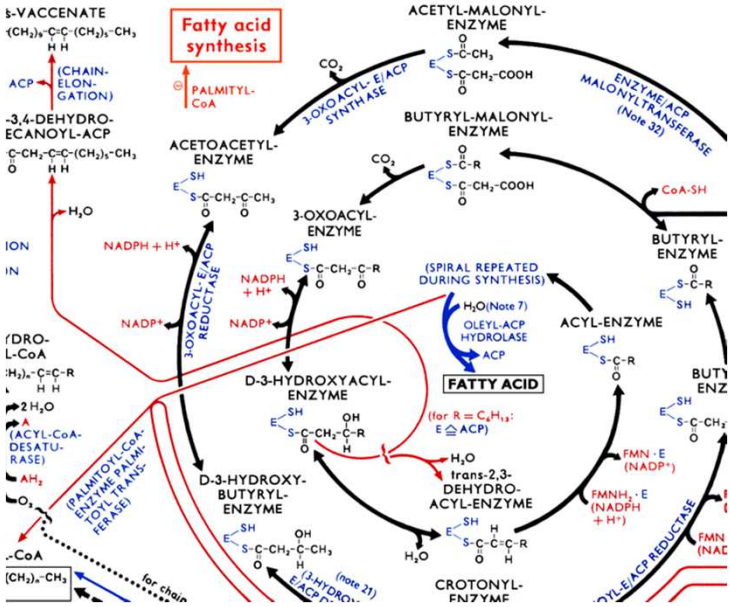
Network biology
(& predicting gene function)

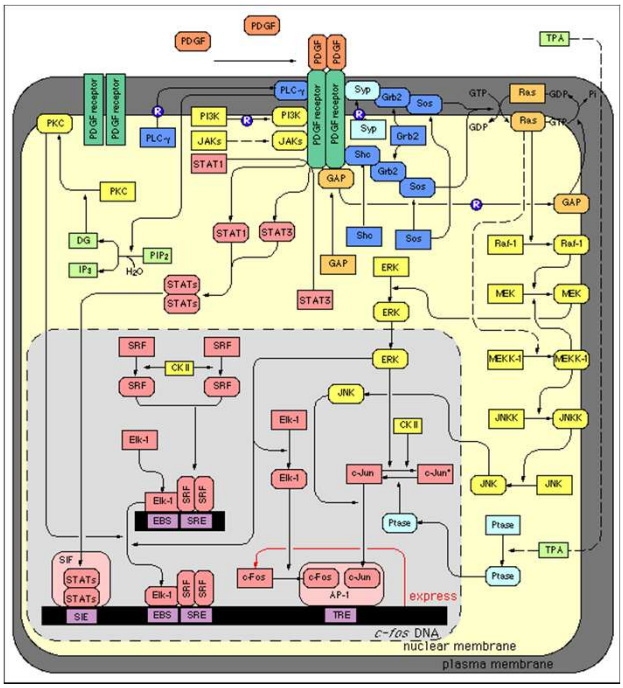
BCH364C/391L Systems Biology / Bioinformatics – Spring 2015

Edward Marcotte, Univ of Texas at Austin

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

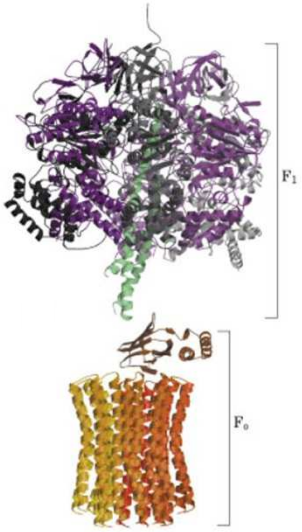


A typical genetic network

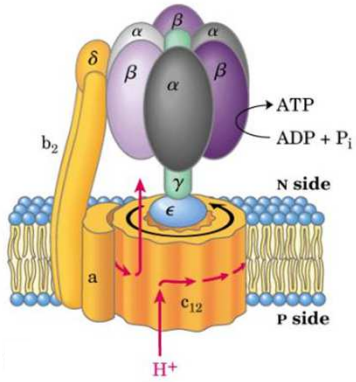


Contacts between proteins define protein interaction networks

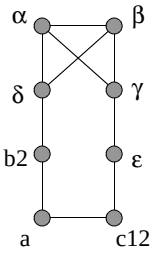
X-ray structure of ATP synthase



Schematic version



Network representation



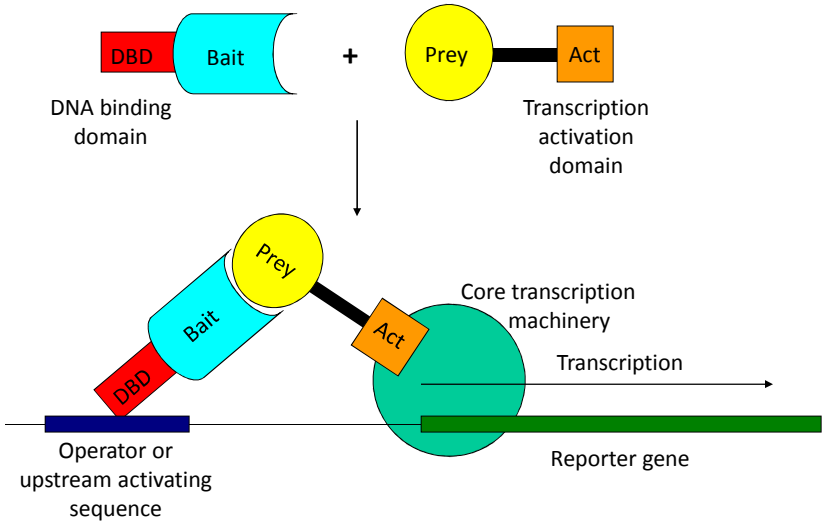
Total set = protein complex
Sum of direct + indirect interactions

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.

