Translational Bioinformatics 2008: The Year in Review

Russ B. Altman, MD, PhD
Stanford University

Thanks!

- Atul Butte
- Larry Hunter
- Dan Masys
- Joel Dudley
- Maureen Hillenmeyer
- Florian Schmitzberger
- Art Toga
- Jill Mesirov
- Alfonso Valencia
- David Pollock
- Faculty of 1000
- Publisher web sites
- Pubmed
- Google Scholar
- Papers v1.7 (for Mac)
- Stanford University Library

Goals

- Provide an overview of the major scientific events, trends and publications in translational bioinformatics
- Create a "snapshot" of what seems to be important in March, 2008 for the amusement of future generations.
- Revel in the richness of opportunities and challenges that we are privileged to approach at this moment in time.

Process: “the think method”

Process

1. Think about what has had impact
2. Think about sources to trust
3. Solicit advice from colleagues
4. Surf online resources
5. Print out abstracts, cut them up, made into piles
6. Selected key papers from piles
7. Tweaked

Caveats

- Focused on human biology and clinical implications (except really important model systems)
- Focused on both data sources and informatics methods
- (Attempted to) Focus on definitive results vs. intriguing preliminary results
- Tried to avoid simply following crowd mentality.

What were the piles?

- Sequencing and sequence analysis
- Genome Wide Association Studies
- Pharmacogenomics & Drugs
- Analysis of high-throughput molecular data for humans
- Neuroscience
- Molecular information for understanding human disease
- Potpourri or "I can't resist..."

Final Results

- 86 semi-finalist papers
- 45 finalist papers
- 27 presented here
- This talk and ENDNOTE files (and other formats) will be made available on the conference website.

- (Yikes!)

Making Life Hard...

"Deja vu--a study of duplicate citations in Medline" (Errami et al, Bioinformatics)

- eTBLAST tool, Deja vu database
- 62,213 medline citations studied
- 0.04% potentially plagiarism
- 1.35% seemed duplicate
- **120,000 duplicate citations in MEDLINE**

Sequencing and Sequence Analysis

"The diploid genome sequence of an individual human"

(Levy et al, PLoS Biology)

- J. Craig Venter's
  - 4.1 million variants vs. standard genome build
  - 22% of variants not SNPs (560K indels, 90 inversions, duplications, CNVs, etc...)
  - Cost: $3B > $70M > $1M > $300K

“Next-generation” Sequencing

“Assembling millions of short DNA sequences using SSAKE” (Warren et al, Bioinformatics)

- Solexa: millions of 25 nt DNA reads per run

“Genome-wide mapping of in vivo protein-DNA interactions”

- Replaces Chromatin-IP + Microarrays (ChIP-chip) with ChIP + Sequencing (ChIP-Seq)
- 1946 targets for NRSF transcription factor found

"Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project" (Birney et al, Nature)

- "Full court press" on all aspects of annotation of 1% of human genome with combination of experimental and computational modalities

- FINDINGS: (1) lots of transcription, (2) chromatin is dominant, (3) 5% conserved/selected, (4) many functional elements not conserved

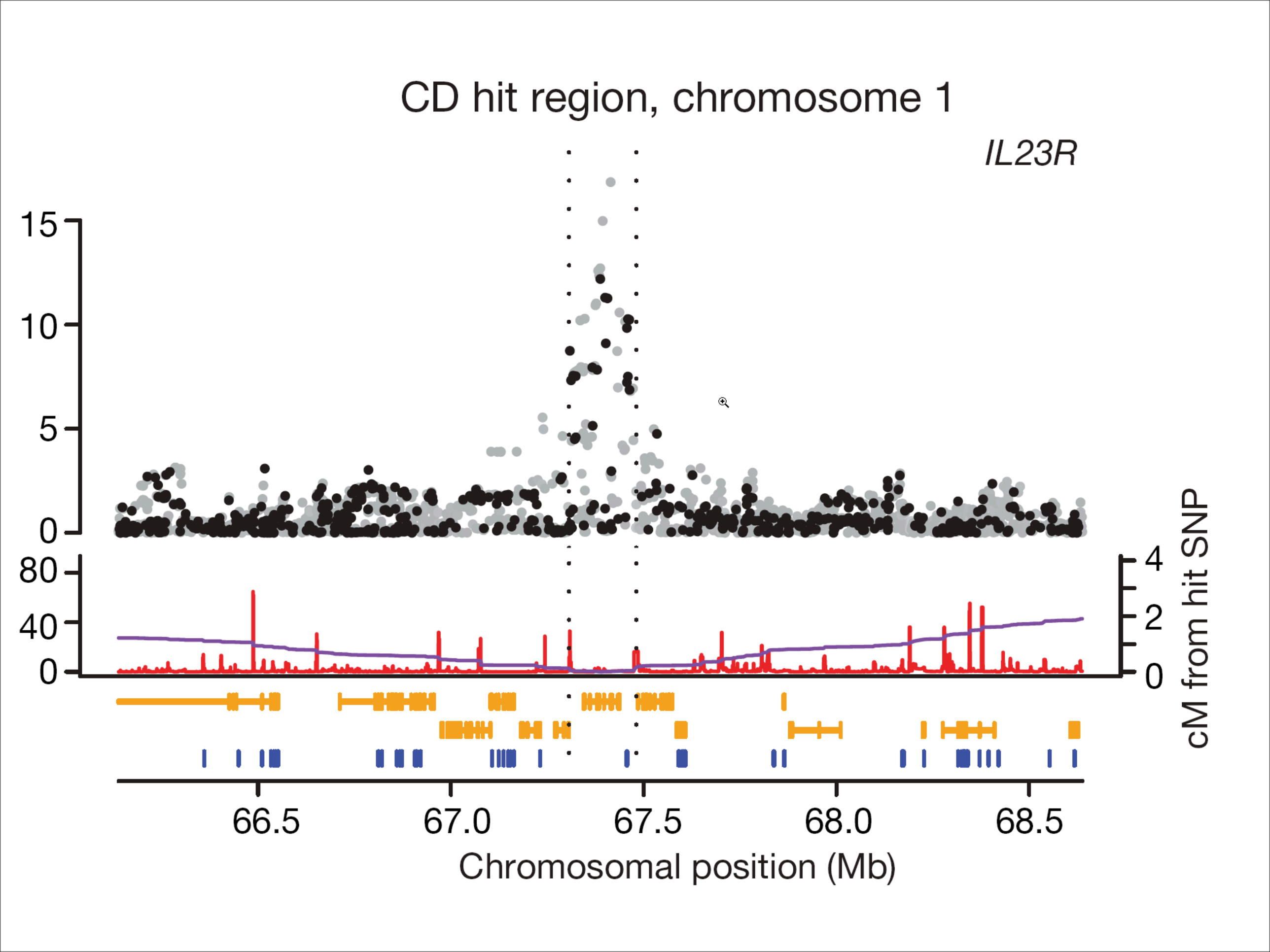
- Creates a market and demo for a "human genome knowledgebase" that annotates the history, significance and function of every base in the genome.

Genome wide associations

"Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls"

(Wellcome Trust Consortium, Nature)

- 500K SNP Affymetrix platform
- 24 independent associations: Bipolar disease (1) Coronary artery disease (1), Crohn's disease (9), RA (3), Type I DM (7), Type II DM (3)
- 58 other SNPs promising...data available.


CD hit region, chromosome 1
IL23R
15
10
5
0
80
40
0
4
2
0
cM from hit SNP
66.5
67.0
67.5
68.0
68.5
Chromosomal position (Mb)

Other GWAS...

- " Psoriasis is associated with increased beta-defensin genomic copy number" (Hollox et al, Nat. Gen)
- "Strong association of de novo copy number mutations with autism" (Sebat et al, Science)
- "Genome wide expression analysis in HPV16 Cervical Cancer: identification of altered metabolic pathways" (Perez-Plasencia et al, Infect Agent Cancer)

Pharmacogenomics and drugs

"Randomized trial of genotype-guided versus standard warfarin dosing in patients initiating oral anticoagulation"

(Anderson et al, Circulation)

- Genotype VKORC1 and CYP2C9 SNPs
- Smaller and fewer dosing changes
- 206 patients
- New gene: CYP4F2 (Caldwell et al, 2008)
- Many awaiting result of major combined analysis of ~6000 patients internationally (IWPC)

More on drugs...

"Drug-target network"

(Yildirim et al, Nature Biotech)

- Mined FDA data to build drug-target binary associations
- Found abundance of "me too" drugs
- New drugs more diverse
- Trend towards rational drug design noted.

