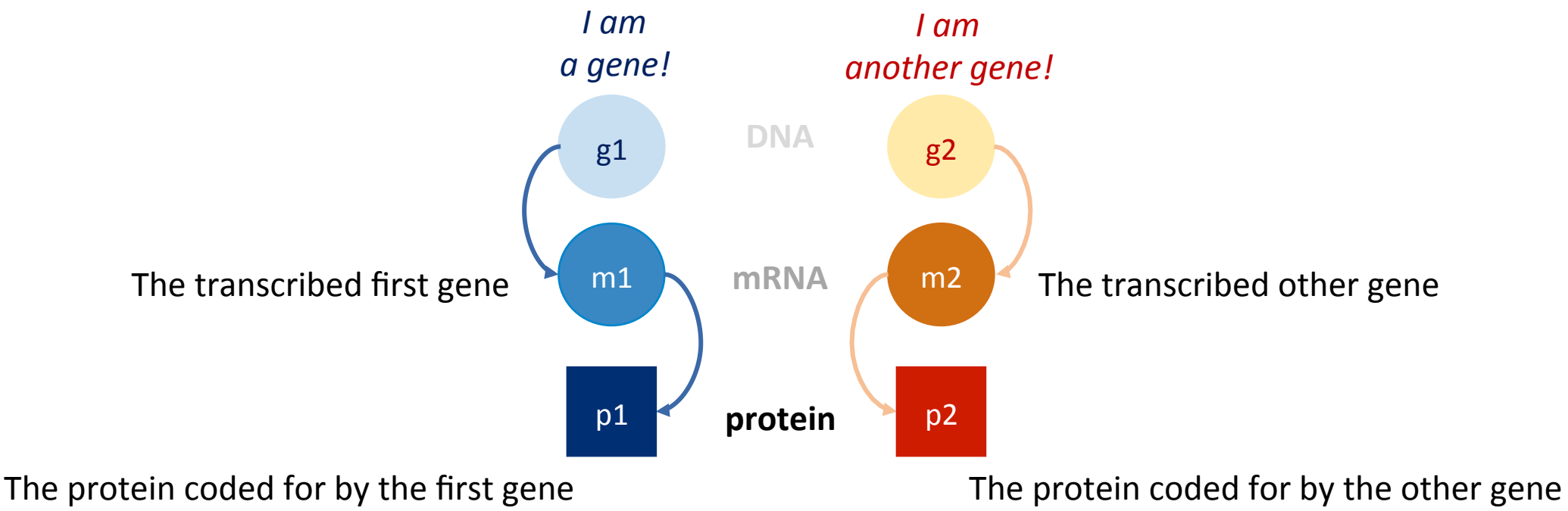
100 genes for schizophrenia

Why are so many genes involved in these traits?

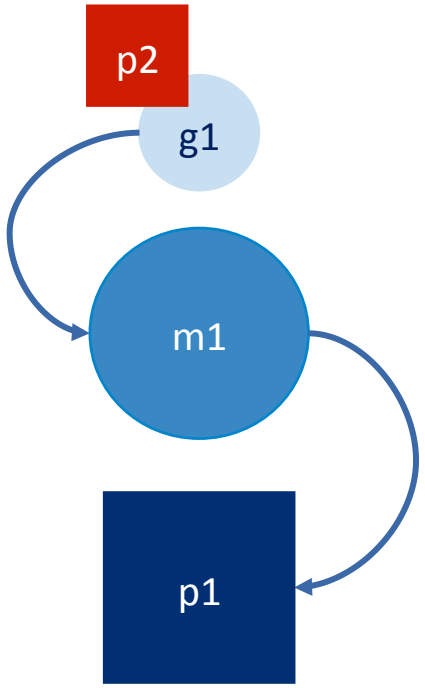


How can genes interact?

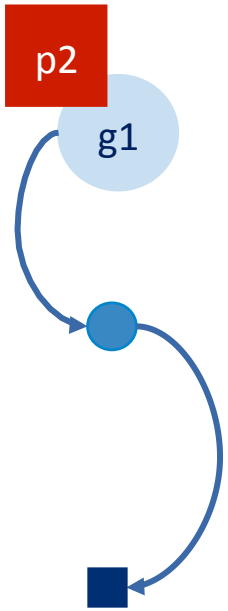
The central dogma of molecular biology



How can genes interact?

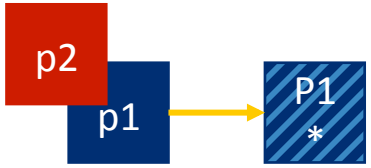


Protein 2 binds to the DNA of gene 1 and facilitates transcription, so lots of protein 1

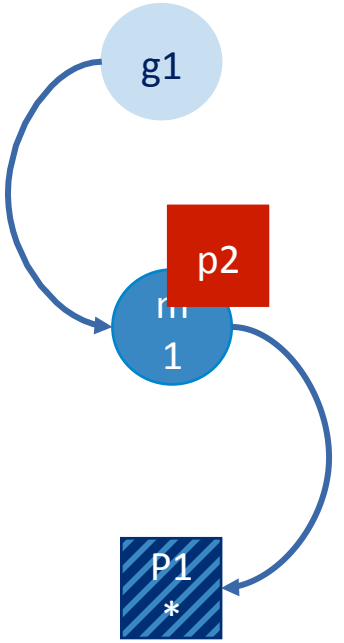


Protein 2 binds to the DNA of gene 1 and suppresses transcription, so very little protein 1

Protein 1 and 2 bind together and form a complex, that does other stuff than protein 1 or 2 alone



Protein 2 modifies protein 1, so protein 1 changes and does different stuff than the original protein 1

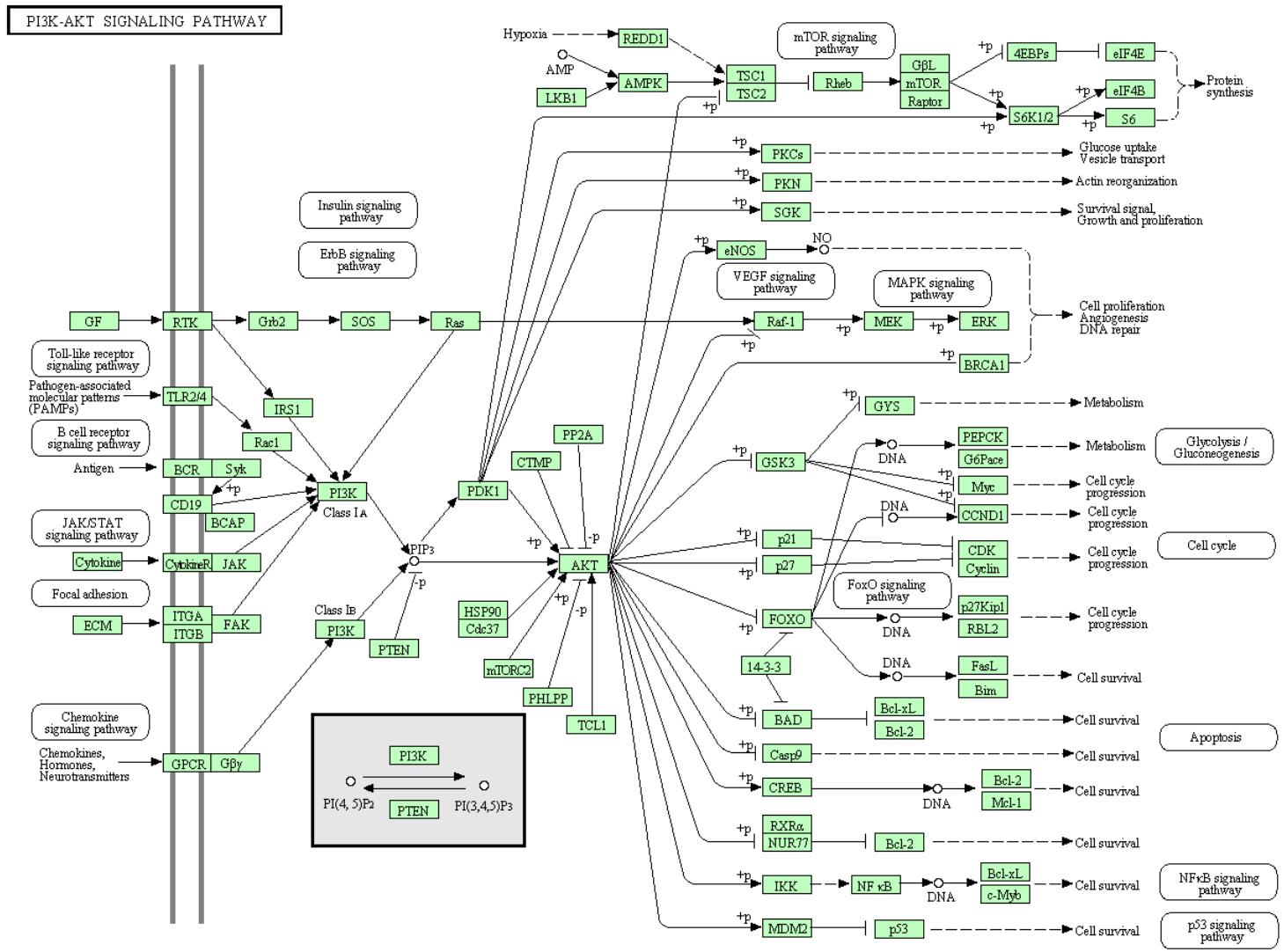


Protein 2 modifies the transcribed gene 1, so protein 1 changes and does different stuff than the original protein 1

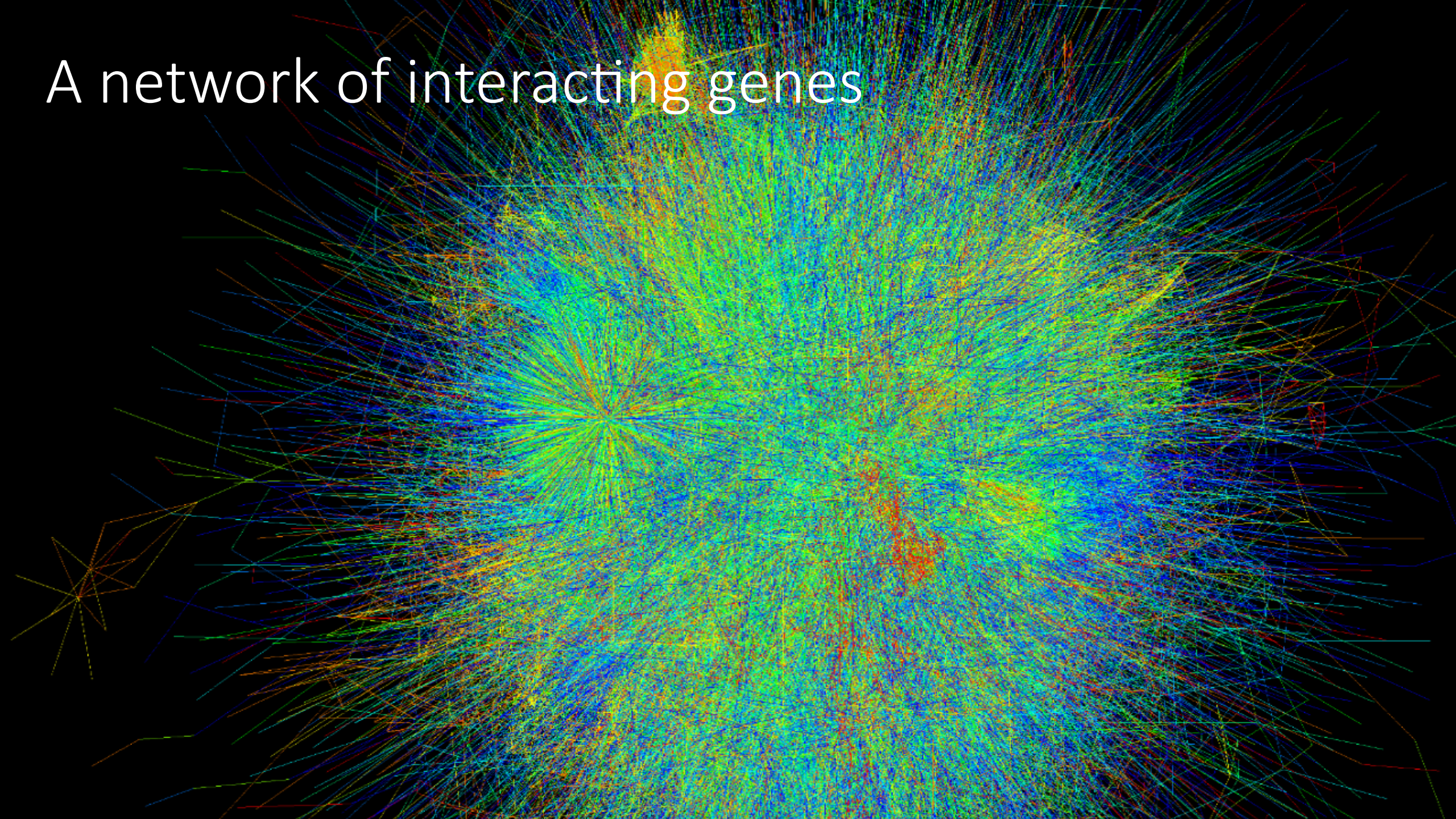
How can genes interact?

