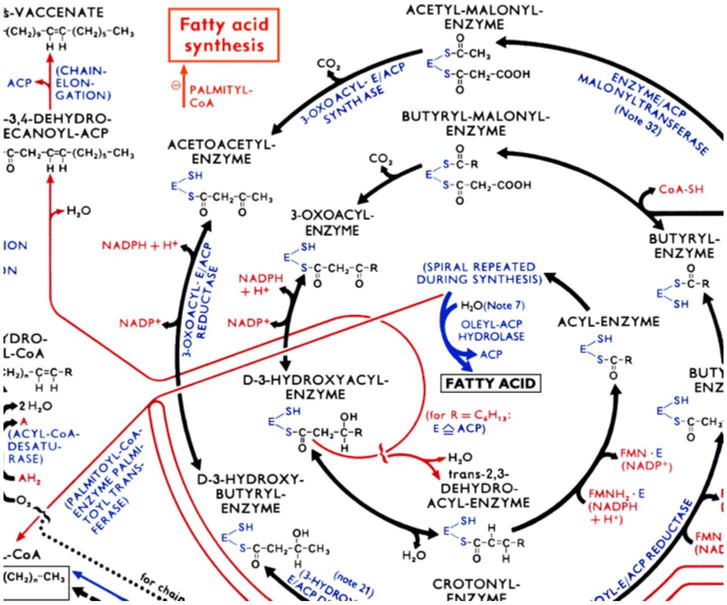
Network biology
(& predicting gene function)

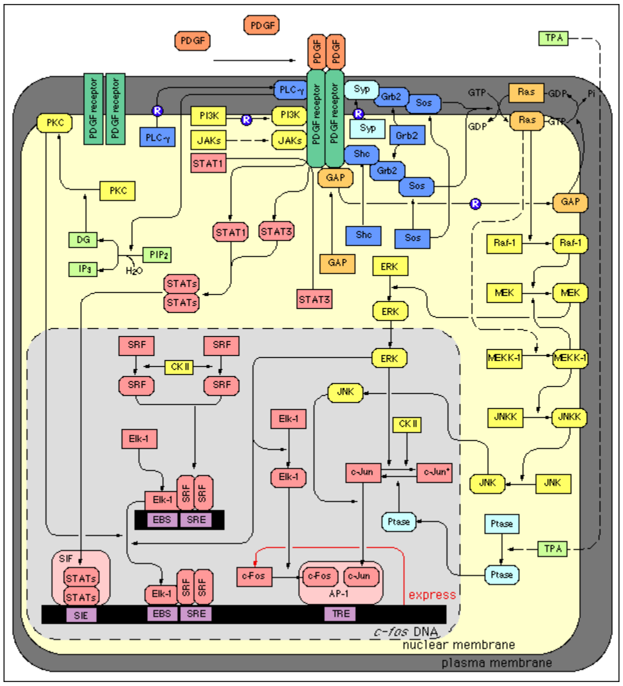
BCH394P/364C Systems Biology / Bioinformatics

Edward Marcotte, Univ of Texas at Austin

There are many types of biological networks.
Here's a small portion of a large metabolic network.

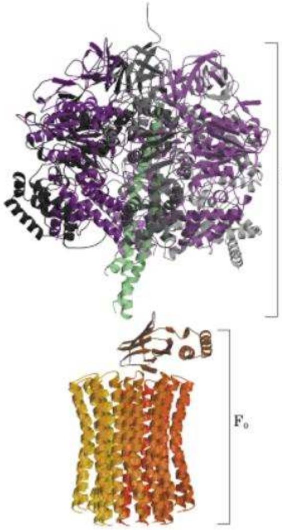


A typical genetic network

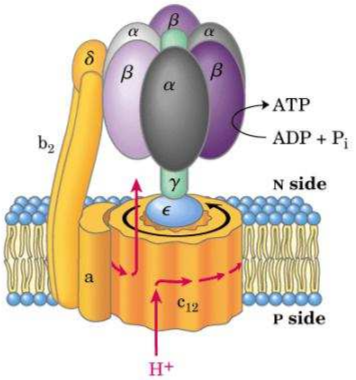


Contacts between proteins define protein interaction networks

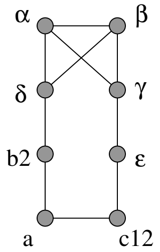
X-ray structure of ATP synthase



Schematic version



Network representation



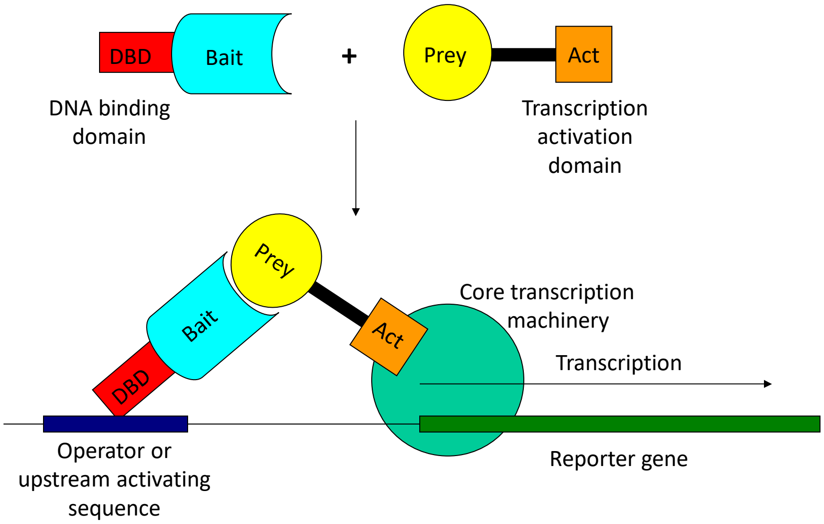
Total set = protein complex
Sum of direct + indirect
interactions

Let’s look at some of the types of interaction data in more detail.

Some of these capture physical interactions, some genetic, some informational or logical.