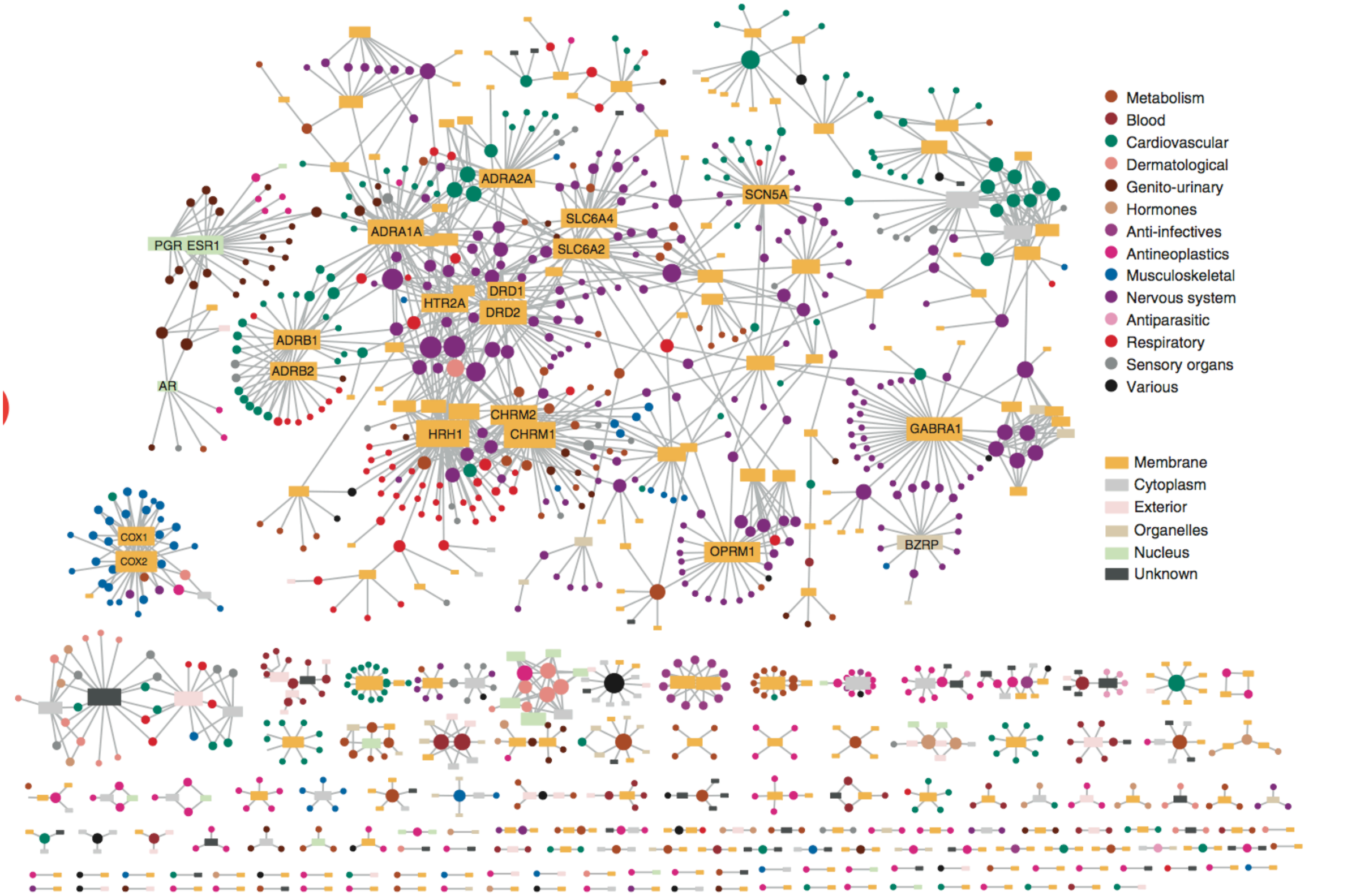
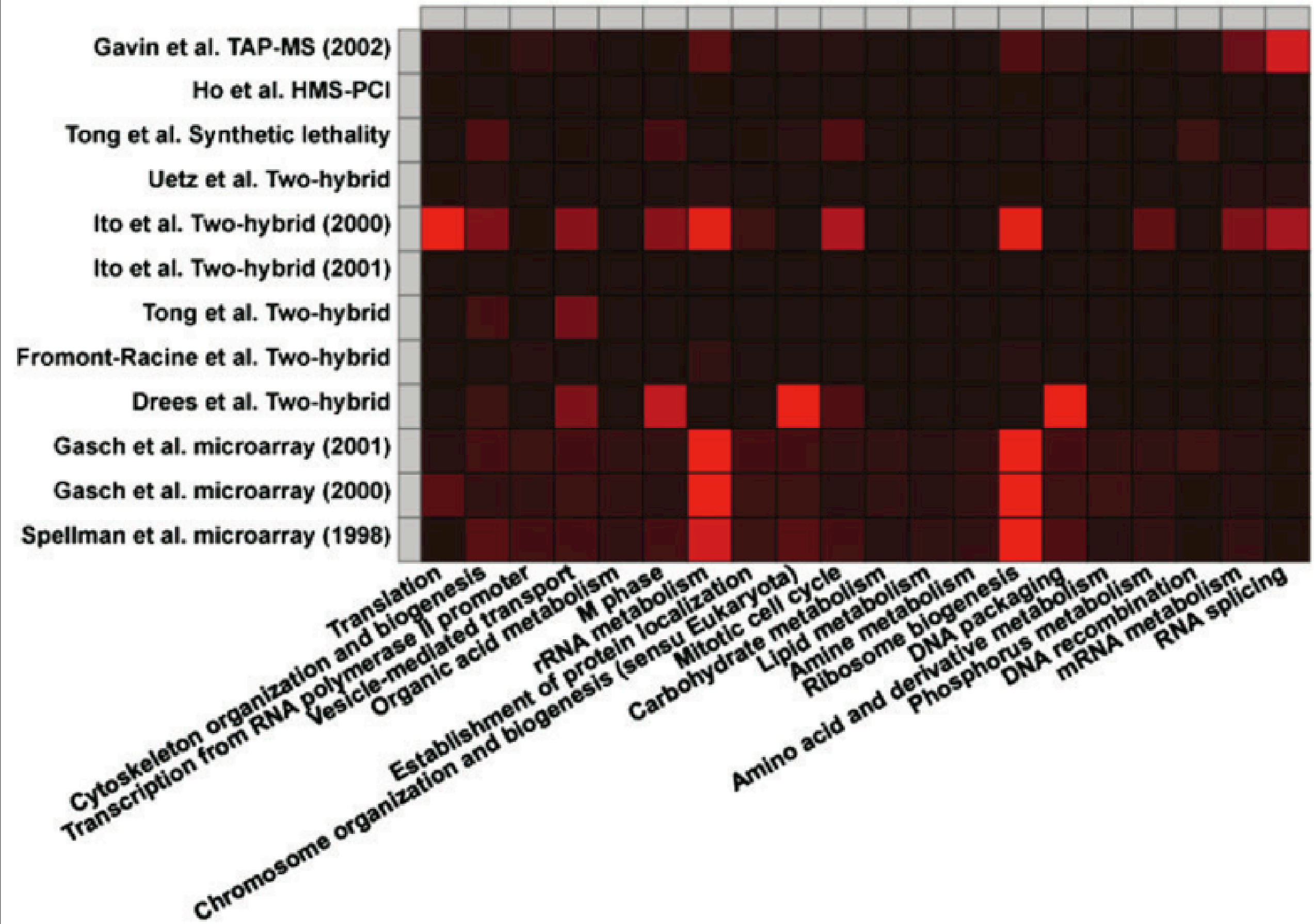


Figure 2 Drug–target network (DT network). The DT network is generated by using the known associations between FDA-approved drugs and their target proteins. Circles and rectangles correspond to drugs and target proteins, respectively. A link is placed between a drug node and a target node if the protein is a known target of that drug. The area of the drug (protein) node is proportional to the number of targets that the drug has (the number of drugs targeting the protein). Color codes are given in the legend. Drug nodes (circles) are colored according to their Anatomical Therapeutic Chemical Classification, and the target proteins (rectangular boxes) are colored according to their cellular component obtained from the Gene Ontology database.

Informatics approaches to high-throughput data analyses

"Context-sensitive data integration and prediction of biological networks" (Myers & Troyanskaya, Bioinformatics)

- Many efforts to integrate across multiple data sets for (re)construction of networks of interactions
- Often plagued by noise in these datasets
- Hypothesis: different high-throughput data sets have different information content based on biological questions asked
- Solution: Use gold standard of known interactions and Bayesian analysis methods to capture relationships between experiments and biological reliability.


Gavin et al. TAP-MS (2002)
Ho et al. HMS-PCI
Tong et al. Synthetic lethality
Uetz et al. Two-hybrid
Ito et al. Two-hybrid (2000)
Ito et al. Two-hybrid (2001)
Tong et al. Two-hybrid
Fromont-Racine et al. Two-hybrid
Drees et al. Two-hybrid
Gasch et al. microarray (2001)
Gasch et al. microarray (2000)
Spellman et al. microarray (1998)
Translation
Cytoskeleton organization and biogenesis
Transcription from RNA polymerase II promoter
Vesicle-mediated transport
Organic acid metabolism
M phase
rRNA metabolism
Establishment of protein localization
Chromosome organization and biogenesis (sensu Eukaryota)
Mitotic cell cycle
Carbohydrate metabolism
Lipid metabolism
Amine metabolism
Ribosome biogenesis
DNA packaging
Amino acid and derivative metabolism
Phosphorus metabolism
DNA recombination
mRNA metabolism
RNA splicing

"Enabling integrative genomic analysis of high-impact human diseases through text mining. (Dudley & Butte, Pac Symp on Biocomp)

- Problem: annotating large data sets, such as available in Gene Expression Omnibus (NCBI)
- Natural language processing (NLP) techniques used to map GEO datasets to UMLS terms for indexing
- Significantly: able to also identify matched control data sets (for 62% of disease data sets)
- Labels cover 30% of US disease mortality

