

Systems Biology Zyklusvorlesung BSc

Computational analysis of omics datasets

Gordana Apic

Bioquant, Cell Networks
University of Heidelberg

What this is about

- Assuming you know something about high-throughput experiments in the life science (e.g. ‘omics’)
- When you run a typical omics platform, you generate some data.
- The next challenge is to understand what these data actually mean.
- There is an entire (quasi-) discipline related to this, essentially, Bioinformatics.



EXZELLENZCLUSTER

CellNetworks

What is Bioinformatics?

Bioinformatics

Bioinformatics  /baɪ.əʊtɪfərmætɪks/ is an interdisciplinary field that develops methods and software tools for understanding **biological** data. As an interdisciplinary field of science, bioinformatics combines computer science, statistics, mathematics, and engineering to analyze and interpret **biological** data.



Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **“informatics” techniques** (derived from disciplines such as applied math, CS, and statistics) to understand and **organize the information associated** with these molecules, **on a large-scale**.

(From Mark Gerstein, Yale)

What is Bioinformatics?

The term dates from 1979. However, it was not widely used until about 1995 when it became a buzzword.

Computational biologists have been around since the 1960s, but it wasn't until they were badly needed in the age of the genome before the term started to be used.

Essentially describes a group of people specialized in analyzing and searching large biological datasets.

The dawn of the information in biology...

CONFIGURATIONS OF POLYPEPTIDE CHAINS WITH FAVORED ORIENTATIONS AROUND SINGLE BONDS: TWO NEW PLEATED SHEETS

BY LINUS PAULING AND ROBERT B. COREY

GATES AND CRELLIN LABORATORIES OF CHEMISTRY,* CALIFORNIA INSTITUTE OF TECHNOLOGY, PASADENA, CALIFORNIA

Communicated September 4, 1951

In recent papers we have described several configurations of polypeptide chains with interatomic distances, bond angles, and other structural features as indicated by the studies in these Laboratories of the structure of crystals of amino acids, simple peptides, and related substances, and have presented evidence for their presence in synthetic polypeptides, fibrous proteins, and globular proteins.¹⁻⁹ The requirements that we have imposed for a satisfactory polypeptide configuration, in addition to the correct bond distances and bond angles, are that each amide group be plane with

Arch Biochem. 1949 Jul;22(3):475.

A method for the determination of amino acid sequence in peptides.

EDMAN P.

PMID: 18134557 [PubMed - indexed for MEDLINE]

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey¹. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical analysis at King's College, London. One of us (J. D. W.) has been aided by a fellowship from the National Foundation for Infantile Paralysis.

J. D. WATSON
F. H. C. CRICK
Medical Research Council Unit for the
Study of the Molecular Structure of
Biological Systems,
Cavendish Laboratory, Cambridge.
April 2.

¹ Pauling, L., and Corey, R. B., *Nature*, 171, 546 (1953); *Proc. U.S. Nat. Acad. Sci.*, 39, 84 (1953).

J Mol Biol. 1975 May 25;94(3):441-8.

A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.

Sanger F, Coulson AR.

PMID: 1100841 [PubMed - indexed for MEDLINE]

To the first whole genome sequence...

Nature. 1976 Apr 8;260(5551):500-7.

Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene.

Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, Min Jou W, Molemans F,
Raeymaekers A, Van den Berghe A, Volckaert G, Ysebaert M.

Abstract

Bacteriophage MS2 RNA is 3,569 nucleotides long. The nucleotide sequence has been established for the third and last gene, which codes for the replicase protein. A secondary structure model has also been proposed. Biological properties, such as ribosome binding and codon interactions can now be discussed on a molecular basis. As the sequences for the other regions of this RNA have been published already, the complete, primary chemical structure of a viral genome has now been established.

PMID: 1264203 [PubMed - indexed for MEDLINE]



EXZELLENZCLUSTER
CellNetworks

To the human genome...

eEnsembl BLAST/BLAT | BioMart | Tools | Downloads | More ▾

Human (GRCh38.p5) ▾

Search Human...



Human

Homo sapiens

Search all categories ▾ Search Human...

e.g. BRCA2 or 17:63973115-6447414 or rs1333049 or osteoarthritis

Genome assembly: GRCh38.p5
(GCA_000001405.20)



Other assemblies

GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.



Find A Gene

Search for from

The NCBI Handbook
 An online guide to the use of NCBI resources. Titles of selected chapters that refer to human genome resources are shown below.

...

Ensembl genomes

UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

move <<< << < > >> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr9:133,252,000-133,280,861 28,862 bp enter position, gene symbol or search terms

scale: chr9 (q34.2) 0.001 10 KB 100 KB 1 MB 10 MB 100 MB 1 GB 10 GB 100 GB 1000 GB

1000 133,255,400 133,276,400 133,278,400 133,279,400 133,280,400 133,281,400

v24 Comprehensive Transcript Set (only Basic displayed by default)

RefSeq Genes OHM Allelic Variant SNPs

Human mRNAs from GenBank mRNA Often Found Near Regulatory Elements on 7 cell lines from ENCODE

se 1 Hypersensitivity Peak Clusters from ENCODE (95 cell types)

100 vertebrates Basewise Conservation by Phylogeny

Multiz Alignments of 100 Vertebrates

ie Nucleotide Polymorphisms (dbSNP 146) Found in > 1% of Samples

Repeating Elements by RepeatMasker

for details. Click or drag in the base position Click side bars for track options. Drag side bars or to reorder tracks. Drag tracks left or right to new tracks below and press refresh to alter tracks displayed

move end < 2.0 >

add custom tracks track hubs configure multi-region reverse resize refresh

NCBI genomes

And beyond... thousands of genomes in many species (e.g. human)



[Home](#) | [About us](#) | [100,000 Genomes Project](#) | [GeCIP](#) | [GENE Con:](#)

[Home](#) > [The 100,000 Genomes Project](#)

The 100,000 Genomes Project

The project will sequence 100,000 genomes from around 70,000 people. Participants are NHS patients with a rare disease, plus their families, and patients with cancer.

The aim is to create a new genomic medicine service for the NHS – transforming the way people are cared for. Patients may be offered a diagnosis where there wasn't one before. In time, there is the potential of new and more effective treatments.

wellcome trust

[Our vision](#) | **Funding** | [Managing a grant](#) | [Education resources](#) | [News](#) | [In](#)
[Sustaining health](#) | **Biomedical science** | [Innovations](#) | [Public engagement](#) | [Medical humanities](#) | [Society and ethics](#)

► [Funding schemes](#)

► [Funded projects](#)

