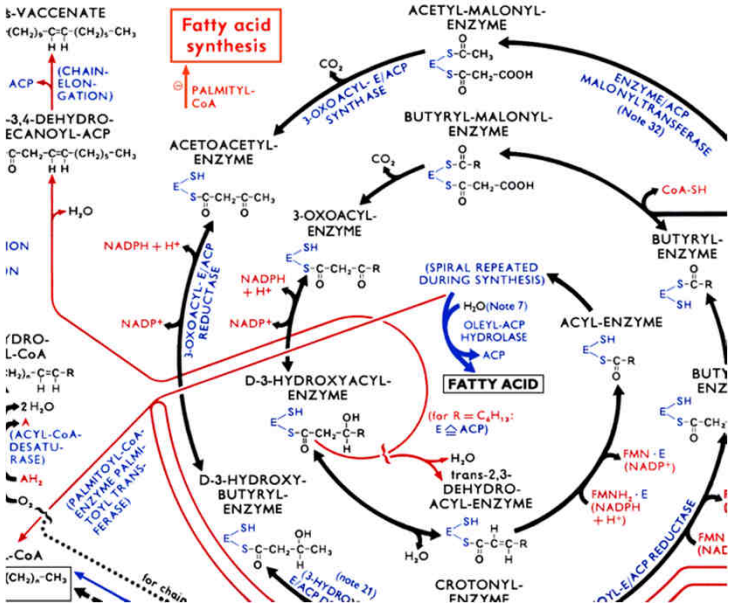
Network biology
(& predicting gene function)

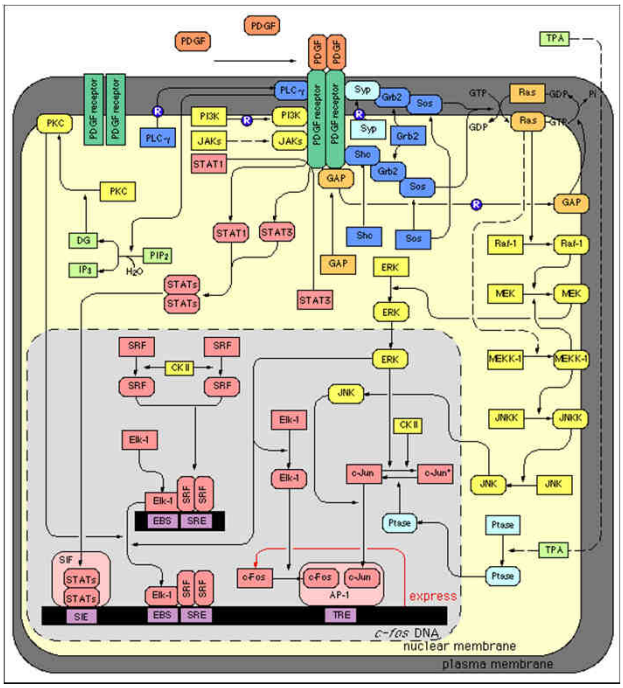
BCH339N Systems Biology / Bioinformatics

Edward Marcotte, Univ of Texas at Austin

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

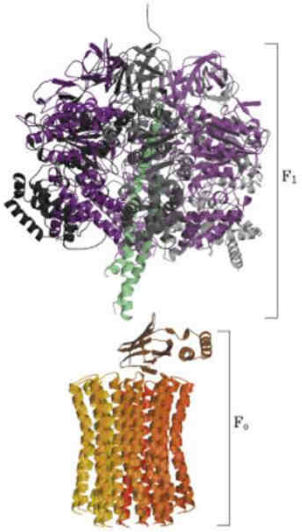


A typical genetic network

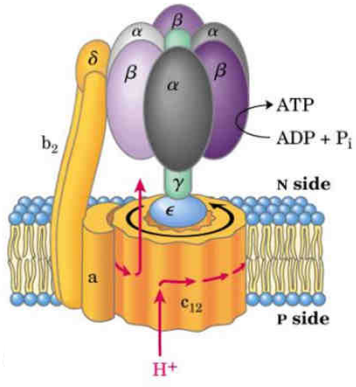


Contacts between proteins define protein interaction networks

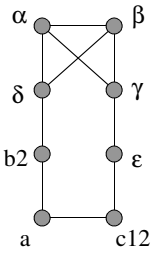
X-ray structure of ATP synthase



Schematic version



Network representation



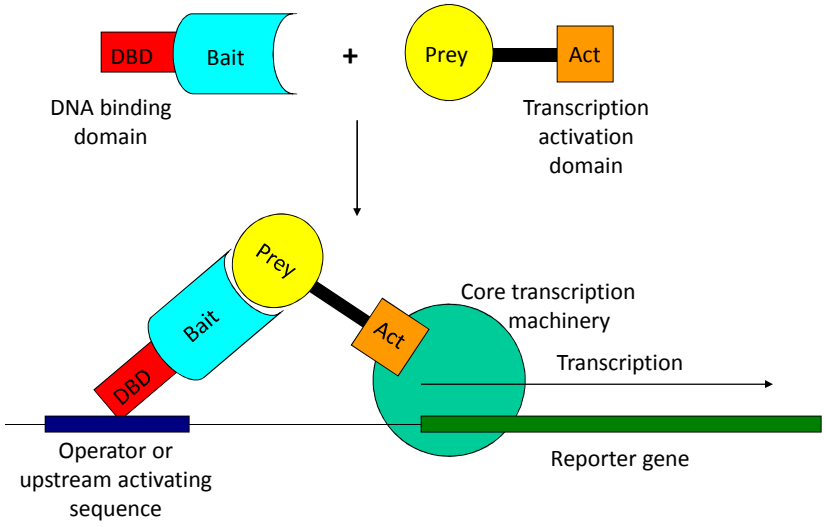
Total set = protein complex
Sum of direct + indirect interactions

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.