- There exist many more mechanisms by which genes (or non-gene elements) can interact, or bu which transcription and translation are modulated
  - Non-gene entities
    - Long Non-coding RNA
    - miRNA
    - Distal regulatory elements
  - Non-coding regions of genes are important too
    - UTRs
    - Alternate splicing + protein variant stability
    - Intron variants?
  - Epigenetics
  - Histone / chromatine modifiers

Pathways of interacting genes



A network of interacting genes

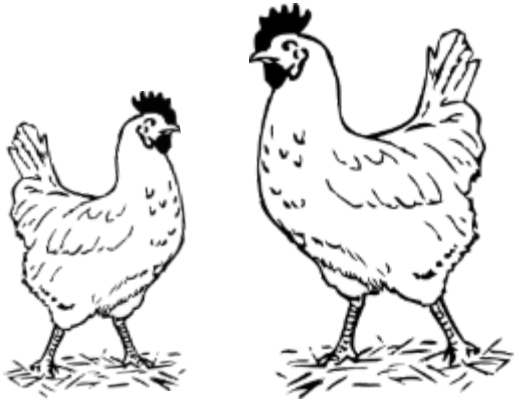


Unfortunately

- A network of gene interactions can not readily be used to infer which genes participate in the same biological processes
- Many of these gene interactions have not been observed, but are predicted using high-throughput methods
- Some gene interactions are only valid under certain conditions
  - In a specific tissue
  - Under certain disease circumstances
  - If the environment changes
  - …

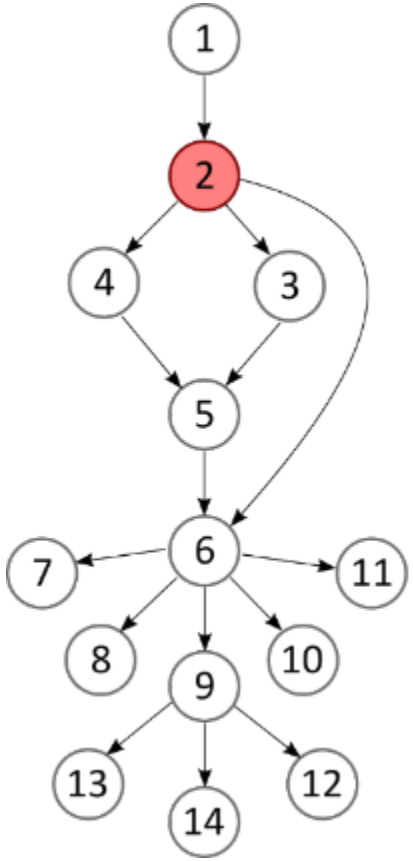
Luckily

- We have devised a way to identify genes that are relevant for a particular phenotype, using the connectivity of the genes in a less-than-perfect gene interaction network.
- All our methods build on the assumption that genes found in the immediate network neighborhood of each other are likely to participate in the same biological processes.
- How can we measure if two genes are close or well connected in a network?



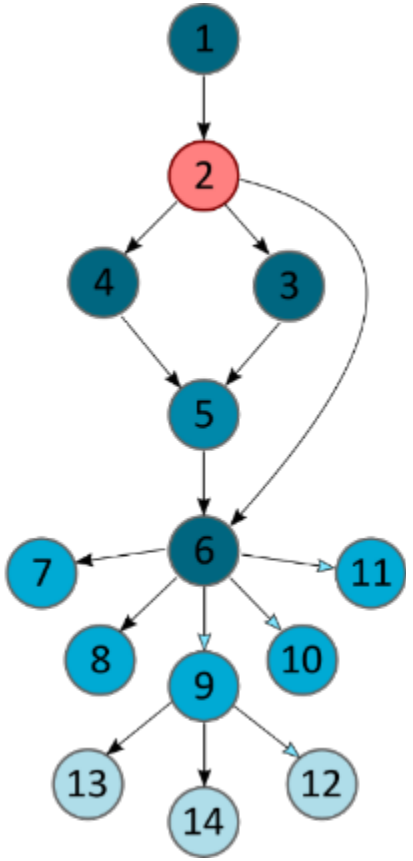
Are we close?

Are we close (genes in a network)?



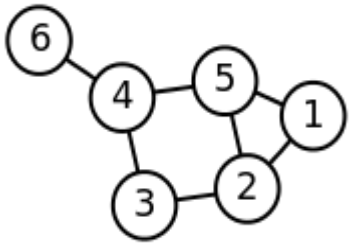
Many possible ways to quantify how well genes are connected

- Neighbors / neighbors of neighbors / neighbors of …
- Shortest paths: problem with distance between genes
- Diffusion techniques: the ink analogy



A tiny bit of graph theory

Labeled graph



Degree matrix D

$$\begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Adjacency matrix A

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

G={E,V}

V=vertices
E=edges

Laplacian matrix L=D-A

$$\begin{pmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}$$

Transition matrix T

$$\begin{bmatrix} 0 & 0.33 & 0 & 0 & 0.33 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0.33 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0.33 & 0.5 \\ 0.5 & 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0 \end{bmatrix}$$

Connectivity measures in a graph

- Shortest path
  - Dijkstra algorithm
  - Needs weighted edges
- Random walks (with restart)
  - $P=L^{-1}$
  - $P_{restart}=(I-aT)^{-1}$
- Diffusion
  - Heat diffusion (e.g., HOTNET, Network based tumour stratifaction)
  - Laplacian diffusion kernel (ink diffusion)
    - $C=e^{-aL}$

3 Applications

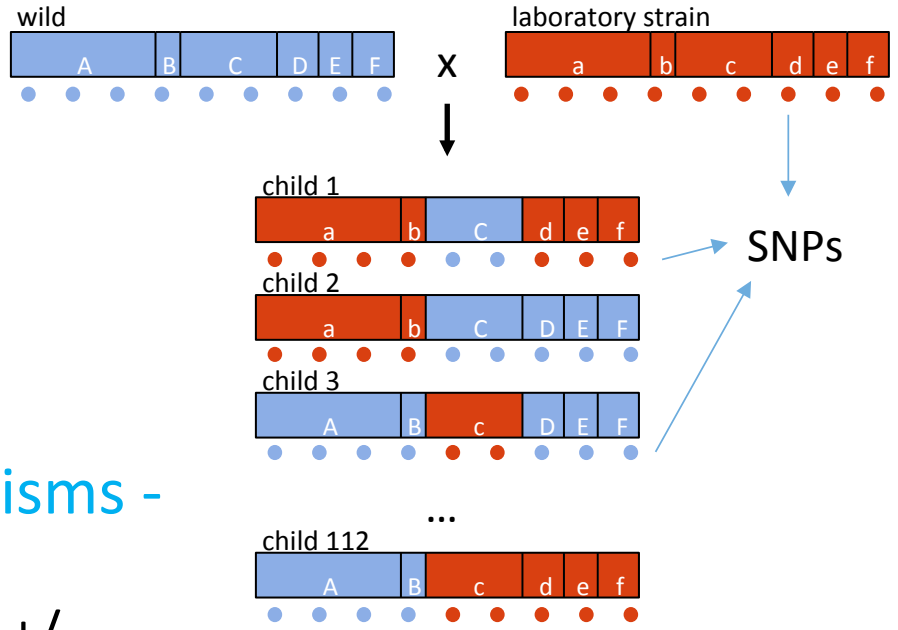
The fact that genes are active in some individuals, and less active in other individuals

1. Prioritize genes (in yeast) whose genetic variation can be linked to differential expression of other genes: **EPSILON**
2. Prioritize genes that can be linked to wood properties in eucalyptus trees: **NBDI**
3. Identify groups of cancer patients that exhibit similar molecular properties, and prioritize genes and pathways that behave abnormally in those patients: **MUNDIS**

Application 1: gene prioritization in yeast

Saccharomyces cerevisiae

- Genome: 12,500,000 base pairs
- +/- 6,000 genes
- Two parent yeast strains were crossed
- 112 children were produced



Genetic data: Single Nucleotide Polymorphisms - SNPs

- The genome of the offspring was sampled at +/- 3000 positions
- Different from whole genome sequencing: SNPs represent an area on a chromosome <-> point mutations

Application 1: gene prioritization in yeast

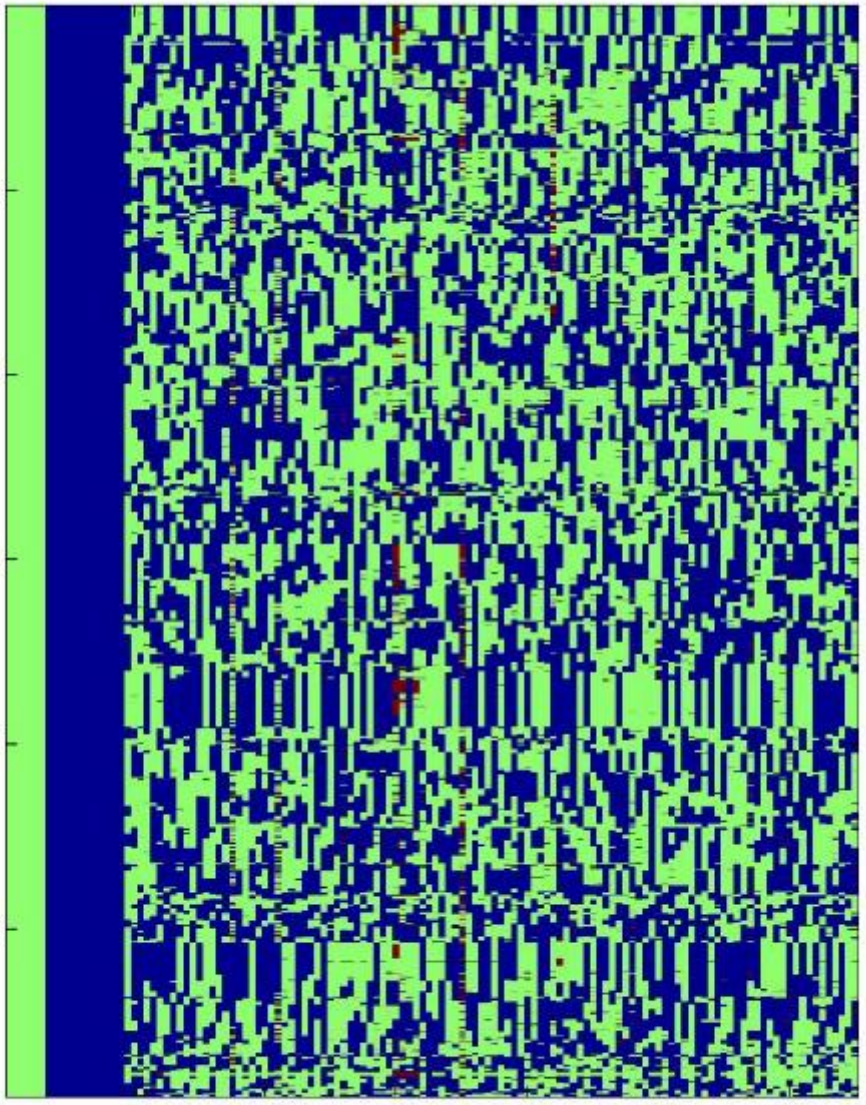
Gene expression data

- mRNA levels for 6000 genes
- 112 samples

Gene interaction network

- Derived from multiple network resources
  - Protein-protein
  - Regulatory
  - Phosporylation
- 4,375 genes
- 35,569 gene interactions

SNP data



SNPs

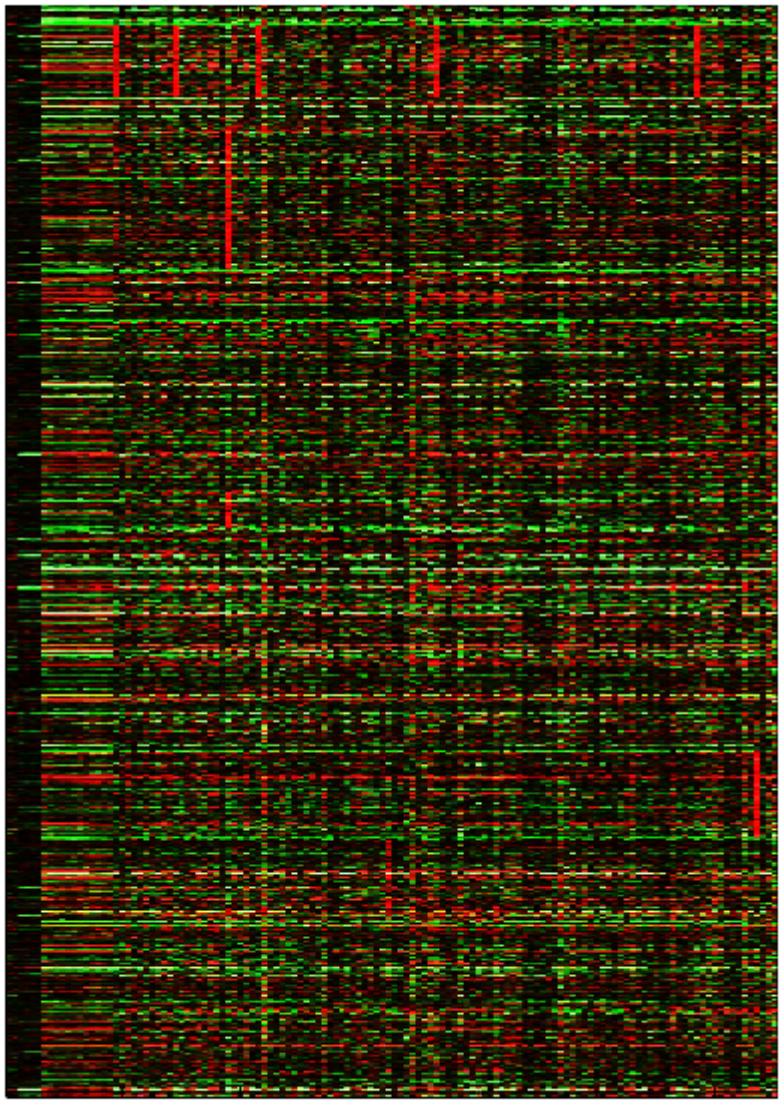
Linkage Desequilibrium

Parent 1

Parent 2

Children

Gene expression data



Genes

Parent 1

Parent 2

Children

Step 1: find eQTL

QTL

quantitative trait locus

- region on a chromosome
- that contains genetic variation
- that can be statistically related to variability of a quantitative trait (phenotype)