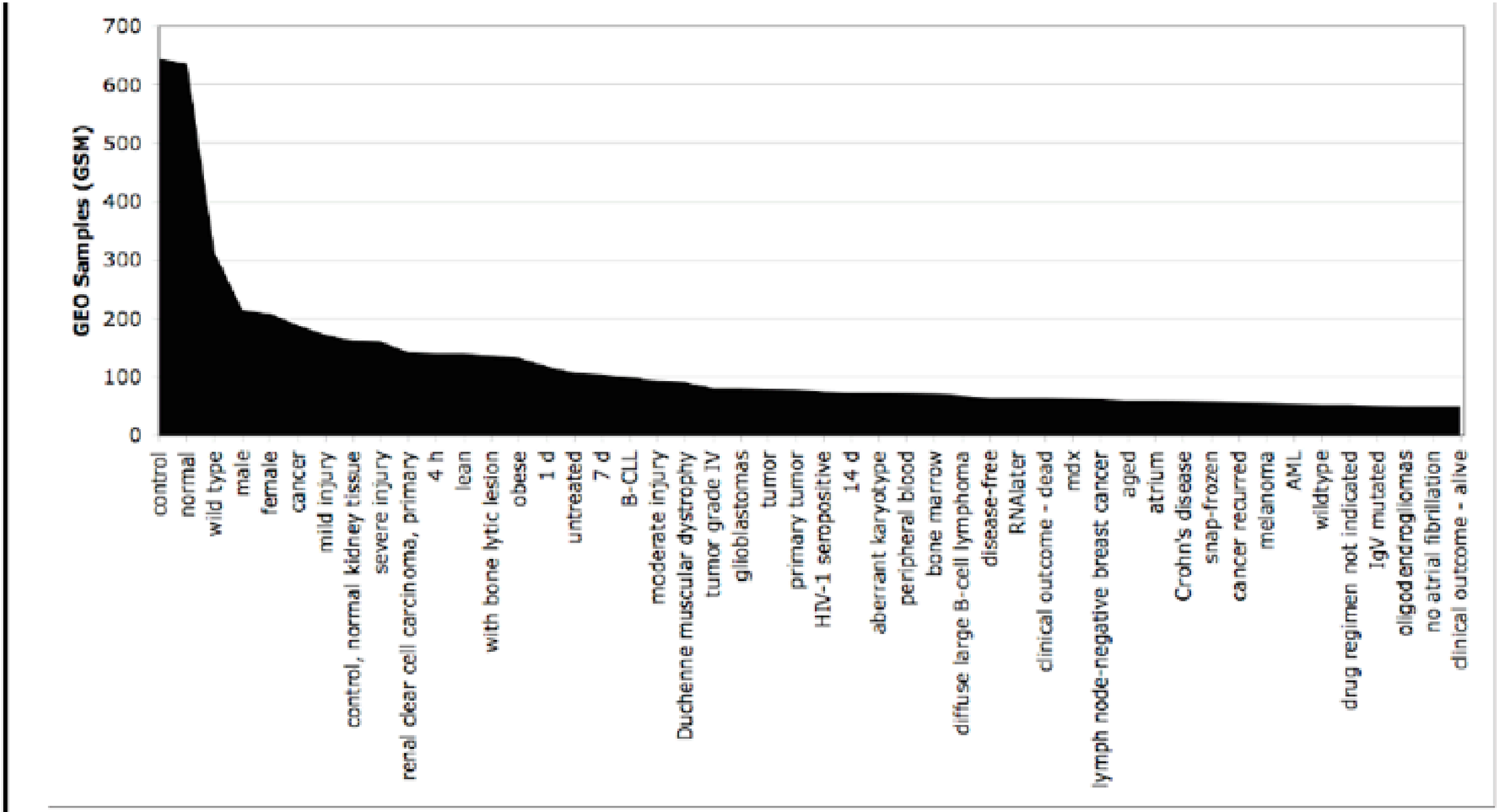
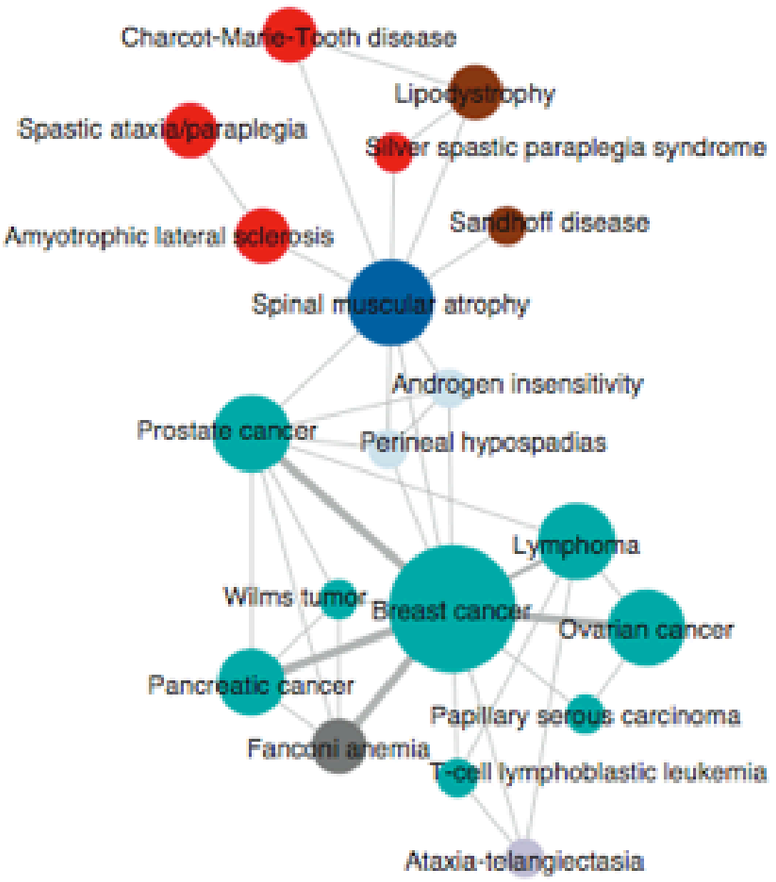


Figure 3. Distribution of GDS subset annotation phrases for all disease-related GDS. The distributions are filtered to terms annotating > 5 GDS and > 50 GSM for display purposes. The distribution shows that the **(a)** majority of disease-related GDS contain subsets annotated with a small set of common control phrases, **(b)** representing a major proportion of samples.

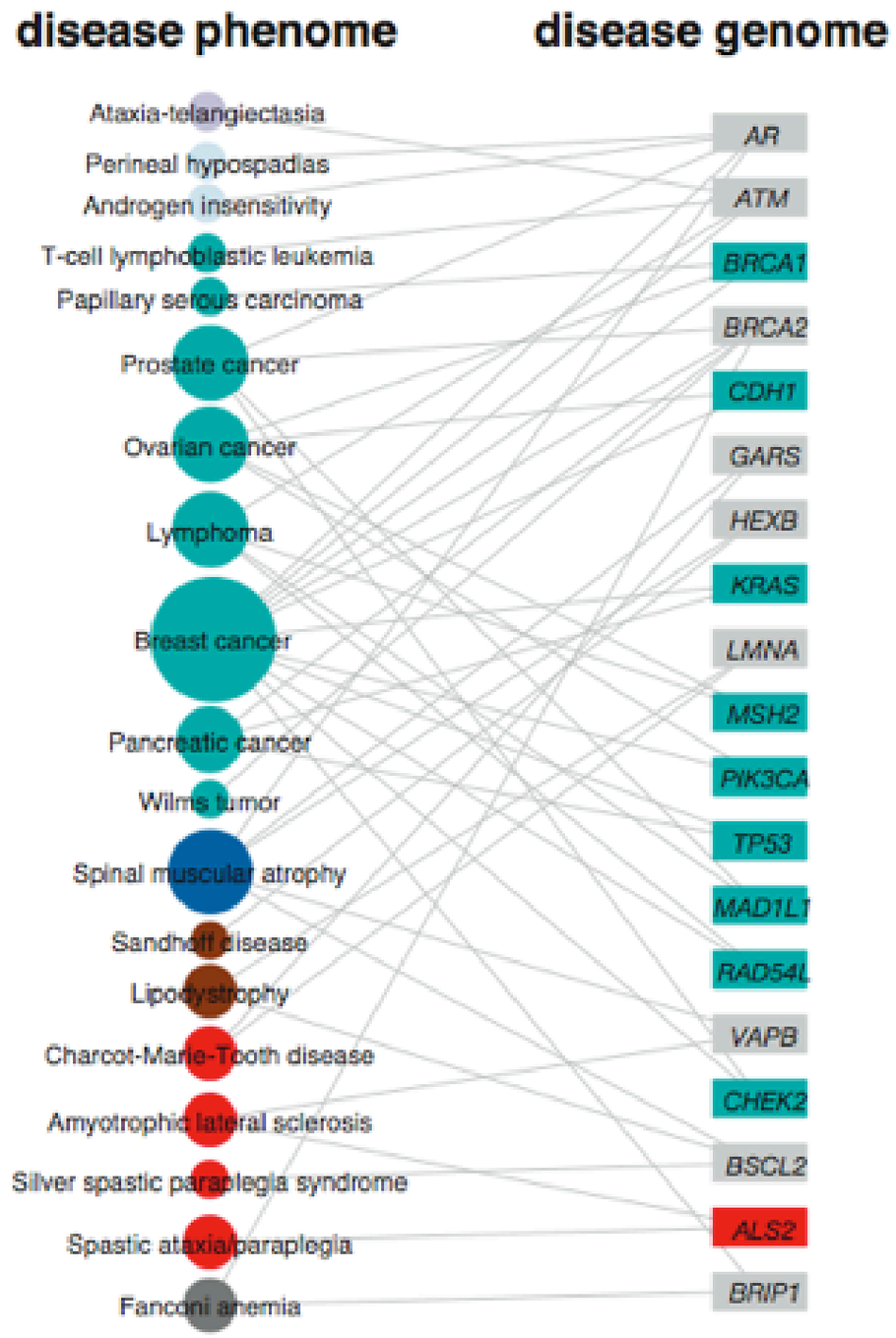
"The human disease network" (Goh et al, PNAS)

- Built a network of diseases and genes from OMIM (Online Mendelian Inheritence in Man, NCBI)
- Genes associated via disease links have higher probability of physical interaction
- Essential genes encode hubs in interaction networks
- Thus, authors predict (and show) that cancer genes will be "central" in network

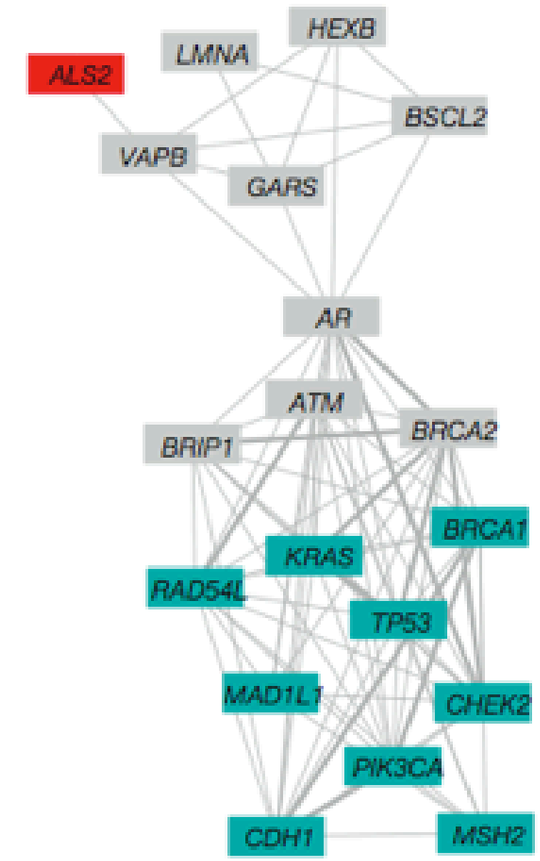
Human Disease Network (HDN)