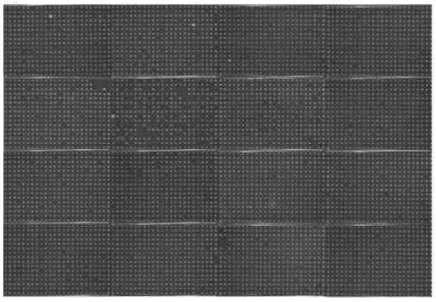
Pairwise protein interactions

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

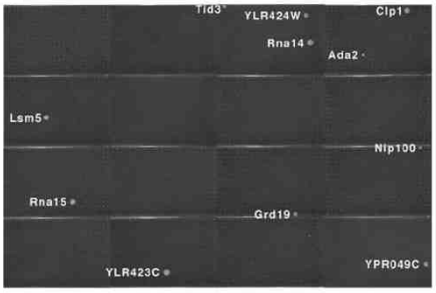


High-throughput yeast two-hybrid assays

Haploid yeast cells expressing activation domain-prey fusion proteins

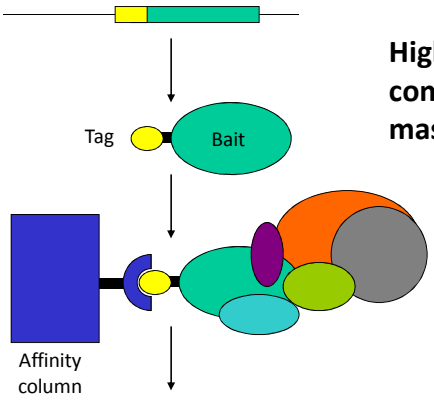


Diploid yeast probed with DNA-binding domain-Pcf11 bait fusion protein

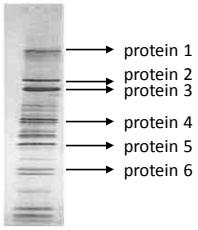


Uetz, Giot, *et al. Nature* (2000)

Protein complexes

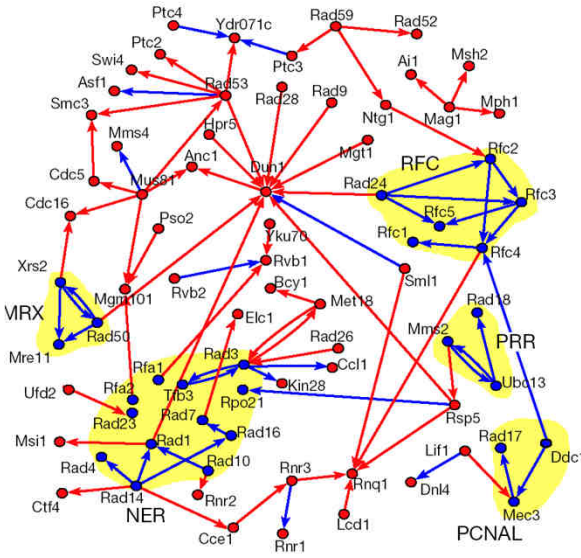


High-throughput complex mapping by mass spectrometry



SDS-page

Trypsin digest, identify peptides by mass spectrometry



493 bait proteins

↓

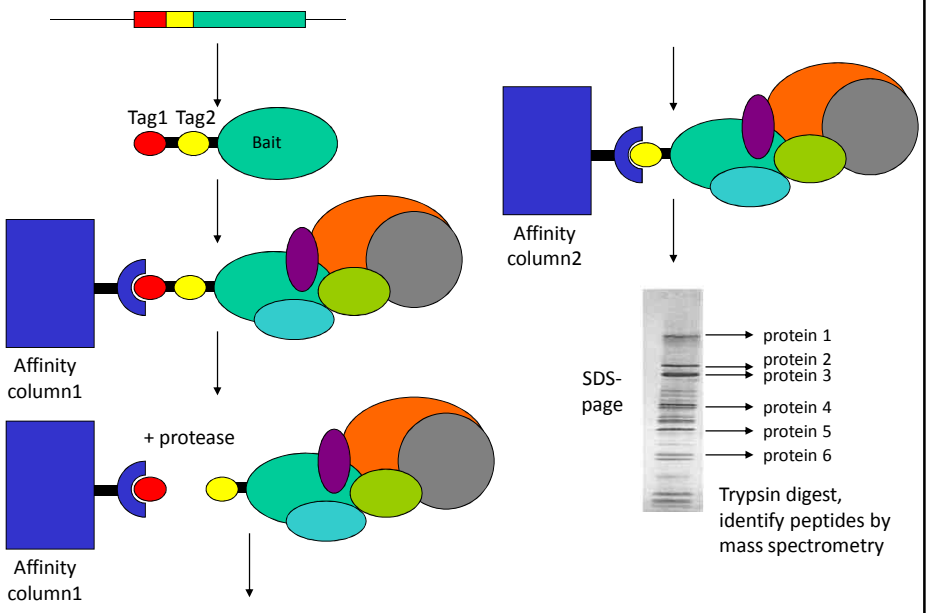
3617 interactions

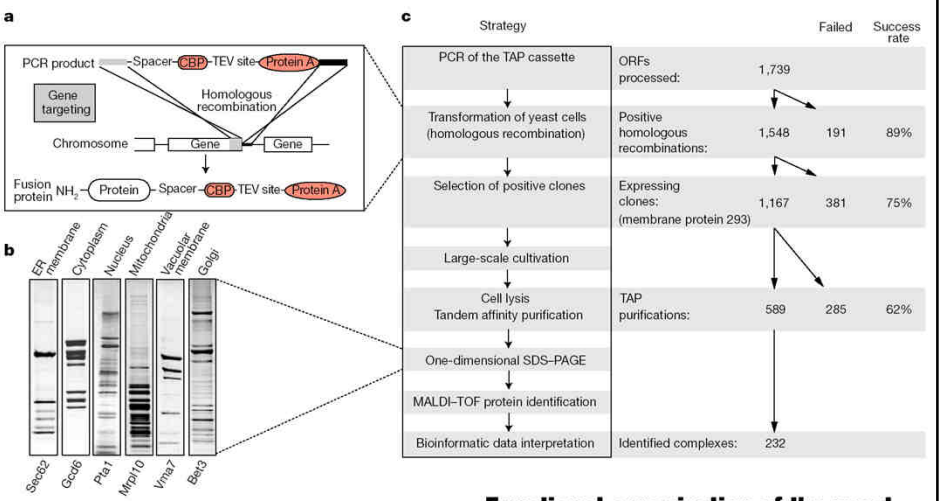
Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry

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

A variant: tandem affinity purification (TAP)



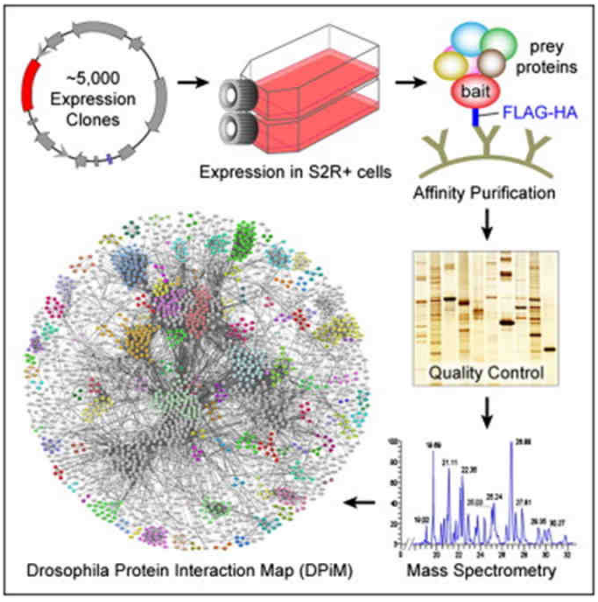


Functional organization of the yeast proteome by systematic analysis of protein complexes

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*, Małgorzata 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

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



~3,500 affinity purification experiments

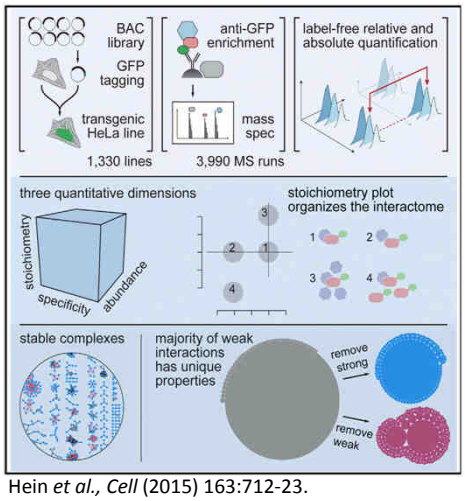
~11K interactions / ~2.3K proteins

→ spans 556 complexes

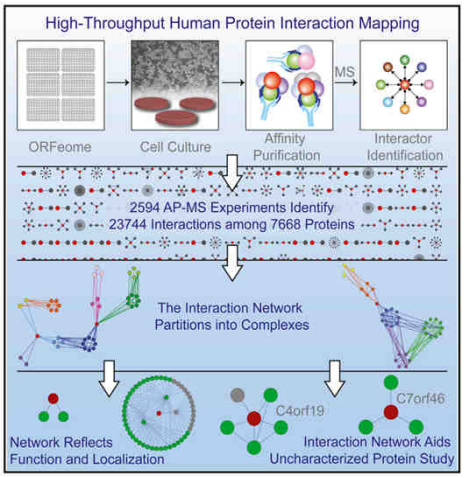
Still daunting for the human proteome, but…

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

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



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



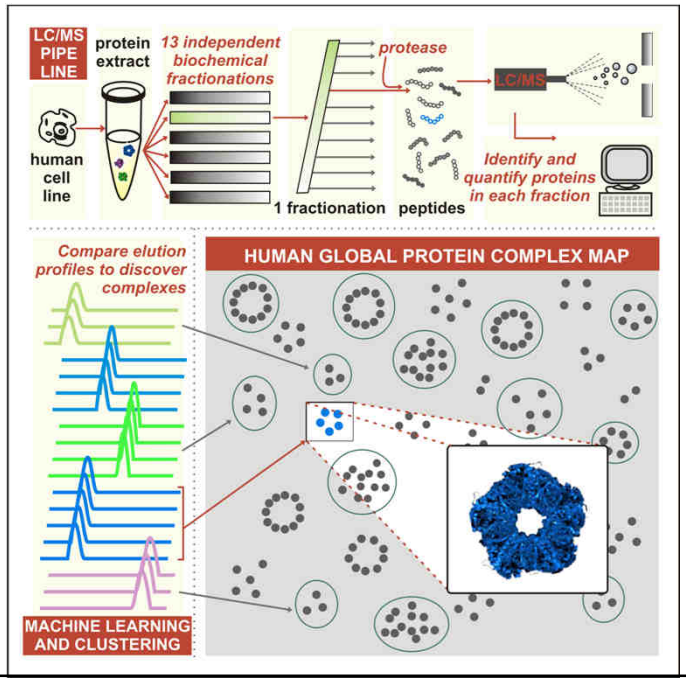
Huttlin *et al., Cell* (2015) 162:425–440
Huttlin *et al., Nature* (2017) 545:505-509

Just in the past 3 years, nearly 6K affinity purification experiments on tagged human proteins expressed in cell lines

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

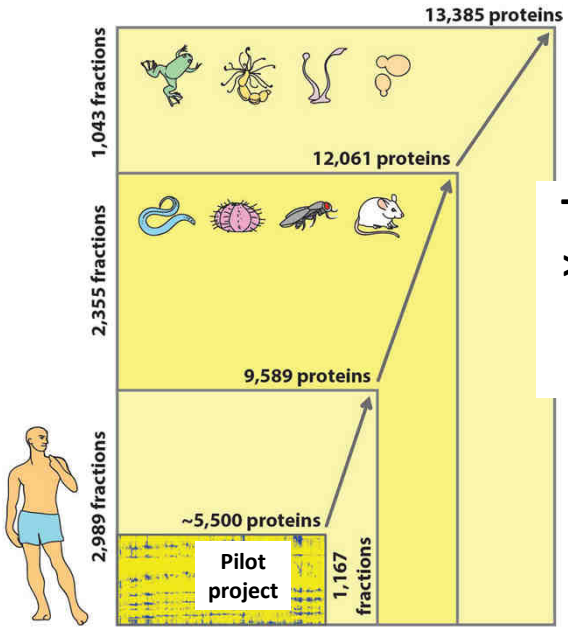
>2,000 biochemical fractions, including replicates