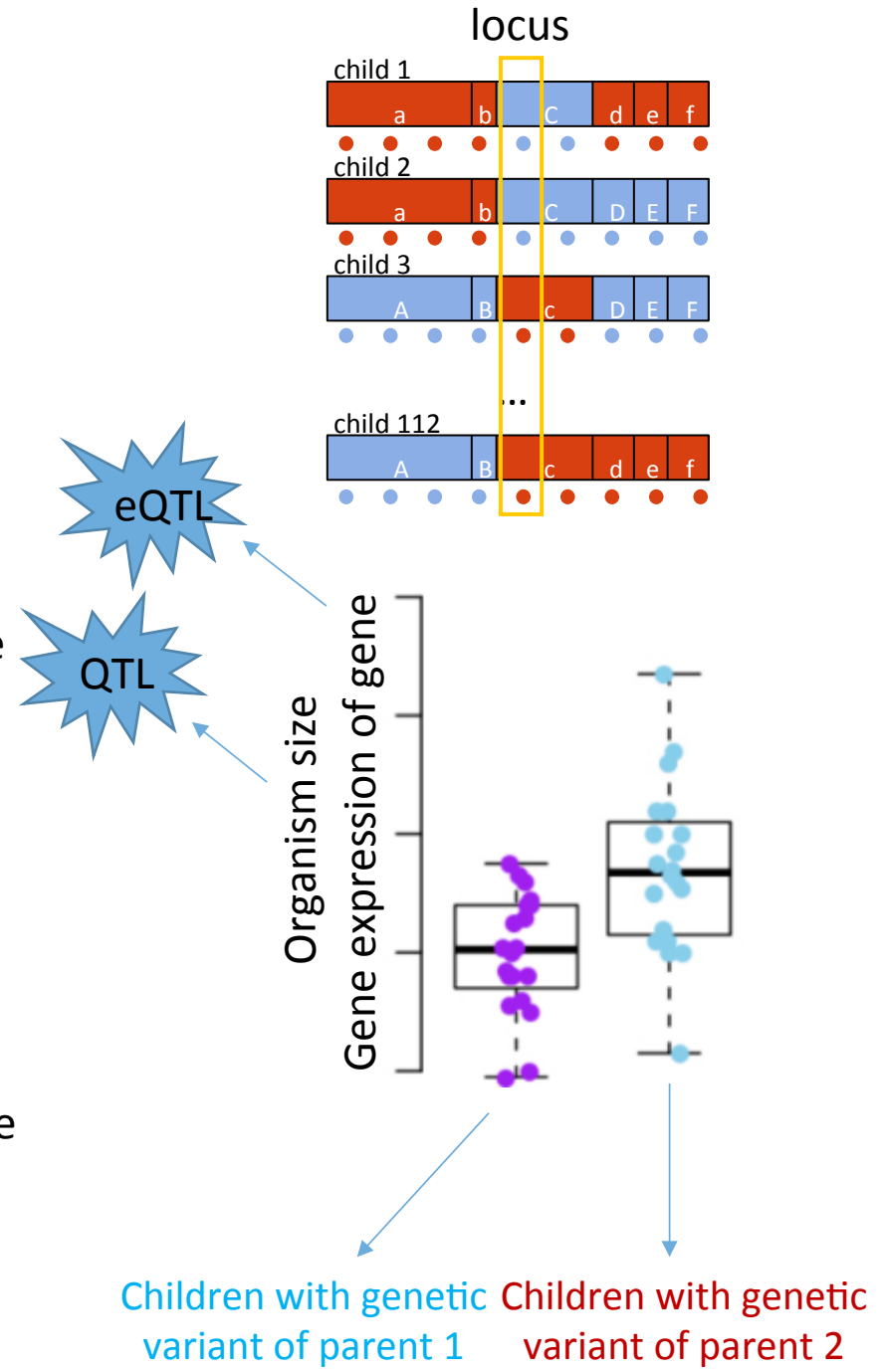
eQTL

expression quantitative trait locus

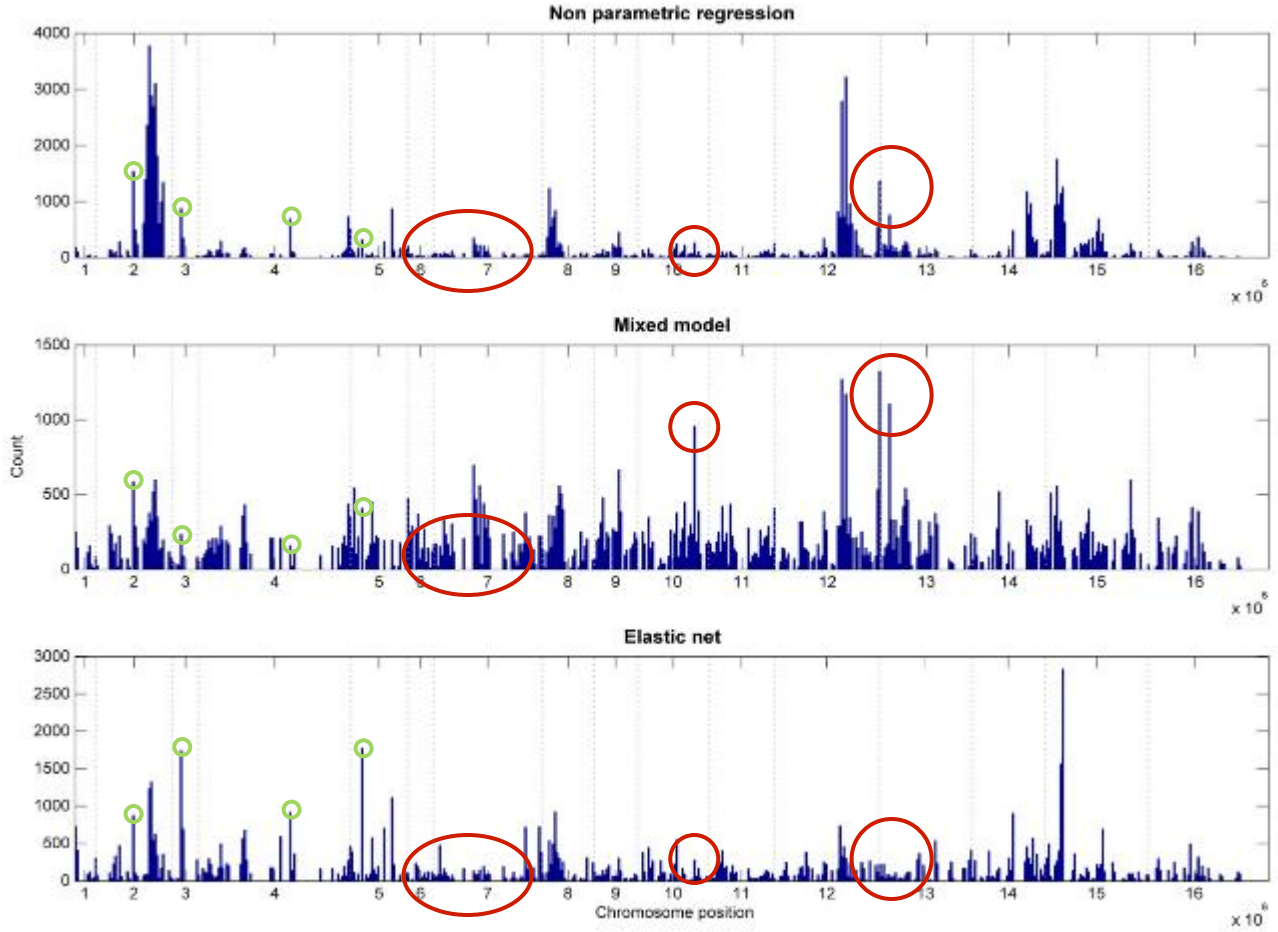
- If we can link variability at a locus to the expression of a particular gene (the target gene)

Why looking for QTL - eQTL?

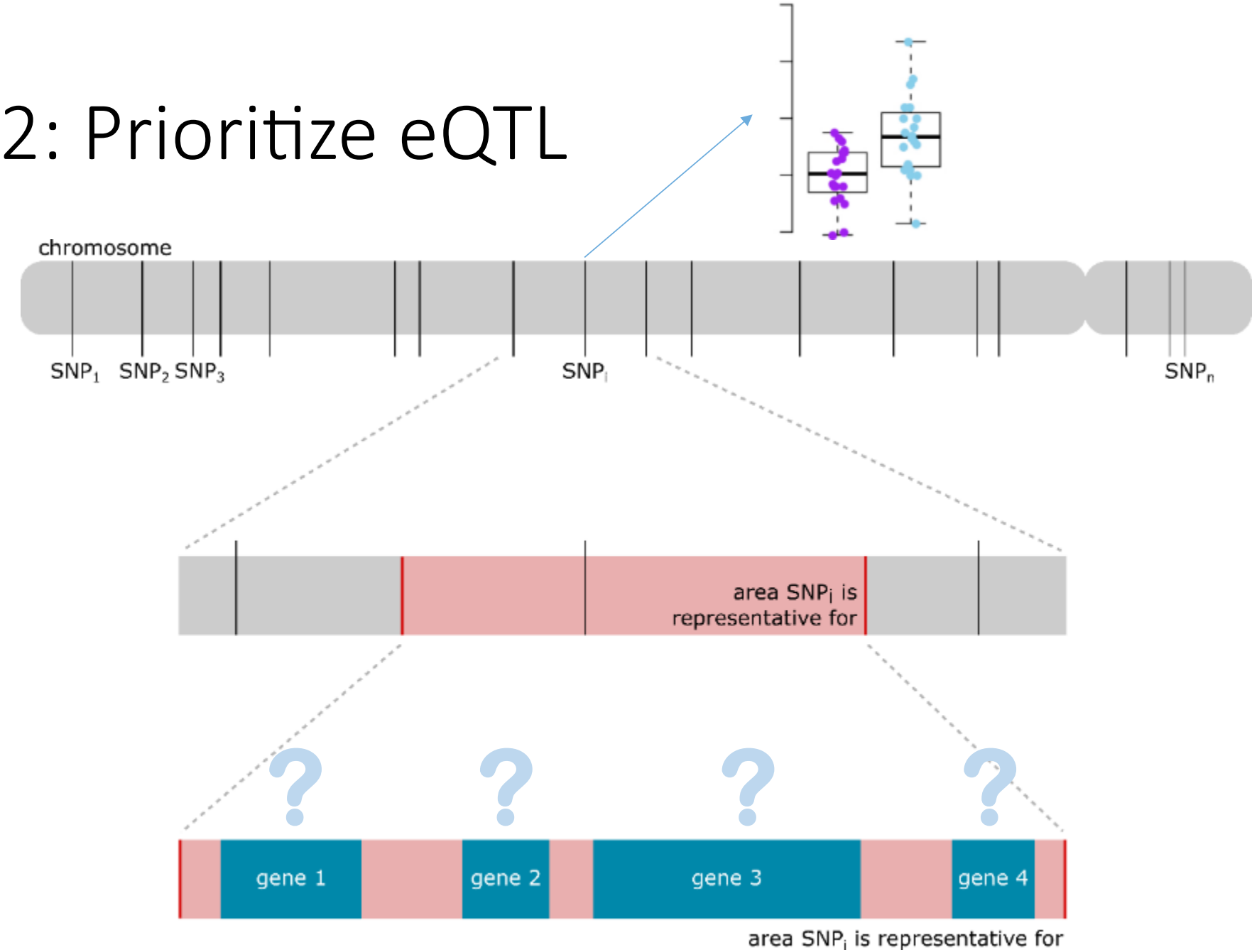
- Reveal mechanics of gene regulation and discover novel gene interactions
- Targeted breeding towards specific properties