- Major initiatives
- Awards made

The 1000 Genomes Project

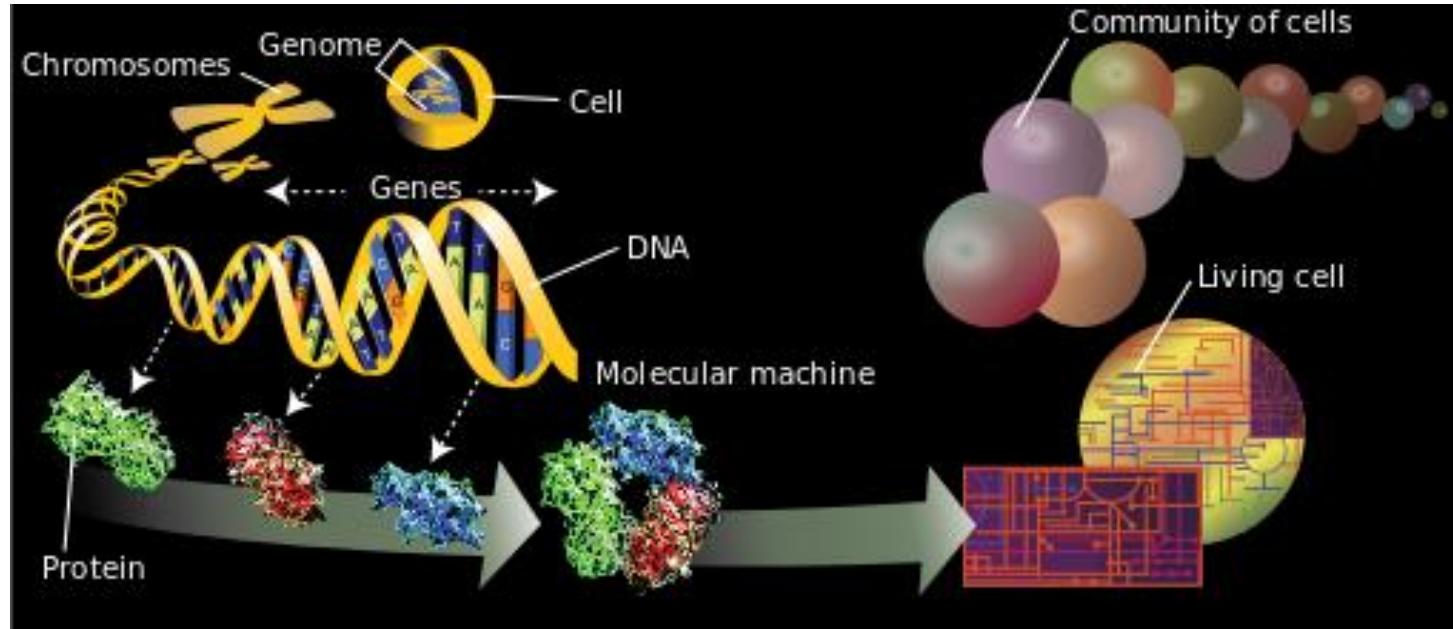


The 1000 Genomes Project is an ambitious effort to sequence the genomes of at least 1000 people to create the most detailed and medically useful catalogue to date of human genetic variation.

1000 Genomes

A Deep Catalog of Human Genetic Variation

Post-genome experiments mean lots of data related to gene or protein function



Transcriptomics – gene expression

Proteomics – protein expression, modification, interaction

Metabolomics – small molecule abundance, flux

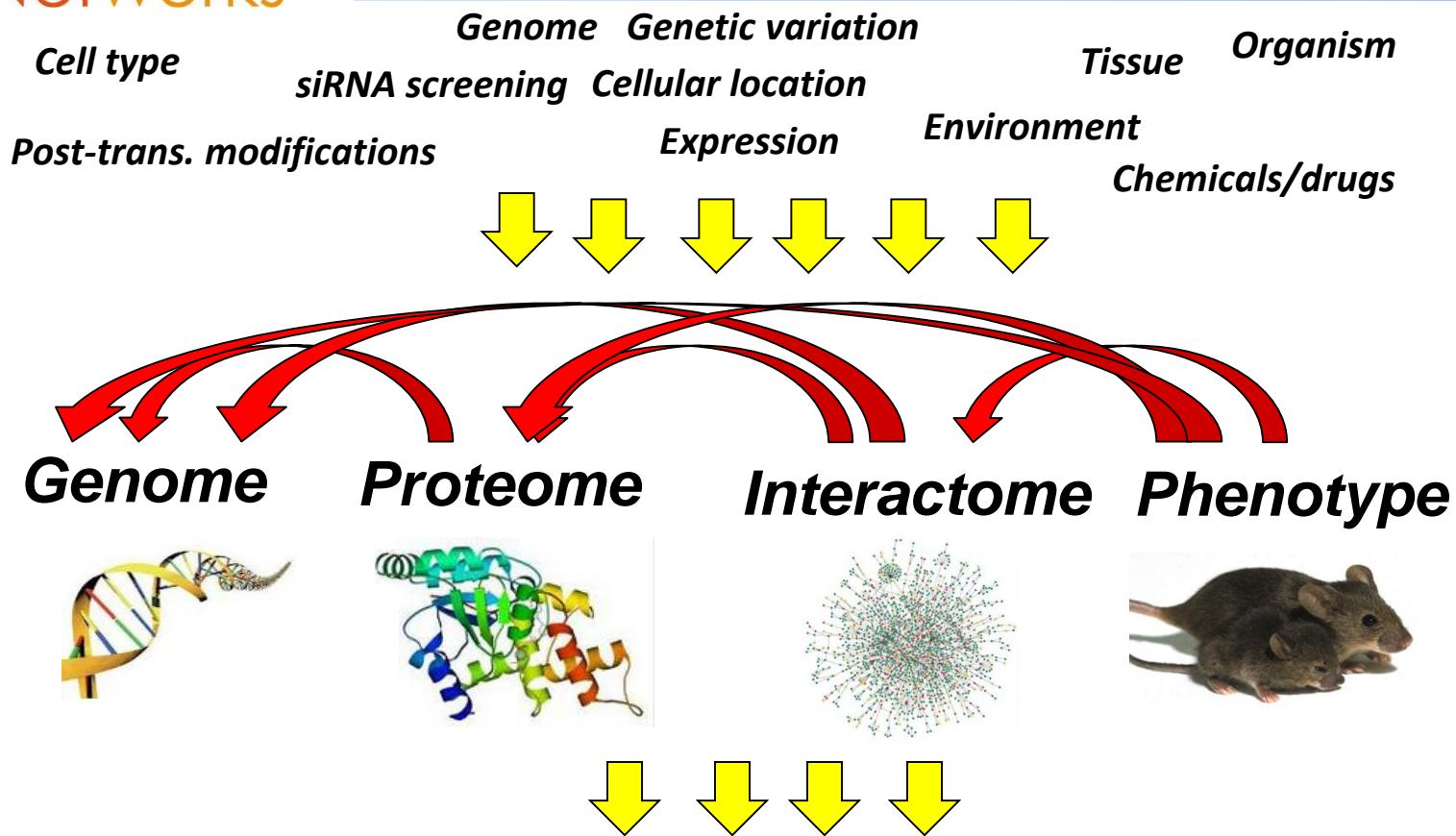
Epigenomics – DNA modifications

High content screening – images & phenotypic consequences

siRNA or CRISPR/Cas – gene silencing (coupled to phenotype)

Etc.

Complex goal of turning multiple observations into meaningful biology

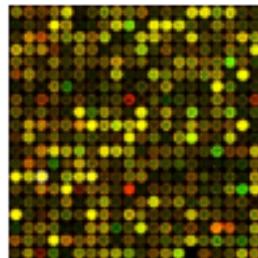


Much of what we need to do first is to study “gene lists” (or “protein lists”)

My screen worked, and identified 326 gene hits.

Now what?

What does this mean in the context of my biological question?