Pairwise protein interactions

In general, purifying proteins one at a time, mixing them, and assaying for interactions is far too slow & laborious. We need something faster! Hence, high-throughput screens, e.g. yeast two-hybrid assays



DNA binding domain

Transcription activation domain

Core transcription machinery

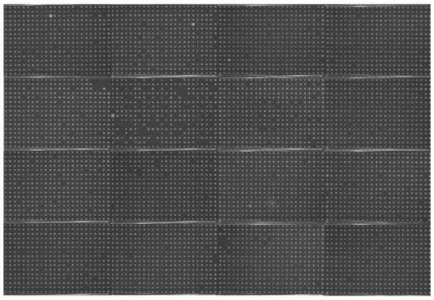
Transcription

Operator or upstream activating sequence

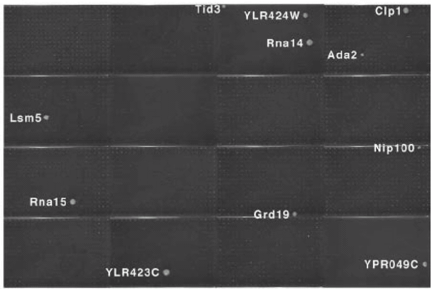
Reporter gene

High-throughput yeast two-hybrid assays

Haploid yeast cells expressing activation domain-prey fusion proteins

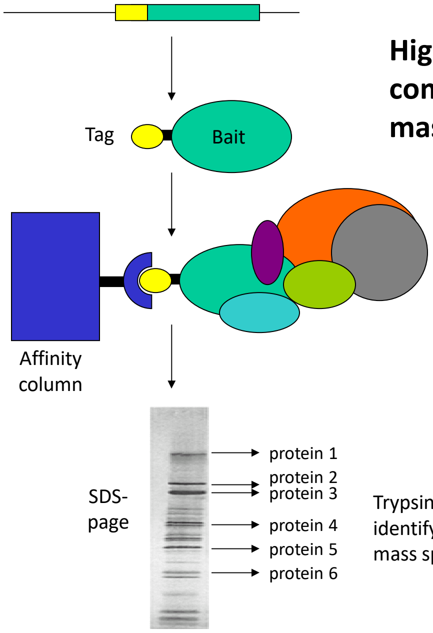


Diploid yeast probed with DNA-binding domain-Pcf11 bait fusion protein



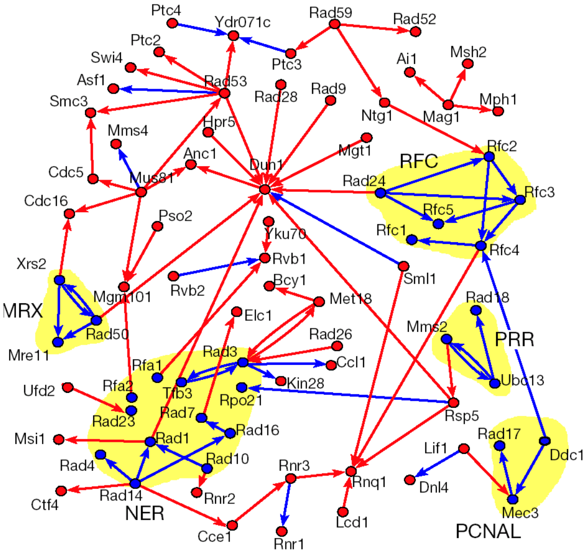
Uetz, Giot, *et al. Nature* (2000)

Protein complexes



High-throughput complex mapping by mass spectrometry

Trypsin digest, identify peptides by mass spectrometry



493 bait proteins
↓
3617 interactions

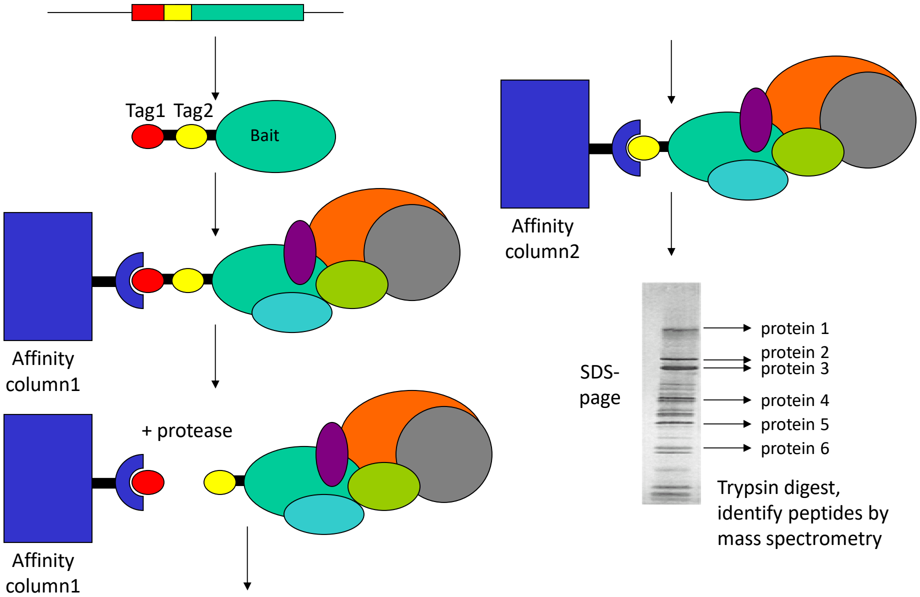
Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry

Yuen Ho*, Albrecht Gruhler*, Adrian Heilbut*, Gary D. Bader†‡, Lynda Moore*, Sally-Lin Adams*, Anna Millar*, Paul Taylor*, Keiryn Bennett*, Kelly Boutilier*, Lingyun Yang*, Cheryl Wolting*, Ian Donaldson*, Søren Schandorff*, Juanita Shewnarane*, Mai Vo*†, Joanne Taggart*†, Marilyn Goudreault*†, Brenda Muskat*, Cris Alfarano*, Danielle Dewar†, Zhen Lin†, Katerina Michalickova†‡, Andrew R. Willems†§, Holly Sassi†, Peter A. Nielsen*, Karina J. Rasmussen*, Jens R. Andersen*, Lene E. Johansen*, Lykke H. Hansen*, Hans Jespersen*, Alexandre Podtelejnikov*, Eva Nielsen*, Janne Crawford*, Vibeke Poulsen*, Birgitte D. Sørensen*, Jesper Matthiesen*, Ronald C. Hendrickson*, Frank Gleeson*, Tony Pawson†§, Michael F. Moran*, Daniel Durocher†§, Matthias Mann*, Christopher W. V. Hogue*†‡, Daniel Figeys* & Mike Tyers†§

NATURE | VOL 415 | 10 JANUARY 2002

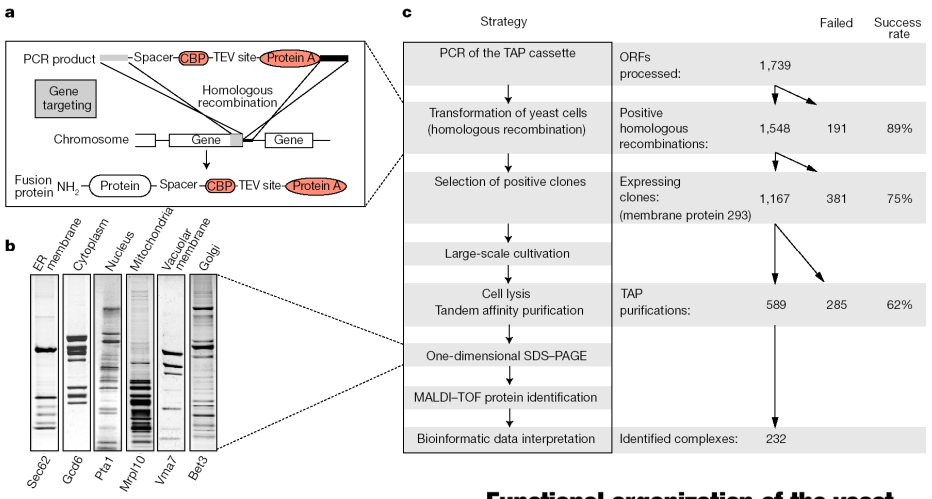
9

A variant: tandem affinity purification (TAP)