Identifying eQTLs

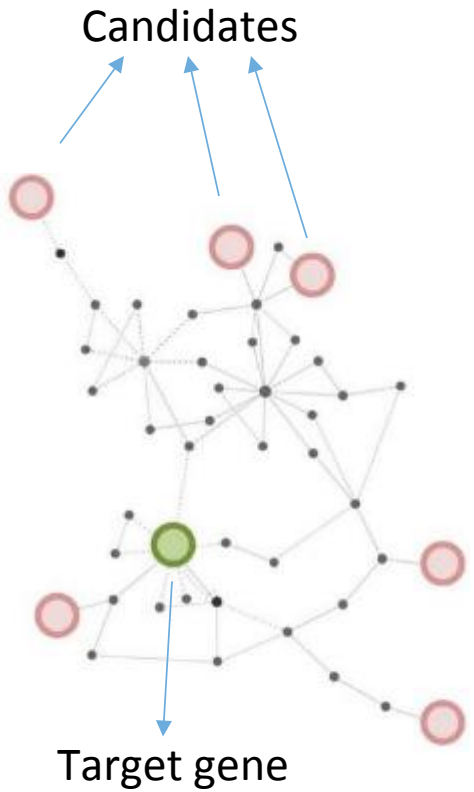


Step 2: Prioritize eQTL

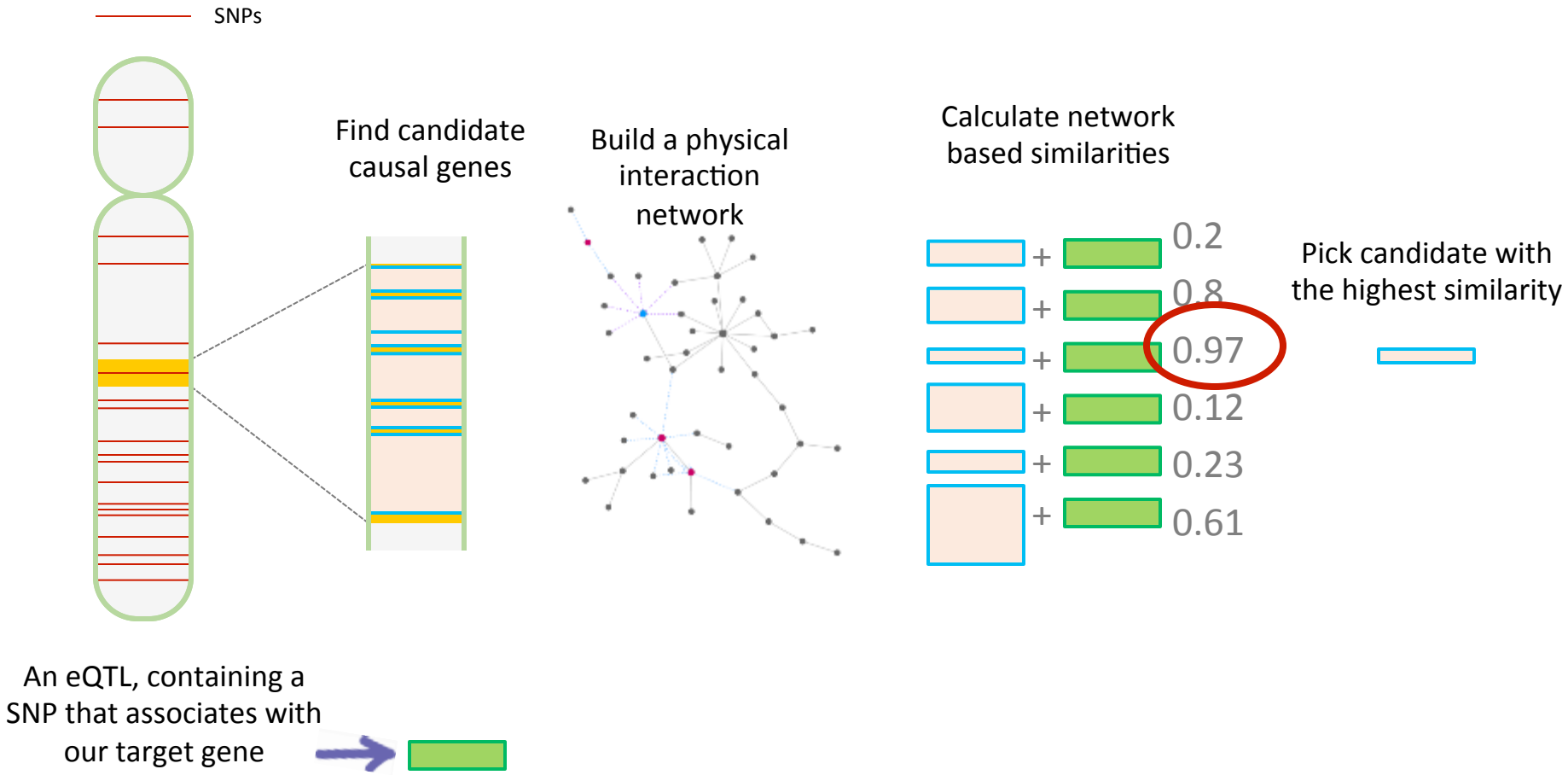


Step 2: Prioritize eQTL

- How to select the best candidate gene?
  - Random assignment
  - Use a network!
    - Take the candidate gene that's closest to the target gene in the network
    - Take the gene that's best connected to the target gene
  - Evaluate using the knockout data
- Method is called **EPSILON**: **E**QTL **P**rioritization using **SI**milarities derived from **LO**cal **N**etworks



Prioritization



Prioritization

- Now what clever similarity measures did we use?
  - Kernels calculated on graph nodes (each gene is a node in the interaction graph), producing node similarity matrices
  - The kernels we use are typically used for recommendation tasks like
    - Customers who bought this also bought …
    - People you may know …
    - Web page importance ranking
  - We are not the first to use kernels for prioritization (see e.g. Nitsch et *al.* 2010) but to our knowledge, this is a new application

Prioritization

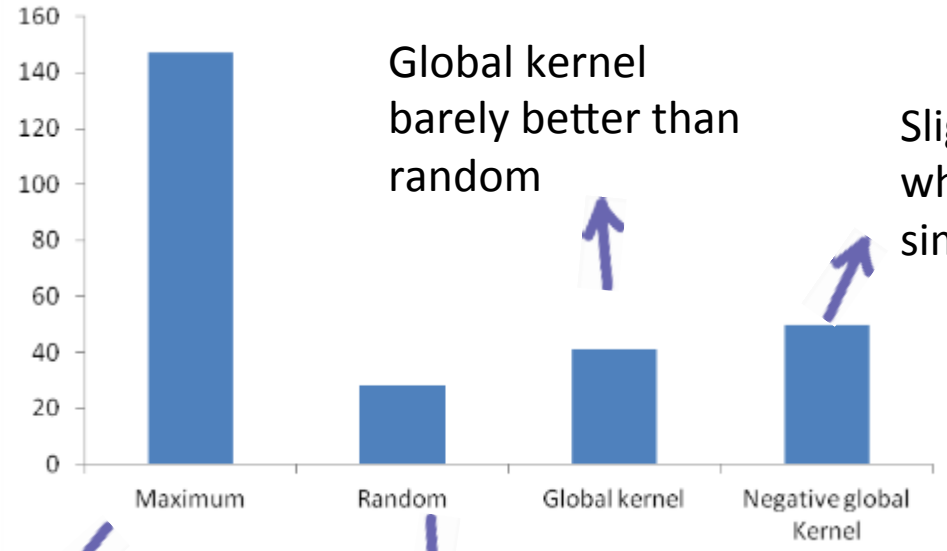
- To evaluate our prioritization, we use an existing compendium of knockout experiments (Hughes et *al.* 2000)
  - Knockout pairs are proved causal relations between genes
  - Aim is to retrieve as much knockout pairs as possible

- Any prioritization method should perform better than randomly picking a candidate

Prioritization

- We have our similarity measure. And an evaluation strategy. Let's try it out!
  - We assembled an interaction network
  - Derived an adjacency matrix from it
  - And calculated a host of kernel matrices
  - All that is left is to use the similarity matrices to do the prioritization
- Unfortunately
  - It Does not work.
  - At least not very well
  - In fact, our results are on par with randomly picking a candidate

Prioritization



Global kernel barely better than random

Slight improvement when picking the least similar candidate

Total number of knockout pairs that a perfect method would be able to retrieve

On average 5-6 candidates per eQTL, so we expect to find +- 30 at random

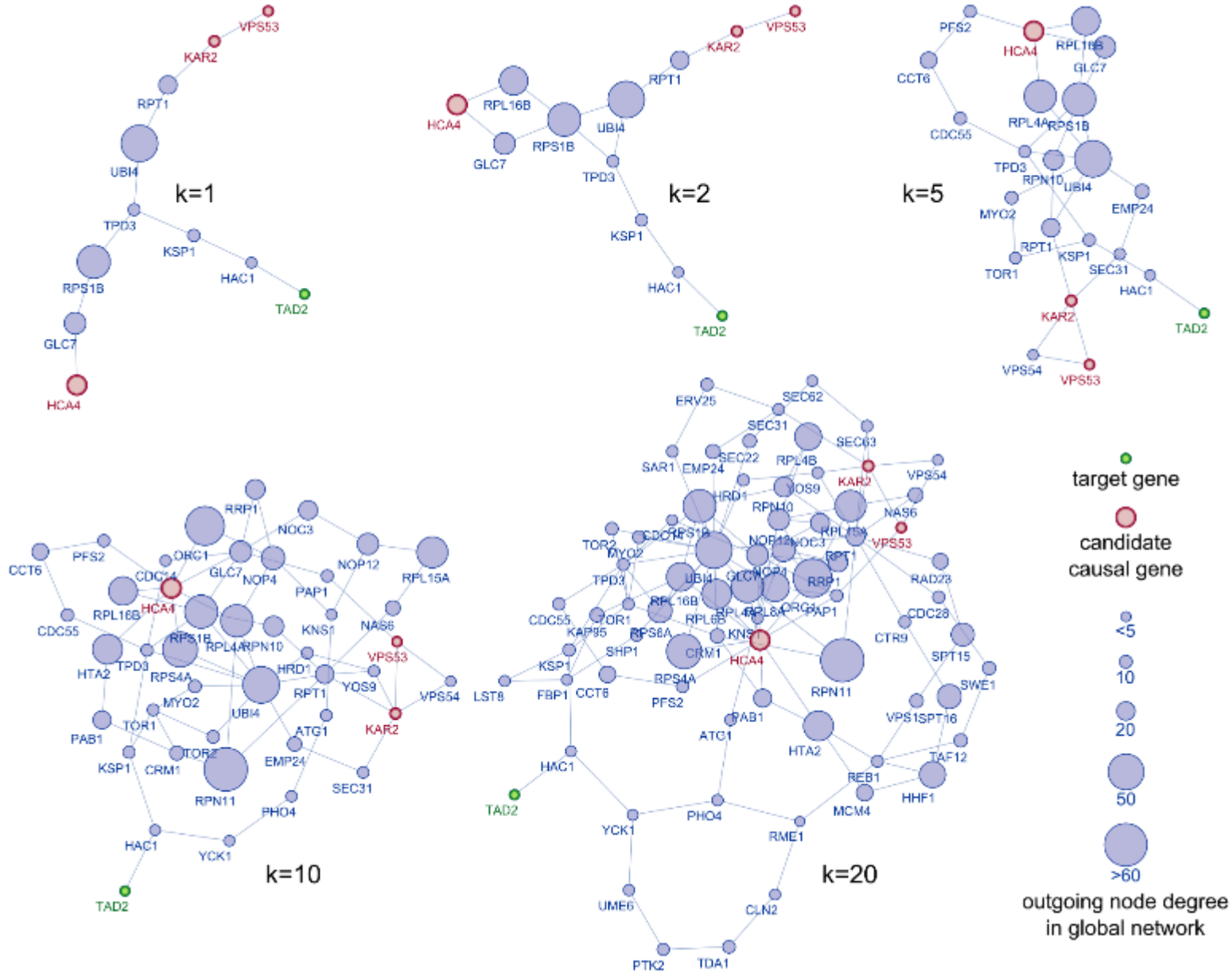
More prioritization

- The hublike structure of our interaction network is causing problems
- Idea:
  - For each eQTL-target gene pair, find a local network connecting the target gene with all candidate causal genes
  - Calculate a similarity measure on this local network
- How to find a local network
  - Take shortest path from candidates to target, and filter network to contain only nodes that are on such a shortest path
- Let's add an extra reference method: take the candidate with the shortest path to the target gene

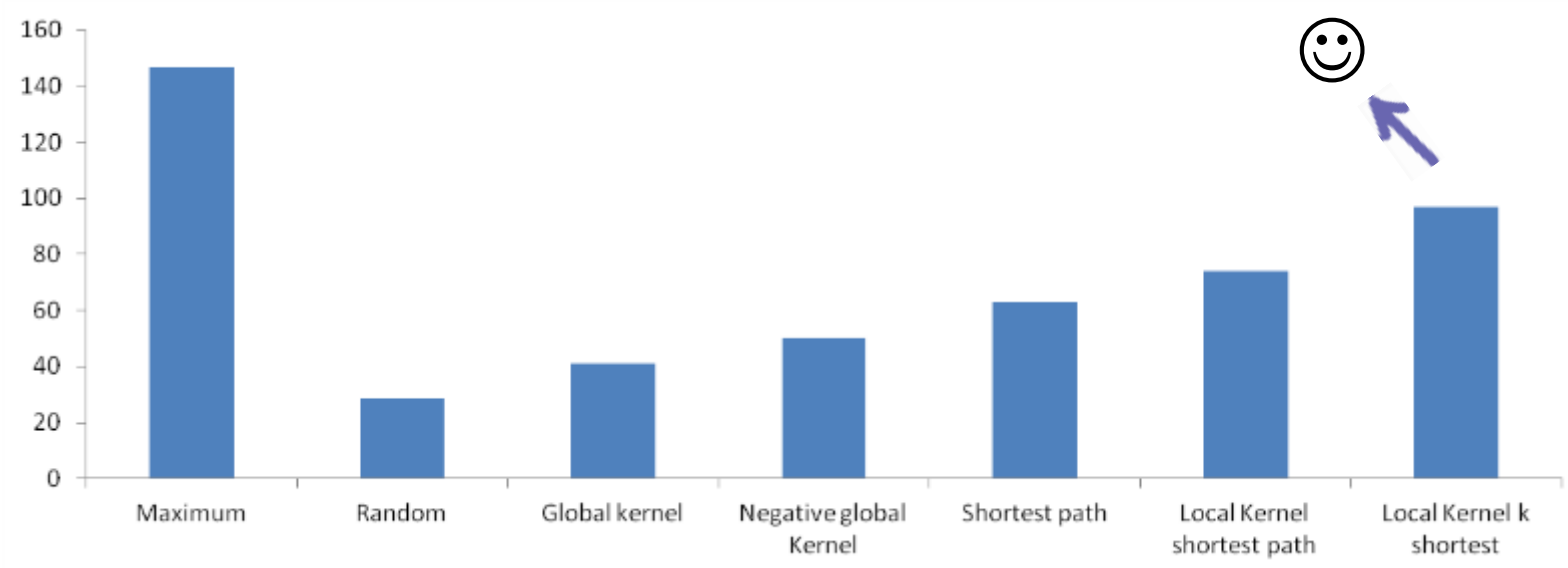
More prioritization

Even more prioritization

- We think we can do even better
  - The shortest path subnetwork is still depending on the hubs in the network
  - Idea: use several alternative paths instead of a single shortest paths
  - => k shortest paths


k=1
k=2
k=5
k=10
k=20
target gene
candidate causal gene
<5
10
20
50
>60
outgoing node degree in global network