Ranking or clustering



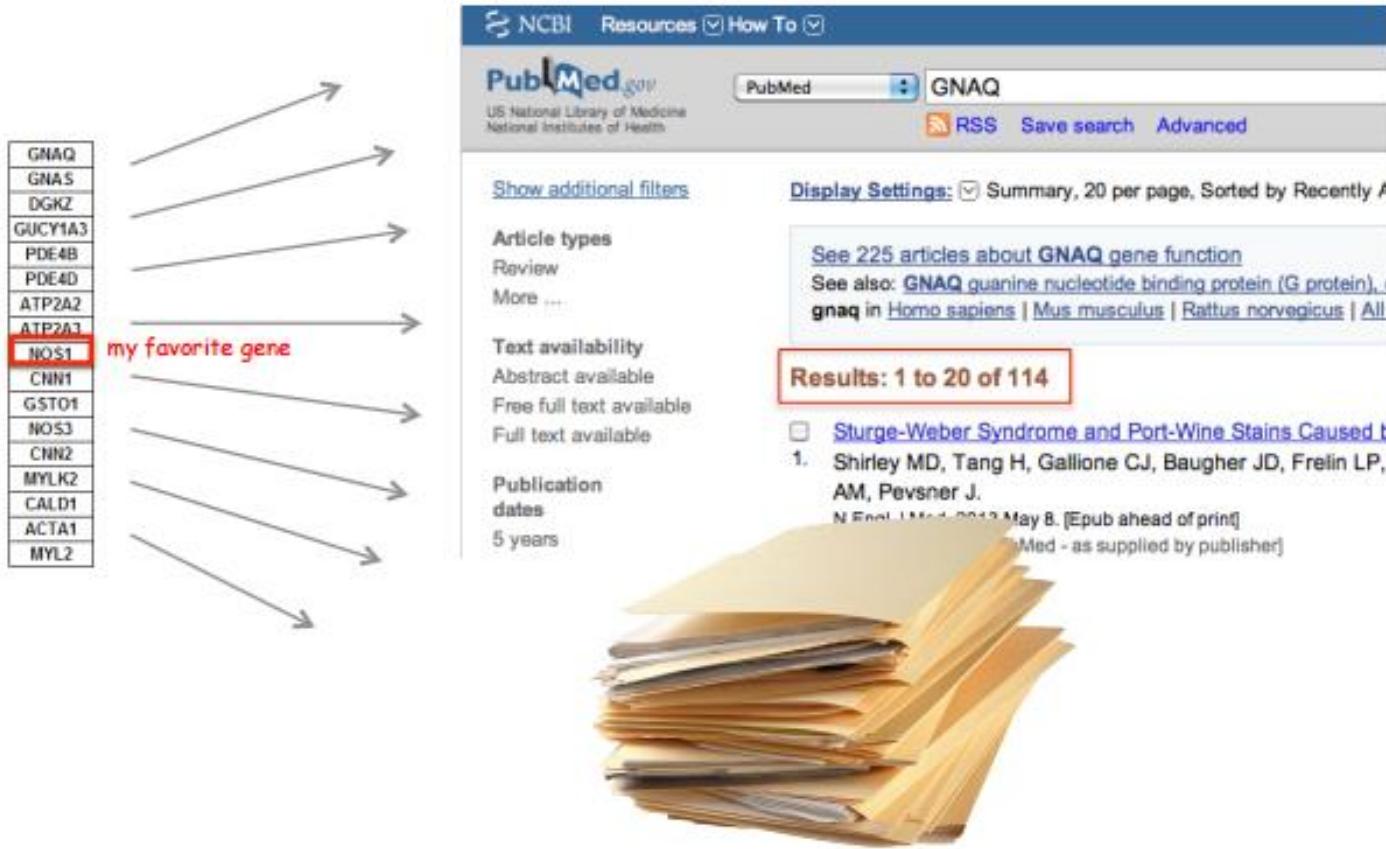
GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2



?

Courtesy Gary Bader, University of Toronto

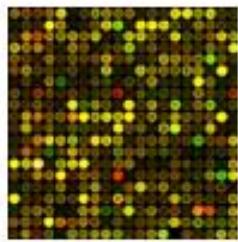
The slow, traditional (and frankly) stupid way...



A list of genes on the left is shown with arrows pointing to a PubMed search results page on the right. The genes listed are: GNAQ, GNAS, DGKZ, GUCY1A3, PDE4B, PDE4D, ATP2A2, ATP2A3, NOS1 (highlighted in red), CNN1, GSTO1, NOS3, CNN2, MYLK2, CALD1, ACTA1, and MYL2. The PubMed search results page shows a search for "GNAQ". The results section displays 1 to 20 of 114 articles. The first result is: "Sturge-Weber Syndrome and Port-Wine Stains Caused by Gain-of-Function in GNAQ" by Shirley MD, Tang H, Gallione CJ, Baugher JD, Frelin LP, AM, Pevsner J. The image also features a stack of papers at the bottom.

Courtesy Gary Bader, University of Toronto

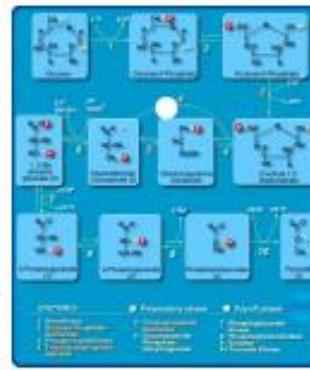
Fortunately there are a variety of tools to study gene lists



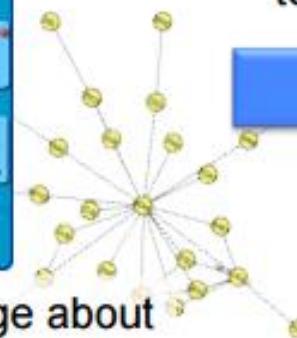
Ranking or clustering



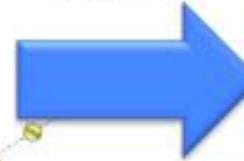
GNAQ
GNA5
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2



Prior knowledge about cellular processes

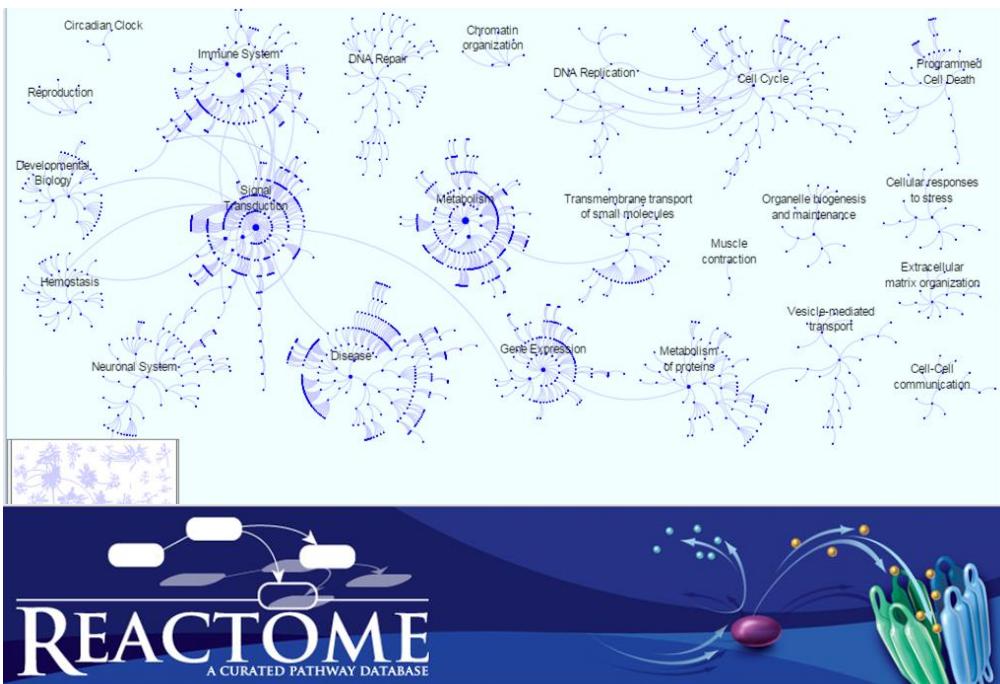


Analysis tools



Great new insight

Databases of pathways



reactome.org

Russell Group, Protein Evolution

KEGG Home
Release notes
Current statistics
Plea from KEGG

KEGG Database
KEGG overview
Searching KEGG
KEGG mapping
Color codes

KEGG Objects
Pathway maps
Brite hierarchies

KEGG Software
KegTools
KEGG API
KGML

KEGG FTP
Subscription

GenomeNet
DBGET/LinkDB
Feedback

Kanehisa Labs

KEGG Overview

1. Genomes to Biological System

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from genomic and molecular-level information. It is a computer representation of the biological system, consisting of molecular building blocks of genes and proteins (genomic information) and chemical substances (chemical information) that are integrated with the knowledge on molecular wiring diagrams of interaction, reaction and relation networks (systems information). It also contains disease and drug information (health information) as perturbations to the biological system.

The KEGG database has been in development by Kanehisa Laboratories since 1995, and is now a prominent reference knowledge base for integration and interpretation of large-scale molecular data sets generated by genome sequencing and other high-throughput experimental technologies.

www.genome.jp/kegg/

Gordana Apic



EXZELLENZCLUSTER

CellNetworks

Databases of gene function/ontology

 Gene Ontology Consortium

Home Documentation Downloads Community Tools About Contact us

Enrichment analysis

Your gene IDs here...

biological process ▾
Homo sapiens ▾

Submit

[Advanced options / Help](#)
Powered by PANTHER

Statistics



Other GOC tools

Explore other GOC tools in the AmiGO software suite.

Gene Ontology Consortium

Search GO data

Search for terms and gene products...