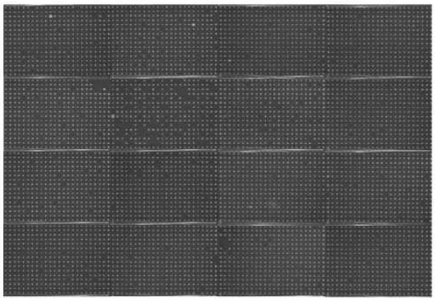
Pairwise protein interactions

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

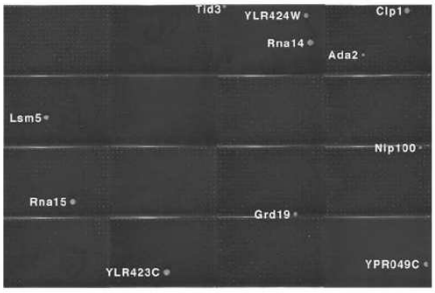


High-throughput yeast two-hybrid assays

Haploid yeast cells expressing activation domain-prey fusion proteins

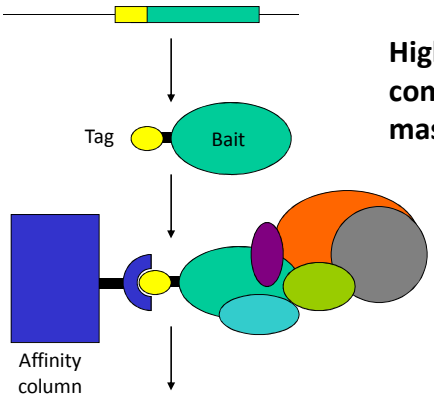


Diploid yeast probed with DNA-binding domain-Pcf11 bait fusion protein

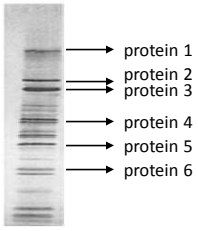


Uetz, Giot, *et al. Nature* (2000)