Trypsin digest, identify peptides by mass spectrometry

10



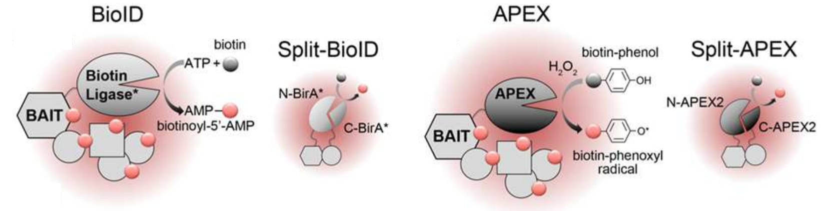
Functional organization of the yeast proteome by systematic analysis of protein complexes

Anne-Claude Gavin*, Markus Bösche*, Roland Krause*, Paola Grandi*, Martina Marzioch*, Andreas Bauer*, Jörg Schultz*, Jens M. Rick*, Anne-Marie Michon*, Cristina-Maria Cruciat*, Marita Remor*, Christian Höfert*, Malgorzata Schelder*, Miro Brajenovic*, Heinz Ruffner*, Alejandro Merino*, Karin Klein*, Manuela Hudak*, David Dickson*, Tatjana Rudi*, Volker Gnau*, Angela Bauch*, Sonja Bastuck*, Bettina Huhse*, Christina Leutwein*, Marie-Anne Heurtier*, Richard R. Copley†, Angela Edelmann*, Erich Querfurth*, Vladimir Rybin*, Gerard Drewes*, Manfred Raida*, Tewis Bouwmeester*, Peer Bork†, Bertrand Seraphin†‡, Bernhard Kuster*, Gitte Neubauer* & Giulio Superti-Furga*†

NATURE | VOL 415 | 10 JANUARY 2002

High-throughput protein interaction mapping by *proximity labeling*

4 variations on a theme:



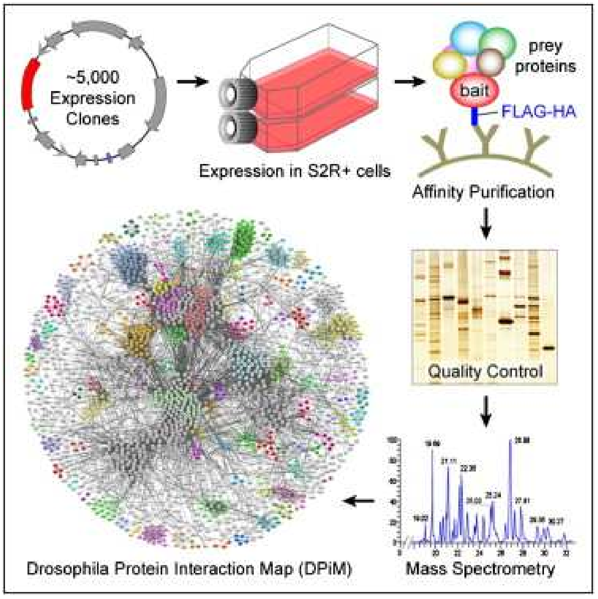
Affinity purify interaction partners with streptavidin beads

↓

ID proteins by mass spectrometry

Figure/review at https://f1000research.com/articles/8-135/v1

The current state-of-the-art in animal PPI maps – AP/MS



~3,500 affinity purification experiments

~11K interactions /
~2.3K proteins

→ spans 556 complexes

Still daunting for the human proteome, but…

Guruharsha *et al.* (2011) *Cell* 147, 690–703

The current state-of-the-art in human PPI maps – Y2H

Human ORFeome (v9.1) → now ~90% of the protein-coding genes!

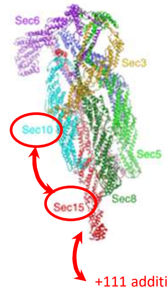
↓

Screened *all x all* (150M pairs!) in 9 Y2H assays

↓

52,569 PPIs involving 8,275 proteins

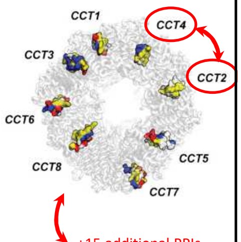
https://www.nature.com/articles/s41594-017-0016-2



+111 additional PPIs

Y2H captures pairwise PPIs that can form when the proteins are expressed out of biological context (e.g., as fusion proteins in a yeast cell nucleus). It can reveal directly contacting proteins but often misses those that require additional molecular context or higher order assemblies,
← the exocyst e.g. the CCT complex→

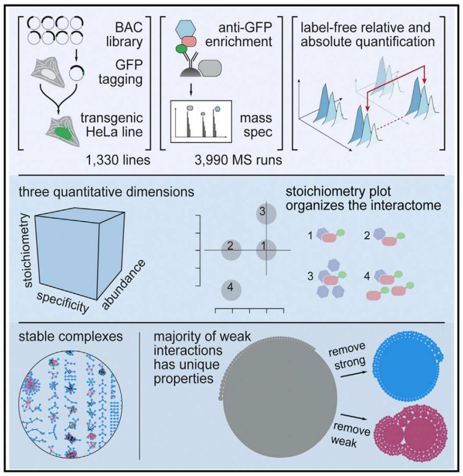
https://www.cell.com/cell/fulltext/S0092-8674(14)01369-5



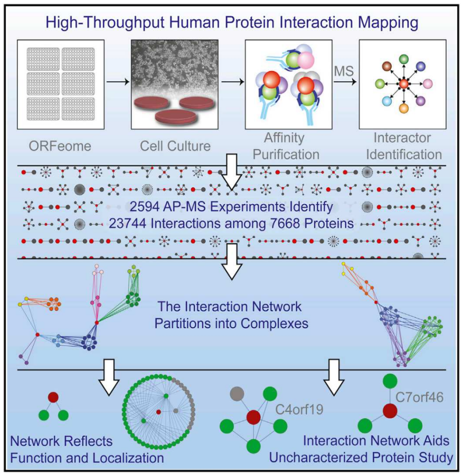
+15 additional PPIs

Luck *et al.*, A reference map of the human protein interactome, *bioRxiv*, posted April 10, 2019
https://www.biorxiv.org/content/10.1101/605451v1, published *Nature*, April 8, 2020

The current state-of-the-art in human PPI maps – large scale AP/MS



Hein *et al.*, *Cell* (2015) 163:712-23.