DISEASOME



Disease Gene Network (DGN)



"Global reconstruction of the human metabolic network based on genomic and bibliomic data" (Duarte et al, PNAS)

- Build 35 of human genome + 50 years of literature
- Used it to:
  - Discover missing information about human metabolism (!)
  - Starting point for numerical model/simulation of metabolism
  - Set context for analysis of high-throughput data sets

Table 1. *H. sapiens* Recon 1 network statistics

Component	Number
Genes	1,496
Transcripts*	1,905
Proteins	2,004
Complex-associated reactions*	248
Isozyme-associated reactions*	946
Intrasystem reactions	3,311
Metabolic	2,233
Transport†	1,078
Exchange reactions†	432
Compartment-specific metabolites	2,712
Cytoplasm	995
Extracellular space	388
Mitochondrion	383
Golgi apparatus	279
Endoplasmic reticulum	231
Lysosome	207
Peroxisome	139
Nucleus	90
Citations	1,587
Primary literature	1,378
Review articles	188
Textbooks	21
Validated metabolic functions	288
Knowledge gaps‡	356

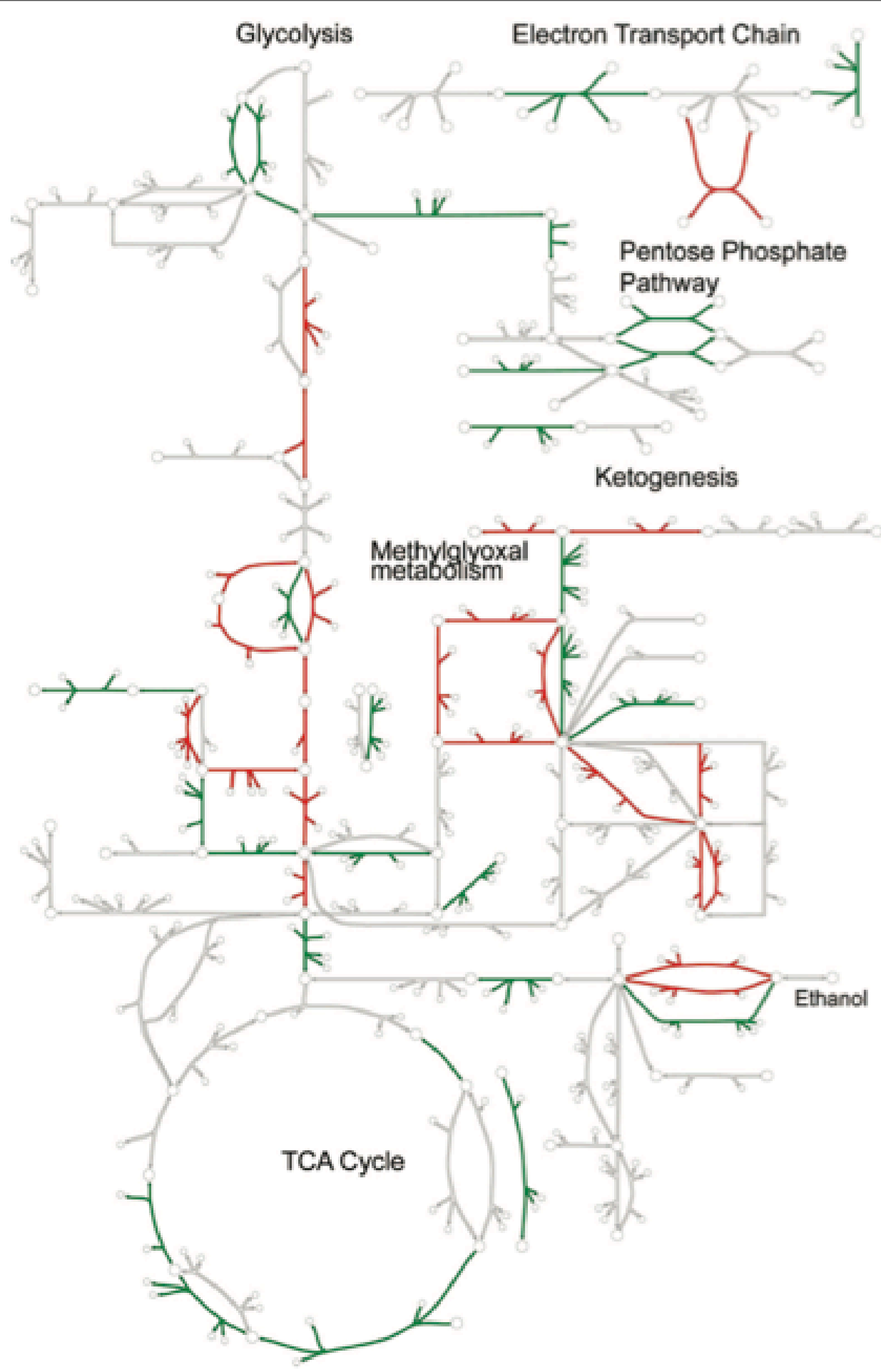


Fig. 4. Integrated analysis of gene expression data from gastric bypass patients before surgery and 1 year afterward. Expression measurements were

"NetworkBLAST: comparative analysis of protein networks" (Kalaev et al, Bioinformatics)

- Addresses problem of taking two networks of genes (protein products) and finding common substructures
- Like BLAST, looks for seed matches of similar connection topology and then extends
- Includes estimates of significance
- One of the authors has previously created Cytoscape, a commonly used network visualization took = Trey Ideker

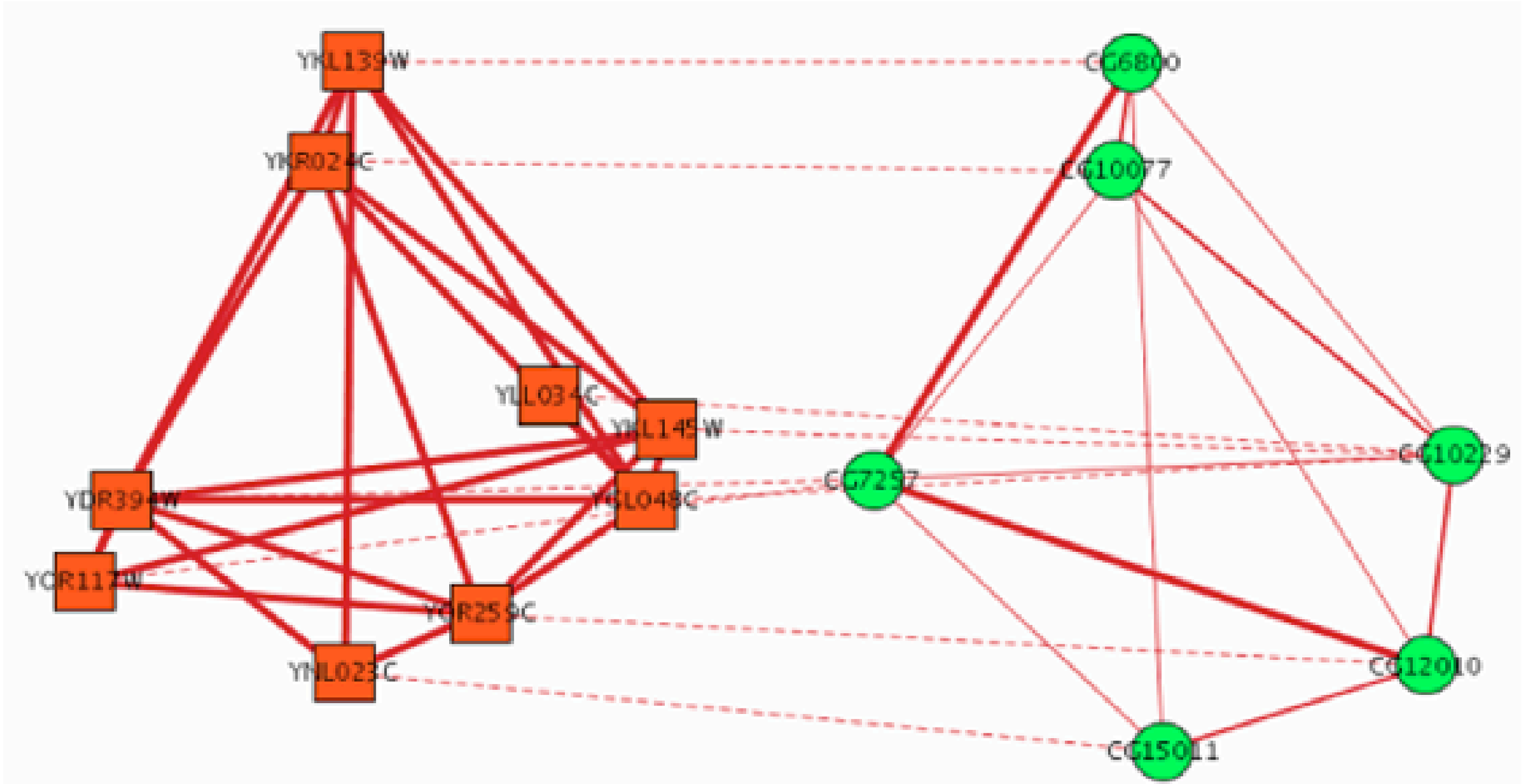


Fig.2. A representative yeast-fly conserved complex from NetworkBLAST's output. Yeast proteins appear in orange; fly proteins appear in green. Sequence-similar proteins are connected with dashed lines. Solid lines represent PPIs, with the line width corresponding to the reliability of the corresponding interaction.