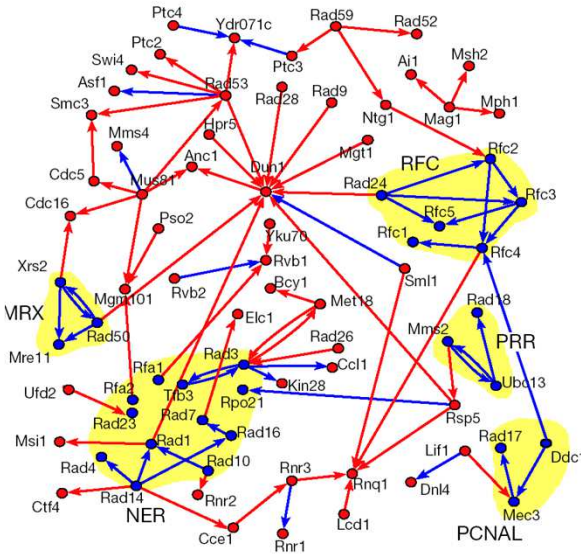
Protein complexes



High-throughput complex mapping by mass spectrometry



Trypsin digest, identify peptides by mass spectrometry



493 bait proteins

↓

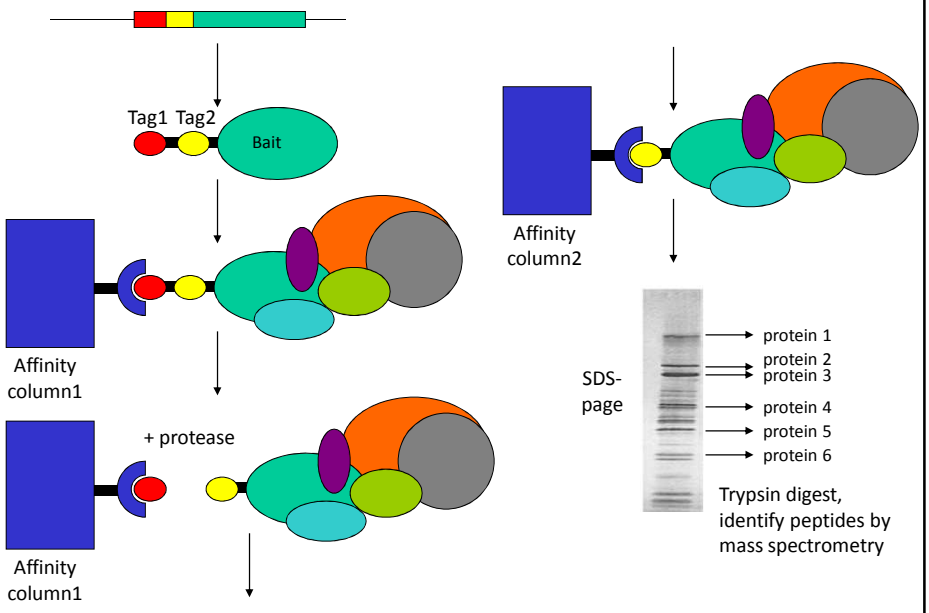
3617 interactions

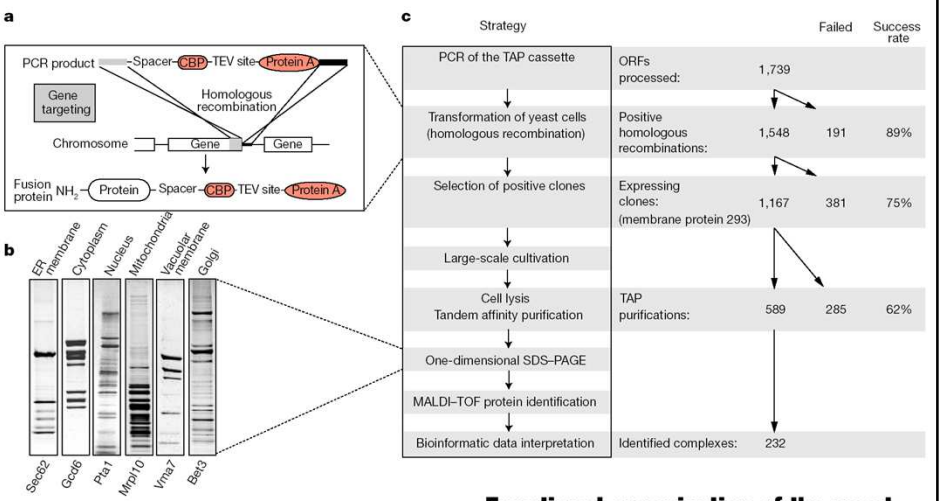
Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry

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)



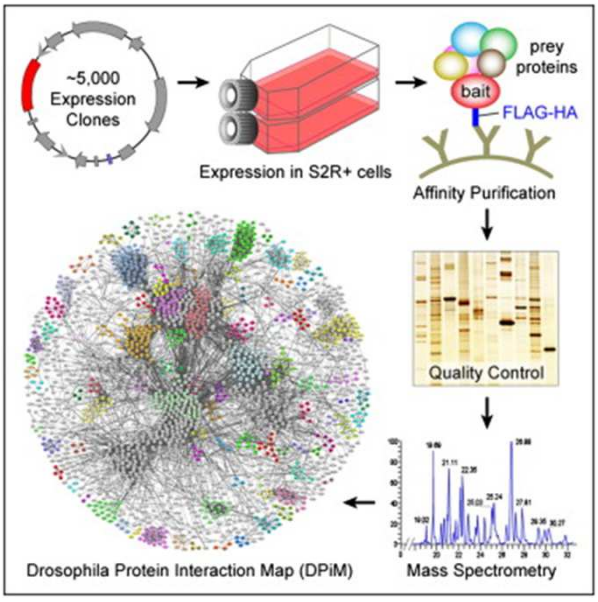


Functional organization of the yeast proteome by systematic analysis of protein complexes

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*, 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 Leuchtenberger*, Marie-Anne Weitzenegger*, 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

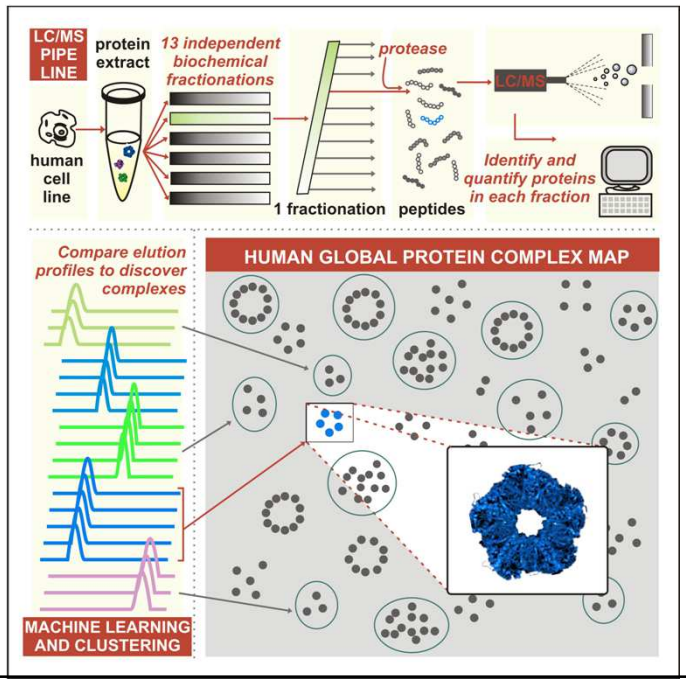
Still daunting for the human proteome

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

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

>2,000 biochemical fractions, including replicates

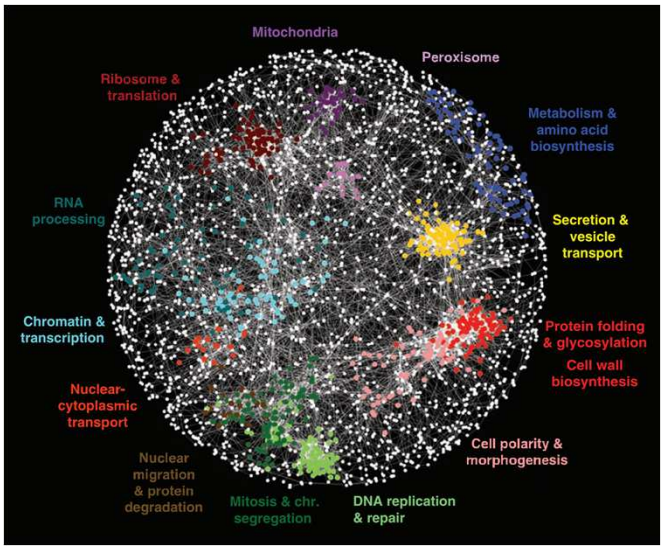
>9,000 hours mass spec machine time



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

Genetic interactions

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



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

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.