Huttlin *et al.*, *Cell* (2015) 162:425–440
Huttlin *et al.*, *Nature* (2017) 545:505-509
Huttlin *et al.*, *bioRxiv* (2020), posted January 19

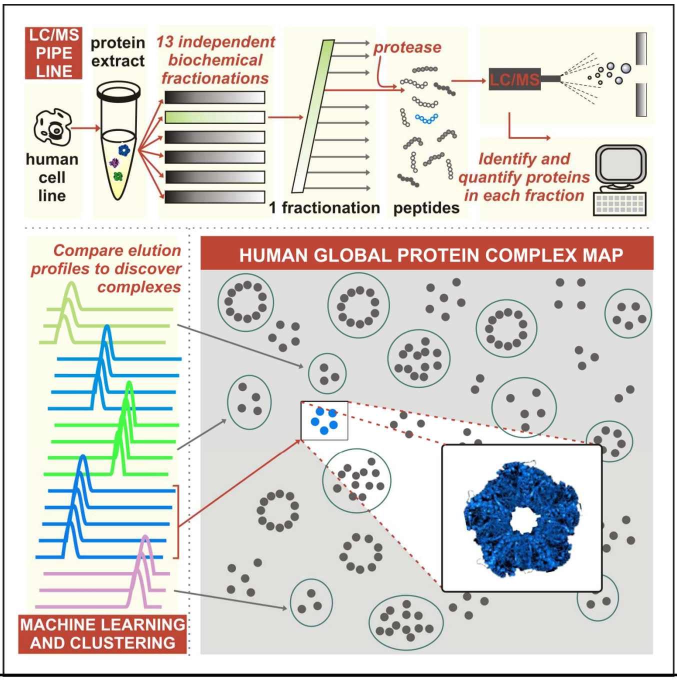
Just in the past 5 years, >16K affinity purification/mass spec experiments on tagged human proteins expressed in cell lines

---

The current state-of-the-art in animal PPI maps – co-fractionation/MS

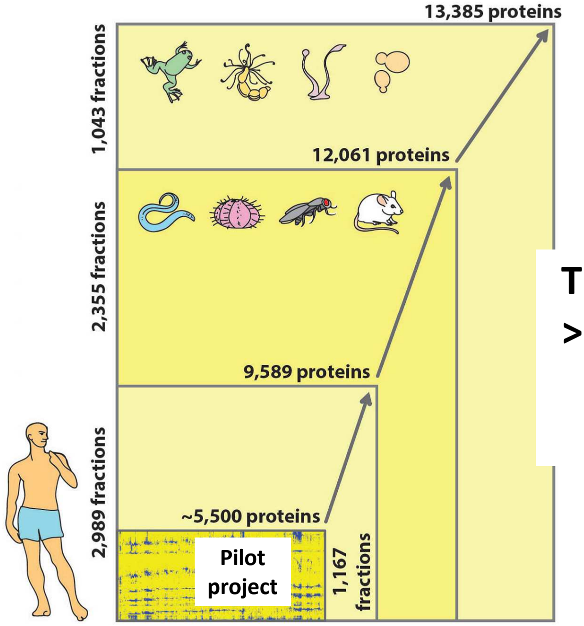
>2,000 biochemical fractions, including replicates

>9,000 hours mass spec machine time



Havugimana, Hart, *et al.*, *Cell* (2012)

Extending the map across animals…

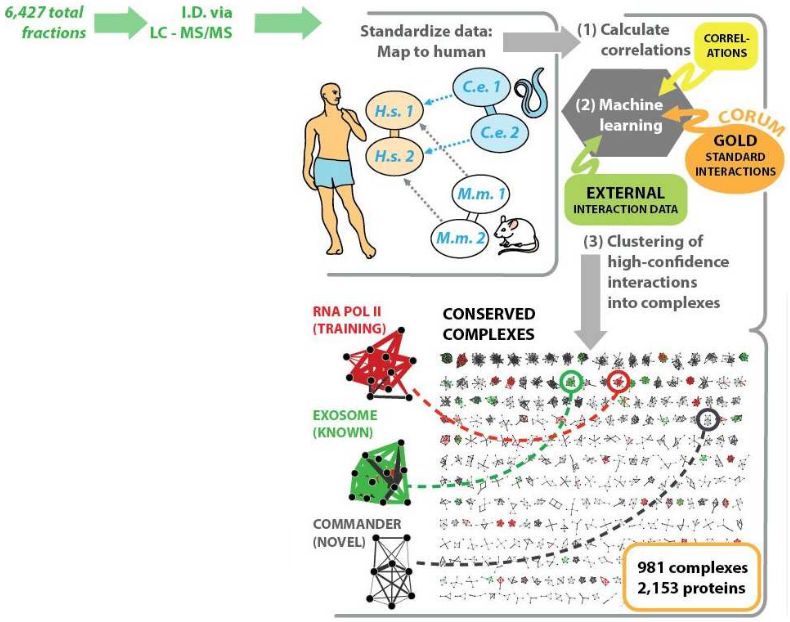


>6,400 CF/MS experiments

These data capture >80 million protein abundance measurements

Wan, Borgeson *et al.* *Nature* (2015)

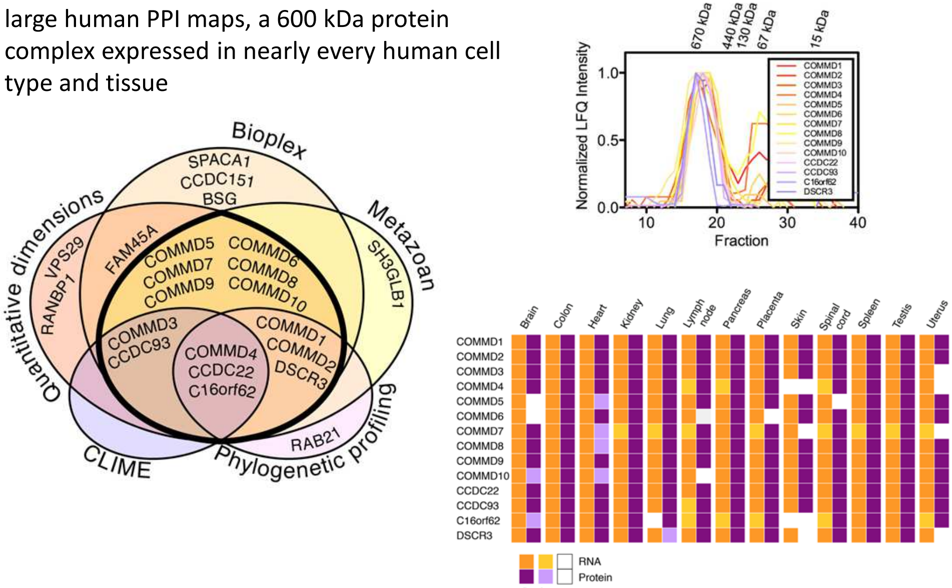
Extending the map across animals…



Wan, Borgeson *et al.* *Nature* (2015)