Search

Ontology

[Filter classes](#)
[Download ontology](#)

Gene Ontology: the framework for the model of biology. The GO defines concepts/classes used to describe gene function, and relationships between these concepts. It classifies functions along three aspects:

- molecular function**
molecular activities of gene products
- cellular component**
where gene products are active
- biological process**
pathways and larger processes made up of the activities of multiple genes

Annotations

[Download annotations \(standard files\)](#)
[Filter and download \(customizable files <10k lines\)](#)

GO annotations: the model of biology. Annotations are statements describing the functions of specific genes, using concepts in the Gene Ontology. The simplest and most common annotation links one gene to one function, e.g. FZD4 + Wnt signaling pathway. Each statement is based on a specified piece of evidence.
[more](#)

The mission of the GO Consortium is to develop an up-to-date, comprehensive, computational model of biological systems, from the molecular level to larger pathways, cellular and organism-level systems. [more](#)

Search documentation



User stories

Explore documentation related to your personal user story.

What is the Gene Ontology?

- An introduction to the Gene Ontology
- What are annotations?
- Ten quick tips for using the Gene Ontology **Important**
- Enrichment analysis
- Downloads

geneontology.org

Challenge to find a significant overlap or enrichment in a gene-list. Just counting probably not good enough

My list

List of (e.g.) 10,299 GO Biological Process terms linked to genes

TP53
RHOA
KRAS
HRAS
CTNNB1
ABL1
ABL2
OR1A1
AKT2

GO:0007596 blood coagulation

ACTN4 ACTB ATP2A1 SCG3 CD44 IFNA4 PRKCE GNAT3 DGKA KCNMA1 ITGAM PRCP L1CAM ITGA5 PRKAR1B DOCK1 PLG VCL PDE9A ADRA2C SLC7A7
LCP2 SERPINE1 ARRB2 KDM1A GP6 MPL SLC7A6 IGF1 MAPK14 ITPK1 SOS1 SERPINB2 TGFB2 IFNA14 LMAN1 PLAU ZFPM2 RAC1 PRKAR2B NFE2 GP1BA
TP53 TUBA4A AMICA1 P2RY1 GNA14 CSK MAFG TLN1 RAP1A GUCY1A3 H3F3A ENTPD1 NOS2 PRKCG NBEAL2 PROZ DTNBP1 LCK TTN KLC1 C1QBP RAF1
AKAP10 HRG KIF26A LYST PECAM1 RCOR1 ENPP4 KIF3A DOCK9 PTPN1 IFNA1 HMG20B EGF ATP2B2 FGR GNA13 DGKZ F2RL2 EHD1 PRKCZ CEACAM6 FYN
RAD51B PDE2A CENPE IRF1 DAGLA PLA2G4A SPN CDK5 ITGB1 IFNA7 CD244 DGKE LAMP2 VEGFC LEFTY2 F13A1 AK3 MGLL GATA3 CD59 HPS1 CRK BRPF3
PRKCQ PLCG2 KIF4A PDGF CAPZA1 SELP KIF2B PLXADL ITGA2B F5 ITGAV JAM3 PSAP GNAI2 SHC1 ESAM ITGA3
PRKAR1A LYN TGF3B SH2B1 SHH DGKG THPO GATA4 PRKCH PRKCB PAPSS2 PIK3R1 JAK2 IFNA17 NOS1 SIRPG A2M PIK3R5
RAB5A PDPK1 PTPRK MAPK1 HIST2H3A IGF2 KVORK C2 COL1A1 SLC8A1 HPS1 GP1BB RAP1B PRKG1 DGKI SLC8A2
RASGRP1 AKAP1 PTPN11 RAB27A ADRA2B ITPR1 KIF5A SLC7A2 NFKB1 KLC2 DGKH HDAC1 OLR1 DAGLB PIK3CG SELL PLAT EHD2 GNAS
PLCG1 HBD RAD51C BCAR1 CD2 GUCY1A2 GNA15 GNA1 C4BYP KCNMB2 PFN1 DGKH HPS5 PRKAR2A IFNA6 P2RY12 CEACAM8 SLC16A8 ITGB3 GP9
POTEKP F2RL3 CDC42 VWF CYP4F2 GP5 SOD1 CAPZA2 GNB1 IFNA2 ITGA6 CABLES1 PIK3CA MAG ATP1B2 DOCK8 GATA6 ANGPT4 PF4 GRB7 WAS FCER1G
BLOC1S6 SIRPA WDR1 ATP1B1 FIGF KIFAP3 F12 F11R ATP2B3 GRB2 ARR8B1 LNP MMRN1 FGA SYK SIN3A F2RL1 SPARC CBX5 PTK2 CALU LRRC16A LAT
IFNA5 DGKQ FGG PDGFA HSPA5 ALB RBSN CEACAM1 CFL1 KIF22 SLC8A3 GUCY1B2 SELPLG SERPINC1 F3 LRP8 SERPING1 RHOB KIF11 GNAQ SLC7A9
CD9 TREM1 SRGN PPIA F13B **HRAS** KCNMB3 APOA1 ACTN1 MAFK SH2B2 CD36 PRKACB GNG2 CLU RACGAP1 PDE3A PRKG2 PABPC4 PIK3CB PDE1B ITGB2
IFNB1 IFNA10 PIK3R6 PROS1 KCNMB4 TF VEGFA HIST1H3A KIF9 ADRA2A CD63 SLC7A11 GNA12 PDE5A TGFB1 PPIL2 TMSB4X VAV2 F7 ITGA2 ADORA2A
PRTN3 HBE1 EHD3 STIM1 SLC7A10 F9 ITPR2 F10 EFEMP2 GAS6 TFP12 MFN2 F2R CAV1 THBS1 CYP4F11 GNA11 GRB14 RAPGEF3 PTGIR **ABL1** PIK3R2
ATP2A2 MAPK3 APBB1IP SERPINF2 RASGRP2 PPBP SLC16A3 TIMP1 JAM2 APOB CALM1 ATP1B3 HDAC2 FGB PLAUR KIF3C ITGA1 YES1 SLC3A2 ALDOA
KRAS GATA5 HABP4 G6B FLNA F8 TRPC6 GNA13 MERTK COL1A2 MAFF SERPIN5 SERPIND1 VAV3 RAPGEF4 HNF4A GLG1 TBXA2R SH2B3 STXBP3 HPS4
KIF2A **RHOA** ANO6 KIFC1 PROCR ACTN2 SLC16A1 PRKCA SLC7A8 PRKCD TRPC7 HGF MMP1 SERPINA1 GGCX KIF18A IFNA21 PRKACG SERPINA10 CFD
PDE1A TRPC3 STX4 ATP2B4 NOS3 KIF15 IFNA16 ITGAL CAPZB VPS45 HBG2 ANXA5 VAV1 DOCK11 KIF3B ANGPT1 NRAS APP PTPN6 ATP2B1 GATA4 P2RX1
PDE10A KIF4B ATP2A3 AP3B1 CD58 CD48 PHF21A VEGFB WEE1 RHOG HBB SRC KNG1 ITGAX SELE CDK2 GATA1 IFNA8 FN1 BSG JMJD1C ABCC4 IRF2
CPB2 HBG1 DGKD ANGPT2 PDE3B PROC PDE11A ITPR3 DOCK6 DGKK INPP5D GUCY1B3 KIF23 THBD YWHAZ TFPI KLKB1 SCUBE1 F11 MYB ITGA10 ITGA4

GO:1901994 negative regulation of meiotic cell cycle phase transition

OVOL1

GO:0043508 negative regulation of JUN kinase activity

DNAJA1 DUSP10 MAPK8IP1 DUSP19 GSTP1 HIPK3 TP73 SERPINB3 PDCD4 ZNF675 AIDA SFRP1 SFRP5 SFRP2

GO:0032287 peripheral nervous system myelin maintenance

AKT1 SOD1 MPP5 FA2H **AKT2** NDRG1 SH3TC2

GO:0060441 epithelial tube branching involved in lung morphogenesis

FOXF1 FOXA2 SOX2 BMP4 LAMA1 RSP02 NKX2-1 NRAS SOX9 **KRAS HRAS** HOXA5 **CTNNB1** DAG1 HHIP FOXA1 DLG5 CTSZ

Overlap of 3

Statistics of gene enrichment

- The simplest way to identify what might be happening in a large gene set is to (say) look for the one process or pathway (or whatever) that has the largest overlap (in terms of number of genes)
- This is over-simplistic as it does not take into account wildly sizes of classes (e.g. “transcription” has roughly 3000 genes, “metabolism of lactate” has fewer than 10).
- Various statistical models exist that try to account for this (e.g. Hypergeometric distribution, Binomial distribution, etc.). Without going into detail, these normally compute:
 - The *expected* overlap (given the number of genes you had in total, the total number of genes in each set, the total number of genes in the genome, the total number of classes, etc.)
 - The enrichment, which is the ratio of the *observed* number and the *expected* number
 - Then a statistical measure, such as a p-value, that allows one to identify *significant* matches and ignore those that (say) are arising only because of large classes (e.g. transcription).