Even more prioritization



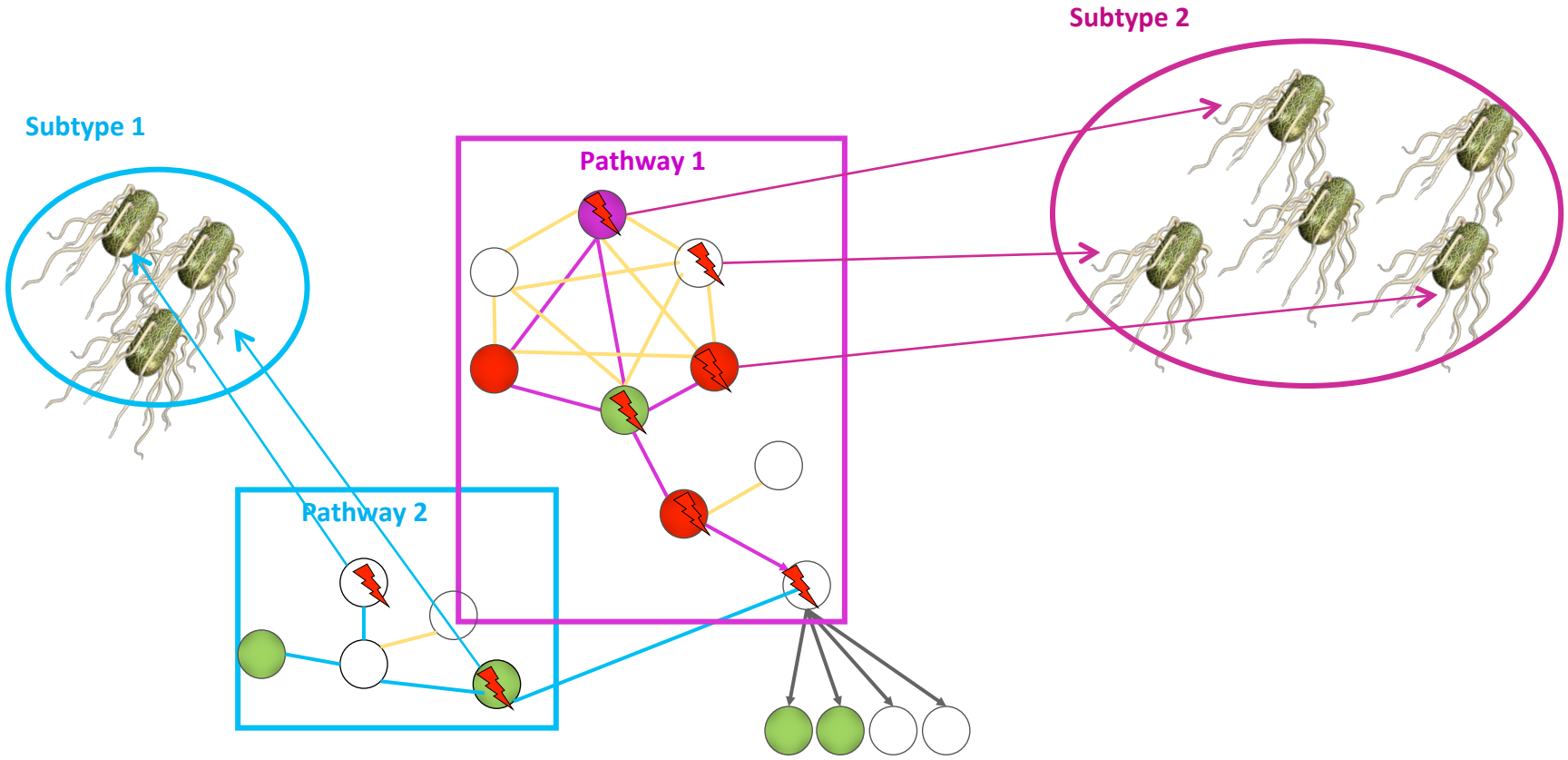
Conclusions

- We have used **SNP** data and **gene expression** data of **S. *cerevisiae*** to detect **eQTLs** using different mapping methods

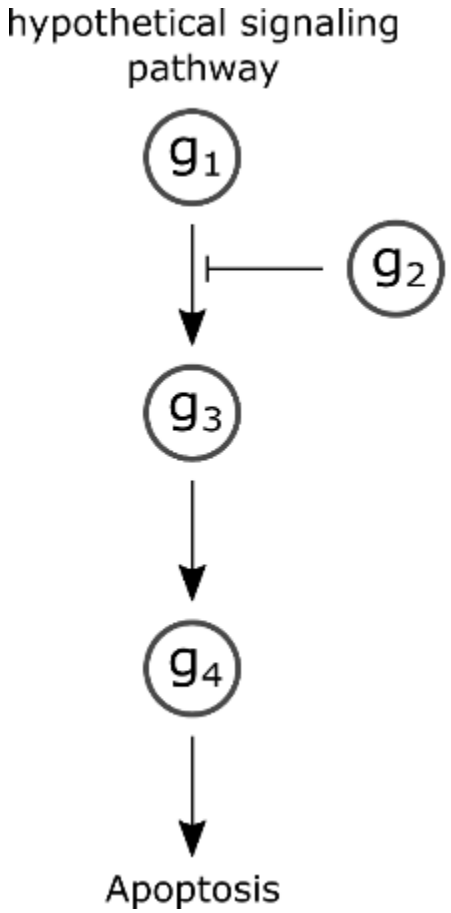
- Using a **physical interaction network**, we **prioritized** eQTLs spanning multiple genes to individual *causal* genes using a **kernel** based approach

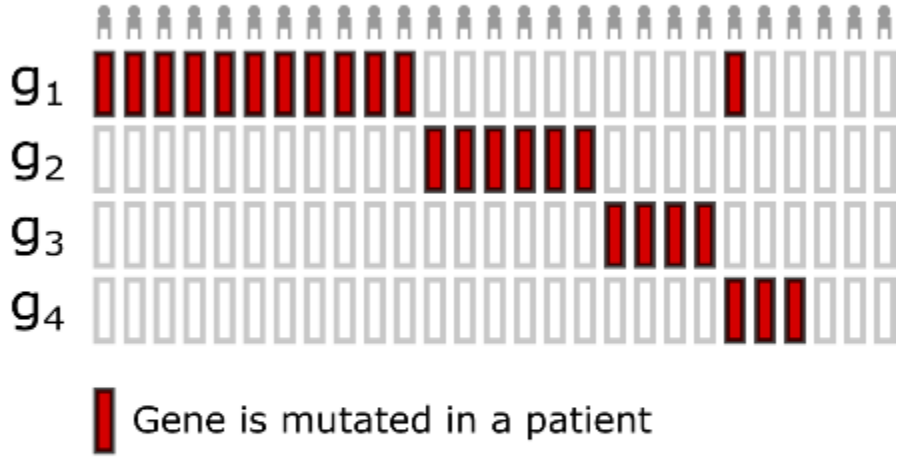
- We obtained superior results when evaluating using **knockout pairs**, and when compared to random assignment or a shortest path approach

Molecular subtypes in clonal systems



Mutual exclusivity of somatic variants within a pathway





Application 3: a data integration framework for tumour analysis

Human tumour samples

- Genome: 3,000,000 base pairs
- +/- 25,000 genes
- Samples were retrieved from the TCGA public repository
- Three different tumour types
  - Breast cancer (BRCA)
  - Glioblastoma multiforme (GBM)
  - Overian cancer (OV)

Gene expression data

- mRNA levels for all genes

Application 3: a data integration framework for tumour analysis

Mutation data

- Somatic mutations only
- Single nucleotide variants

Copy number data

- Structural variants
- Quantifies the number of copies of a gene are present in a tumour sample
- Will influence gene expression

Methylation data

- Epigenetic data
- Quantifies the methylation status of a gene
- In general, exciessive methylation will prevent gene expression

Application 3: a data integration framework for tumour analysis

Network data

- Derived from different public repositories
- In total, 12,000 genes are present in the network, with +/- 100,000 gene interactions

Clinical data

- Information of patients
- Contains age, sex, …
- Contains time of diagnosis, treatment
- Contains survival data

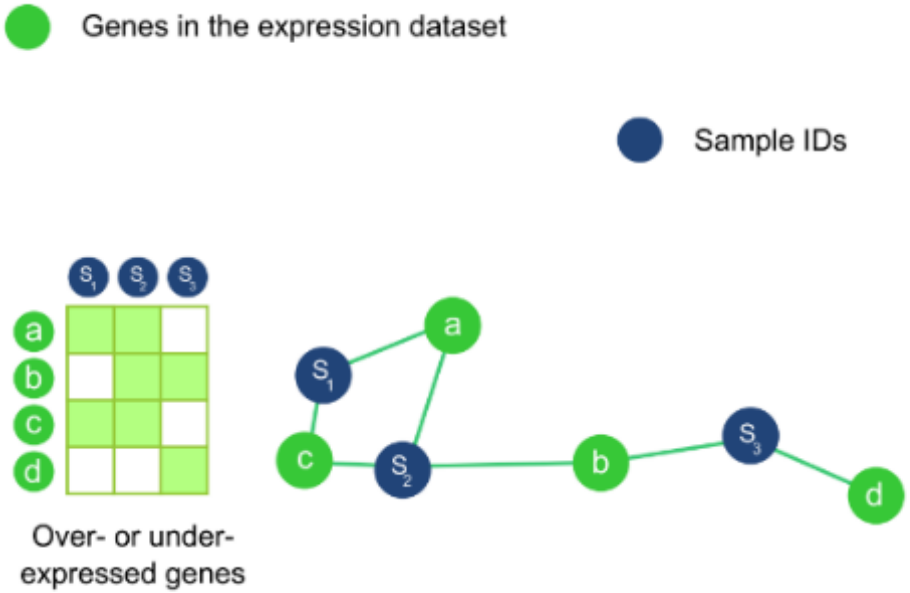
The problems we want to solve

? Find groups of patients that exhibit similar molecular properties

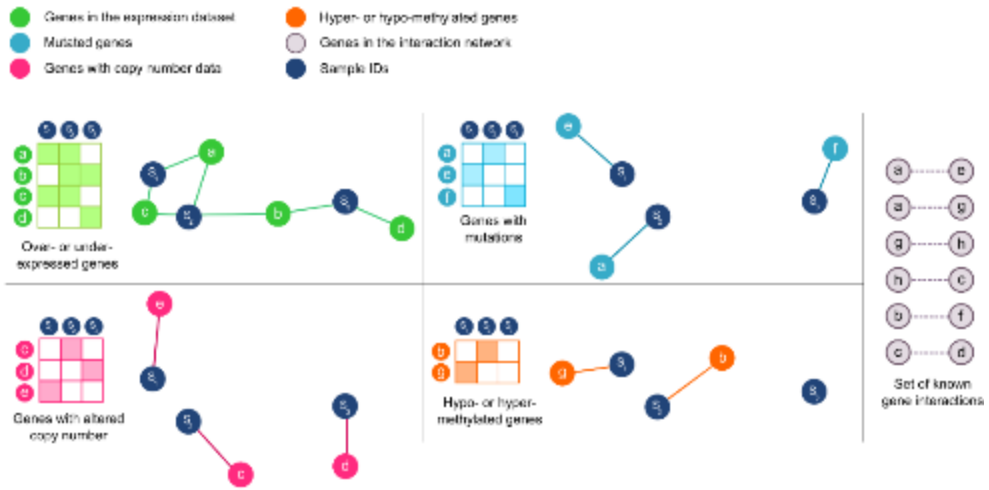
? Find out which genes and pathways are disturbed in a homogeneous set of patients

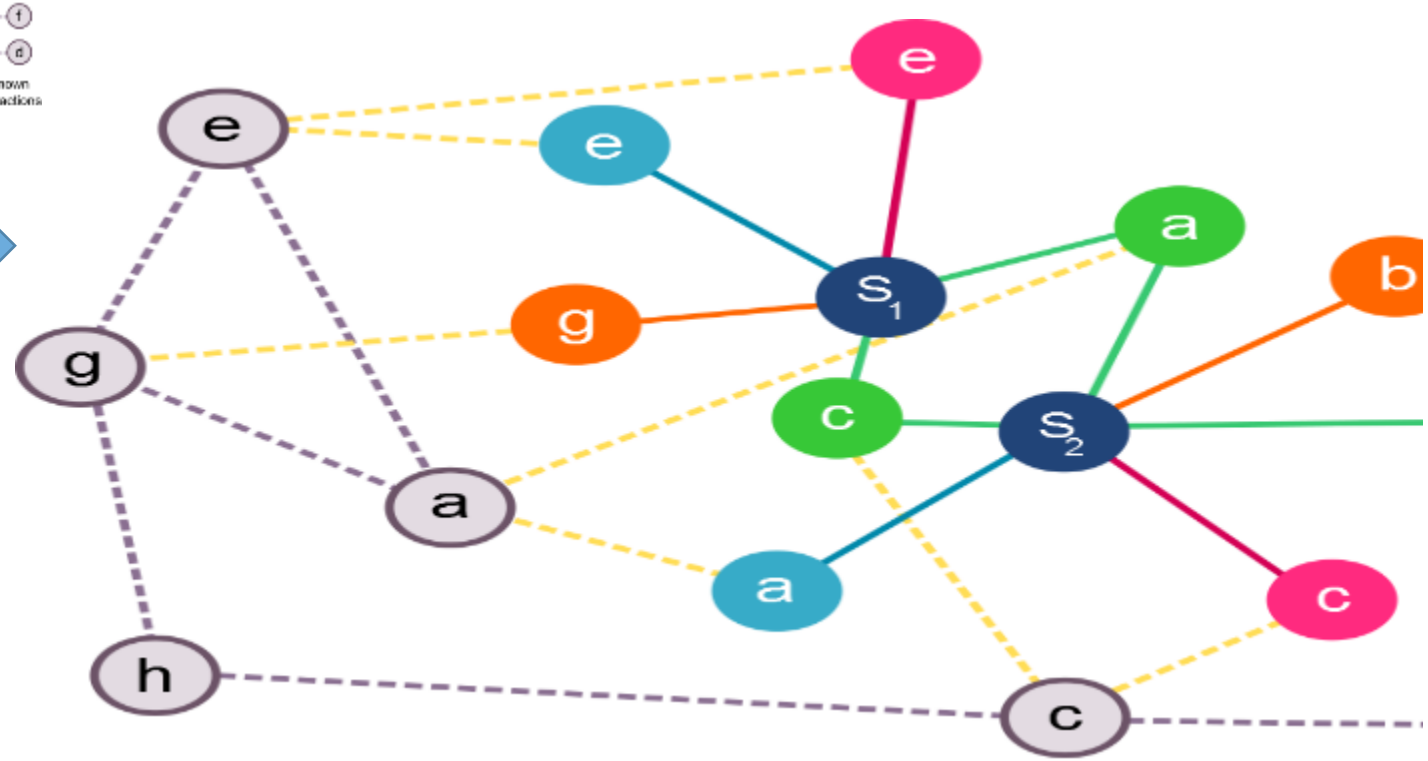
Solved using a method called **MUNDIS**: **MU**lti purpose **N**etwork-based **D**ata-**I**ntegration **S**trategy