"Universally sloppy parameter sensitivities in systems biology models" (Gutenkunst et al, PLoS Comput Biol)

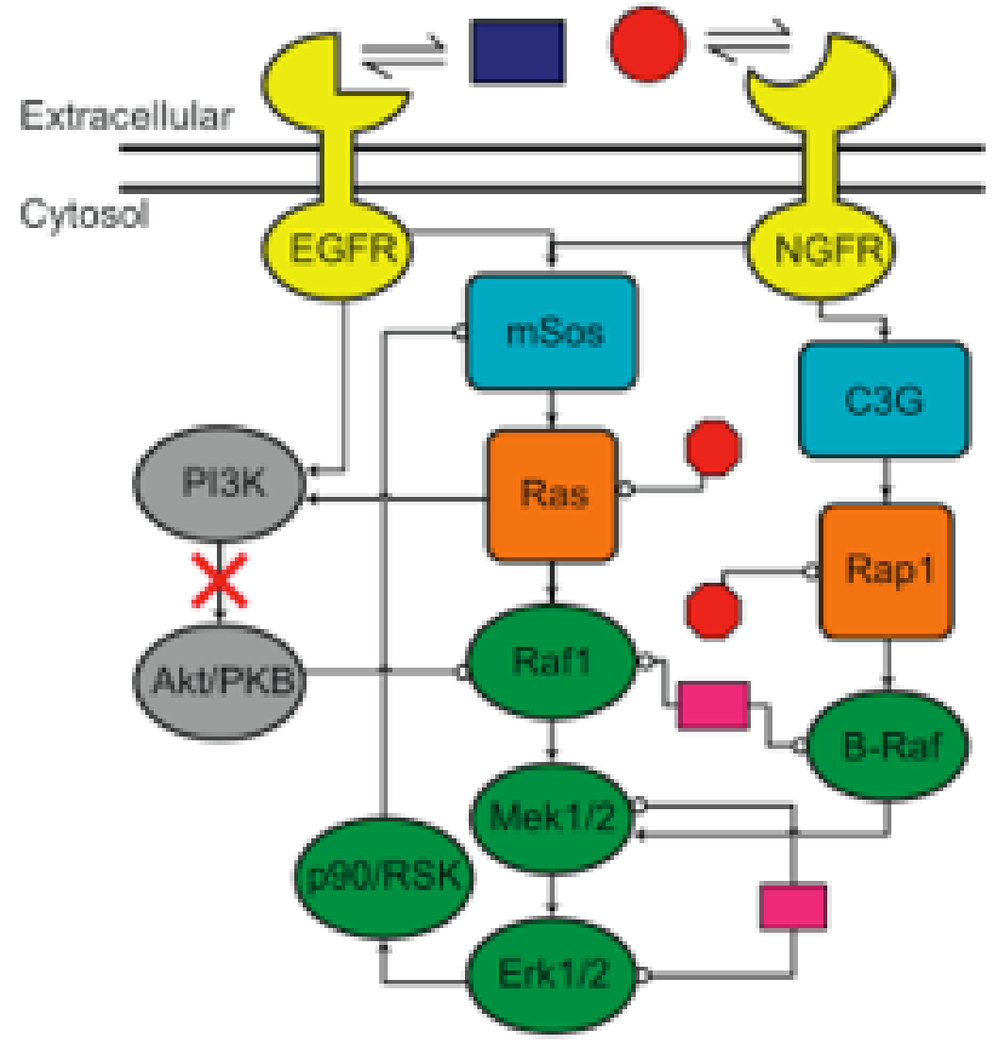
- Systems biology mathematical models have "sloppy" sensitivities--model depends on hard-to-predict combinations of parameters. Authors investigated this in 17 models.
- Sloppy sensitivities pervasive
- Even very large data sets will not yield accurate estimates of parameters (bad news)
- Modelers should focus on prediction sensitivity with estimated parameters as way to guide experiments (good news)

Red = tight parameter estimates ($$$)

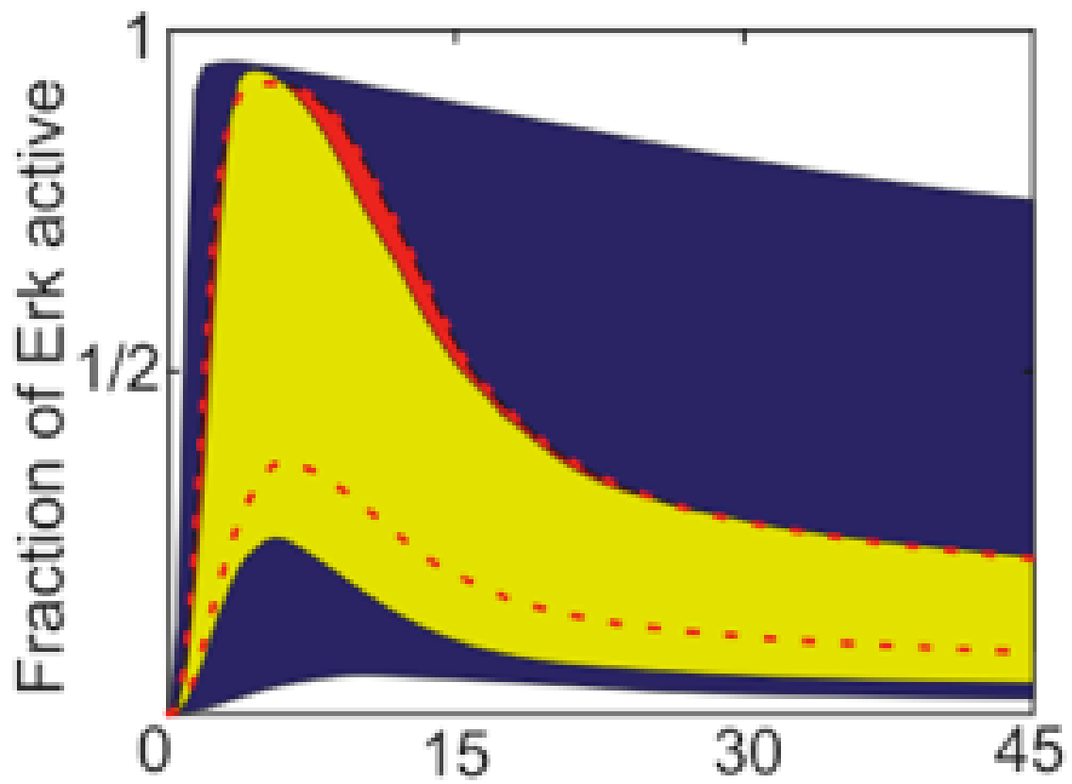
Blue = getting the key parameters inaccurately

Yellow = collectively fitting all parameters over multiple experiments

A



B



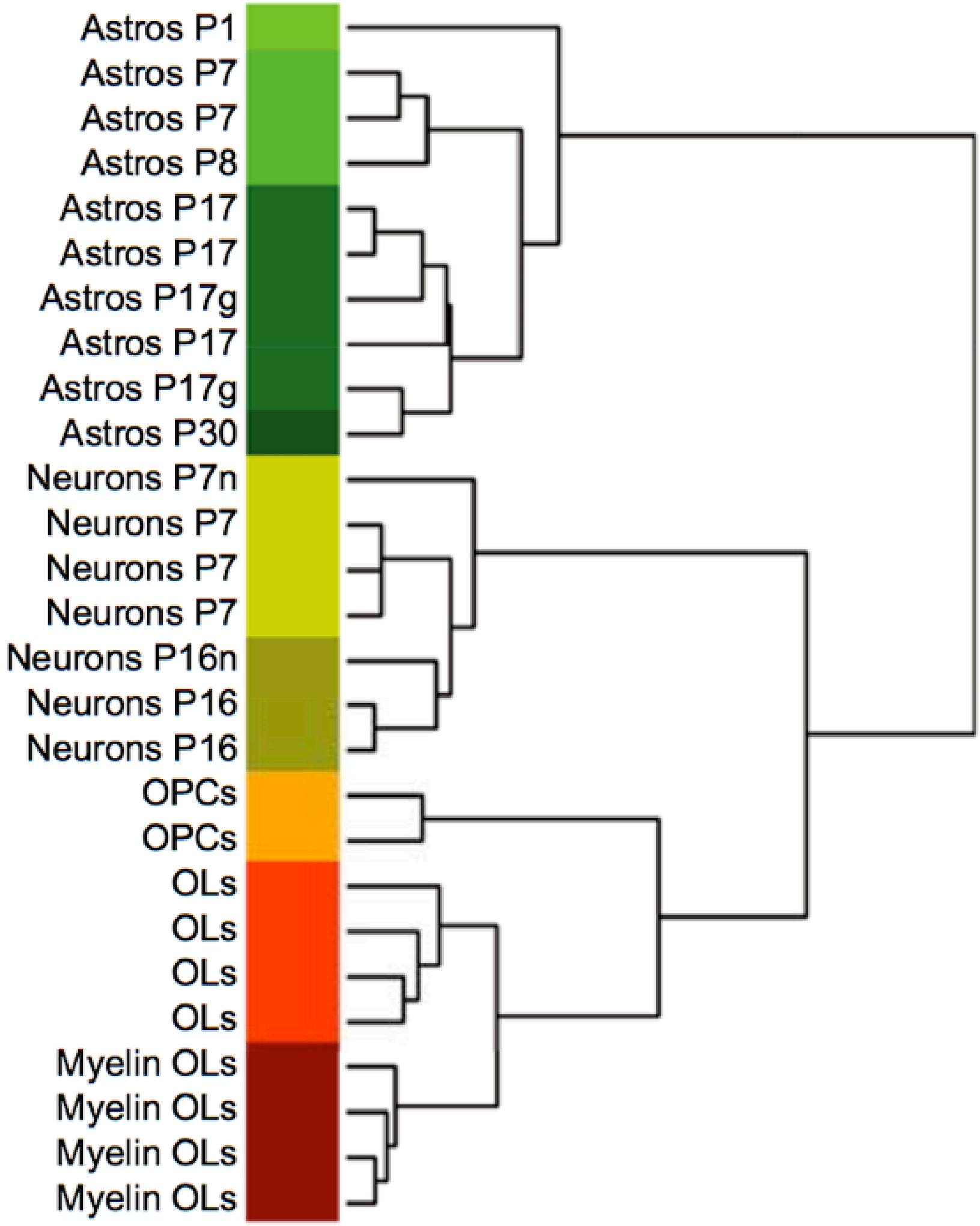
"Advancing translational research with the Semantic Web" (Ruttenberg et al, BMC Bioinformatics)

- Outlines a vision for semantic web and biomedical research. Very useful primer on the subject.
- Goal: general infrastructure for aggregation, annotation and integration of diverse biological datatypes
- HCLSIG within W3C
- RDF, OWL and other structured representations hold exciting promise of data integration infrastructure. Challenges remain...

