The complexity of diseases

Traditionally:
1 gene ↔ 1 disease (hemoglobin beta-chain ↔ sickle cell disease)

In reality:
• Many genes ↔ 1 disease (locus heterogeneity)
• 1 gene ↔ many diseases (allelic heterogeneity)
• Many genes ↔ many diseases

More generally:
• Diseases → complex disorders/traits
• Disease-causing mutations in genes → genetic variations/mutations that confer susceptibility (increased risk) to disorders
Explain the nodes and edges, colors and widths! What features do you observe?
How would the network look like if most diseases were monogenic?
What network features do you observe? What do giant nodes mean? Hubs?
Cancers vs metabolic disorders?
How does a random network look like? How to generate a “reasonable” random network?
How does this differ from our old friend, the PPI network? A form of genetic interaction network (Lecture 10)
Clustering patterns
How to interpret those?

More edges between nodes of the same class
What does that mean?
Is that surprising?
Validate your results (on orthogonal data) and flesh out your paper!

If two genes are involved in the same disorder, are they more likely to...

- Share GO terms?
- Physically interact?
- Be expressed at the same time & space?
- What constitutes “reasonable” random control for comparison?
Hubs, party hubs, disordered hubs...
Apply our template: **Hubs are more X...**

Are disease genes more likely to be hubs?
Weak correlation!

“If your experiment needs statistics, you ought to have done a better experiment.”

- Different types of hubs
- Different types of disease genes
- Heterogeneous dumpbin
- Side effect?

**Classify disease genes – essential vs nonessential**

- What constitutes essential genes in humans? (single deletion strain in yeast)
- Some truly critical genes are not associated with diseases (or you won't be born!)
- Are some disease genes really that serious (e.g. those leading to deafness)?
- Evolutionary pressure?
Be careful with your figure!
Cancer Classification

- Given a disorder and patients (cases):
  - Discover if there are subtypes of the disorder (**class discovery**)
  - Assign new cases to subtypes (**class prediction**)

- Start with cases with known subtypes (**supervised learning**):
  - Select features with most distinguishing power (feature selection)
  - Combine those features into a predictor and fine-tune parameters (training)
  - Evaluate its performance (testing, cross-validation)

- Start with cases with unknown subtypes (**unsupervised learning**):
  - Assign “subtypes” yourself based on clustering
  - Follow the same steps as in supervised learning
Dataset: 38 bone marrow samples (27 ALL, 11 AML)

- Classification traditionally done by “experience”
- Subtypes can have similar morphology but drastically different outcomes
- Need to target therapies to max efficacy and min toxicity
- Want to do this systematically through molecular markers

- What molecular markers can you think of?
- Why did they focus on mRNA data?
- Why didn't they explore the interconnections among marker genes?
Combine the most informative genes (2 types) into a predictor.

In a sample, is the expression of a predictor gene high or low (and how much)? Weighted vote.

Further weighted by the rank of predictor gene.
If the data were “random”, you would also get “marker genes”!

How many such “marker genes” would you get?

How strong (in terms of correlations with phenotypes) would those be?

What constitutes a “reasonable” random control?
Famous figure!

Top 25 for ALL and top 25 for AML

- Should you fix the number of marker genes?
- Should you have equal numbers of marker genes for the two classes?

How to explore the interconnections among those marker genes?
Disease Classification

Looking for molecular markers distinguishing between:

- Disease and healthy
- Subtypes of disease (AML vs ALL)
- Clinical courses of disease (drug response, metastasis, etc)

So far looked at the marker genes as a set

- How reproducible are those marker genes?
- Are known disease genes among the top hits?
- Biological context of novel marker genes?

Want to explore the interconnections among marker genes

- Clustering of marker genes is more significant (even if they are individually weak)
- Connector of marker genes looks promising (even if it is not significant itself)
- Better understanding of mechanisms and hypothesis generation
- More reproducible and better performance
Datasets: expression profiles of two cohorts of breast cancer patients
From different studies
Used different technologies

Vijver: 78 (M) + 217 (ok) = 295
Wang: 106 (M) + 180 (ok) = 286

Each study found ~70 marker genes
Prediction accuracy 60-70% on their own data (compare this with ALL-AML)
Performance got worse on the other dataset
Those 2 sets of marker genes share only 3 in common

Use marker subnetworks in place of marker genes!
Vijver: 149 marker subnetworks (618 genes), 88% associated with M
Wang: 243 marker subnetworks (906 genes), even split between M and ok

Those marker subnetworks are GO-enriched (what does that say?)
But proteins in any subnetworks would be more likely to share GO terms...
Marker subnetworks instead of marker genes!
How to score a marker subnetwork?
If the data were random, how many “marker subnetworks would you see and how strong would they correlate to “phenotypes”?
What constitutes “reasonable” random control in this case? (more than one!)

How to find marker subnetworks?
Combine the most informative marker subnetworks into a predictor (logistic regression)
Explain the figure!
More reproducible

Better performance

Biological insight

Still a long way to go!
A graph shows the percentage of known susceptibility genes for subnetwork markers and single-gene markers. The percentage decreases as the significance increases, with subnetwork markers having a higher percentage of known susceptibility genes compared to single-gene markers.

B graph also shows a decrease in percentage of known susceptibility genes as significance increases, with subnetwork markers again having a higher percentage.

C-E diagrams depict molecular interactions with proteins such as LRCH4, F11R, RTN3, RTN4, PDHA1, PDHX, EWSR, RAD23A, NLE1, GNL3I, GTPBP4, FN1, MAPK1, NRAS, SHC1, PSMC6, FRS2, ARHGEF7, and others. The diagrams indicate expression levels ranging from downregulated to upregulated, with specific cell and tissue remodeling, circulation and coagulation, apoptosis, signaling of cell growth and survival, and cell proliferation and replication pathways highlighted.