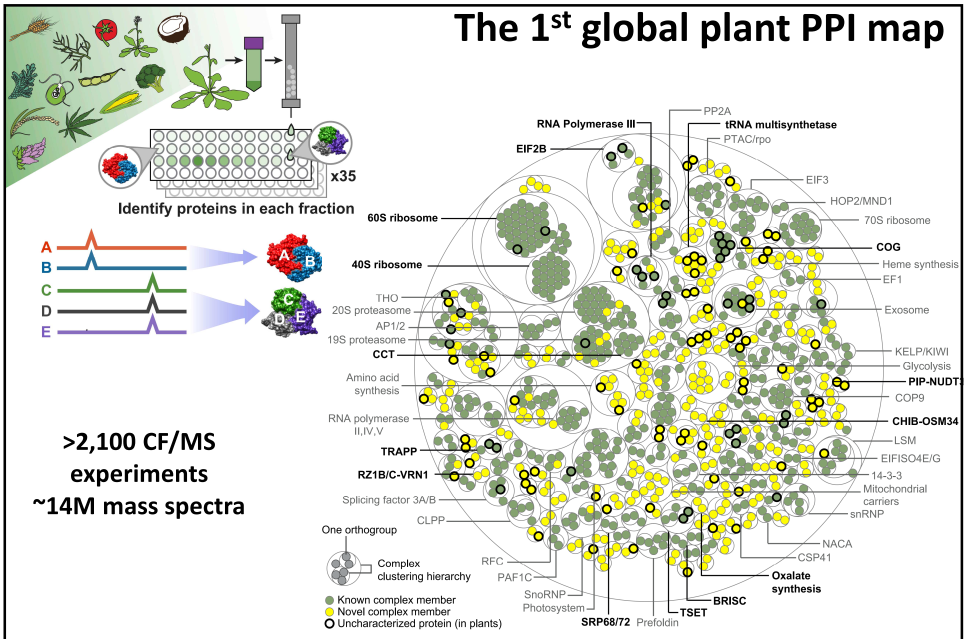
There are still lots of cellular machines left to find

e.g. the “Commander” complex, found in all 3 large human PPI maps, a 600 kDa protein complex expressed in nearly every human cell type and tissue



Reviewed in Mallam & Marcotte, *Cell Systems* (2017)

The 1st global plant PPI map



>2,100 CF/MS experiments
~14M mass spectra

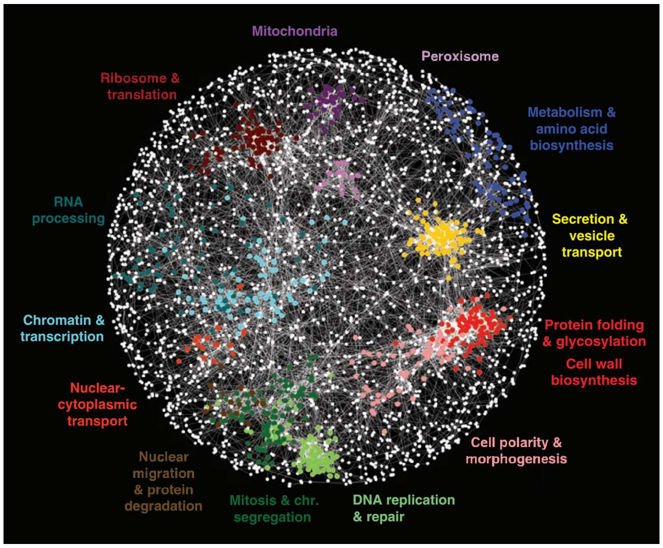
A Pan-plant Protein Complex Map Reveals Deep Conservation and Novel Assemblies

Claire D. McWhite,[1,6] Ophelia Papoulas,[1,6] Kevin Drew,[1] Rachael M. Cox,[1] Viviana June,[1] Oliver Xiaoou Dong,[2,3] Taejoon Kwon,[4] Cuihong Wan,[1,5] Mari L. Salmi,[1] Stanley J. Roux,[1] Karen S. Browning,[1] Z. Jeffrey Chen,[1] Pamela C. Ronald,[2,3] and Edward M. Marcotte[1,7,*]

Cell 181(2) P460-474.e14, April 16, 2020

Genetic interactions

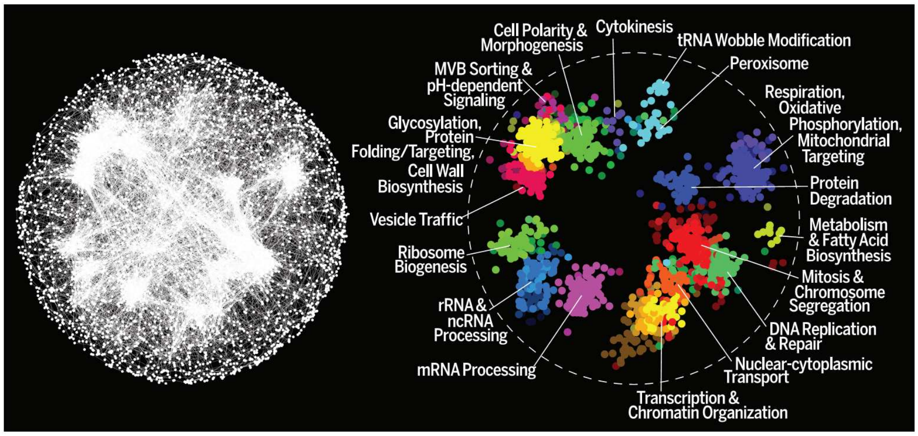
5.4 million gene-gene pairs assayed for synthetic genetic interactions in yeast



Costanzo *et al., Science* 327: 425 (2010)

Genetic interactions, the 2016 version

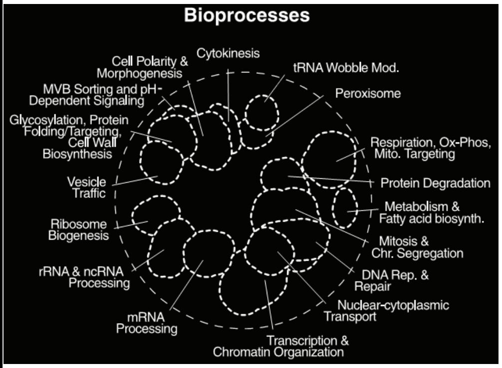
23 million gene-gene pairs assayed for synthetic genetic interactions in yeast, identifying ~550,000 negative and ~350,000 positive genetic interactions

