Orthologs, Paralogs, and Phenologs
Using bioinformatics to find new genes for genetic traits

BCH394P/364C Systems Biology / Bioinformatics
Edward Marcotte, Univ of Texas at Austin

Are you a research parasite?

“The aerial view of the concept of data sharing is beautiful.”

[but!]

A ... concern ... is that a new class of research person will emerge...the system will be taken over by ... “research parasites.”

“[they might] use another group’s data for their own ends... or even use the data to try to disprove what the original investigators had posited”
Today’s lecture gives a case study of using bioinformatics and public data to make new discoveries, specifically to discover new genes for genetic traits.
Let’s think in the abstract for a moment:

How are mouse models useful for studying human disease? Are some models better than others?

What’s the worm equivalent of breast cancer?

Are there plant versions of human diseases? Why would they possibly be useful?

Conserved systems intermediate between organism-specific genotype and phenotype

The systems are conserved; their mutant phenotypes often differ
A quick aside on terminology:
Comparative evolution studies rely on finding orthologs

**Orthologs** = genes from different species that derive from a single gene in the last common ancestor of the species

**Paralogs** = genes that derive from a single gene that was duplicated within a genome
Phenologs = significantly overlapping sets of orthologous genes, such that each gene in a given set gives rise to the same phenotype in that organism.

E.g., 'high incidence of male' *C. elegans* genes predict human breast/ovarian cancer genes

<table>
<thead>
<tr>
<th>Human/Worm Ortholog</th>
<th>Linked to breast cancer in humans</th>
<th>Linked to male disease in worms</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATM/atm-1</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>BRI1/dog-1</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>KRAS/let-50</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>PHB/phb-1</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>PIK3CA/age-1</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>RAD51/rad-51</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>RAD54L/rad-54</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SLC22A18/C53B4.3</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>TSG101/tag-1010</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>BARD1/brd-1</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>BRCA1/brc-1</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>CHEK2/ck2-2</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>FAM82B/F33H2.6</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>GCC2/hcp-1,hcp-2</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>HMG20A/B/W02D9.3</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>HORMA2D1/him-3,htp-1,2</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>KIF15/kip-10.18</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>MRE11A/mre-11</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>PIGA/D205S.6</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>RAD1/mrt-2</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>RAD21/coh-1</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SEH1L/npp-18</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SVIL/vin-1</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>TSP-O, BZRPL1/C41G7.3</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>WDHD1/F17C11.10</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

4649 orthogroups total

- Human breast/ovarian high incidence cancer male progeny

9 3 13

\[ p \leq 7.2 \times 10^{-6} \]

Includes BRCA1
Building & searching a collection of phenotypes

Mining available databases + manual collection from the primary literature

# gene-phenotype

<table>
<thead>
<tr>
<th>Organism</th>
<th>associations</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>1,923</td>
</tr>
<tr>
<td>mouse</td>
<td>74,250</td>
</tr>
<tr>
<td>worm</td>
<td>27,065</td>
</tr>
<tr>
<td>yeast</td>
<td>86,383</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>22,921</td>
</tr>
</tbody>
</table>

Spanning ~300 human diseases, >7,000 model organism mutational phenotypes

Computational scan phenotypes for novel models of a disease of interest, identify significant phenologs using permutation tests

Discovering phenologs

Measure \( p(overlap \geq k | n_1, n_2, N) \) for each disease-phenotype pair, considering only human-yeast orthologs

Identify all significant phenologs by permutations or reciprocal best hits
Computationally, we find many genes shared between human diseases and mouse, yeast, worm, and even plant traits.

Waardenburg syndrome accounts for ~2-5% of cases of deafness.
Plants sense and respond to gravity → gravitropism

Fukaki et al., The Plant Journal 14, 425–430 (1998)

Plant gravitropism genes predict Waardenburg syndrome, a human congenital deafness syndrome

Waardenburg syndrome genes (with plant orthologs)

Defective gravitropism genes (with human orthologs)

Human versions of these plant genes are candidate Waardenburg genes


Waardenburg syndrome

Gravitropism defects

Waardenburg syndrome is a defect of neural crest cells

Some WS correlates in other animals:
Deafness in Dalmatian dogs (22% unilaterally deaf)

Variations in the Blenheim spot of Cavalier King Charles Spaniels

Association between white blue-eyed cats and deafness (noted by Darwin in 1859)

White forelock and deafness/bowel blockage in foals & many more...