Integrate all data into a single model

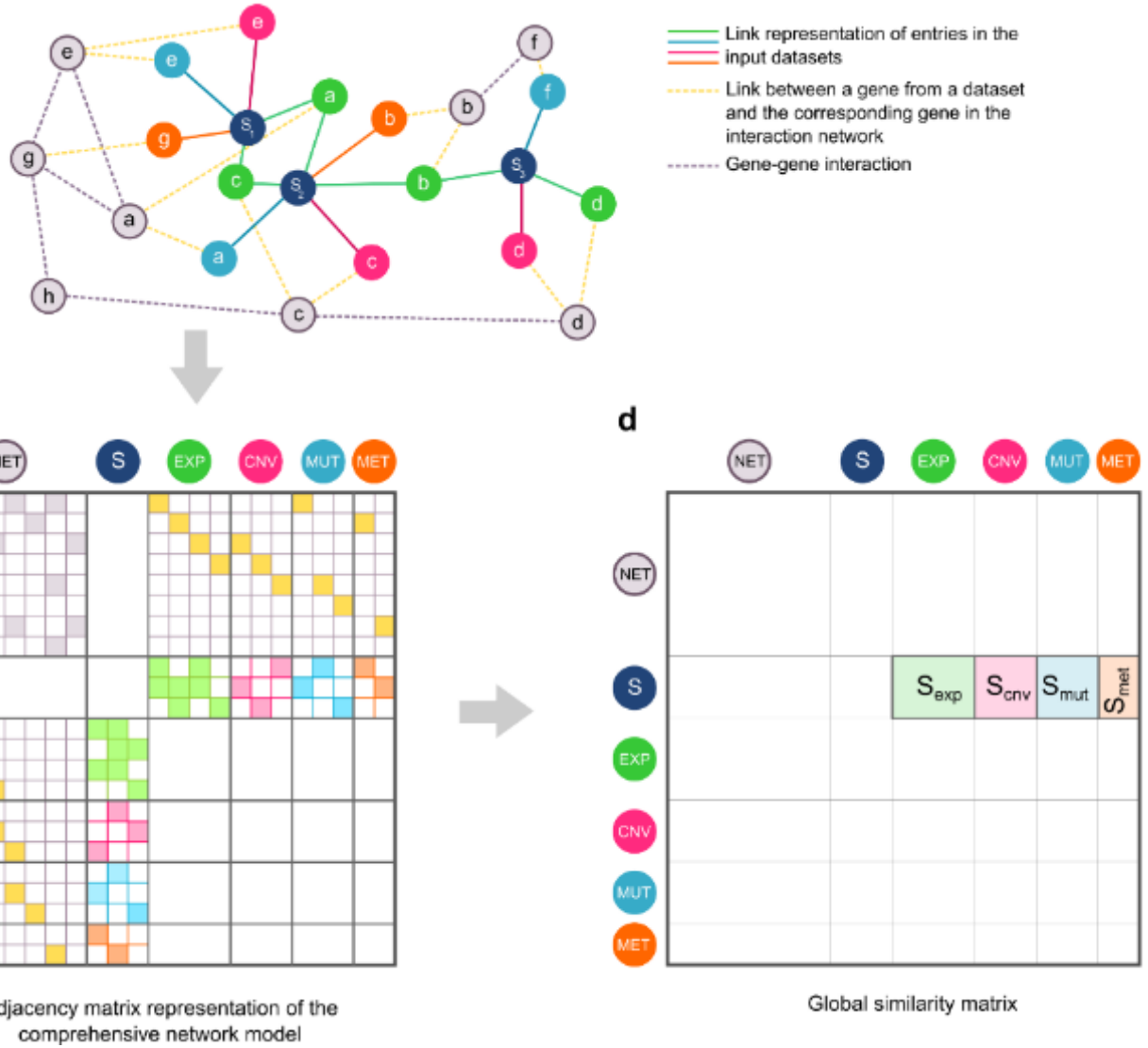


Integrate all data into a single model

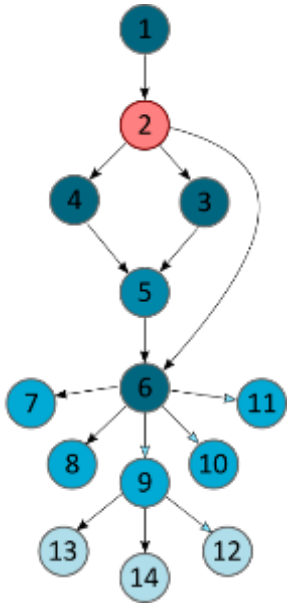




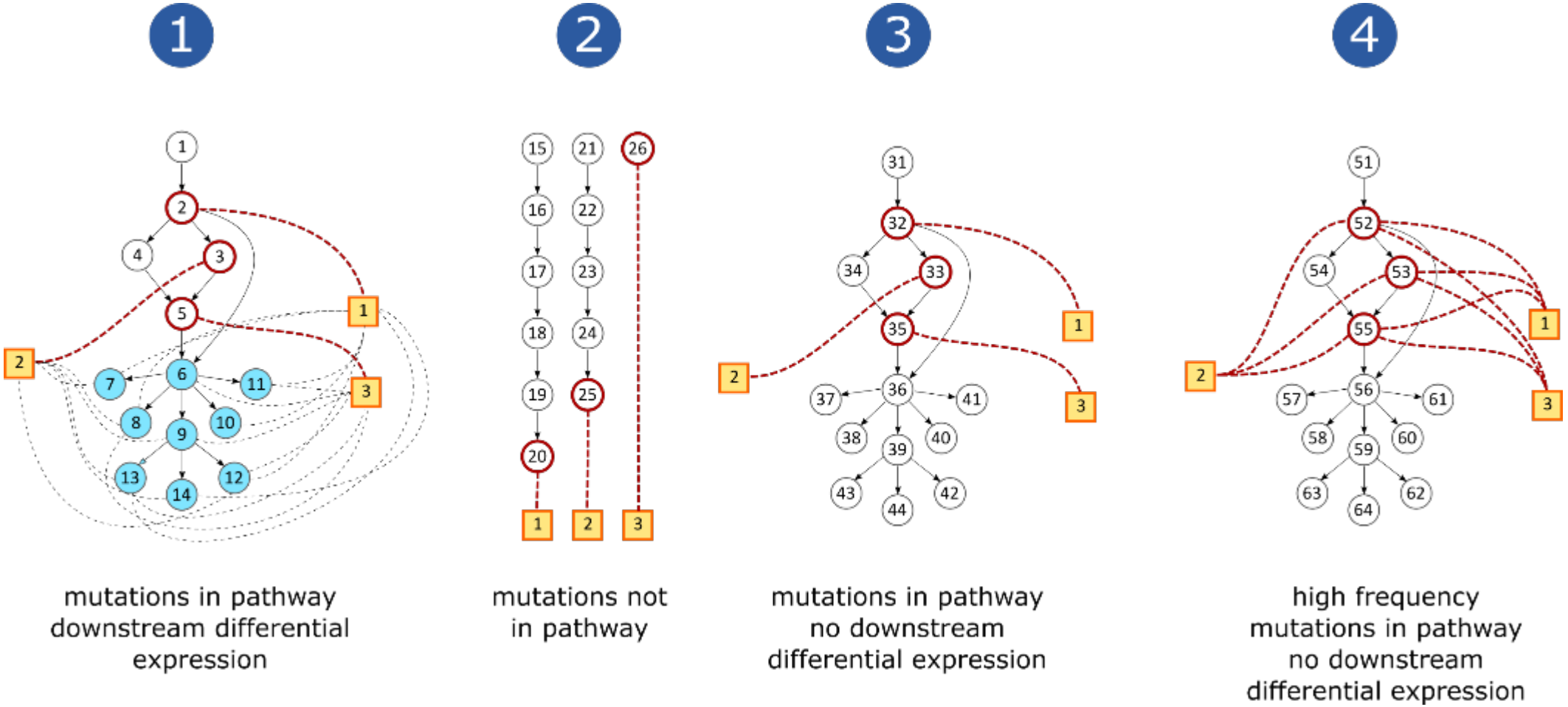
Calculate connectivity metrics



Remember this?



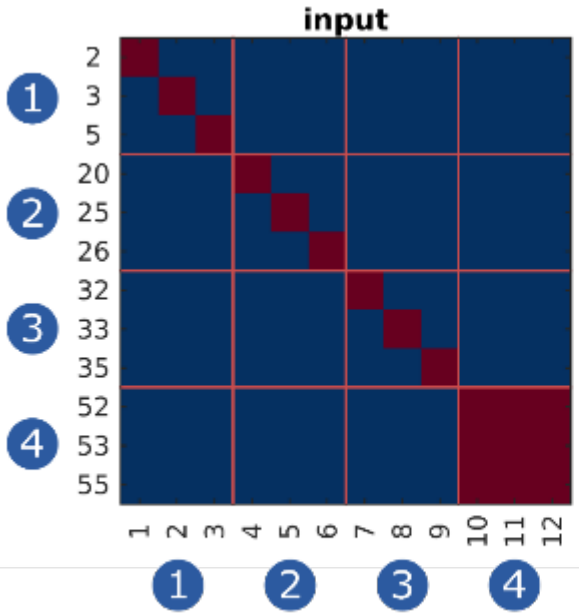
An intuition for the diffusion method



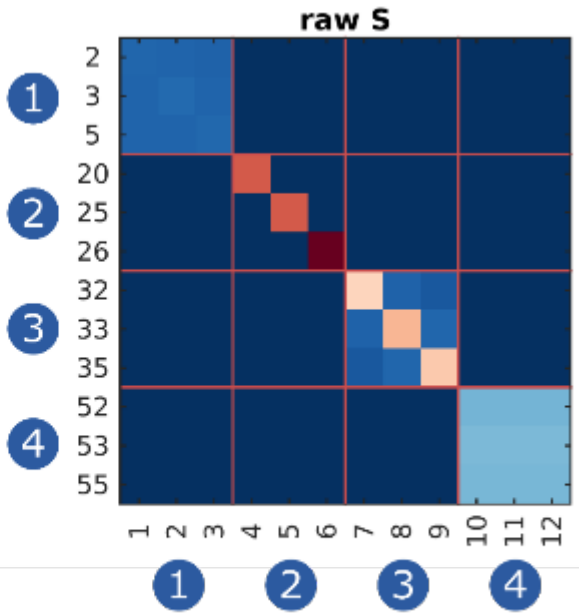
An intuition for the diffusion method

E

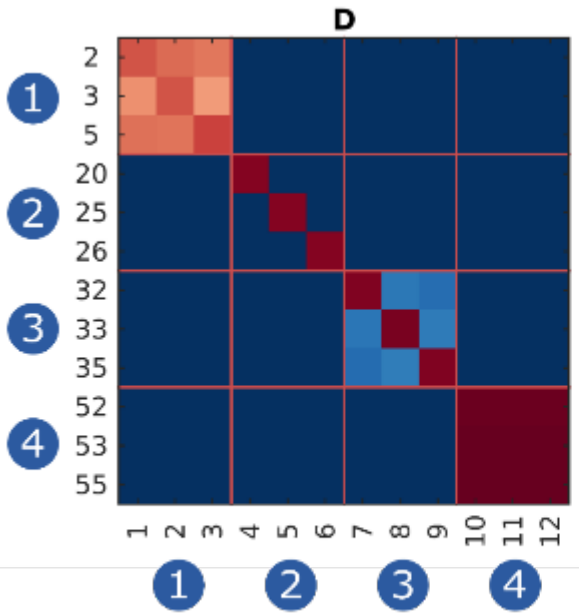
1. mutations in pathway downstream differential expression
2. mutations not in pathway
3. mutations in pathway no downstream differential expression
4. high frequency mutations in pathway no downstream differential expression



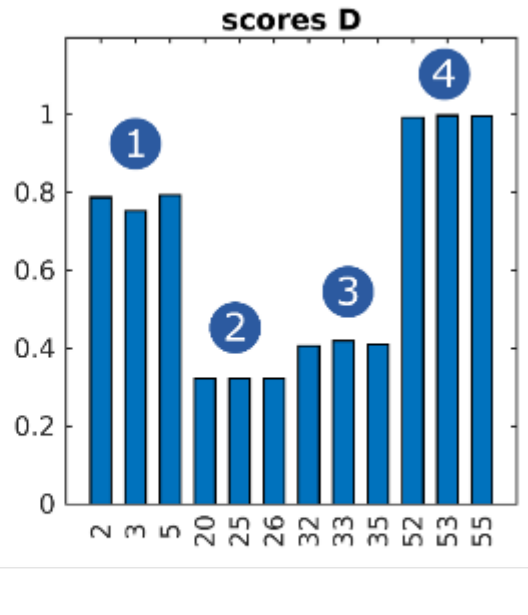
input mutation data



(part of) the network connectivity matrix
Laplacian exponential diffusion kernel



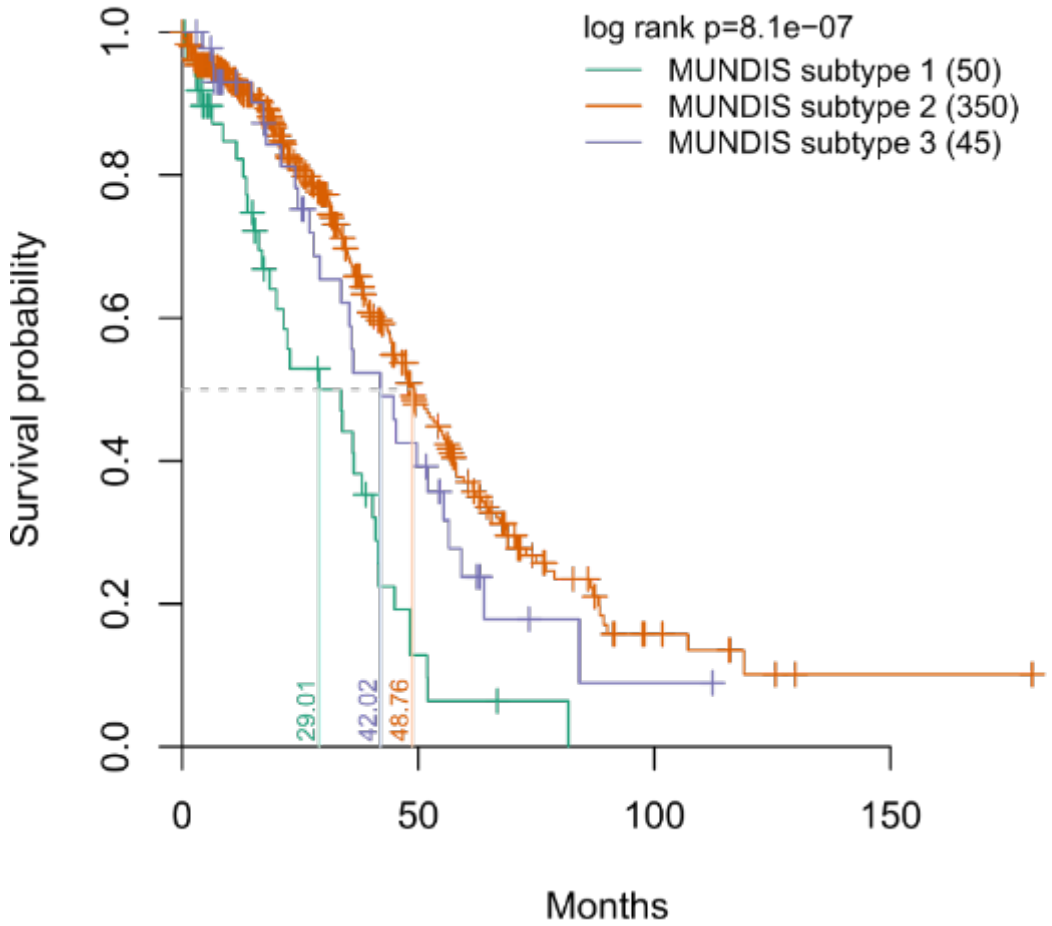
(part of) the normalized network connectivity matrix