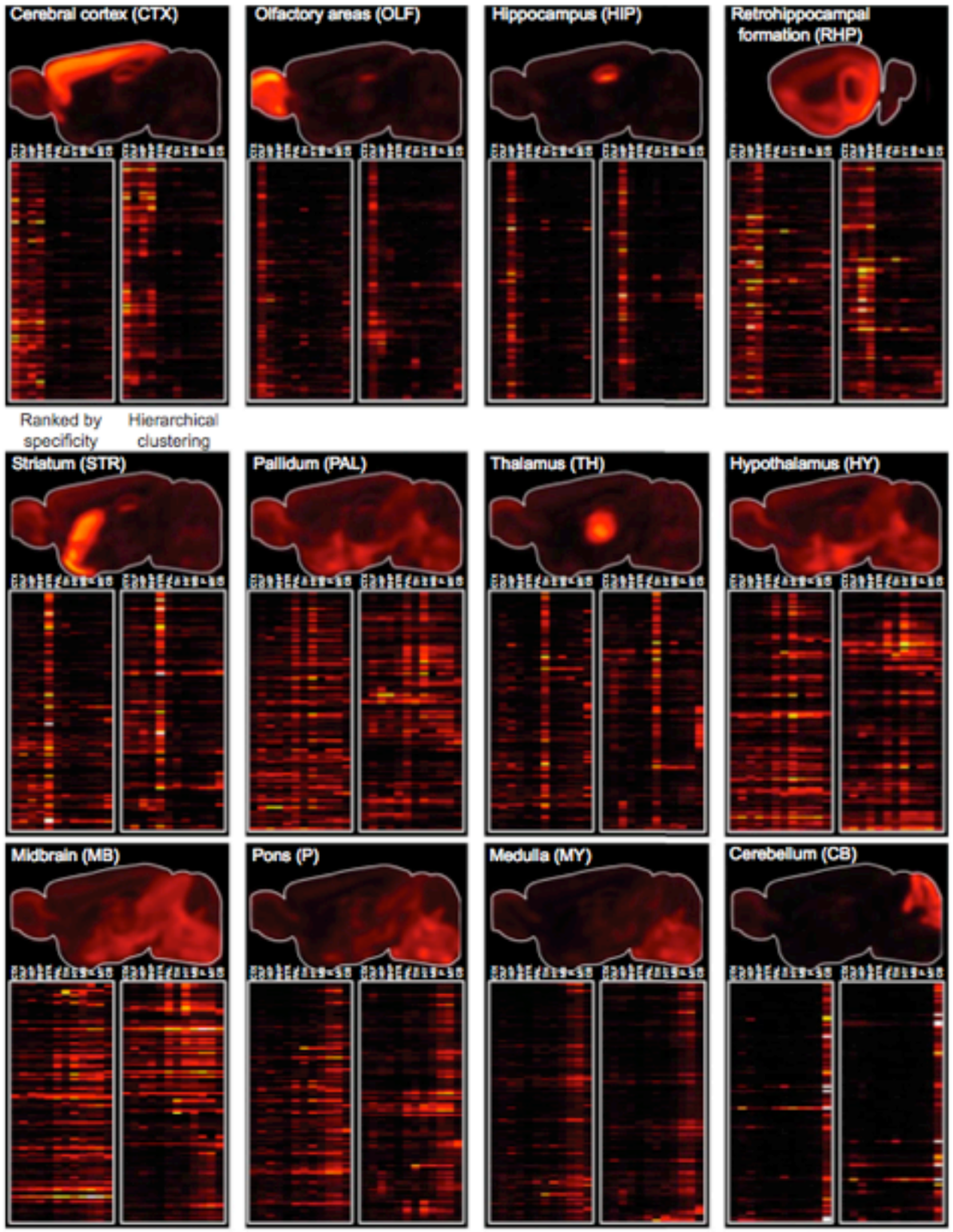
Blockbuster neuroscience datasets

"A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function" (Cahoy et al, J. Neurosci)

- Brain development and expression is one of the great challenges of our era. Sometimes making the data available can catalyze an entire field.
- Isolated populations of mouse brain cells using FACs
- Affymetrix expression analysis on 20K genes
- Ingenuity pathway analysis tool to find markers
- Differences in gene expression now driving experiments


Astros P1
Astros P7
Astros P7
Astros P8
Astros P17
Astros P17
Astros P17g
Astros P17
Astros P17g
Astros P30
Neurons P7n
Neurons P7
Neurons P7
Neurons P7
Neurons P16n
Neurons P16
Neurons P16
OPCs
OPCs
OLs
OLs
OLs
OLs
Myelin OLs
Myelin OLs
Myelin OLs
Myelin OLs

"Genome-wide atlas of gene expression in the adult mouse brain" (Lein et al, Nature)

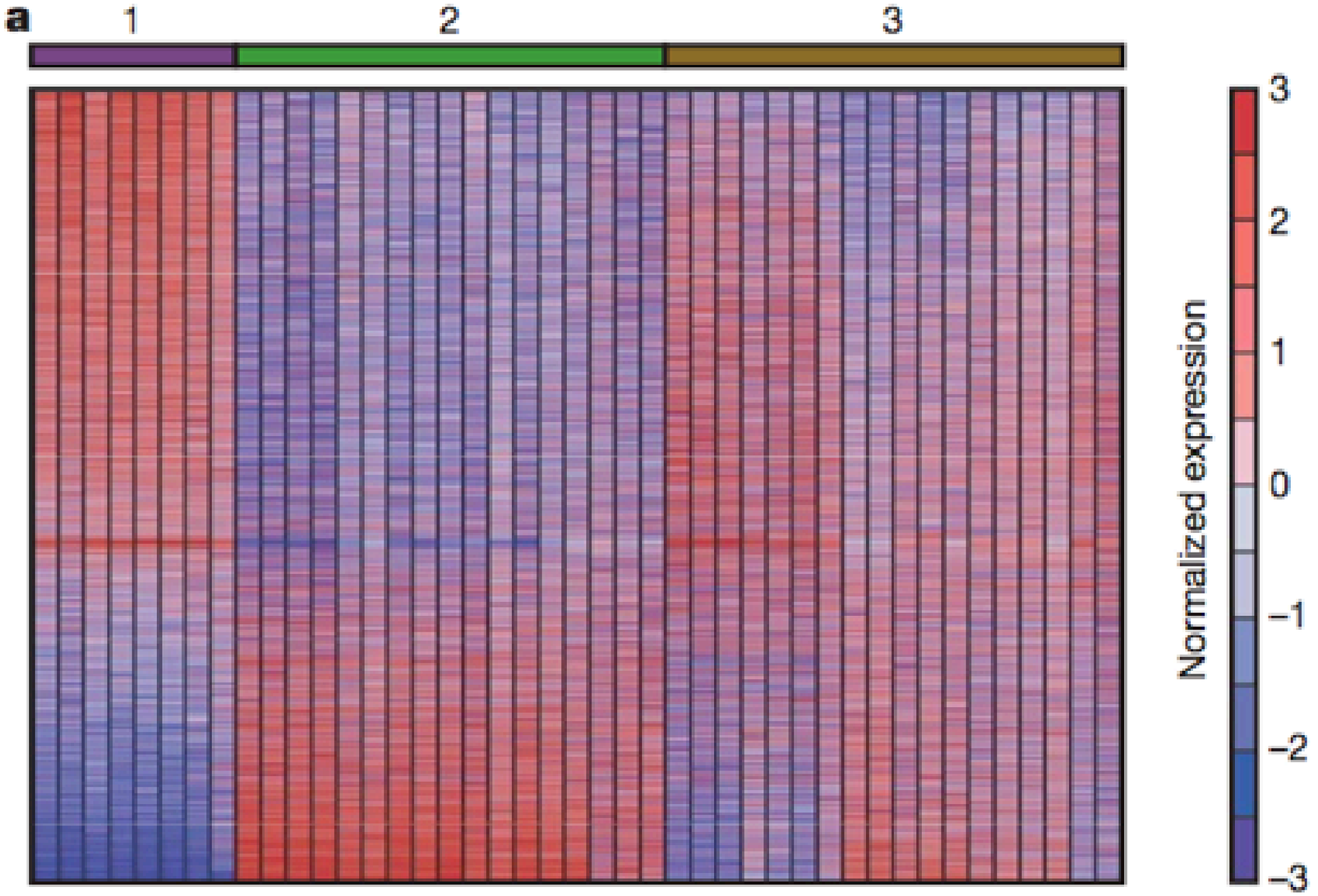


Molecular information improving detection, prognosis and treatment of disease

"Distinct physiological states of Plasmodium falciparum in malaria-infected patients" (Daily et al, Nature)

- In vitro analysis of malaria showed little difference in gene expression profiles
- BUT what about in the context of hosts who are different genetically (humans)
- Took 43 blood samples and analyzed gene expression
- Found significant differences over the sample, and these correlated to severity of disease and molecular response to malaria.

a



b

Variable	Cluster 1 (*n* = 8)	Cluster 2 (*n* = 17)	Cluster 3 (*n* = 18)
Age (years)	5 (2–15)	6 (3–6)	7 (4–15)
Parasitaemia (%)	2 (1–8)	4 (2–6)	3 (2–9)
Glucose (mg dl^{-1})	108 (94–113)	99 (88–121)	101 (54–13)
Haematocrit	39 (30–45)	33 (29–36)	30 (27–35)
Days ill	3 (3–3)	3 (2–3.8)	4 (3–6)*
Temperature (°C)	38 (37–40)	38 (37–39)	39 (38–40)*
IL-6 (pg ml^{-1})	53 (11–108)	29 (19–180)	190 (53–1,051)
IL-10 (pg ml^{-1})	561 (157–1,853)	216 (113–2,354)	4270 (1,349–8,088)*
TNF-α (pg ml^{-1})	69 (35–119)	52 (25–82)	67 (51–95)
TGF-α (pg ml^{-1})	29 (19–60)	33 (22–43)	87 (50–383)*
Tissue factor (pg ml^{-1})	169 (147–231)	181 (151–200)	345 (248–539)*
VCAM-1 (μg ml^{-1})	6 (3–8)	5 (3–7)	10 (6–18)*
Lymphotactin (pg ml^{-1})	197 (189–204)	200 (181–236)	265 (219–336)*

"Blood gene expression signatures predict exposure levels" (Bushel et al, PNAS)

- Can we detect exposure to toxins based on expression changes BEFORE frank signs of toxicity emerge
- Tested on acetominophen (APAP) exposure in rats, then humans
- Conclusion: the expression profile of liver cells shows signs of response well before clinical symptoms emerge. This may be model for many other exposures.