>9,000 hours mass spec machine time



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

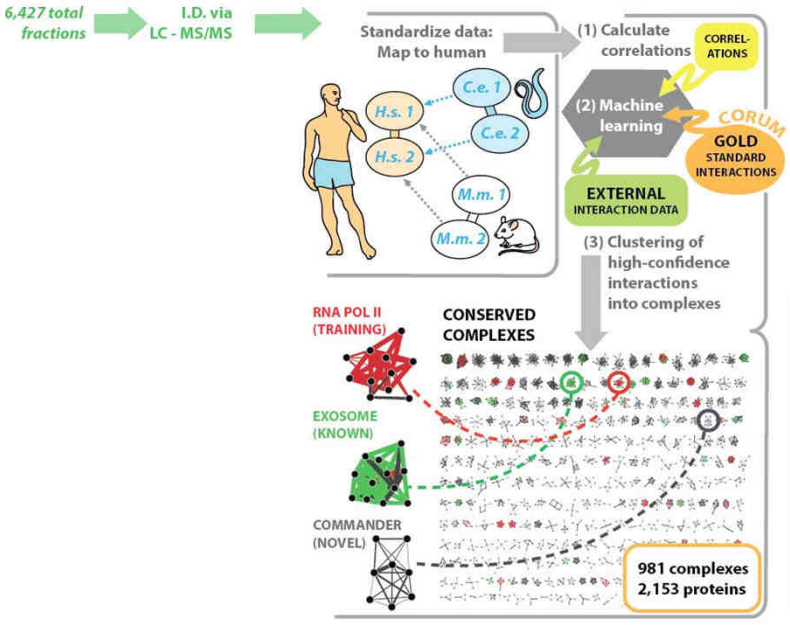
Now >6,400 CF/MS experiments across animals



These data capture >80 million protein abundance measurements

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

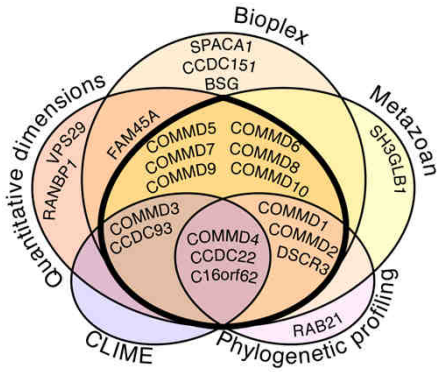
Extending the map across animals…

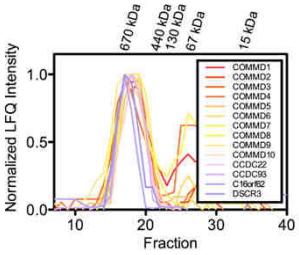


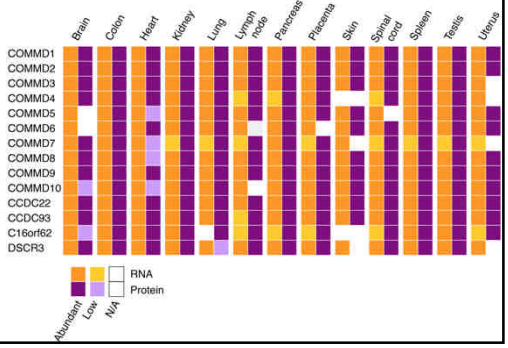
Wan, Borgeson *et al. Nature* (2015)

There are still lots of cellular machines left to find

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



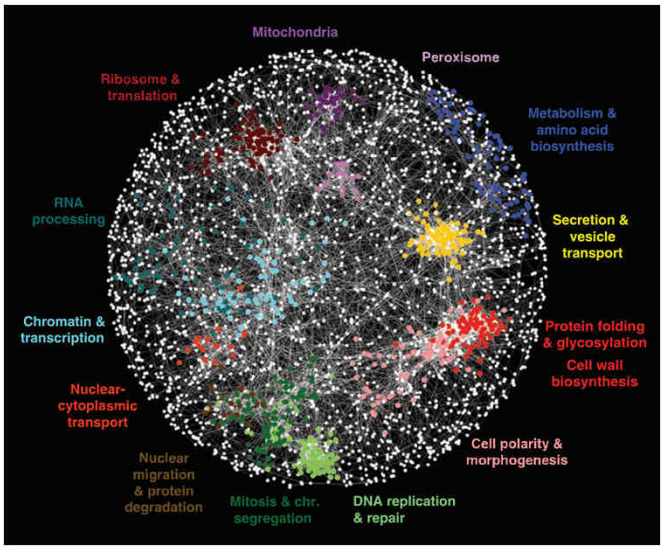




Reviewed in Mallam & Marcotte, *Cell Systems* (2017)