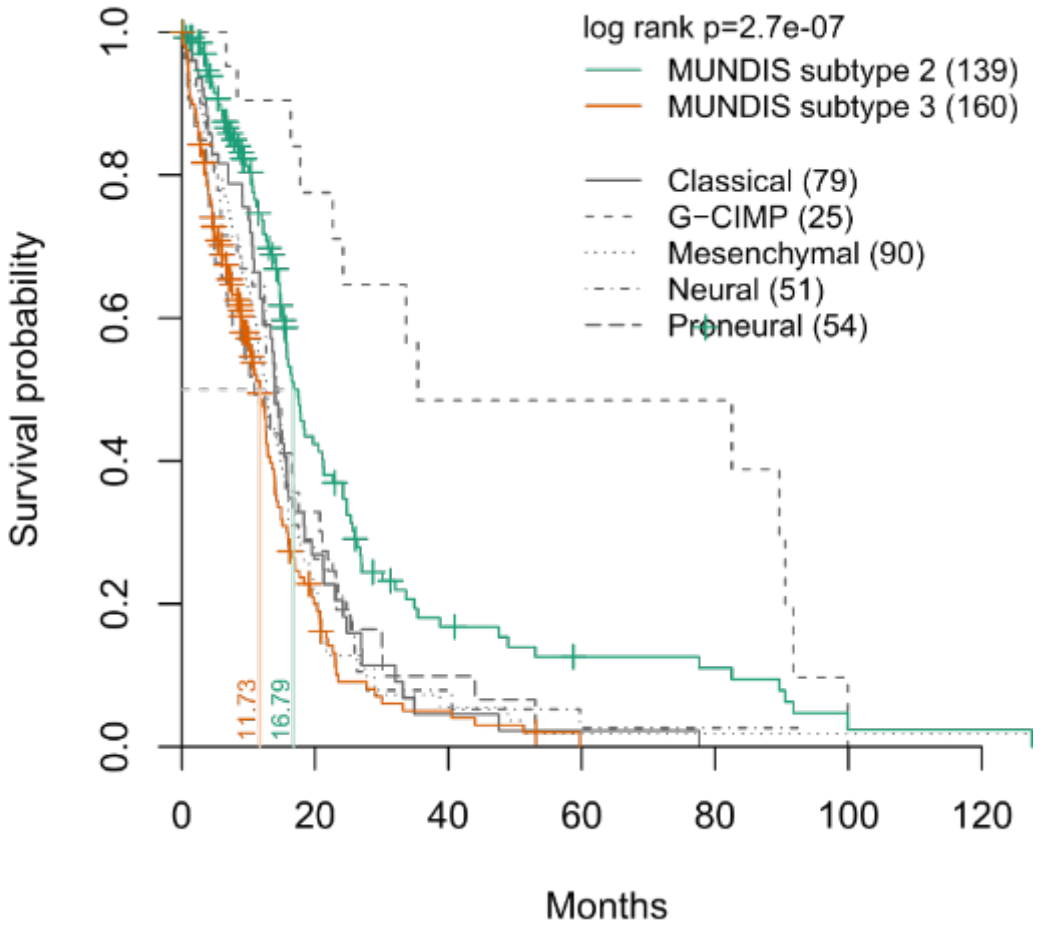
connectivity scores of the mutated genes averaged over the samples

Results: patient subtyping

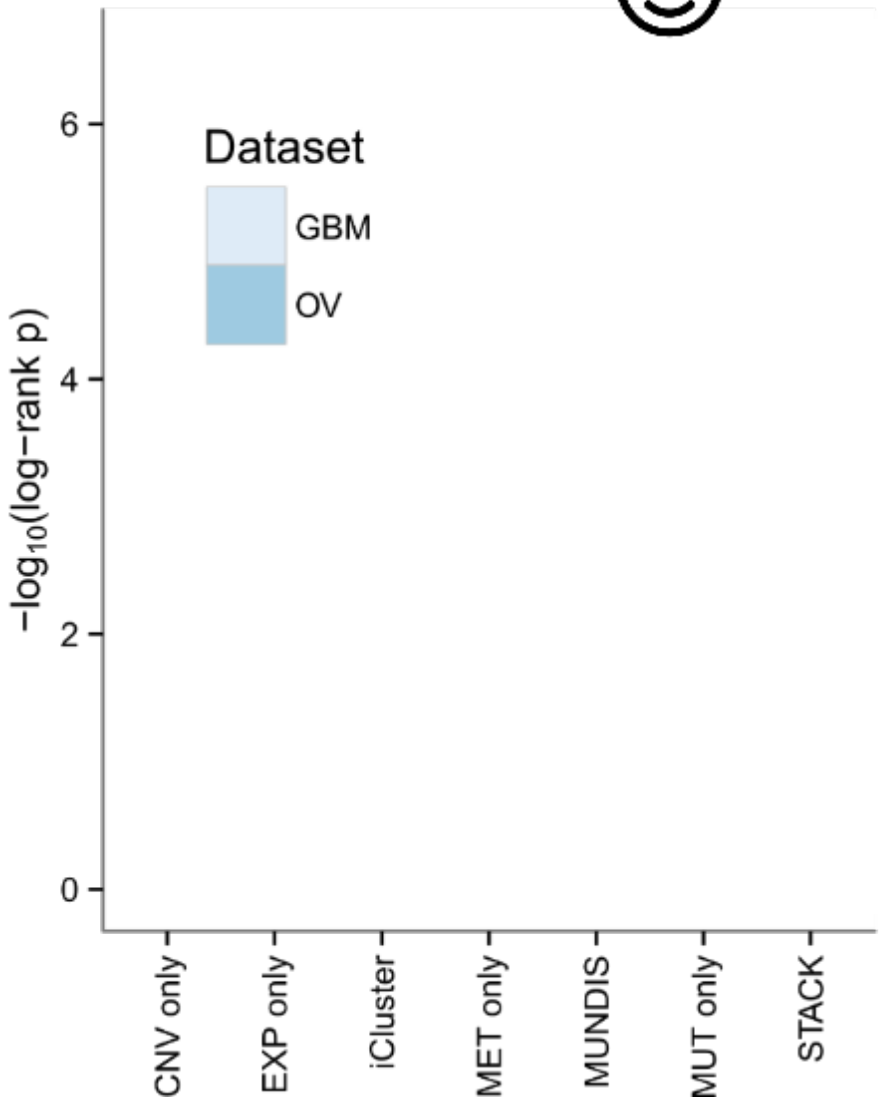
Ovarian cancer



Glioblastoma



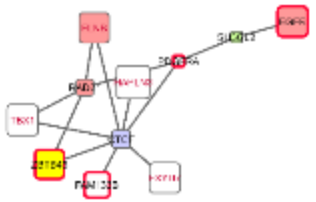
Results: patient subtyping



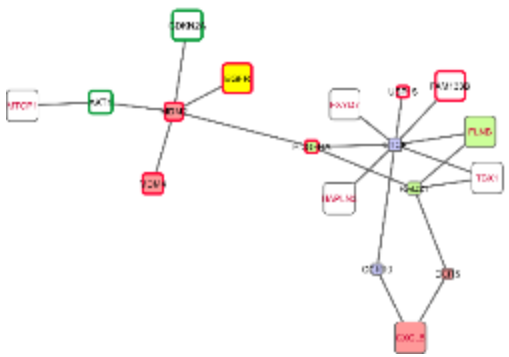
This says how good our patient groups correlate with patient survival

Results: driver networks for subtypes

Good outcome

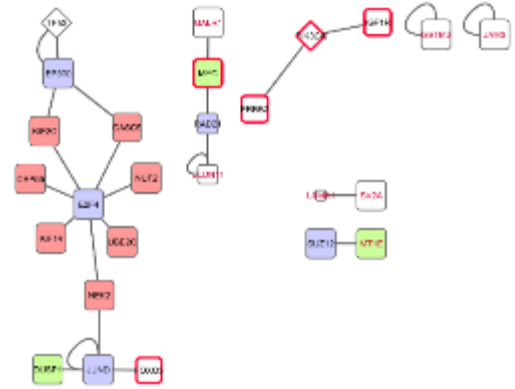


Bad outcome



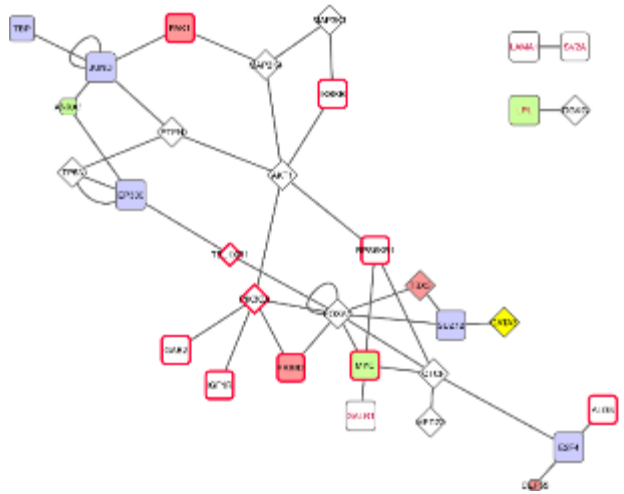
under-expressed
over-expressed
copy number loss
copy number amplification
mutated
methylated
selected network gene
selected in 20 percent of patients
selected in 100 percent of patients

Basal

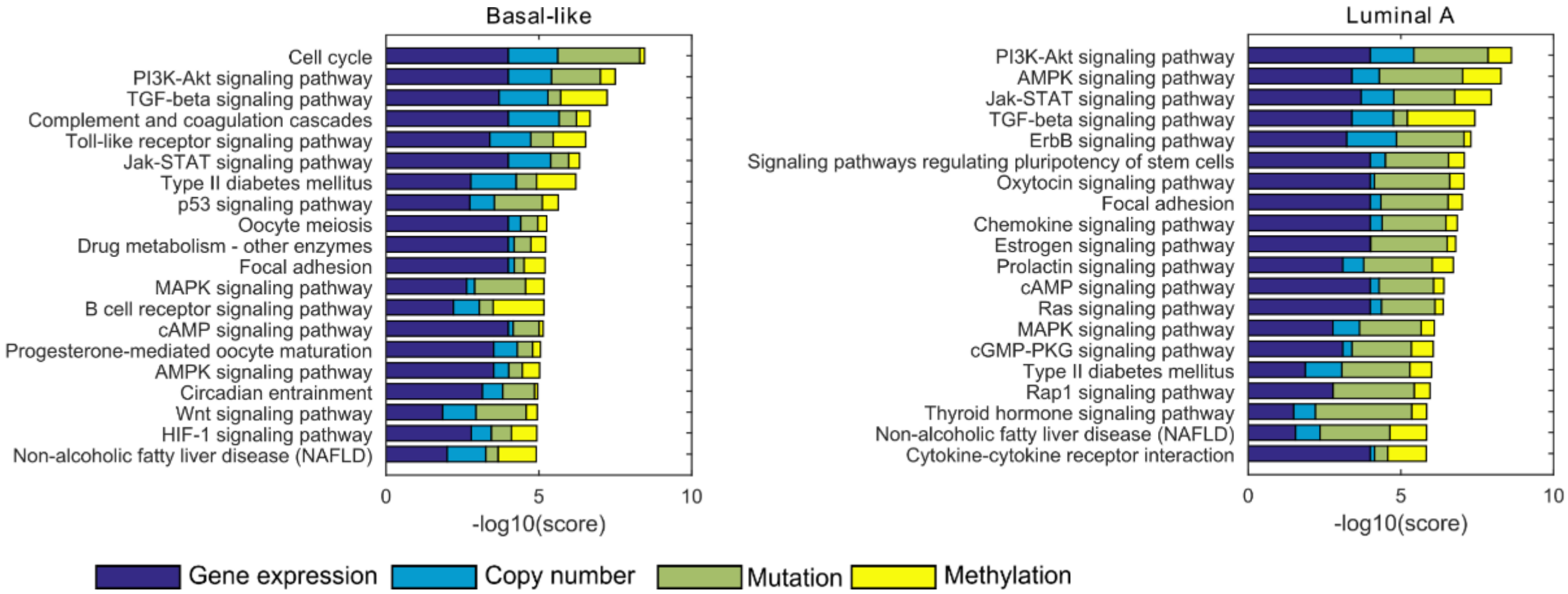


under-expressed
over-expressed
copy number loss
copy number amplification
mutated
methylated
selected network gene
selected in 20 percent of patients
selected in 100 percent of patients