Sure enough, inactivating one of the genes—predicted from plants—in a tadpole disrupts neural crest cells, consistent with Waardenburg syndrome
**Conservation and “repurposing” of core cell machinery across deep evolutionary time**

Last common ancestor

Set of genes in LCA

Plant

Genes now used to direct polarized growth in gravitropism

Orthologous genes

Human

Genes now used to direct neural crest cell migration

**Another example: Yeast genes linked to statin sensitivity predict blood vessel defects**

Angiogenesis abnormal in mice

Lovastatin sensitive in yeast

3 5 62

The human versions of these yeast genes are candidate angiogenesis genes

Can these really tell us about these?

Disrupting the SOX13 gene causes strong blood vessel defects


Last common ancestor
Gene module in LCA

A yeast model of angiogenesis = example of a deeply conserved, but “repurposed” gene module

McGary, Park et al. PNAS (2010)
The yeast/angiogenesis gene module

Chemicals that interact genetically with this module are candidate angiogenesis inhibitors
TBZ = thiabendazole
FDA-approved antifungal drug with 40 years of safety data

Approved by U.S. Food and Drug Administration in 1967

- Fungicide and parasiticide
- Not mutagenic or carcinogenic; 2 year dog safety trials
- Off-patent, marketed as a generic

Screening for drugs that interact genetically with this yeast module led us to identify a new angiogenesis inhibitor

TBZ = thiabendazole
FDA-approved antifungal drug with 40 years of safety data

Imaging the blood vessels of a living, transgenic tadpole in a dish of water

kdr:GFP transgenic Xenopus laevis
TBZ disrupts vascular integrity, making vascular endothelial cells retract & round up

Control (DMSO carrier) + TBZ

TBZ specifically inhibits just 1 of the 9 human beta tubulins, TUBB8, needed for angiogenesis

Steric clashes

Found 5 more, including benomyl & 3 more FDA approved antifungal drugs
Let’s talk about how such projects play out in practice.

How are discoveries made? How do you computationally explore ideas?

Let’s step through this particular discovery process:

1. We had an idea, based on a puzzling observation:

   Why do mutations in worm retinoblastoma genes induce ectopic vulva while a mutation in the human ortholog causes eye cancer?

   We weren’t interested in specific mechanism here, but rather the impact of organismal context on conserved systems. In particular, how do ever-more distant evolutionary models inform us about human disease?
Let’s step through this particular discovery process:

2. We thought about how this might be part of a large trend—does it illustrate a general principle? Could we look for new cases systematically?

3. We thought about other examples, mentally assembling what could serve as positive and negative control cases. *i.e.* how do we decide if a systematic approach is working?

This might be the single most important lesson in the entire class: Computational analyses need controls, just like wet lab experiments.

4. A grad student (Kris McGary) started assembling relevant datasets. We took heavy advantage of existing resources: model organism databases that had already painstakingly curated relevant data, large-scale screens reporting easy-to-process data.
Let’s step through this particular discovery process:

5. We started inventing/evaluating statistical models/algorithms, exploring the data and thinking about how to search for the relevant trends. We iterated these steps until we thought we understood the problem better.

6. At some point, the lab bet a 6 pack of beer on the outcome:
   Can we discover plant models of human disease?

7. The algorithms predicted some remarkable and crazy results. We had no option but to test or reject the new predictions, so began testing, thanks to collaborators in the Wallingford lab willing to sink a few weeks into high-risk experiments.

8. Some tests worked, some didn’t. We went back & thought about the ones that didn’t and refined how we prioritized the results.

9. Iterate, iterate. Jackpot! A plant model of deafness! Shouting in the halls...
Try it out yourself!
http://www.phenologs.org

You can start by rediscovering the plant model of Waardenburg syndrome:

Search known diseases for "Waardenburg", or enter the human genes linked to Waardenburg (Entrez gene IDs 4286, 5077, 6591, 7299) to start.

Tools for finding orthologs are linked on the class website