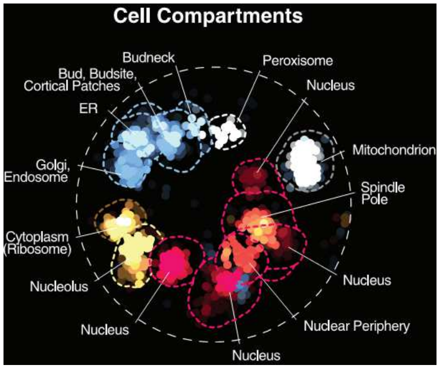


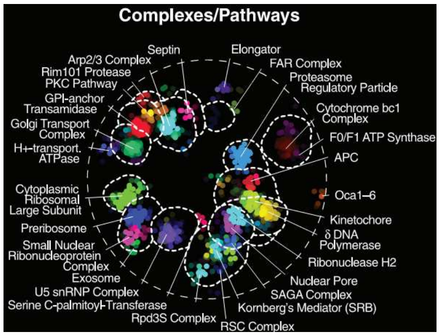
A global network of genetic interaction profile similarities. (**Left**) Genes with similar genetic interaction profiles are connected in a global network, such that genes exhibiting more similar profiles are located closer to each other, whereas genes with less similar profiles are positioned farther apart. (**Right**) Spatial

Costanzo *et al., Science* 353: 1381 (2016)

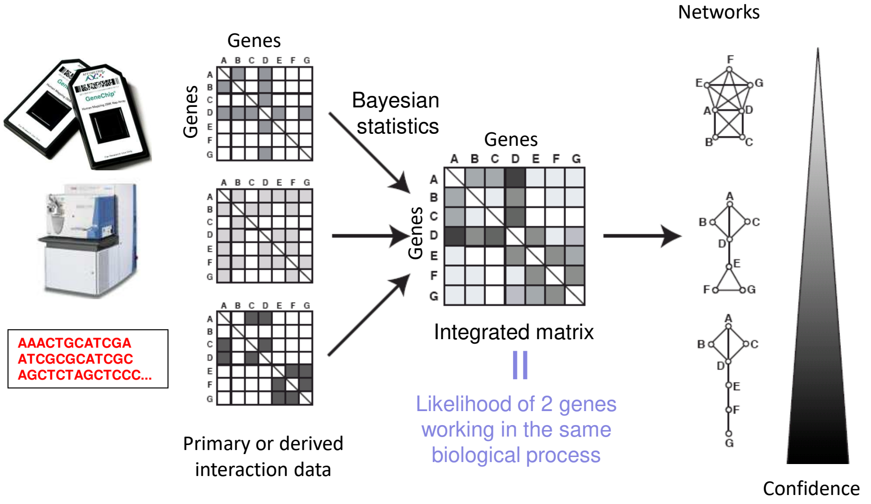
The global genetic interaction profile similarity network reveals a hierarchy of cellular function.







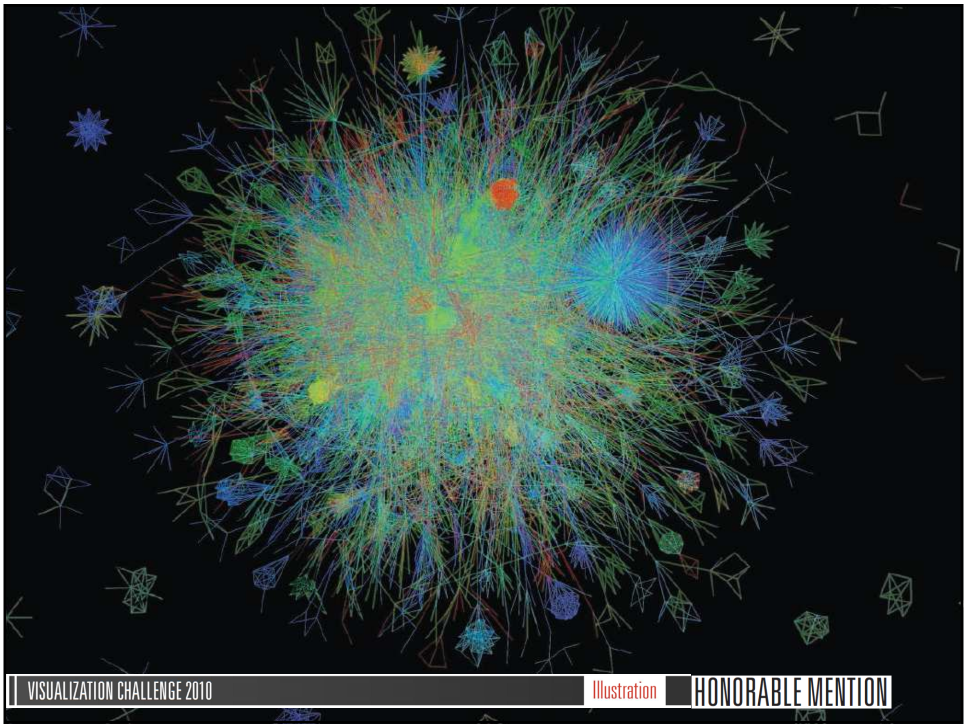
These sorts of data can be combined into functional gene networks



Primary or derived interaction data

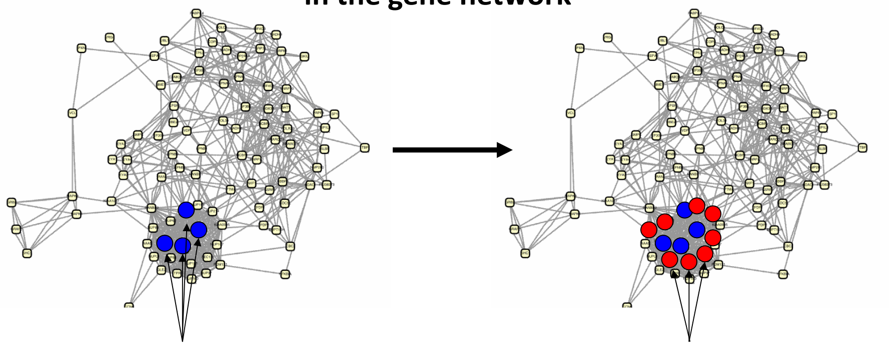
Integrated matrix || Likelihood of 2 genes working in the same biological process

Adapted from Fraser & Marcotte, *Nature Genetics* (2004)



These networks are hypothesis generators.
Given a gene, what other genes does it function with?
What do they do?