Luminal A



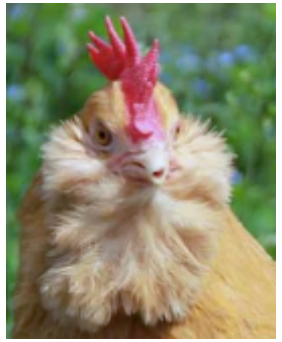
Results: pathway ranking: BRCA



Conclusions and take-home message

ACGCCTACC**GCA**ATGCTGAAA

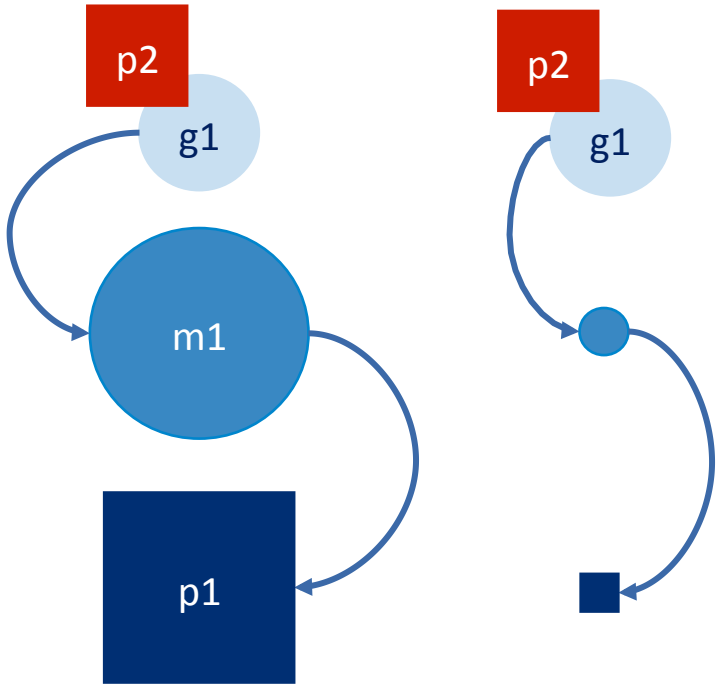
ACGCCTACC**CCT**ATGCTGAAA



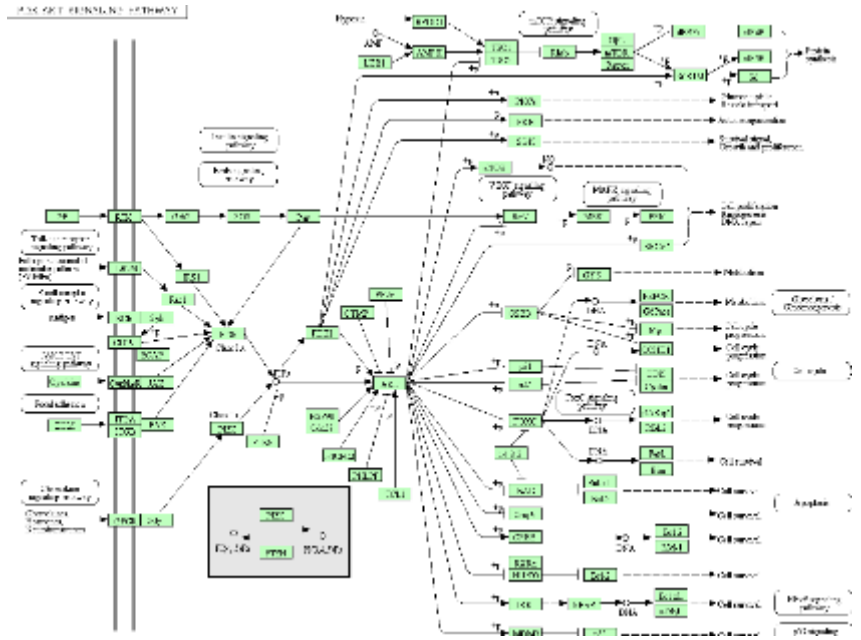
Genetic variability drives phenotypic variation

Most traits are influenced by many genes

Conclusions and take-home message

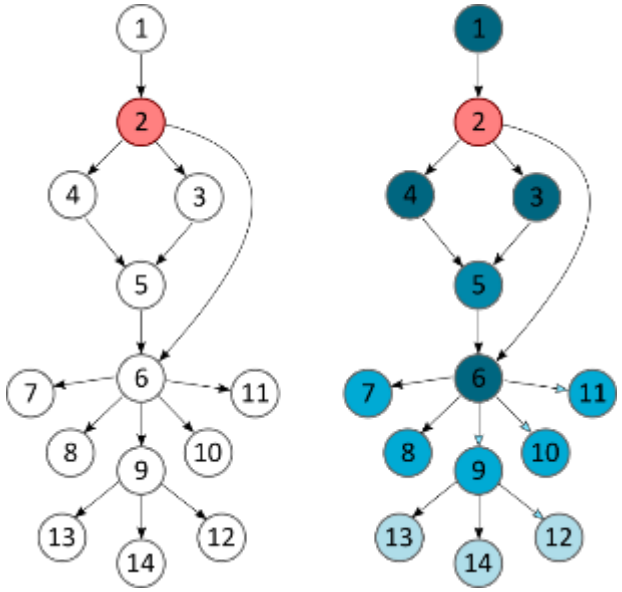


Genes can interact with each other

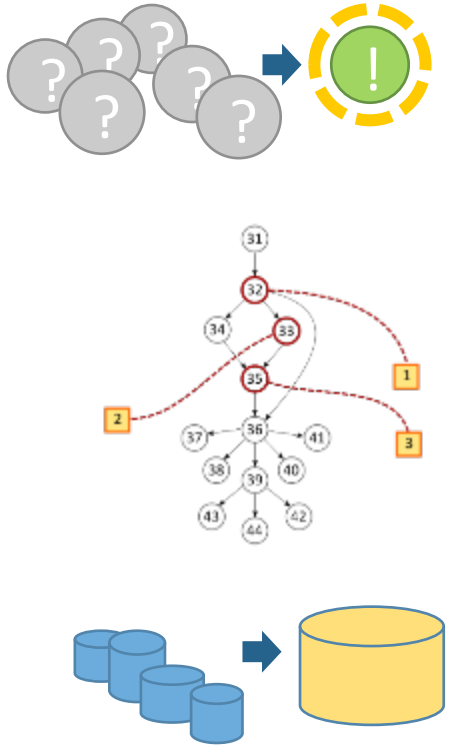


Interacting genes constitute pathways and networks

Conclusions and take-home message

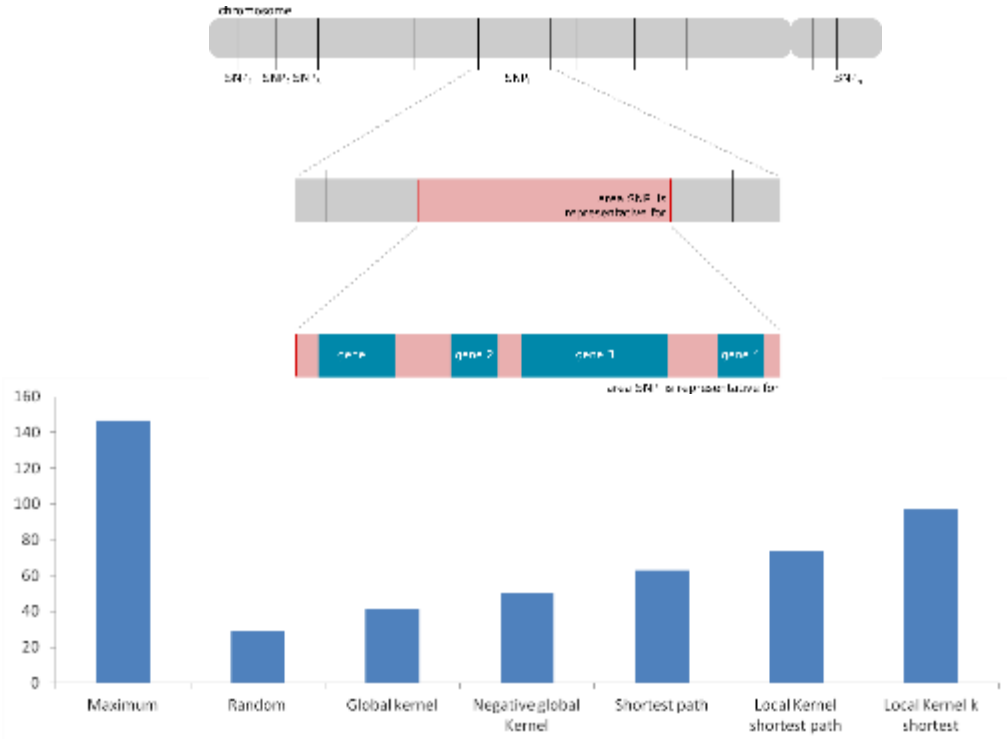


The better genes are connected in a network, the more likely they participate in the same biological processes

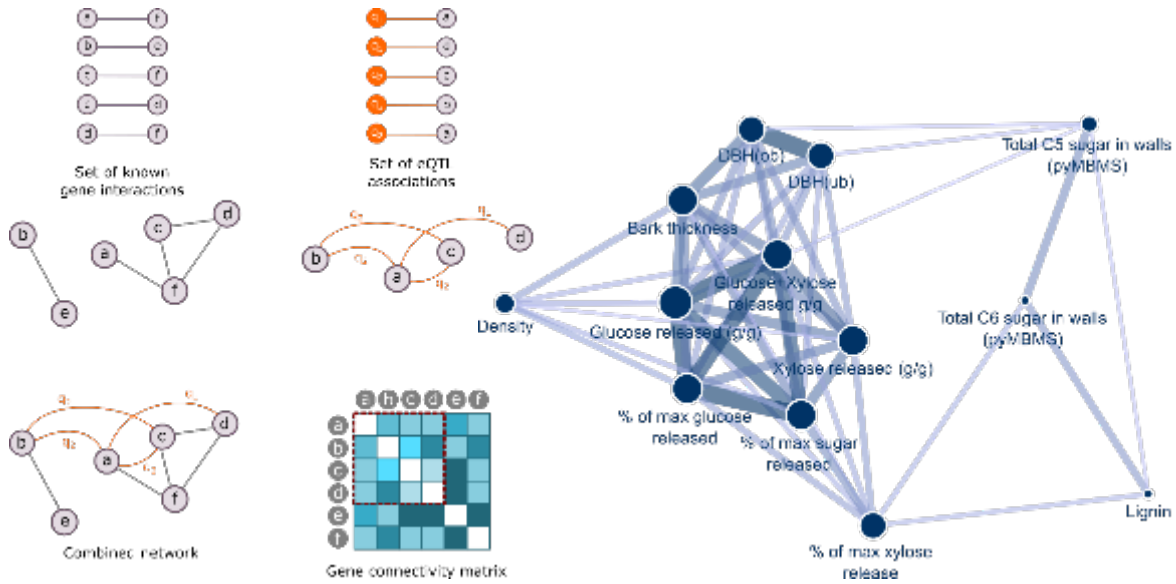


We developed several network-based methods for data-integration and gene prioritization

Conclusions and take-home message

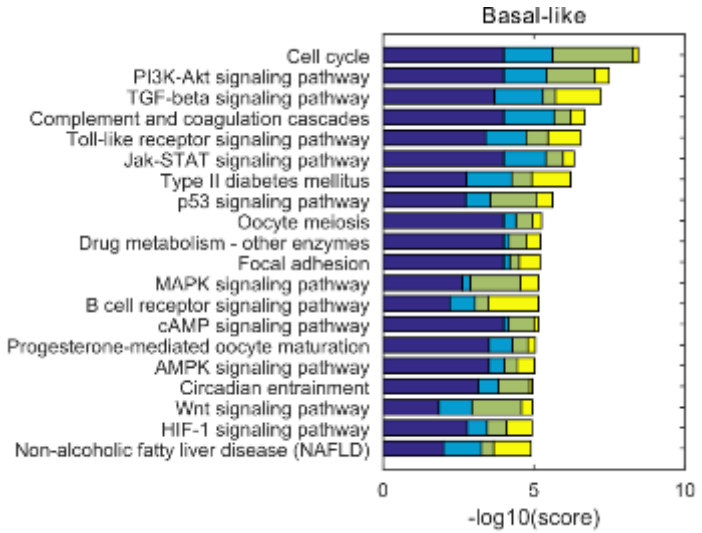


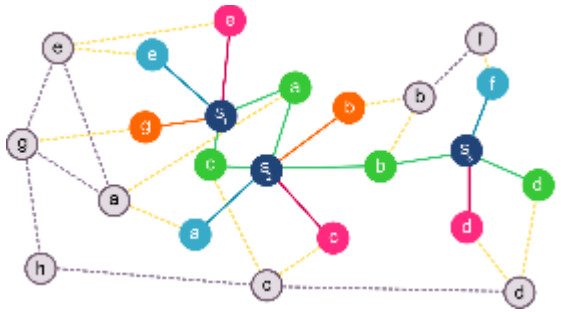
We developed an eQTL prioritization strategy

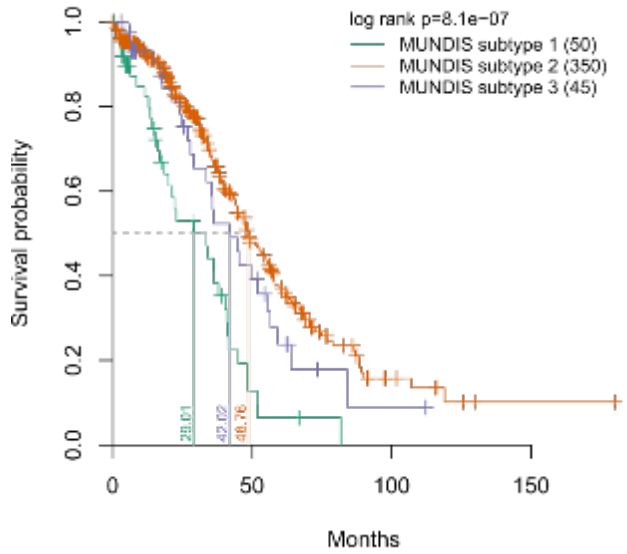


We found genes related to wood properties in Eucalyptus

Conclusions and take-home message







We could rank pathways according to their relevance for tumour samples

We could identify groups of patients with similar survival and molecular characteristics


Questions?