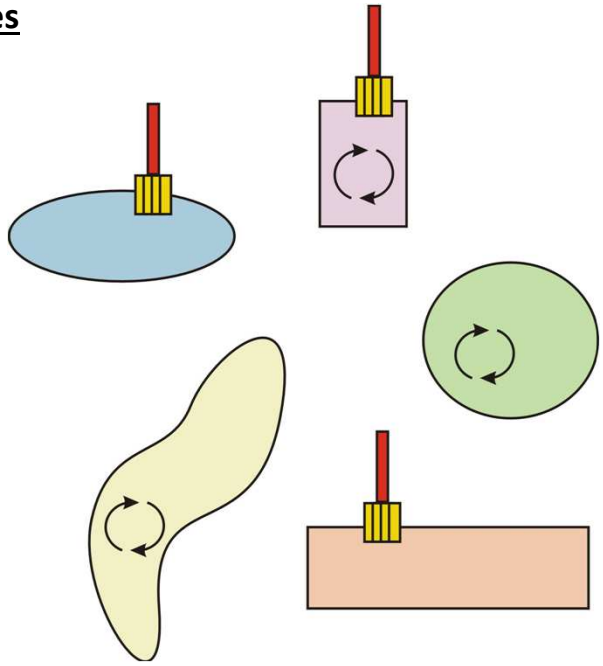
Linkages can be found from aspects of gene context, including:

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

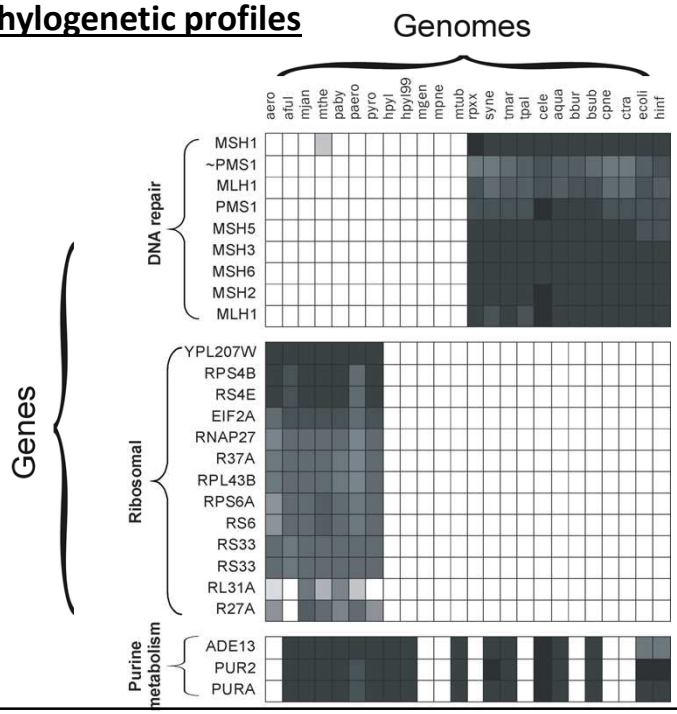
Organisms with e.g. a flagellum have the necessary genes; those without tend to lack them.

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



PNAS 96, 4285-4288 (1999)

Phylogenetic profiles

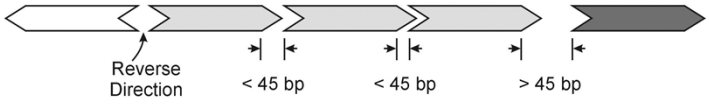


Grayscale indicates sequence similarity to closest homolog in that genome

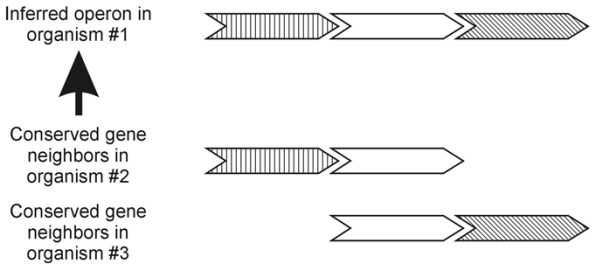
Nature 402, 83-86 (1999)

Operons and evolutionary conservation of gene order

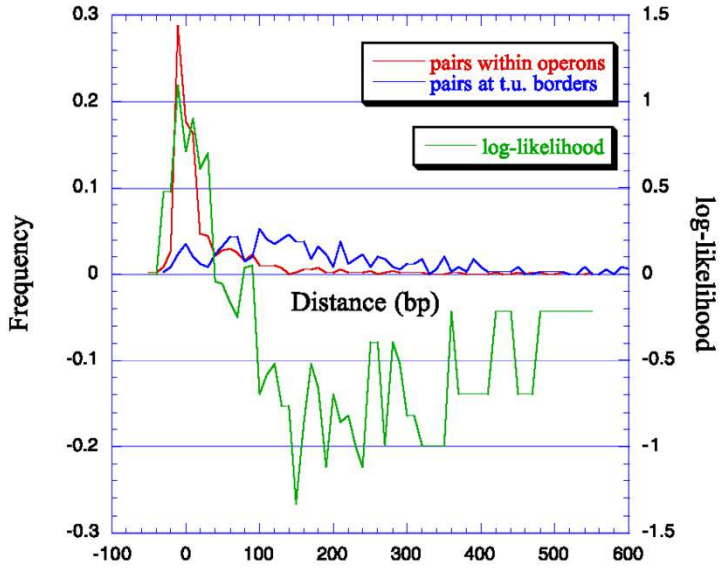
Prokaryotic operons tend to favor certain intergenic distances



Conserved gene neighbors also reveal functional relationships



Again, such observations can be turned into pairwise scores:



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†§

To summarize so far:

Data about gene interactions comes many sources but is dominated by several major ones:

- mRNA co expression. Historically microarrays & ESTs, increasingly RNAseq. Typically very high coverage data.
- Comparative genomics. Available for free for all organisms (typically phylogenetic profiles & operons)
- Protein interactions, especially co-complex interactions from mass spectrometry
- Genetic interactions (more so matching profiles of interaction partners than the interactions themselves)
- Transfer from other species

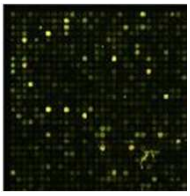
More abstractly, we might consider all of these as indicating "functional linkages" between genes

- Protein-protein interactions
- Participating in consecutive metabolic reactions
- Sharing genetic interactors
- Forming the same protein complex
- Giving rise to similar mutational phenotypes
- Exhibiting similar biological function

and so on…

These sorts of data can be combined into functional gene networks

Including measurements of...

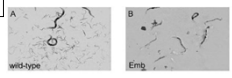


Gene expression (RNA-seq/arrays)

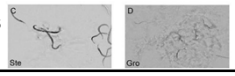
Protein expression and interactions (Mass spectrometry)

AAACTGCATCGA
ATCGCGCATCGC
AGCTCTAGCTCCC...

Gene organization (Genome sequences)



Gene-gene interactions (Genetic assays)