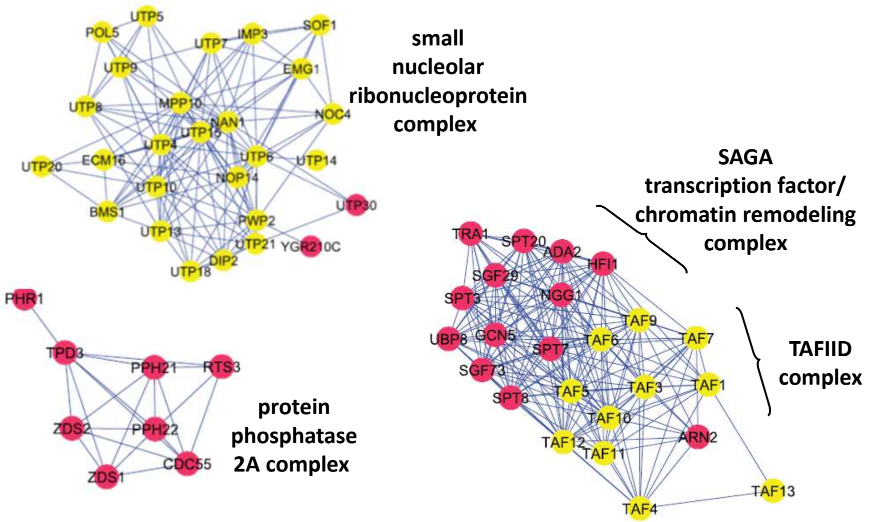
Guilt-by-association
in the gene network



Genes already linked
to a disease or function

New candidate genes
for that process

Gene networks frequently reflect functions, pathways, & phenotypes, e.g., lethality in yeast is linked to the molecular machine, not the gene



small nucleolar ribonucleoprotein complex

SAGA transcription factor/ chromatin remodeling complex

TAFIID complex

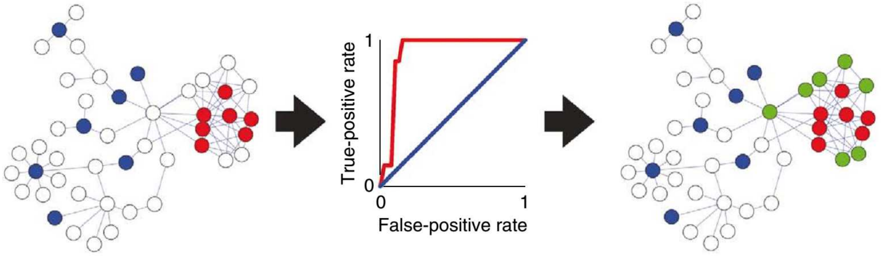
protein phosphatase 2A complex

● **Essential gene**
● **Nonessential gene**

Hart, Lee, & Marcotte, *BMC Bioinformatics* **8**:236 (2007)

We can propagate annotations across the graph to infer new annotations for genes (<u>network "guilt-by-association", or GBA</u>).

Testing how well this works on hidden, but known, cases let's us measure how predictive it will be for new cases.



Query with genes already linked to a disease or function, e.g. the red or blue function

Assess the network's predictive ability for that function using cross-validated ROC or recall/precision analysis

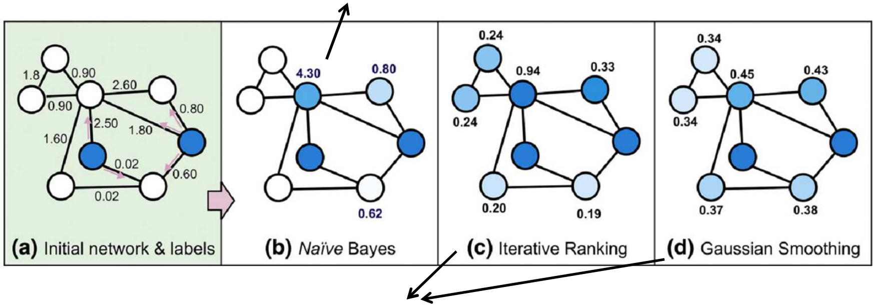
Infer new candidate genes for that process (e.g. predicting the green genes for the red function)

Lee, Ambaru *et al. Nature Biotechnology* 28:149-156 (2010)

Numerous algorithms exist for network GBA

Similar to Google's personalized PageRank

Naïve Bayes assigns scores to neighboring nodes based on edges

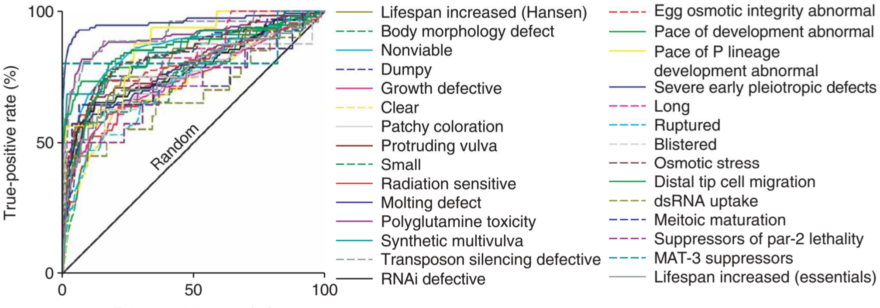


Network diffusion algorithms start with initial annotations and the graph topology,
then propagate initial scores across the network,
e.g. Gaussian smoothing tries to find scores:

$$f^{final} = argmin_f \alpha \Sigma_i \underbrace{\left(f_i - f_i^0\right)}_{\text{minimizing the difference between final and initial scores of a protein}}{}^2 + (1-\alpha)\Sigma_i\Sigma_j w_{ij} \underbrace{\left(f_i - f_j\right)}_{\text{\& between a protein's score and that of each of its neighbors}}{}^2$$

Reviewed in Wang & Marcotte, *J Proteomics* (2010)

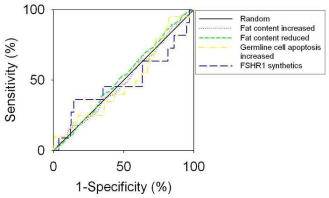
For example, predicting genes linked with worm phenotypes in genome-wide RNAi screens



ROC curves! Here, indicating the likely predictive power of the network for a system of interest, independent of how big the system is.

A poor ROC → no better than random guessing.