Genetic interactions

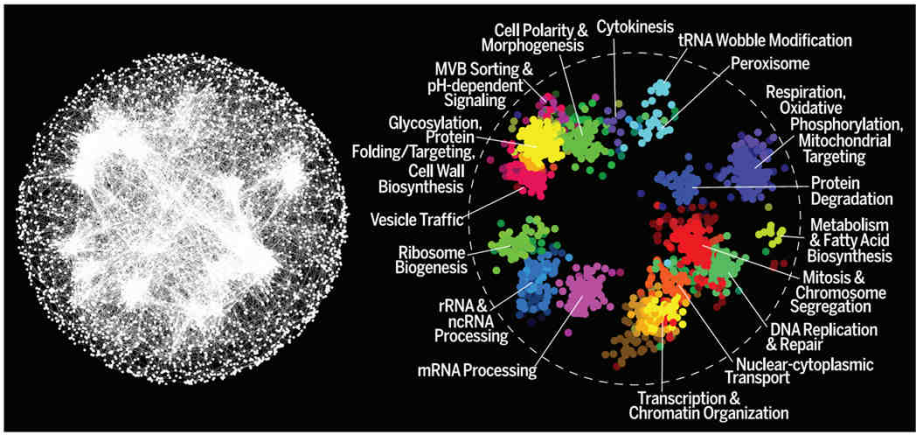
5.4 million gene-gene pairs assayed for synthetic genetic interactions in yeast



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

Genetic interactions, the 2016 version

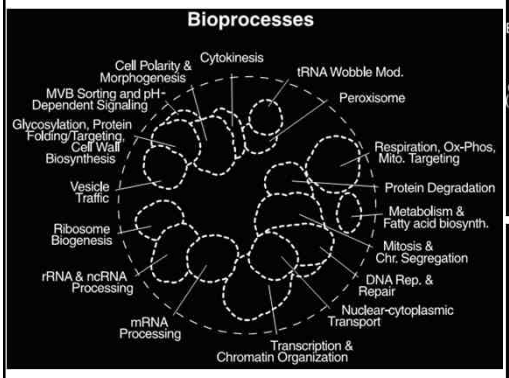
23 million gene-gene pairs assayed for synthetic genetic interactions in yeast, identifying ~550,000 negative and ~350,000 positive genetic interactions

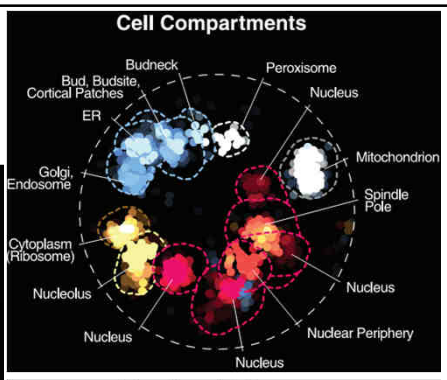


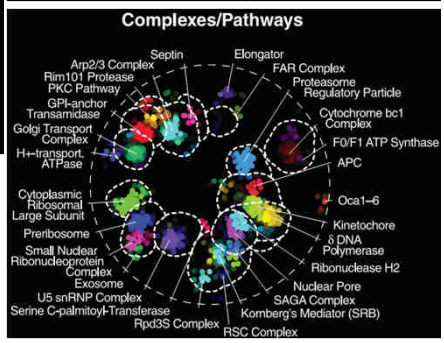
A global network of genetic interaction profile similarities. (**Left**) Genes with similar genetic interaction profiles are connected in a global network, such that genes exhibiting more similar profiles are located closer to each other, whereas genes with less similar profiles are positioned farther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

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

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







Comparative genomics

Functional relationships between genes impose subtle constraints upon genome sequences. Thus, genomes carry intrinsic information about the cellular systems and pathways they encode.

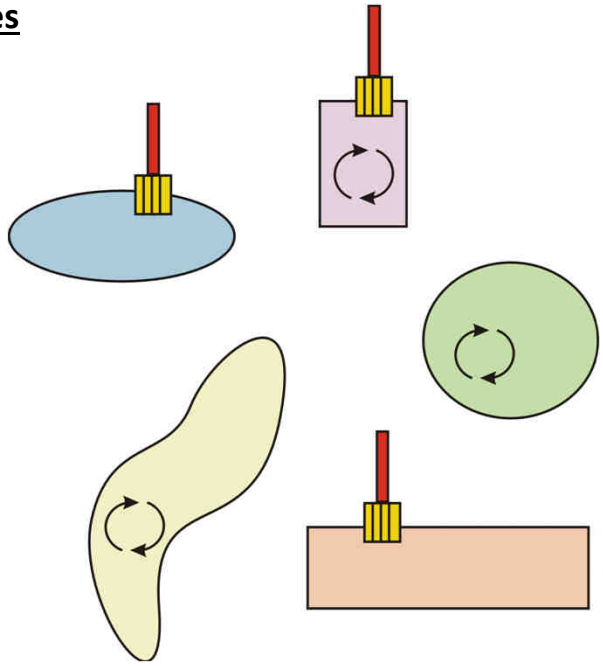
Linkages can be found from aspects of gene context, including:

Distances between sequence elements
Order of sequences
Variation in order between organisms
Regulatory sequences near genes
Gene content of an organism
Variation in gene content between organisms
Fusions between genes from different organisms

Phylogenetic profiles

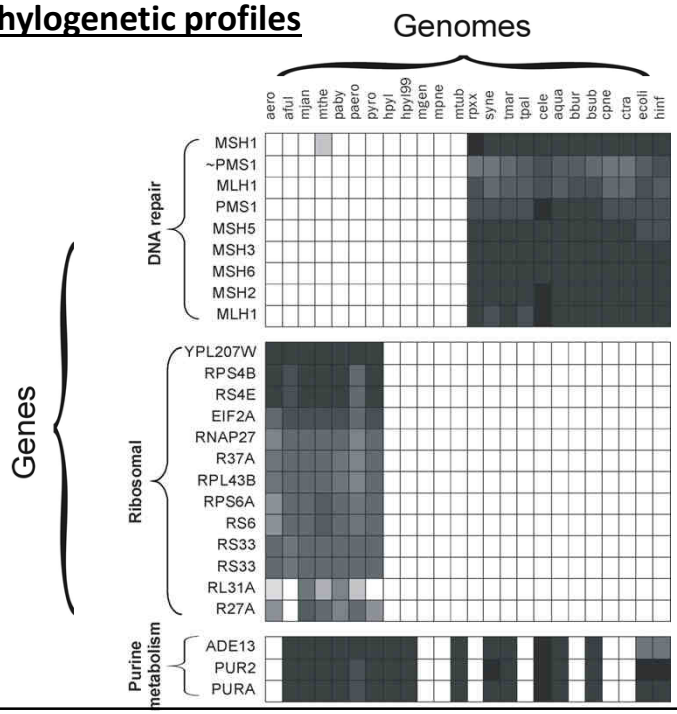
Organisms with e.g. a flagellum have the necessary genes; those without tend to lack them.