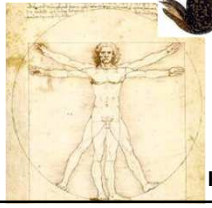
Yeast

Bacteria

C. elegans nematodes

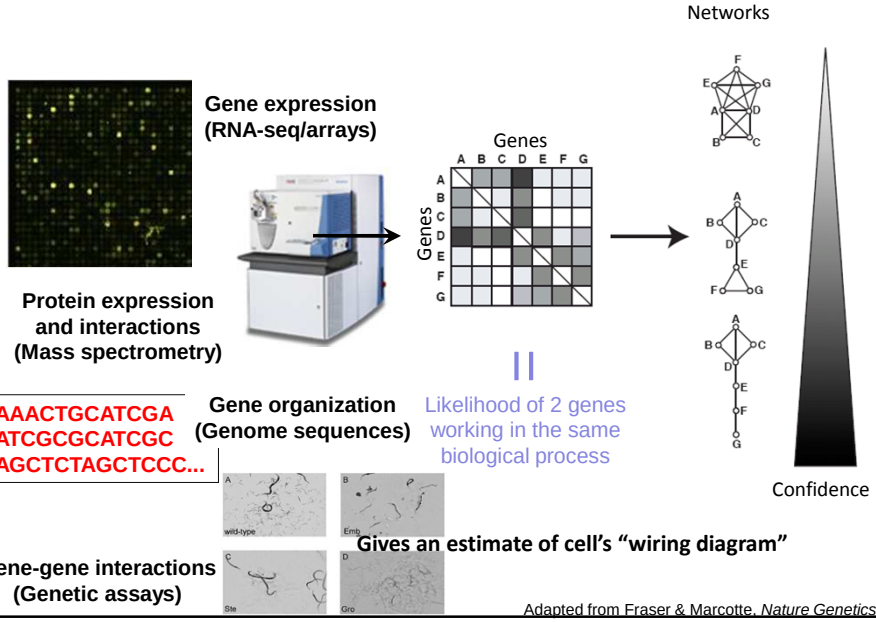
Mouse

***Xenopus* frogs**



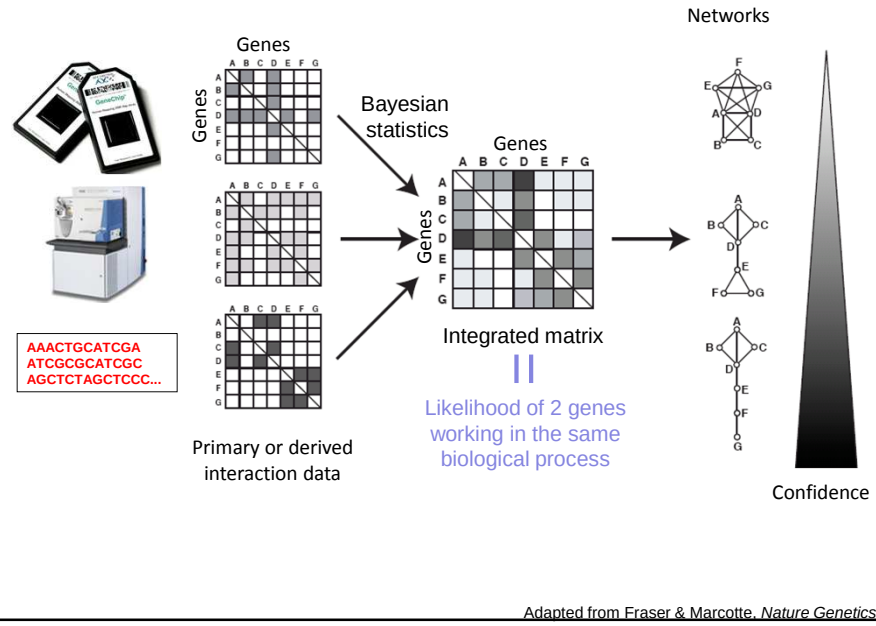
Humans

These sorts of data can be combined into functional gene networks



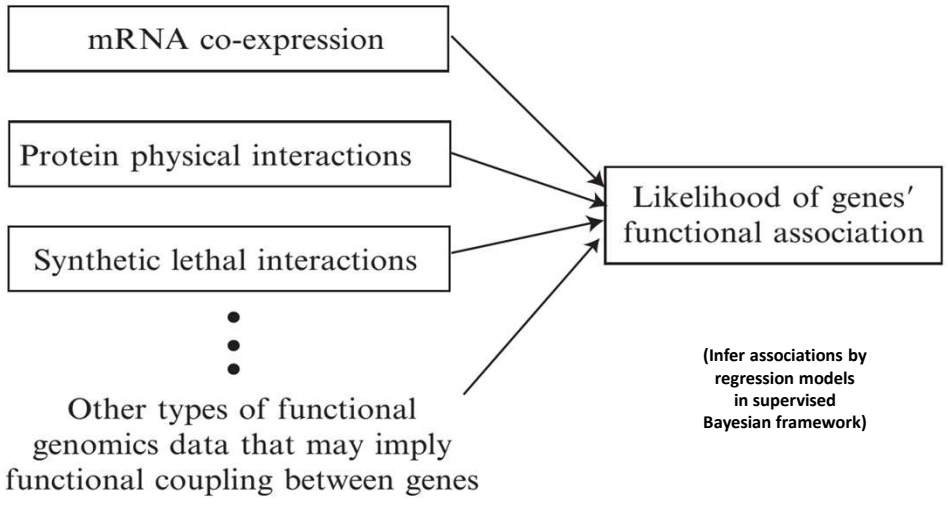
Adapted from Fraser & Marcotte, *Nature Genetics* (2004)

These sorts of data can be combined into functional gene networks



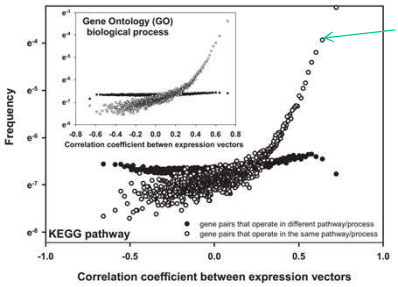
Adapted from Fraser & Marcotte, *Nature Genetics* (2004)

In more detail: Constructing a functional gene network



Lee & Marcotte, *Methods Mol Biol*. 453:267-78. (2008)

e.g., inferring functional linkages from mRNA co-expression across a given set of conditions

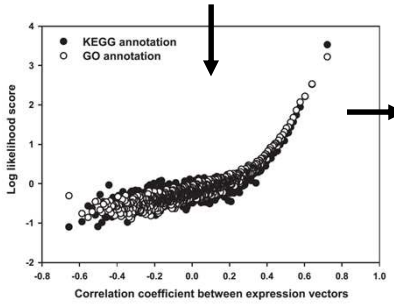


bins of 20K gene pairs

Frequencies in the dataset (D) of gene pairs sharing pathway annotations (I) and not sharing annotations (~I), calculated per bin

$$\mathrm{LLS} = \ln\left(\frac{P(I|D)/P(\sim I|D)}{P(I)/P(\sim I)}\right)$$

Background frequencies of gene pairs sharing and not sharing annotations

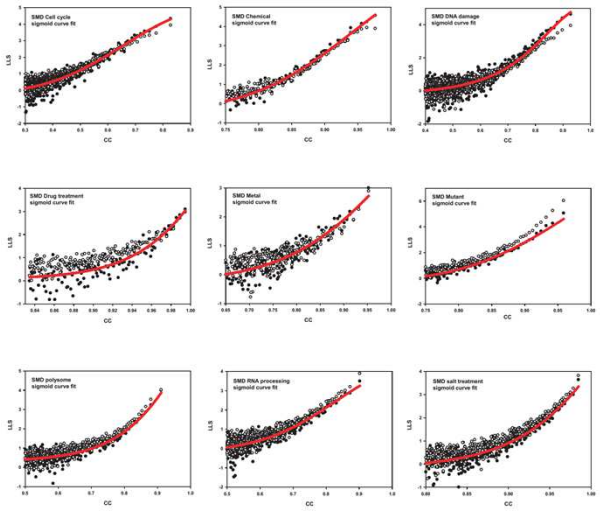


Fit regression model, score all expressed gene pairs (annotated + unannotated)

Repeat for other datasets
Integrate scores for each link
Assess performance by cross-validation or bootstrapping

Lee *et al. Science*, 306:1555-8 (2004)

Typically calculated for many different sets of experiments sampling many different conditions…



Each represents a different set of mRNA abundance profiling experiments (here, DNA microarrays) interrogating a given set of conditions.

Different gene pairs may be correlated in each condition.

… and so on

Lee *et al. Science*, 306:1555-8 (2004)

21 evidence types contribute to the HumanNet human gene network

Note: many are not human!

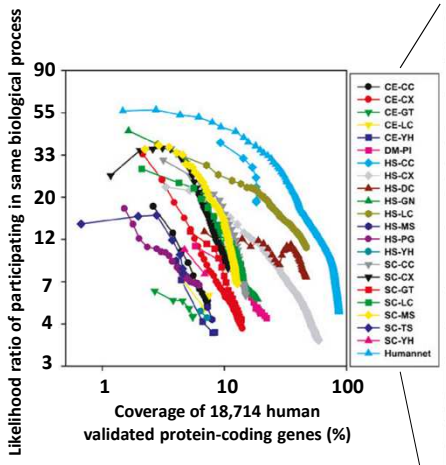


Table S1. Twenty-one different lines of evidence supporting HumanNet linkages.