Statistics of gene enrichment, e.g.

GO:0007596 blood coagulation

overlap 5 in 337 genes, expected is $(337/20,202) \times 9 = 0.15$, ratio = $5/0.15 = 33.3$

ACTN4 ACTB ATP2A1 SCG3 CD44 IFNA4 PRKCE GNAT3 DGKA KCNMA1 ITGAM PRCP L1CAM ITGA5 PRKAR1B DOCK1 PLG VCL PDE9A ADRA2C SLC7A7 LCP2 SERPINE1 ARRB2 KDM1A GP6 MPL SLC7A6 IGF1 MAPK14 ITPK1 SOS1 SERPINB2 TGFB2 IFNA14 LMAN1 PLAU ZFPM2 RAC1 PRKAR2B NFE2 GP1BA **TP53** TUBA4A AMICA1 P2RY1 GNA14 CSK MAFG TLN1 RAP1A GUCY1A3 H3F3A ENTPD1 NOS2 PRKCG NBEAL2 PROZ DTNBP1 LCK TTN KLC1 C1QBP RAF1 AKAP10 HRG KIF26A LYST PECAM1 RCOR1 ENPP4 KIF3A DOCK9 PTPN1 IFNA1 HMG20B EGF ATP2B2 FGR GNAI3 DGKZ F2RL2 EHD1 PRKCZ CEACAM6 FYN RAD51B PDE2A CENPE IRF1 DAGLA PLA2G4A SPN CDK5 ITGB1 IFNA7 CD244 DGKE LAMP2 VEGFC LEFTY2 F13A1 AK3 MGLL GATA3 CD59 HPS1 CRK BRPF3 PRKCQ PLCG2 KIF4A PDGFB CAPZA1 SELP KIF2B PLEK CD84 KCNMB1 ANXA8 ZFPM1 CXADR ITGA2B F5 ITGAV JAM3 PSAP GNAI2 SHC1 ESAM ITGA3 PRKAR1A LYN TGFB3 SH2B1 SHH DGKG THPO GATA2 DOK2 AKT1 SERPINE2 CD177 PRKCH PRKCB PAPSS2 PIK3R1 JAK2 IFNA17 NOS1 SIRPG A2M PIK3R5 RAB5A PDPK1 PTPRJ MAPK1 HIST2H3A IGF2 VKORC1 TEK ORA1 CD47 PRKACA F2 RAC2 COL1A1 SLC8A1 HPS6 GP1BB RAP1B PRKG1 DGKI SLC8A2 RASGRP1 AKAP1 PTPN11 RAB27A ADRA2B ITPR1 KIF5A SLC7A5 CAP1 KIF2C MRV11 KLC2 DGKH HDAC1 OLR1 DAGLB PIK3CG SELL PLAT EHD2 GNAS PLCG1 HBD RAD51C BCAR1 CD2 GUCY1A2 GNA15 GNA11 C4BPB KCNMB2 PFN1 DGKB HPS5 PRKAR2A IFNA6 P2RY12 CEACAM8 SLC16A8 ITGB3 GP9 POTEKP F2RL3 CDC42 VWF CYP4F2 GP5 SOD1 CAPZA2 GNB1 IFNA2 ITGA6 CABLES1 PIK3CA MAG ATP1B2 DOCK8 GATA6 ANGPT4 PF4 GRB7 WAS FCER1G BLOC1S6 SIRPA WDR1 ATP1B1 FIGF KIFAP3 F12 F11R ATP2B3 GRB2 ARR8 LNP MMRN1 FGA SYK SIN3A F2RL1 SPARC CBX5 PTK2 CALU LRRC16A LAT IFNA5 DGKQ FGG PDGFA HSPA5 ALB RBSN CEACAM1 CFL1 KIF22 SLC8A3 GUCY1B2 SELPLG SERPINC1 F3 LRP8 SERPING1 RHOB KIF11 GNAQ SLC7A9 CD9 TREM1 SRGN PPIA F13B **HRAS** KCNMB3 APOA1 ACTN1 MAFK SH2B2 CD36 PRKACB GNG2 CLU RACGAP1 PDE3A PRKG2 PABPC4 PIK3CB PDE1B ITGB2 IFNB1 IFNA10 PIK3R6 PROS1 KCNMB4 TF VEGFA HIST1H3A KIF9 ADRA2A CD63 SLC7A11 GNA12 PDE5A TGFB1 PPIL2 TMSB4X VAV2 F7 ITGA2 ADORA2A PRTN3 HBE1 EHD3 STIM1 SLC7A10 F9 ITPR2 F10 EFEMP2 GAS6 TFPI2 MFN2 F2R CAV1 THBS1 CYP4F11 GNA11 GRB14 RAPGEF3 PTGIR **ABL1** PIK3R2 ATP2A2 MAPK3 APBB1IP SERPINF2 RASGRP2 PPBP SLC16A3 TIMP1 JAM2 APOB CALM1 ATP1B3 HDAC2 FGB PLAUR KIF3C ITGA1 YES1 SLC3A2 ALDOA **KRAS** GATA5 HABP4 G6B FLNA F8 TRPC6 GNA13 MERTK COL1A2 MAFF SERPINA5 SERPIND1 VAV3 RAPGEF4 HNF4A GLG1 TBXA2R SH2B3 STXBP3 HPS4 KIF2A **RHOA** ANO6 KIFC1 PROCR ACTN2 SLC16A1 PRKCA SLC7A8 PRKCD TRPC7 HGF MMP1 SERPINA1 GGCX KIF18A IFNA21 PRKACG SERPINA10 CFD PDE1A TRPC3 STX4 ATP2B4 NOS3 KIF15 IFNA16 ITGAL CAPZB VPS45 HBG2 ANXA5 VAV1 DOCK11 KIF3B ANGPT1 NRAS APP PTPN6 ATP2B1 GATA4 P2RX1 PDE10A KIF4B ATP2A3 AP3B1 CD58 CD48 PHF21A VEGFB WEE1 RHOG HBB SRC KNG1 ITGAX SELE CDK2 GATA1 IFNA8 FN1 BSG JMJD1C ABCC4 IRF2 CPB2 HBG1 DGKD ANGPT2 PDE3B PROC PDE11A ITPR3 DOCK6 DGKK INPP5D GUCY1B3 KIF23 THBD YWHAZ TFPI KLKB1 SCUBE1 F11 MYB ITGA10 ITGA4

GO:0060441 epithelial tube branching involved in lung morphogenesis

overlap 4 in 18 genes, expected is $(18/20,202) \times 9 = 0.008$, ratio = $4/0.008 = 500$

FOXF1 FOXA2 SOX2 BMP4 LAMA1 RSPO2 NKX2-1 NRAS SOX9 **KRAS HRAS HOXA5 CTNNB1** DAG1 HHIP FOXA1 DLG5 CTSZ

Most significant

Tools to compute gene enrichment

 Gene Ontology Consortium

Home Documentation Downloads Community Tools About Contact us

Enrichment analysis

Your gene IDs here...

biological process
Homo sapiens
Submit
Advanced options / Help
Powered by PANTHER

Gene Ontology Consortium

Search GO data

Search for terms and gene products...
Search

Ontology
Filter classes
Download ontology

Annotations
Download annotations (standard files)
Filter and download (customizable files <10k lines)

Gene Ontology: the framework for the model of biology. The GO defines concepts/classes used to describe gene function, and relationships between these concepts. It classifies functions along three aspects:
molecular function
molecular activities of gene products
cellular component
where gene products are active
biological process
pathways and larger processes made up of the activities of multiple genes

GO annotations: the model of biology. Annotations are statements describing the functions of specific genes, using concepts in the Gene Ontology. The simplest and most common annotation links one gene to one function, e.g. FZD4 + Wnt signaling pathway. Each statement is based on a specified piece of evidence.
more

Search documentation

Search

User stories

Explore documentation related to your personal user story.