Specific trends of gene presence/absence thus inform about biological processes.



PNAS 96, 4285-4288 (1999)

Phylogenetic profiles

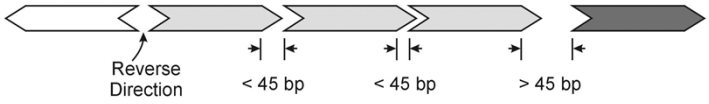


Grayscale indicates sequence similarity to closest homolog in that genome

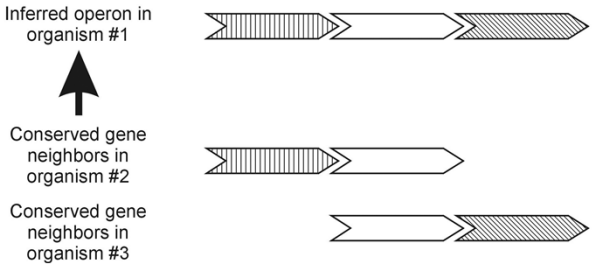
Nature 402, 83-86 (1999)

Operons and evolutionary conservation of gene order

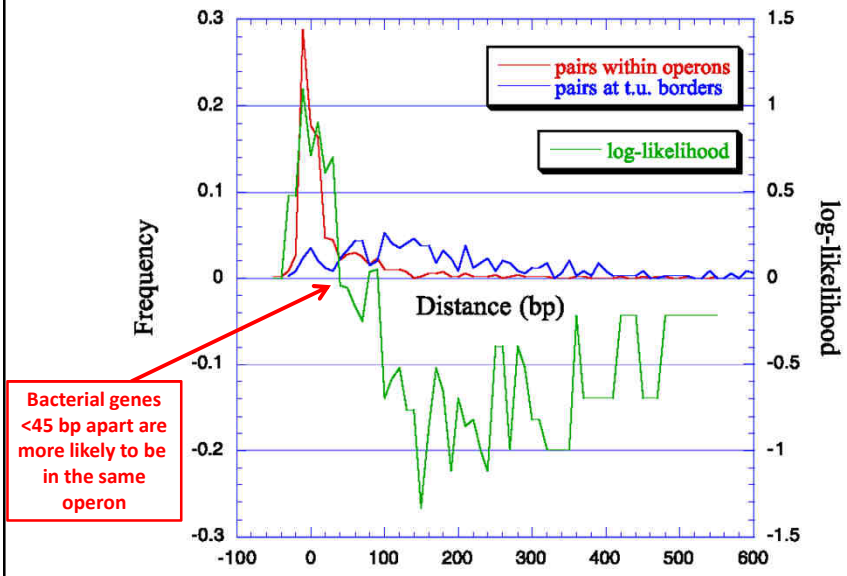
Prokaryotic operons tend to favor certain intergenic distances



Conserved gene neighbors also reveal functional relationships



Again, such observations can be turned into pairwise scores:



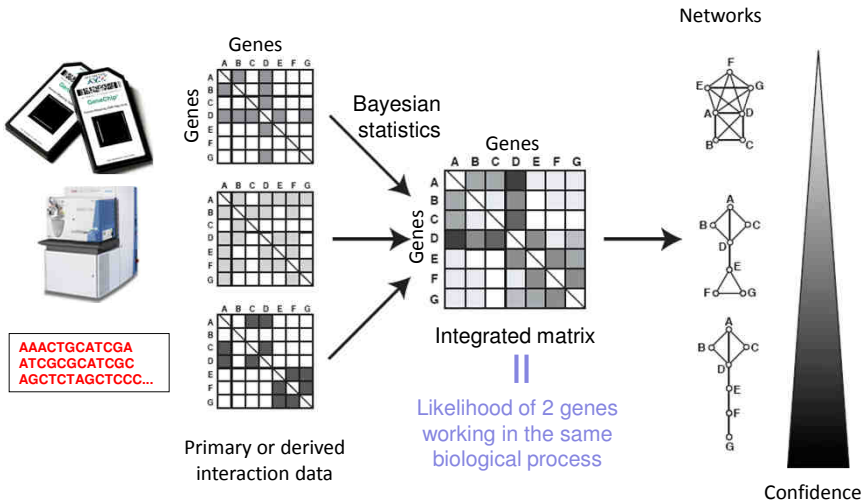
Bacterial genes <45 bp apart are more likely to be in the same operon