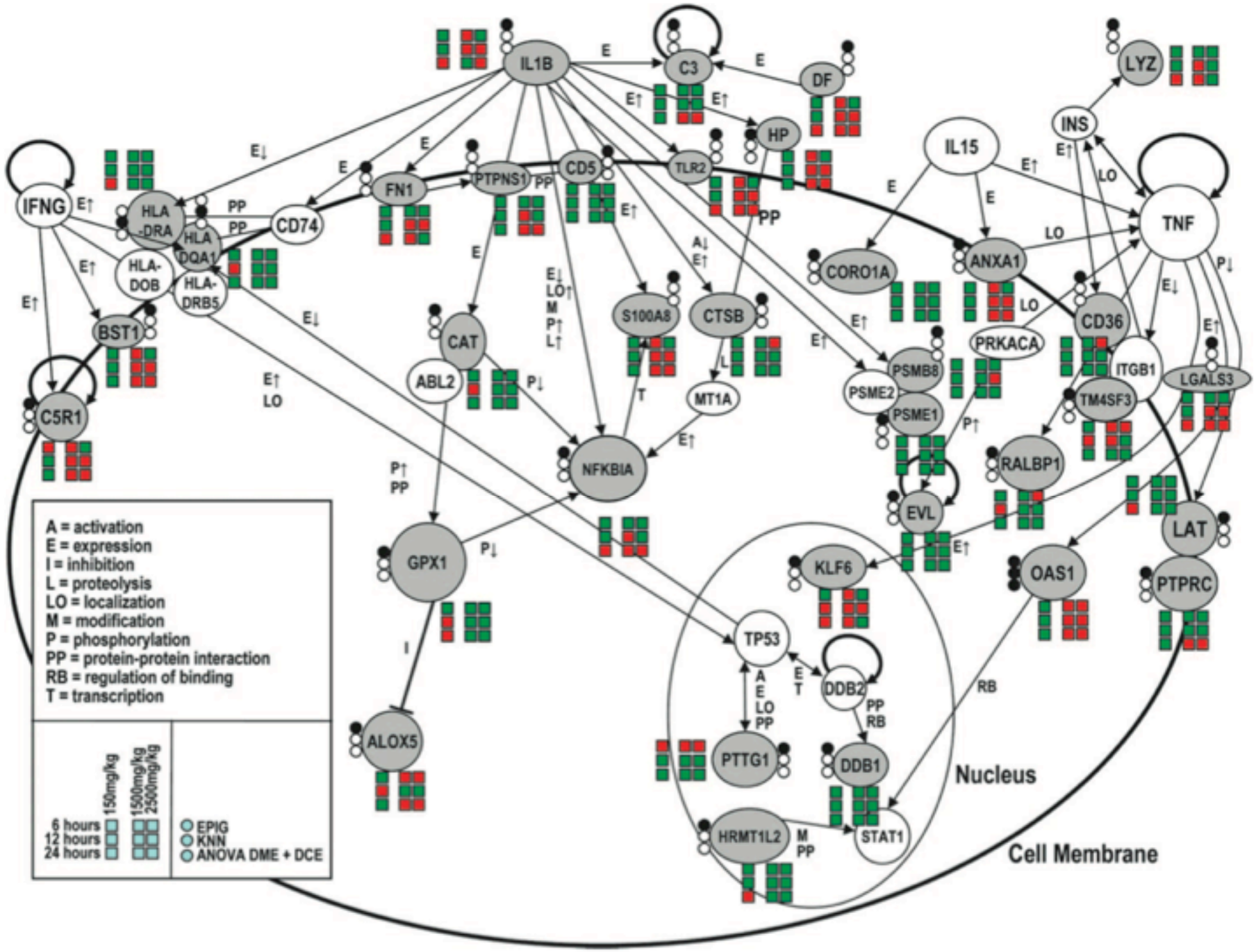


Fig. 2. Differentially expressed genes in blood that discriminate between exposure to subtoxic/nontoxic or toxic dose of APAP. Pictured is a subgroup of genes involved in immune response and inflammation. Gray filling of circles beside genes indicates identification of genes in the list of discriminatory genes (top circle, EPIG; middle circle, *k*-NN, bottom circle, ANOVA DME and DCE), and coloring of squares signals direction of change (red, up-regulation; green, down-regulation in comparison with mock-treated control).

"Social regulation of gene expression in human leukocytes" (Cole et al, Genome Biol)

- Can our social situation influence our gene expression? YES
- Compared gene expresion in 14 subjects who were subjectively socially isolated with those who were not
- 209 genes found to be differentially expressed
- Anti-inflammatory = DOWN, Pro-inflammatory = UP
- Major transcription control pathways showed changes...

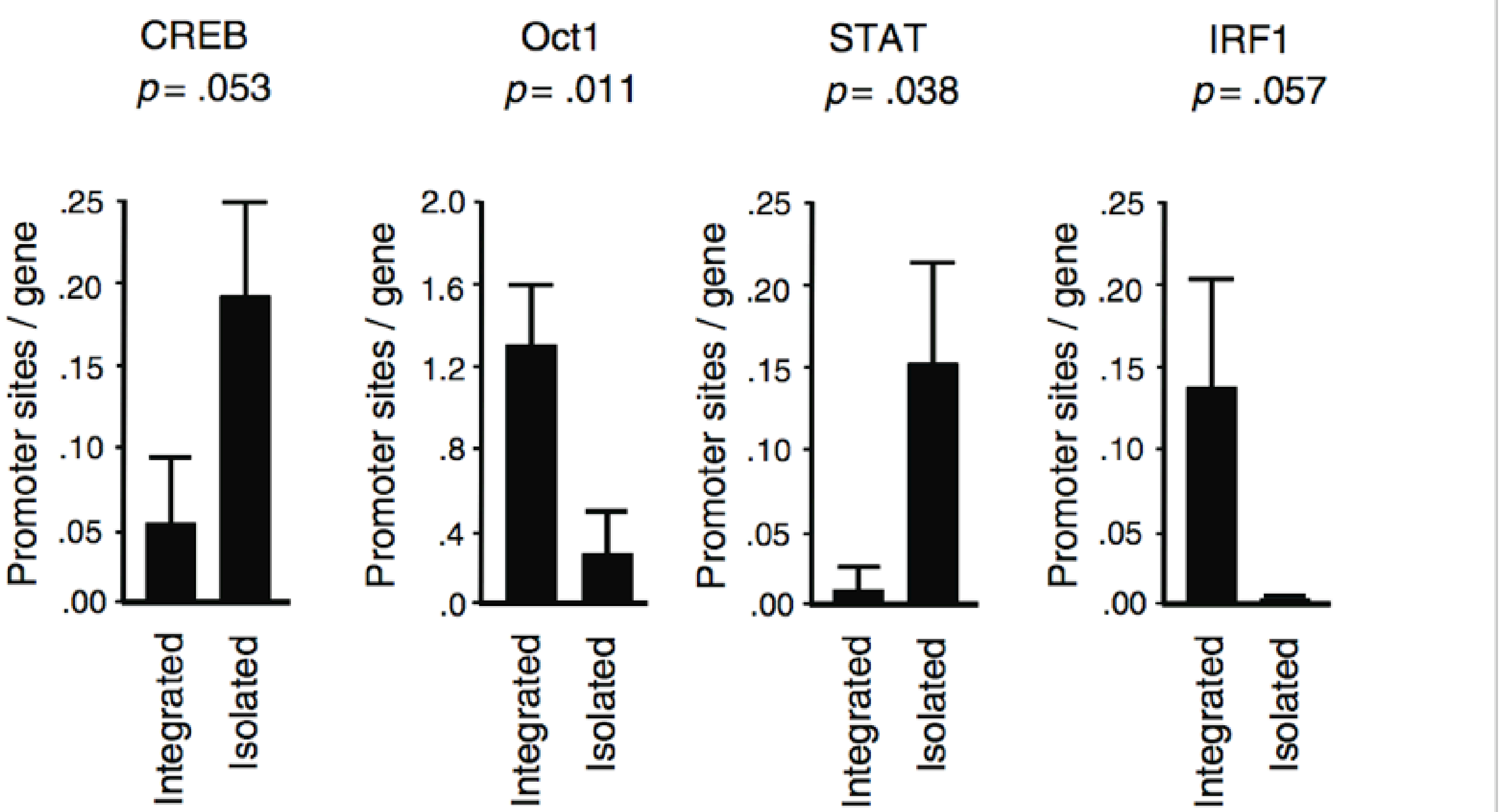


Figure 3
Transcription control pathways differentially active in high- versus low-lonely individuals. Data represent the mean (± standard error) prevalence of transcription factor-binding motifs in primary TELiS bioinformatics analysis of genes over-expressed in leukocytes from high- versus low-lonely individuals.

"Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting" (Dupuy & Simon, J Natl Cancer Inst)

- Do investigators make routine avoidable mistakes repeatedly in analysis of microarrays? YES
- Analyzed 90 studies, many in "good" journals, listed 40 (!) dos and don'ts...50% of studies did one of:

1. Inadequate control for multiple hypothesis testing
2. Spurious claim of cluster-outcome correlation
3. Biased estimate of performance based on x-validation

Great work skipped...

- Ultraconserved sequences across genomes
- Biomarket statistics
- miRNA promotors conserved in vertebrates
- Rapid RNA 3D structure determination
- Genome-wide analysis of cell-cycle transcription
- SNP and CNV impact
- Gene expression across ethnic groups
- Paired-end sequencing to identify major genome rearrangements
- Reconstructing regulatory networks
- Genes for regulating lifespan
- Modeling avian flu spread
- Histone methylation patterns in humans
- Positive selection in monkeys vs. humans

I can't resist

"Conversion of amino-acid sequence in proteins to classical music: search for auditory patterns" (Takahashi & Miller, Genome Biology)

- Map amino acids to 13 base notes
- Variations of chords distinguish similar amino acids
- Codon distribution for rhythm
- Help listeners find patterns: thyA and huntingtin

I can't resist

"A single IGF 1 allele is a major determinant of small size in dogs." (Sutter et al, Science)

- Domestic dog ~15,000 years old
- GWAS on Portugese water dog (size variation)
- IGF 1 = insulin-like growth factor
- Looked at SNPs, found two haplotypes (I and B)
- B have smaller size
- Validated on a wider section of dogs...