What is the Gene Ontology?

- An introduction to the Gene Ontology
- What are annotations?
- Ten quick tips for using the Gene Ontology **Important**
- Enrichment analysis
- Downloads

geneontology.org

Russell Group, Protein Evolution



DAVID Bioinformatics Resources 6.7
National Institute of Allergy and Infectious Diseases (NIADDK), NIH

Home Start Analysis Shortcut to DAVID Tools Technical Center Downloads & APIs Term of Service Why DAVID? About Us

*** Announcing DAVID 6.8 Beta with updated Knowledgebase ([more info](#)). You may explore the new version at david-d.ncifcrf.gov.

Recommending: A paper published in *Nature Protocols* describes step-by-step procedure to use DAVID!

Welcome to DAVID 6.7

2003 - 2016

What's Important in DAVID?

- Current (v 6.7) release note
- New requirement to cite DAVID
- IDs of Affy Exon and Gene arrays supported
- Novel Classification Algorithms
- Pre-built Affymetrix and Illumina backgrounds
- User's customized gene background
- Enhanced calculating speed

Shortcut to DAVID Tools

Functional Annotation
Gene annotation enrichment analysis, functional annotation clustering , BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and more

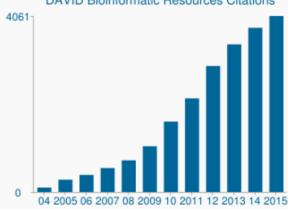
Gene Functional Classification
Provide a rapid means to reduce large lists of genes into functionally related groups of genes to help unravel the biological content captured by high throughput technologies. More

Gene ID Conversion
Convert list of gene ID/accessions to others of your choice with the most comprehensive gene ID mapping repository. The ambiguous accessions in the list can also be determined semi-automatically. More

Gene Name Batch Viewer
Display gene names for a given gene list; Search functionally related genes within your list or the entire database and get to enriched detailed information. More

Statistics of DAVID

DAVID Bioinformatic Resources Citations



Year	Citations
04	~100
05	~200
06	~400
07	~600
08	~1000
09	~1500
10	~2000
11	~3000
12	~4000
13	~5000
14	~6000
15	~40,000

- > 21,000 Citations
- Average Daily Usage: ~2,600 gene

david.ncifcrf.gov

Gordana Apic



EXZELLENZCLUSTER

CellNetworks

E.g. running DAVID on a gene expression sample

Effect of Valproic acid (VPA) on expression (SEURAT-1 Project)

	A	B	C	D	E	F
1	symbol	FCd2-0.5	FCd8-0.15	FCd8-0.5	FCd14-0.15	FCd14-0.5
2	GNMT	-3.4200512	-1.5056477	-2.0550076	-1.4837006	-2.43474587
3	CCL20	-3.0082003	-1.7569313	-2.0010222	1.27835876	-1.28902101
4	IGSF23	-3.007573	-1.4323363	-2.3809579	-1.4008533	-2.85662217
5	TAT	-2.6461464	-1.1485478	-1.1461061	1.09276156	-1.13326837
6	AHSG	-2.3255954	-1.3480667	-2.2125185	-1.6927436	-3.34501294
7	ABCB11	-2.210286	-1.2933891	-2.0223599	-1.2689865	-2.5839639
8	CHAC1	-2.1852995	-1.3072288	-2.3030681	-1.2740291	-1.42586597
9	INHBE	-2.149874	-1.3668306	-2.0658698	-1.3681003	-2.07231396
10	AKR1D1	-2.1396615	-1.1987528	-1.3082941	-1.1836023	-1.3347069
11	UBD	-2.1057916	-1.4296438	-1.8328095	-1.4606478	-2.61794341
12	FETUB	-2.0378259	-1.2604623	-1.9377909	-1.4817262	-3.11805694
13	UBD	-2.0153373	-1.3874231	-1.744188	-1.4774916	-2.5001375
14	UBD	-2.0153373	-1.3874231	-1.744188	-1.4774916	-2.5001375
15	UBD	-2.0153373	-1.3874231	-1.744188	-1.4774916	-2.5001375
16	CYP7A1	-1.9917766	-1.0285462	-2.0432567	-1.1549422	-2.66780126
17	UBD	-1.964659	-1.364476	-1.7437102	-1.4207546	-2.29640835
18	PSAT1	-1.9619694	-1.5143786	-1.5348269	1.06968099	1.04682857
19	UBD	-1.9543502	-1.47444535	-1.9544974	-1.544179	-2.77409556
20	CRP	-1.9532213	-1.6989478	-3.9198775	-2.4439166	-4.62963202
21	GLYATL1	-1.8873292	-1.3431879	-1.878024	-1.3145902	-2.30649382
22	PFKFB1	-1.887076	-1.1417868	-1.5341713	-1.2051444	-2.11780676
23	GBP7	-1.8821852	-1.2685911	-1.6025085	-1.3950364	-2.22057136
24	IL8	-1.8662524	-1.2505138	-1.7161424	1.69668459	1.20143748
25	UBD	-1.8638358	-1.3660403	-1.732867	-1.3953078	-2.24111663

Figure legend:

FC – Fold Change

d – day (2, 8 and 15)

Tested two concentrations:

0.5 and 0.15 mM

E.g. running DAVID on a gene expression sample

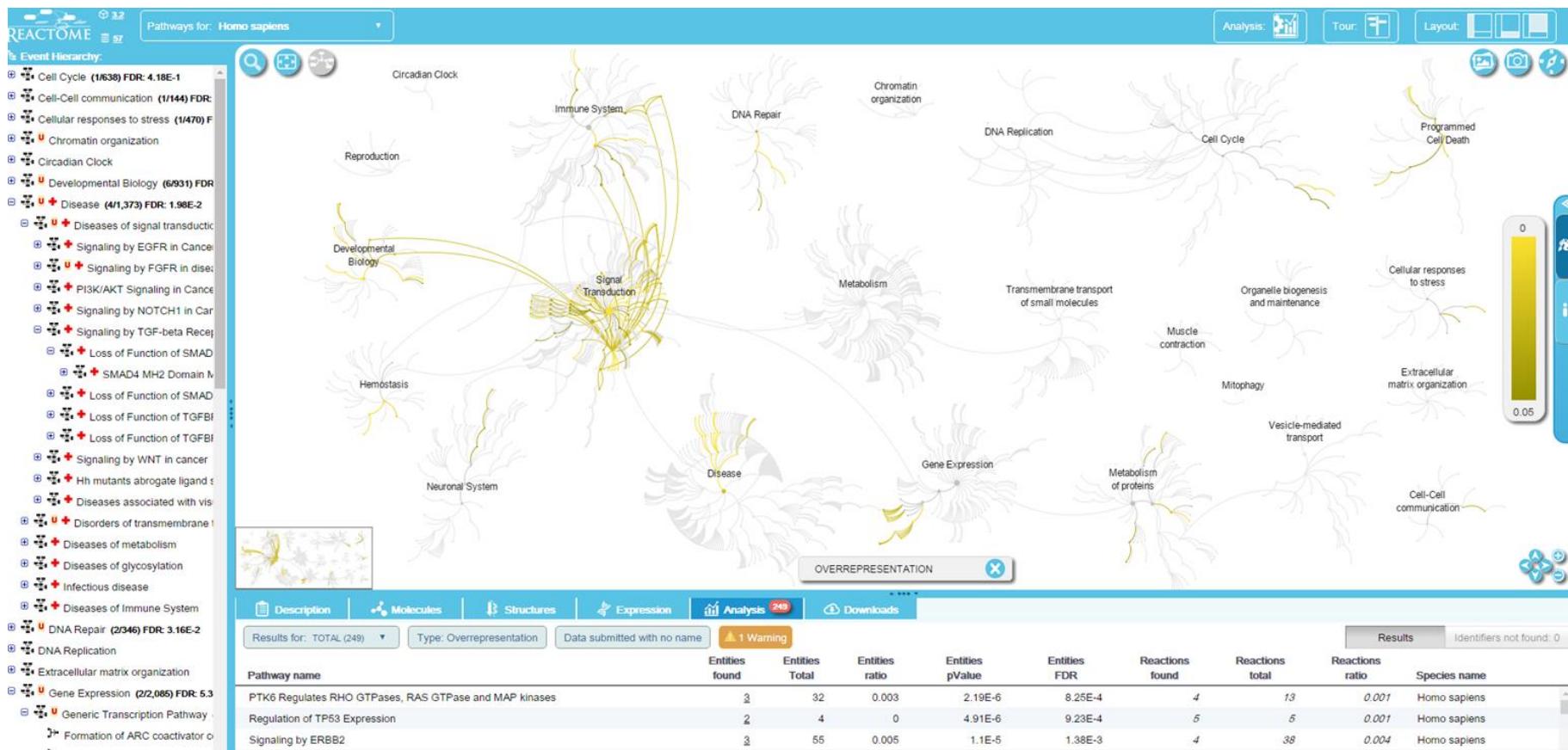
Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_BP_FAT	acute inflammatory response	RT	1	12	0.7	5.9E-9	8.2E-6
	GOTERM_BP_FAT	oxidation reduction	RT	1	25	1.4	4.8E-8	3.3E-5
	GOTERM_BP_FAT	response to wounding	RT	1	22	1.3	1.5E-7	6.9E-5
	GOTERM_BP_FAT	inflammatory response	RT	1	17	1.0	3.0E-7	1.0E-4
	GOTERM_BP_FAT	acute-phase response	RT	1	7	0.4	3.4E-6	9.5E-4
	GOTERM_BP_FAT	defense response	RT	1	20	1.2	2.3E-5	5.3E-3
	GOTERM_BP_FAT	regulation of inflammatory response	RT	1	7	0.4	1.4E-4	2.8E-2
	GOTERM_BP_FAT	regulation of response to external stimulus	RT	1	9	0.5	2.7E-4	4.6E-2
	GOTERM_BP_FAT	cellular hormone metabolic process	RT	1	6	0.3	3.8E-4	5.6E-2
	GOTERM_BP_FAT	positive regulation of cell adhesion	RT	1	6	0.3	4.1E-4	5.5E-2
All related somehow: CRP, KLF6, S100A8, AHSG, APCS, ANXA8L2, CCL20, CXCL10, F13B, C8A, C9, CFHR1, IL8, LBP, SPP1, SERPIND1, SAA1, SAA2, SAA4, TLR4, TGFB2, VNN1								
Most often down-regulated (apart from SPP1)								
Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_MF_FAT	electron carrier activity	RT	1	15	0.9	5.8E-8	2.2E-5
	GOTERM_MF_FAT	heme binding	RT	1	11	0.6	4.4E-7	8.3E-5
	GOTERM_MF_FAT	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen	RT	1	7	0.4	5.0E-7	6.3E-5
	GOTERM_MF_FAT	tetrapyrrole binding	RT	1	11	0.6	7.9E-7	7.5E-5
	GOTERM_MF_FAT	aromatase activity	RT	1	6	0.3	4.7E-6	3.6E-4
	GOTERM_MF_FAT	oxygen binding	RT	1	6	0.3	7.4E-5	4.6E-3
	GOTERM_MF_FAT	iron ion binding	RT	1	12	0.7	3.2E-4	1.7E-2
Drug metabolising enzymes plus a few others								



EXZELLENZCLUSTER

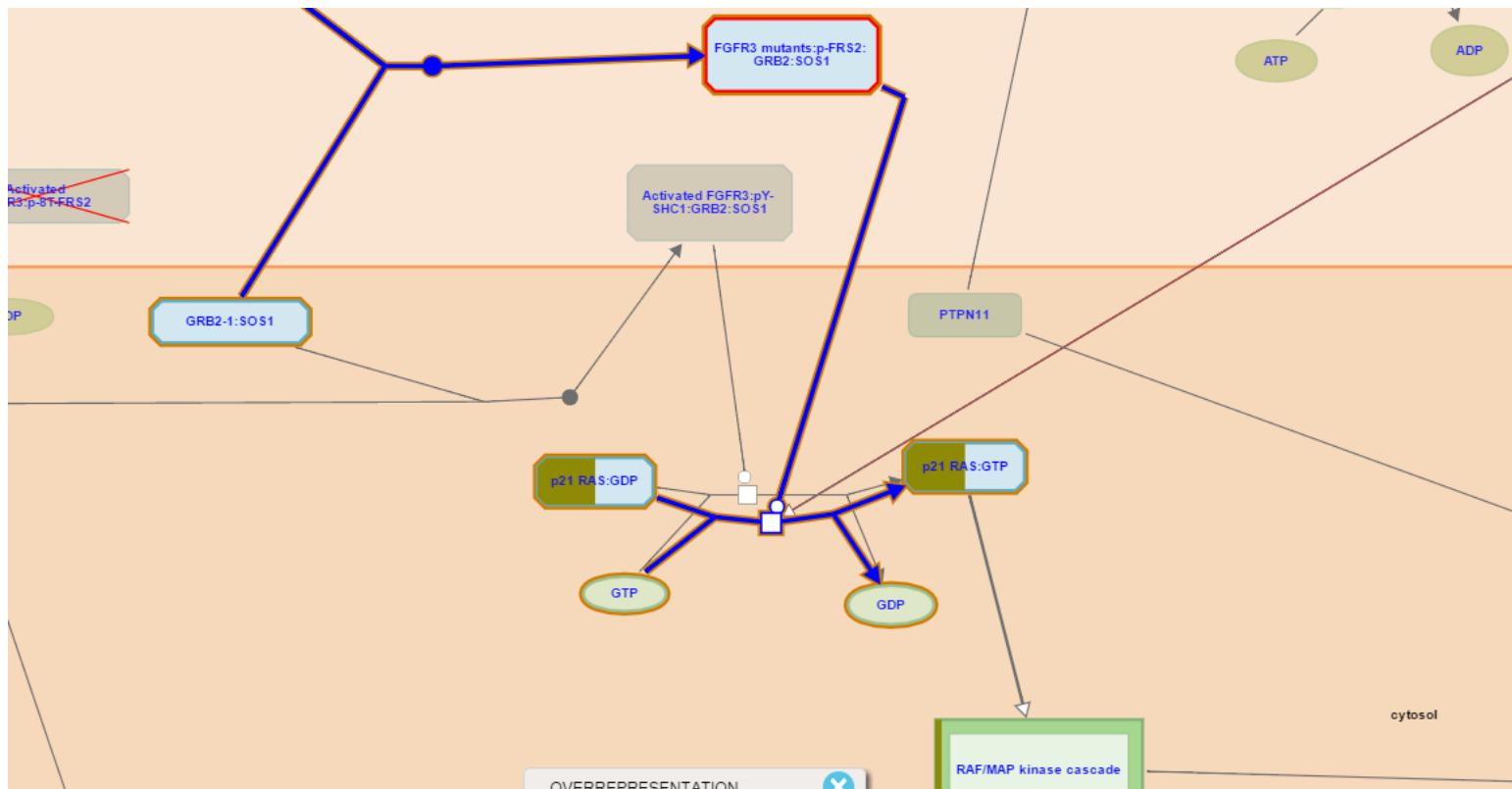
CellNetworks

My list TP53 RHOA KRAS HRAS CTNNB1 ABL1 ABL2 OR1A1
AKT2



Interrogating pathways in Reactome (zoomed in to one pathway element)