Operons in *Escherichia coli*: Genomic analyses and predictions

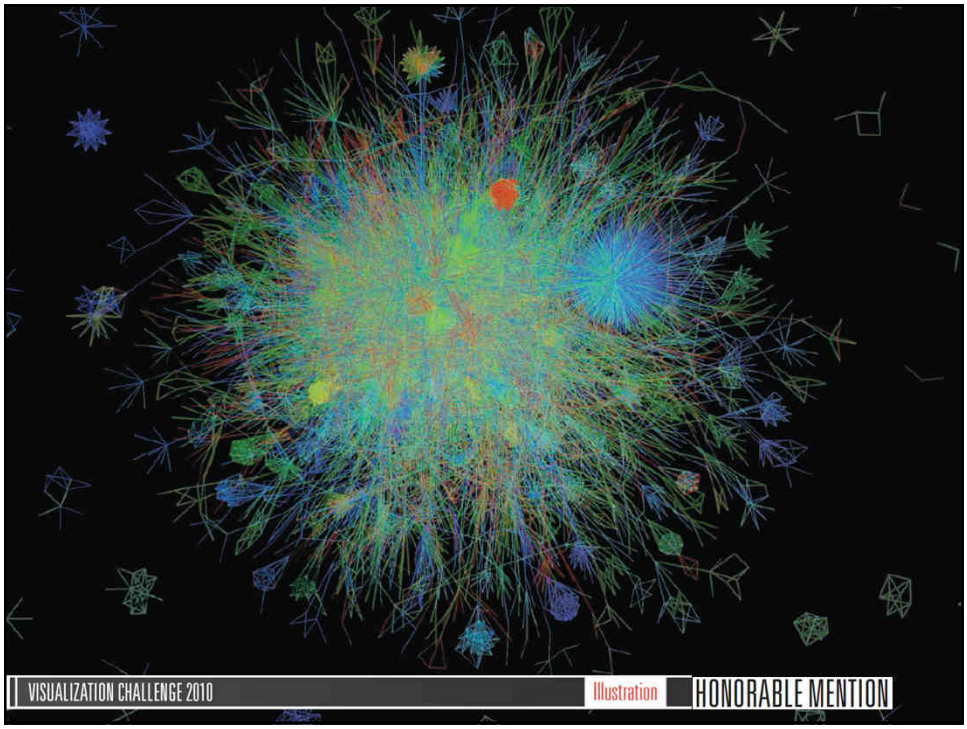
6652–6657 | PNAS | June 6, 2000 | vol. 97 | no. 12

Heladia Salgado*†, Gabriel Moreno-Hagelsieb*†, Temple F. Smith‡, and Julio Collado-Vides†§

These sorts of data can be combined into functional gene networks

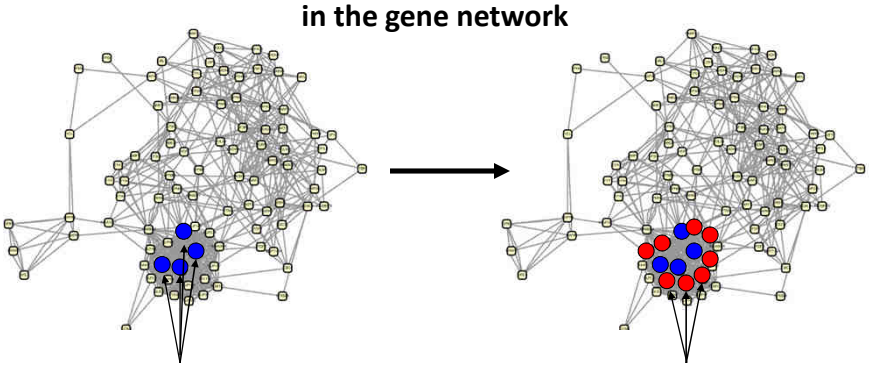


Adapted from Fraser & Marcotte, *Nature Genetics* (2004)



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

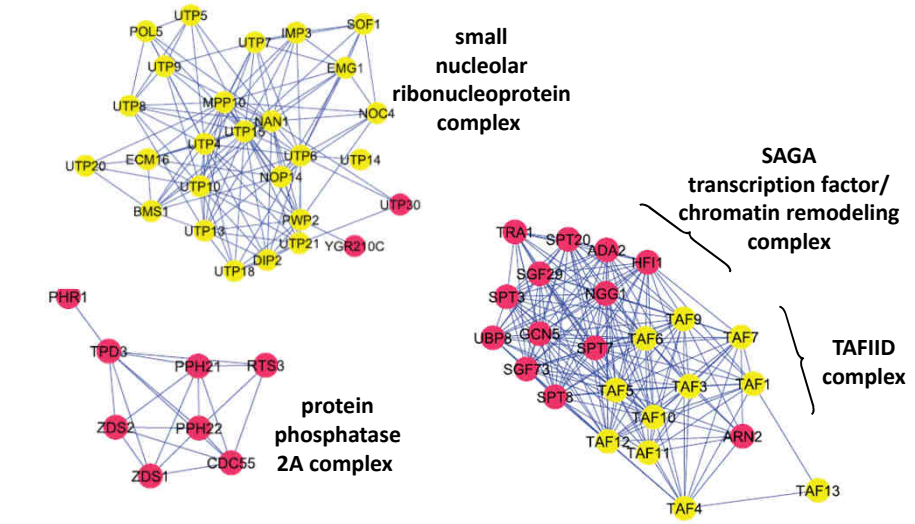
Guilt-by-association
in the gene network



Genes already linked to a disease or function

New candidate genes for that process

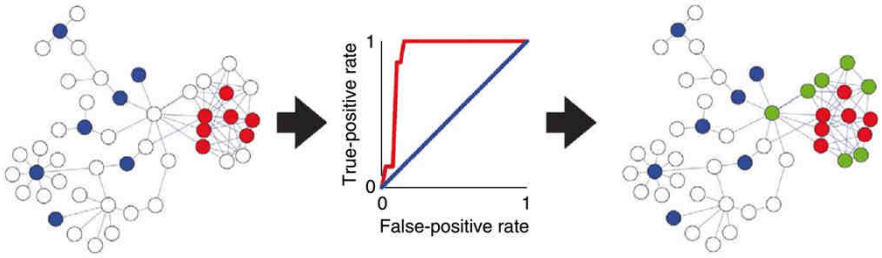
Gene networks frequently reflect functions, pathways, & phenotypes, e.g., lethality in yeast is linked to the molecular machine, not the gene



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

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

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

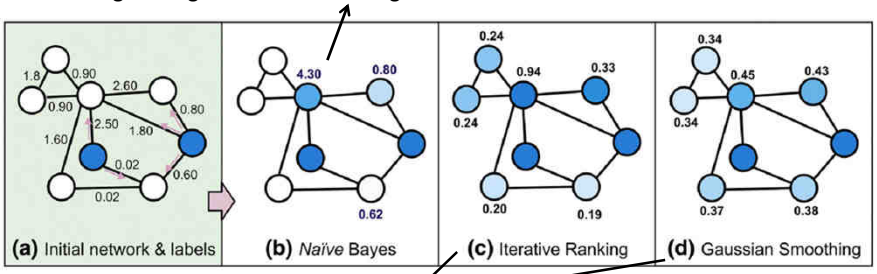
Infer new candidate genes for that process (e.g. predicting the green genes for the red function)