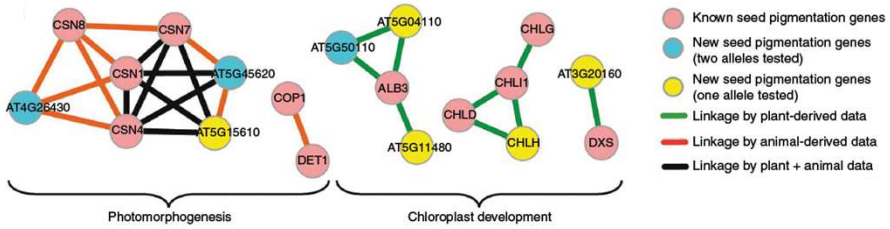
Data set	Description	# genes	# gene pairs
CE-CC	Co-citation of worm genes	1,370	12,928
CE-CX	Co-expression among worm genes	2,633	41,645
CE-GT	Worm genetic interactions	1,040	5,430
CE-LC	Literature-curated worm protein physical interactions	1,402	2,640
CE-YH	High-throughput yeast 2-hybrid assays among worm genes	1,561	3,254
DM-PI	Fly protein physical interactions	4,153	15,738
HS-CC	Co-citation of human genes	3,423	6,172
HS-CX	Co-expression among human genes	11,050	156,317
HS-DC	Co-occurrence of domains among human proteins	8,737	38,797
HS-GN	Gene neighborhoods of bacterial and archaeal orthologs of human genes	3,504	36,487
HS-LC	Literature-curated linkages from protein-protein interaction DBs (HPRD, BIND, BioGRID, IntAct, MINT) and Rual *et al.*, and pathway DB (Reactome)	8,783	56,505
HS-MS	Human protein complexes from affinity purification/mass spectrometry	1,485	3,575
HS-PG	Co-inheritance of bacterial and archaeal orthologs of human genes	1,170	18,868
HS-YH	High-throughput yeast 2-hybrid assays among human genes	1,358	1,365
SC-CC	Co-citation of yeast genes	2,798	31,353
SC-CX	Co-expression among yeast genes	2,001	48,423
SC-GT	Yeast genetic interactions	2,584	17,678
SC-LC	Literature-curated yeast protein physical interactions	2,661	17,280
SC-MS	Yeast protein complexes from affinity purification/mass spectrometry	2,382	65,986
SC-TS	Yeast protein interactions inferred from tertiary structures of complexes	859	6,270
SC-YH	High-throughput yeast 2-hybrid assays among yeast genes	1,292	1,801

Lee, Blom *et al*., *Genome Research* 21:1109-21 (2011)

Evolutionary information is usually a key predictor– e.g., predictions for plant-specific traits often use fungal & animal data

For example, new seedling pigmentation genes…

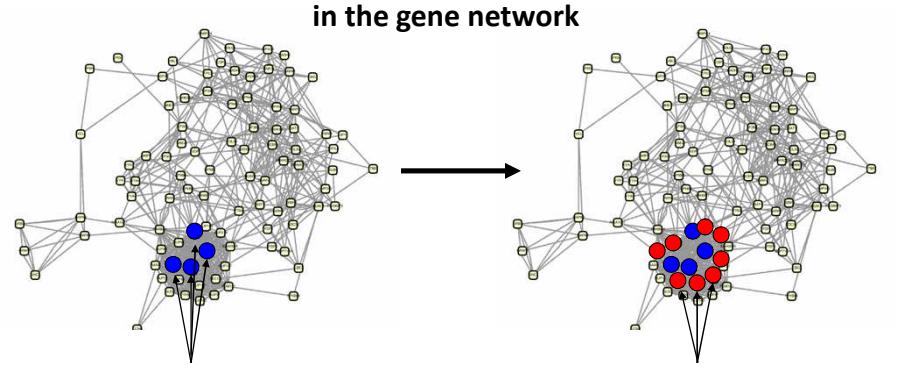
…were predicted from both plant and animal data:



Lee, Ambaru, *et al. Nature Biotech* 28(2):149-156 (2010)

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

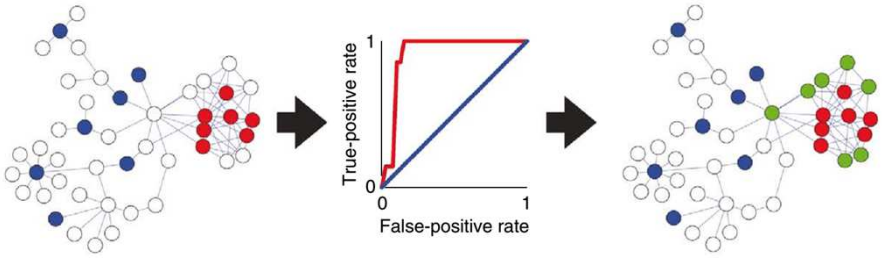
Guilt-by-association
in the gene network



Genes already linked to a disease or function

New candidate genes for that process

We can propagate annotations across the graph to infer new annotations for genes (network "guilt-by-association", or GBA). Measuring how well this works on hidden, but known, functions gives us an idea 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)

Calculating ROC curves

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

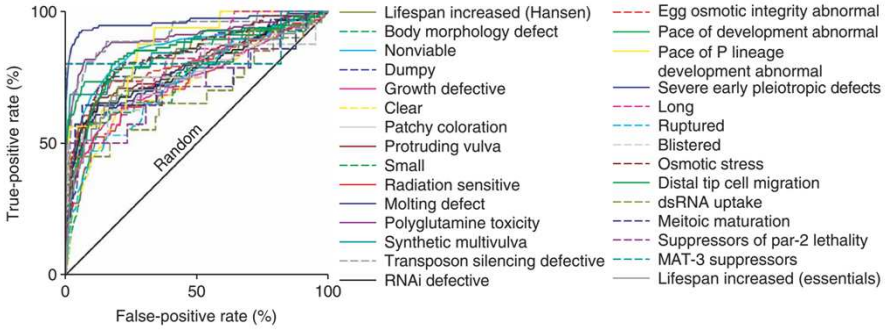
Basic idea: sort predictions from best to worst, plot TPR vs. FPR as you traverse the ranked list