CATERS


© AP

I can't resist

"A mutation in the myostatin gene increases muscle mass and enhances racing performance in heterozygote dogs." (Mosher et al, PLOS Genetics)

- Two-base-pair delection in 3rd exon of MSTN associated with "double muscle" phenotype
- Associated with increased athletic performance
- 'performance-enhancing polymorphisms..."

A

+/+

B

mh/+

C

mh/mh

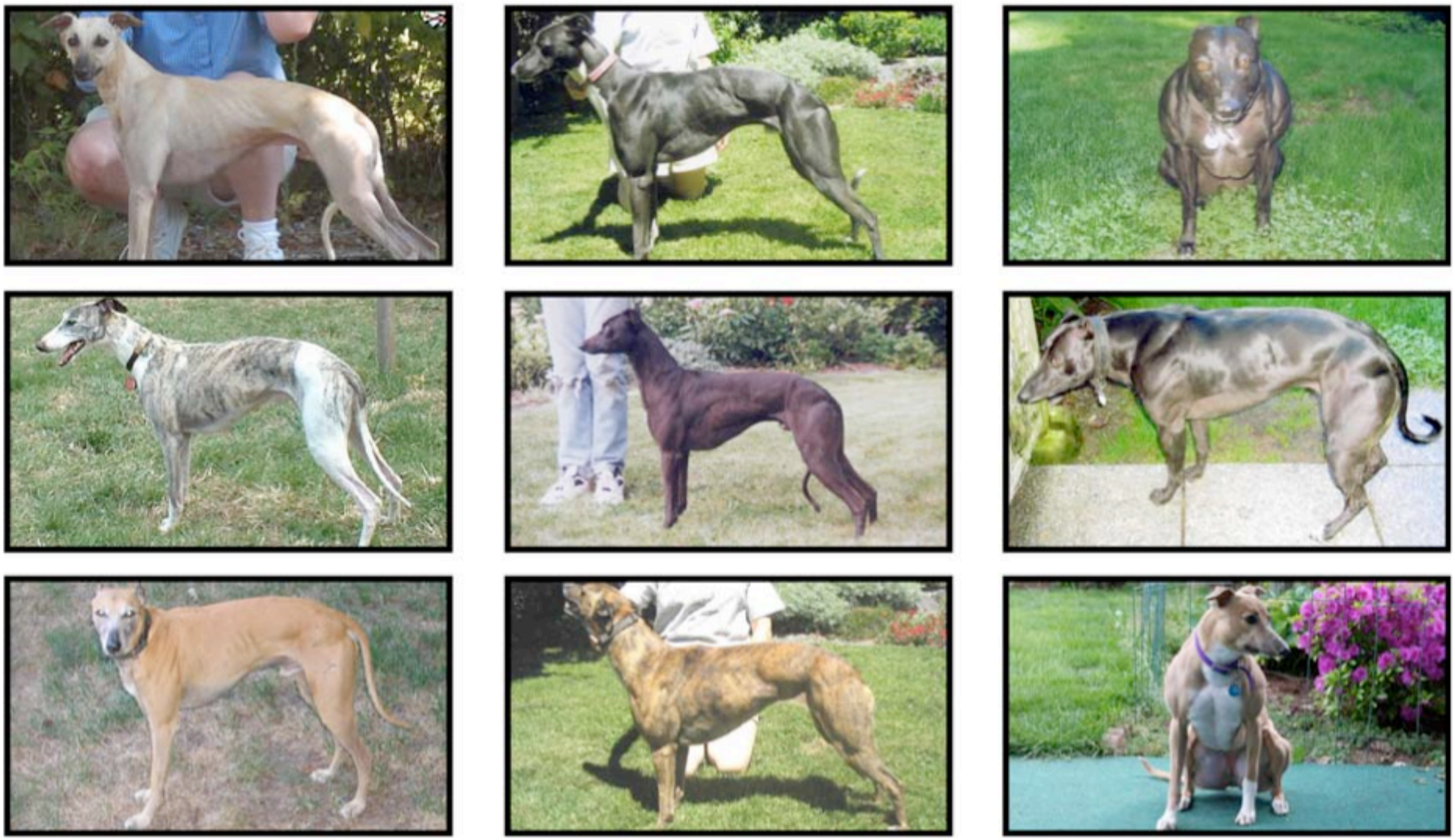


Figure 1. Comparison of Whippets with Each of the Three Potential Genotypes

(A) Dogs have two copies of the wild-type allele (+/+).
(B) Dogs are heterozygous with one wild-type allele and one mutant cys → stop allele (*mh/+*).
(C) Dogs are homozygous for the mutant allele with two copies of the cys → stop mutation *(mh/mh)*.
All photos represent unique individuals except for the top and middle panels in the righthand column.
doi:10.1371/journal.pgen.0030079.g001

I can't resist...

"Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, ASPM and Microcephalin" (Dediu & Ladd, PNAS)

- Most languages are unambiguously tonal or not (Chinese =YES, English = NO)
- Authors hypothesize a relationship between distribution of tonal languages and distribution of certain variants of ASPM and MCPH
- Population preferences for language may stem from their genetics.

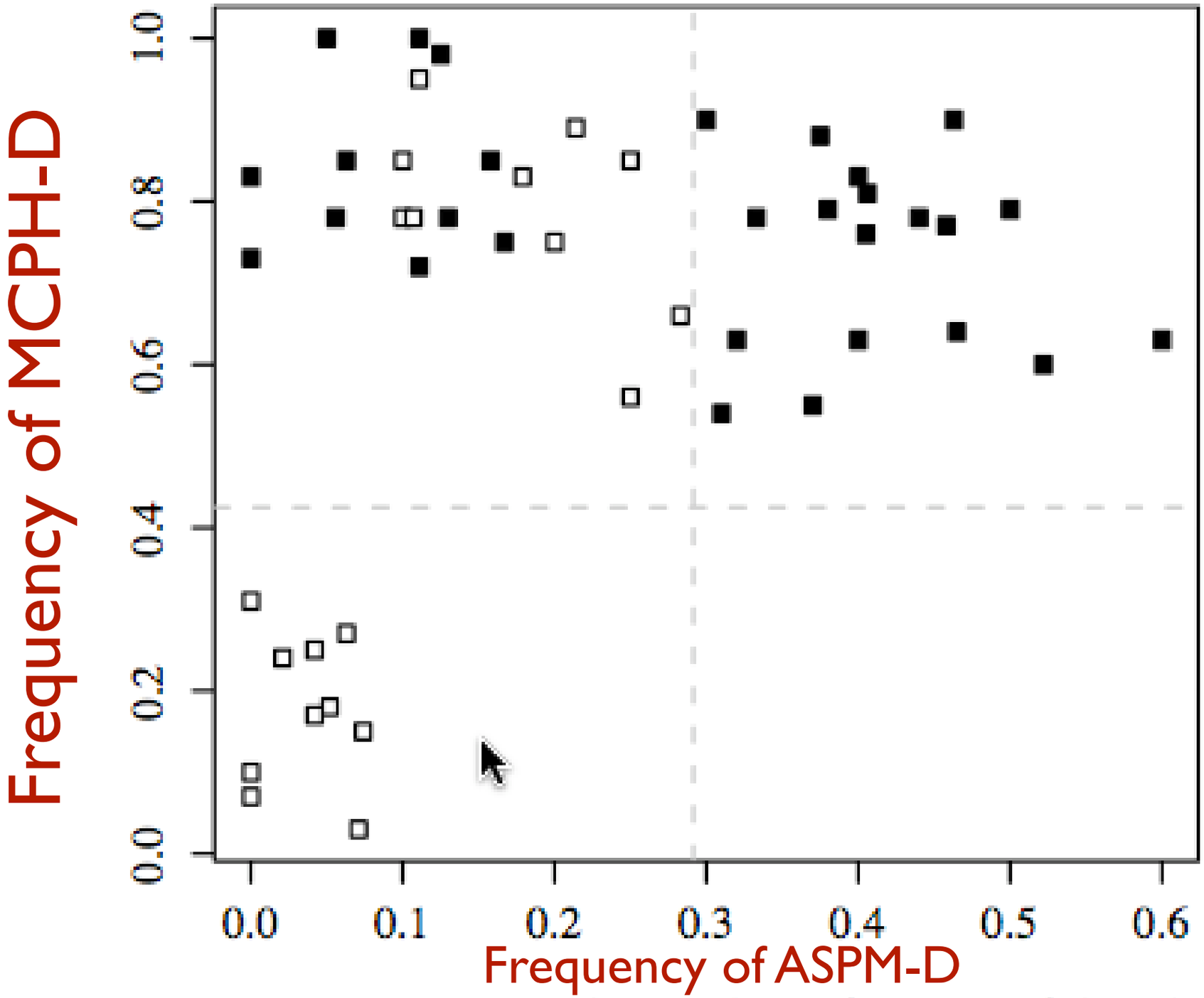


Fig. 1. Linguistic tone versus the population frequency of the adaptive haplogroups of *ASPM* and *Microcephalin*. The horizontal axis represents the frequency of *ASPM-D*, whereas the vertical axis represents the frequency of *MCPH-D*. Filled squares represent nontonal languages and open squares tonal languages. Gray dashed lines correspond to 0.292 *ASPM-D* and 0.425 *MCPH-D*. See last paragraph of *Results* for details.

Crystal ball...

- Sequencing makes a comeback (watch out microarrays....)
- Translational science projects will create astounding data sets (hopefully available) to catalyze research
- GWAS will continue to proliferate
- Consumer-oriented genetics will create demand for online resources for interpretation
- Difficult decisions about when/how to bring new molecular diagnostics to practice.

Thanks.

See you in 2009!

russ.altman@stanford.edu