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:

$$f^{final} = argmin_f \alpha \Sigma_i (f_i - f_i^0)^2 + (1-\alpha)\Sigma_i \Sigma_j w_{ij} (f_i - f_j)^2$$

minimizing the difference between final and initial scores of a protein

& between a protein's score and that of each of its neighbors

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

Calculating ROC curves

Prediction \ Actual	P	N
P'	True Positive	False Positive
N'	False Negative	True Negative

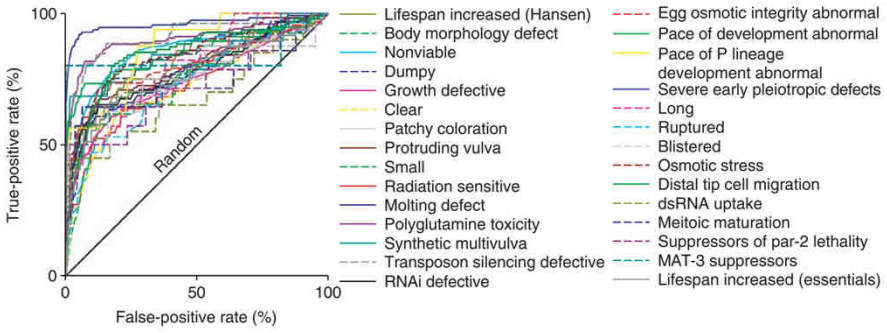
Basic idea: sort predictions from best to worst, plot TPR vs. FPR as you traverse the ranked list

TPR = TP / P = TP / (TP + FN)
= True Positive Rate
= Sensitivity, Recall

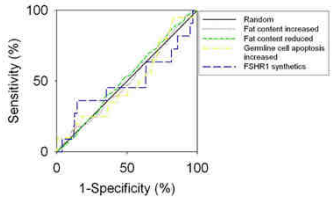
FPR = FP / N = FP / (FP + TN)
= False Positive Rate
= 1 - Specificity

Also useful to plot Precision [= TP / (TP + FP)] vs. Recall (= TPR)

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



Some very poorly predicted pathways:



ROC analysis indicates the likely predictive power of the network for a system of interest.

A poor ROC → no better than random guessing.

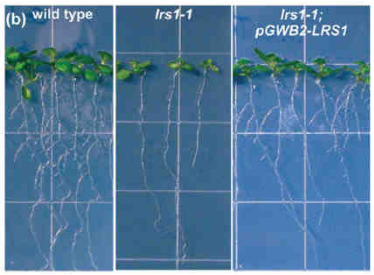
Lee, Lehner *et al.*, *Nat Genet*, 40(2):181-8 (2008)

Remarkably, this strategy works quite well

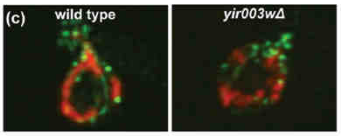
Some examples of network-guided predictions:

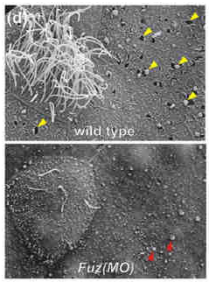
In worms:
Genes that can reverse 'tumors' in a nematode model of tumorigenesis
Lee, Lehner *et al.* *Nature Genetics* (2008)

In *Arabidopsis*:
New genes regulating root formation
Lee, Ambaru *et al.* *Nature Biotech* (2010)

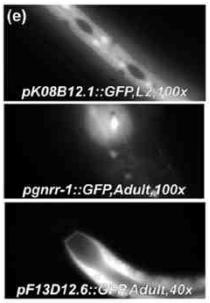


In yeast: **New mitochondrial biogenesis genes**
Hess *et al.*, *PLoS Genetics* (2009)





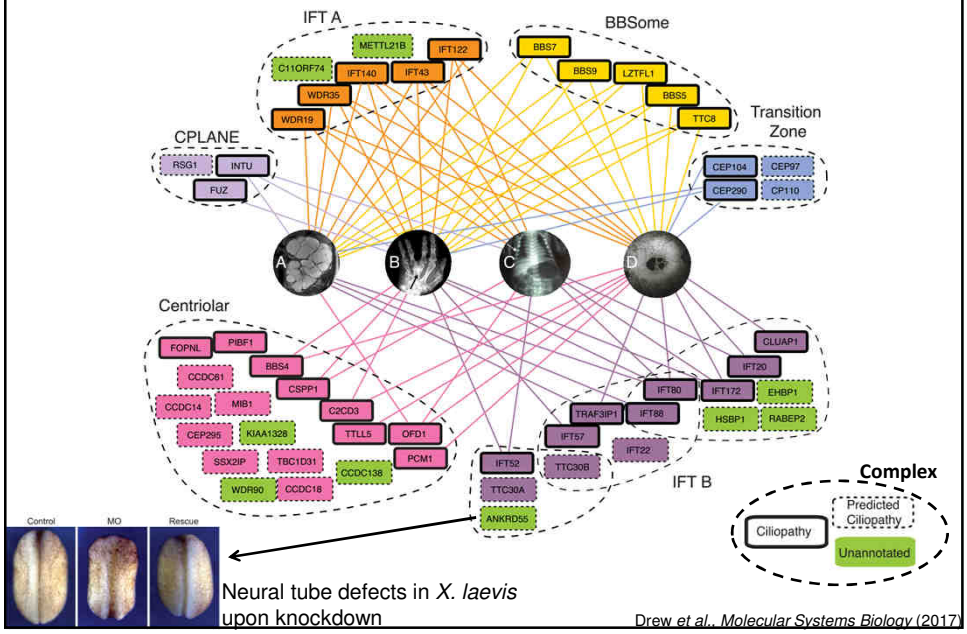
In mice/frogs:
Functions for a birth defect gene
Gray *et al.*, *Nature Cell Biology* (2009)



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, 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)

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