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

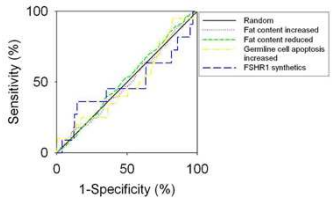
FPR = FP / N = FP / (FP + TN)
= False Positive Rate
= 1 - Specificity

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:

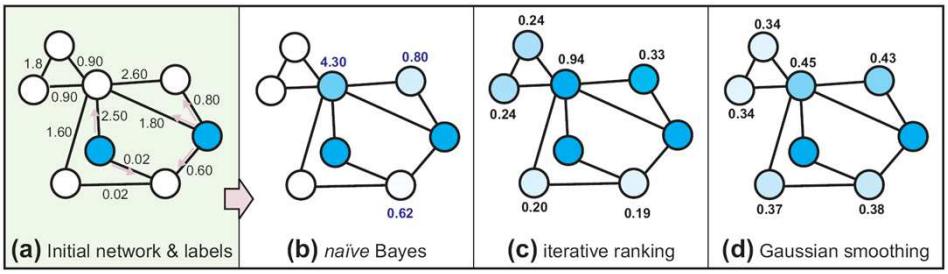


ROC analysis indicates the likely predictive power of the network for a system of interest.

A poor ROC → no better than random guessing.

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

A variety of algorithms have been developed for GBA



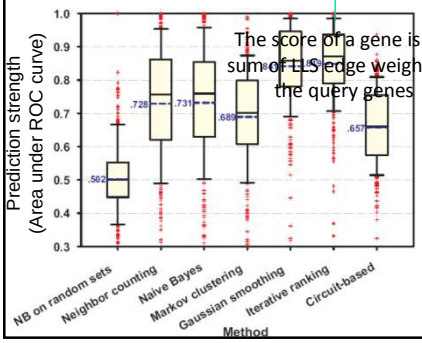
McGary, Lee, Marcotte *Genome Biol.* (2007)

Related to Google's PageRank Ramakrishnan *et al. Bioinformatics* (2009)

GeneMANIA Mostafavi *et al., Genome Biol.* (2008)

The score of a gene is the sum of LLS edge weights to the query genes

The score of a gene is the combination of the initial seeds and the weighted average of scores of the protein's neighbors, defined iteratively.



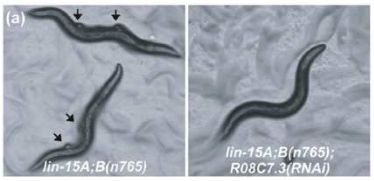
Predicting genes for 318 *C. elegans* RNAi phenotypes

Wang & Marcotte, *J. Proteomics* (2010)

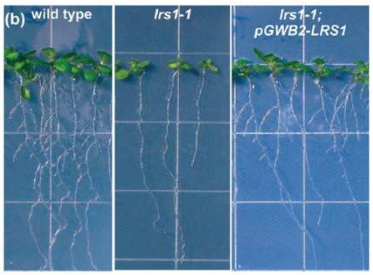
Remarkably, this strategy works quite well

Some examples of network-guided predictions:

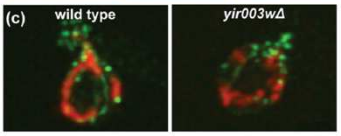
In worms:
Genes that can reverse 'tumors' in a nematode model of tumorigenesis
Lee, Lehner *et al. Nature Genetics* (2008)

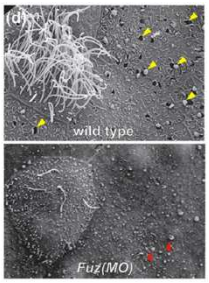


In *Arabidopsis*:
New genes regulating root formation
Lee, Ambaru *et al*. *Nature Biotech* (2010)

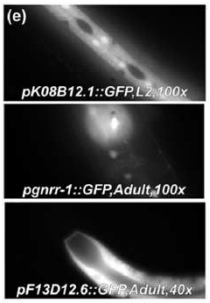


In yeast: **New mitochondrial biogenesis genes**
Hess *et al*., *PLoS Genetics* (2009)





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)

Applicable to non-model organisms:

In rice: Identifying genes regulating resistance to *Xanthomonas oryzae* infection …

Rice is the primary food source for >2 billion people worldwide

- >500 million tons rice/year are grown
- Bacterial rice blight destroys up to 10-50% / year in Africa/Asia

Normal **Infected**

Source: Russell, Hertz, McMillan *Biology: The Dynamic Science*

A new gene promoting resistance to rice blight

Infection with *Xoo*

wild type resistant resistant + RNAi

Lee, Seo, *et al*. *PNAS* 108:18548–18553 (2011)

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

Summary of the major themes

- Gene networks serve as general frameworks for studying gene function
- Functional gene networks can be (re)constructed based upon millions of experimental observations via integrating these data into statistical models of functional connectivity among genes
- Guilt-by-association in such a network allows association of genes with functions, and genes with phenotypic traits, even highly polygenic ones
- Functional gene networks have been constructed for yeast, *C. elegans*, *Arabidopsis*, rice, mice, humans, many prokaryotes, and many other organisms…

Live demo of functional networks and Cytoscape