Some very poorly predicted pathways:



Lee, Lehner *et al., Nat Genet*, 40(2):181-8 (2008)

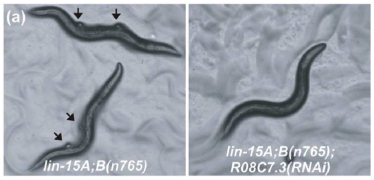
Remarkably, this strategy works quite well

Some examples of network-guided predictions:

In worms:

Genes that can reverse 'tumors' in a nematode model of tumorigenesis

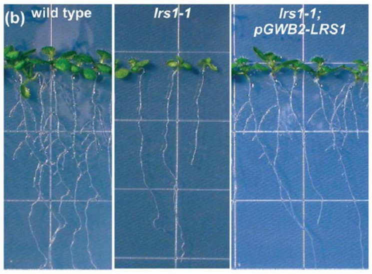
Lee, Lehner *et al.* *Nature Genetics* (2008)



In *Arabidopsis*:

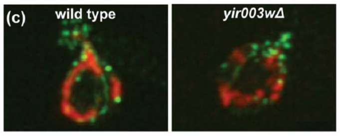
New genes regulating root formation

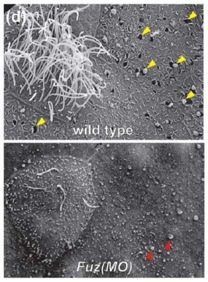
Lee, Ambaru *et al.* *Nature Biotech* (2010)



In yeast: **New mitochondrial biogenesis genes**

Hess *et al.*, *PLoS Genetics* (2009)

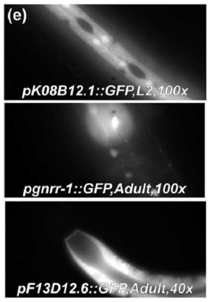




In mice/frogs:

Functions for a birth defect gene

Gray *et al.*, *Nature Cell Biology* (2009)



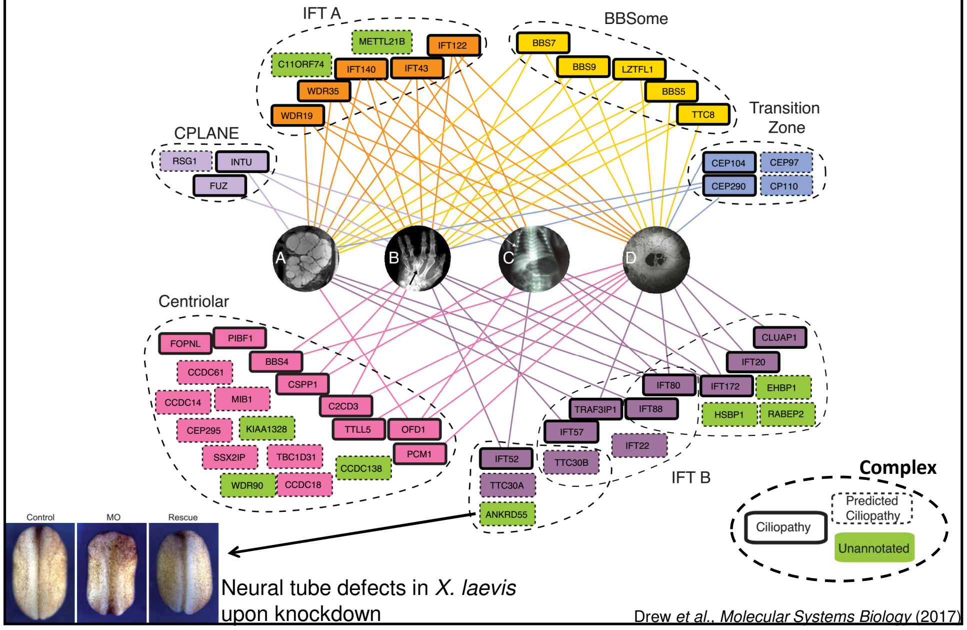
In worms:

Predicting tissue specific gene expression

Chikina *et al.*, *PLoS Comp Biology* (2009)

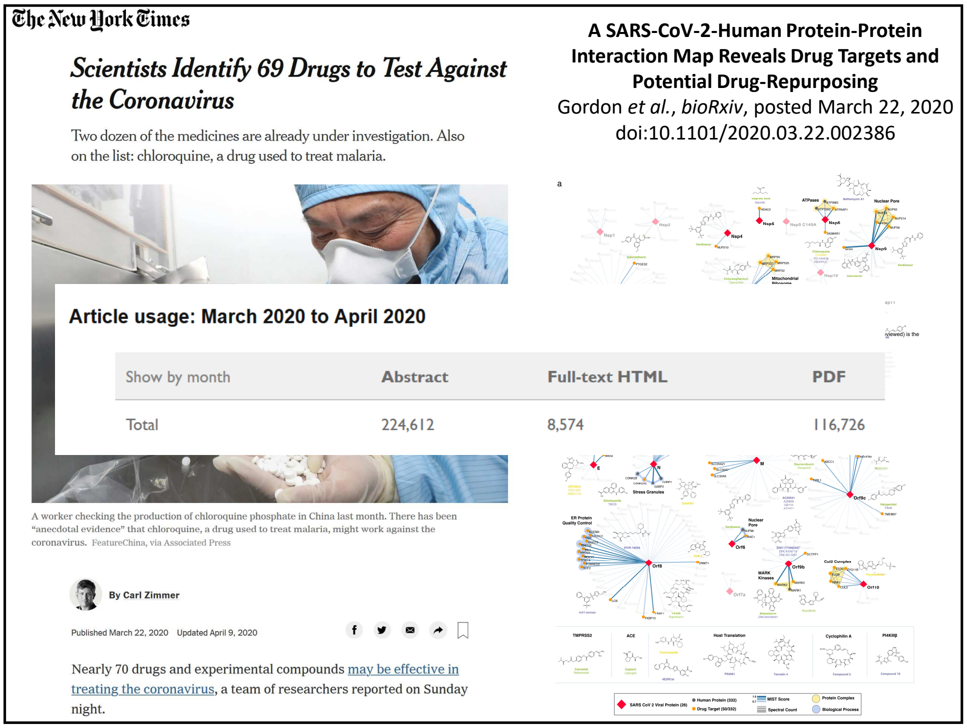
Reviewed in Wang & Marcotte, *J Proteomics* (2010)

We use this approach routinely in the lab, e.g. a recent example predicting new ciliopathy genes from protein complexes



Neural tube defects in *X. laevis* upon knockdown

Drew *et al.*, *Molecular Systems Biology* (2017)



The New York Times

Scientists Identify 69 Drugs to Test Against the Coronavirus

Two dozen of the medicines are already under investigation. Also on the list: chloroquine, a drug used to treat malaria.

A worker checking the production of chloroquine phosphate in China last month. There has been "anecdotal evidence" that chloroquine, a drug used to treat malaria, might work against the coronavirus. FeatureChina, via Associated Press

By Carl Zimmer

Published March 22, 2020 Updated April 9, 2020

Nearly 70 drugs and experimental compounds may be effective in treating the coronavirus, a team of researchers reported on Sunday night.

A SARS-CoV-2-Human Protein-Protein Interaction Map Reveals Drug Targets and Potential Drug-Repurposing
Gordon *et al.*, *bioRxiv*, posted March 22, 2020
doi:10.1101/2020.03.22.002386

Article usage: March 2020 to April 2020

Show by month	Abstract	Full-text HTML	PDF
Total	224,612	8,574	116,726

Live demo of STRING, BioGRID, GeneMania, functional networks and Cytoscape