My list TP53 RHOA KRAS HRAS CTNNB1 ABL1 ABL2 OR1A1 AKT2

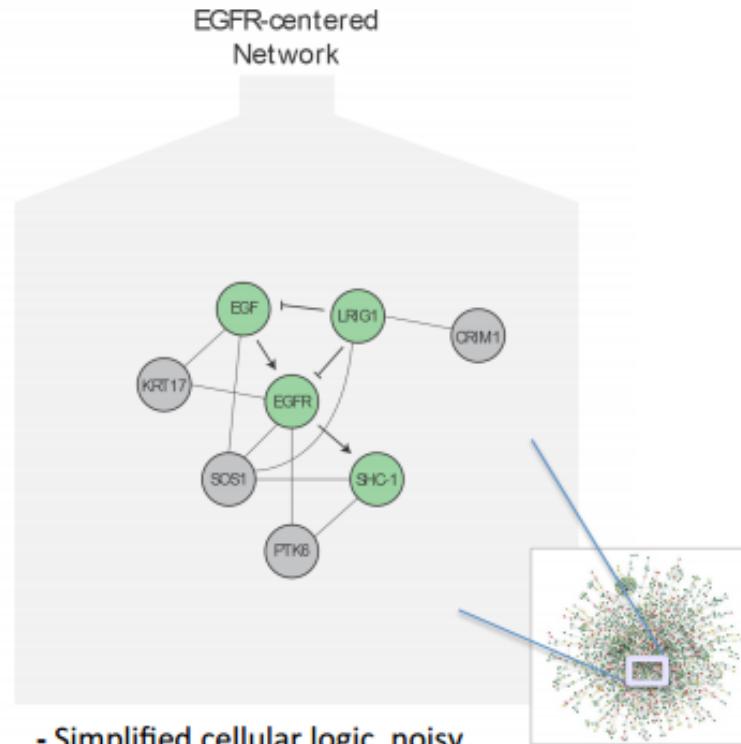
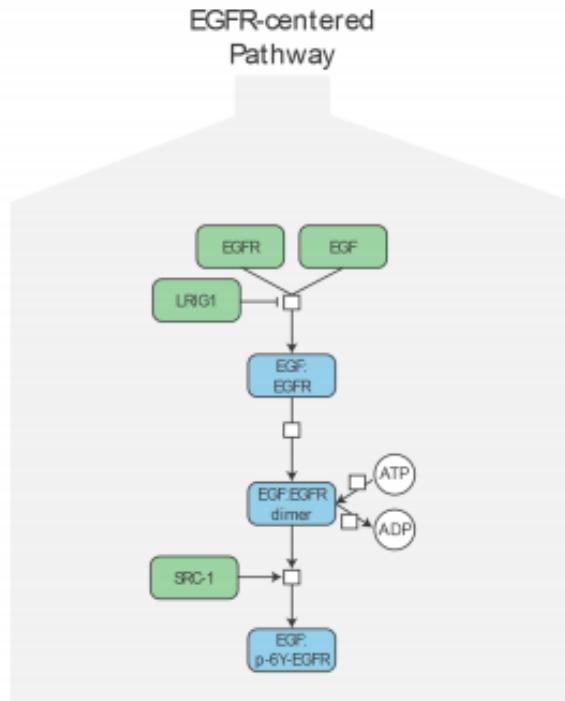


Pathways maybe not enough?

- For well studied proteins (e.g. My List), one can profit from many very carefully annotated sets (e.g. Reactome), but this currently only covers about 30-40% of the human genome.
- For less well studied (e.g. novel or uncharacterized) genes, one can still get a view of the community of genes or proteins related to it by exploiting high-throughput data resources.
- In other words, your genes might not be in a pathway, but they can still be in a *network*.

What's the difference?

Pathways vs. Networks

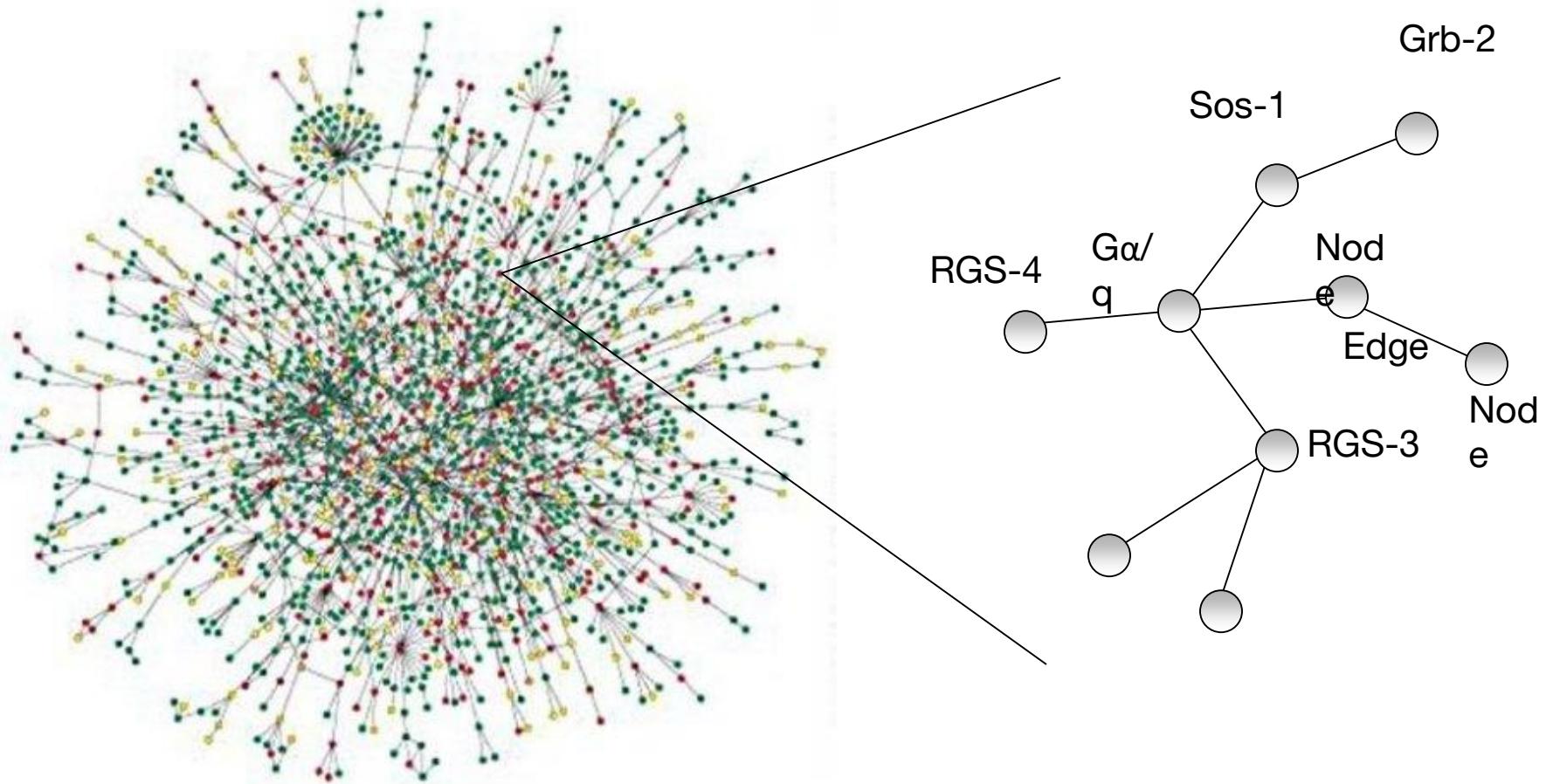


- Detailed, high-confidence consensus
- Biochemical reactions
- Small-scale, fewer genes
- Concentrated from decades of literature

- Simplified cellular logic, noisy
- Abstractions: directed, undirected
- Large-scale, genome-wide
- Constructed from *omics* data integration

Courtesy Gary Bader, University of Toronto

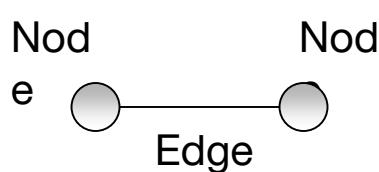
Interaction networks



Physical & functional interactions between genes & proteins.

- There are now thousands and thousands of associations between proteins that have been determined over decades of research
- Interactions can be *physical* i.e. that an experiment has determined that they molecules are actually in contact or in a complex
- Or they can be looser associations, indicating just a functional relationship, such as a common pathway, biological process, a common expression pattern, or genetic interactions.
- These data are most typically presented as networks

Biological interaction networks



Nodes:

- Proteins
- Genes
- Chemicals
- Effects
- Diseases

Edges:

- Physical interaction (e.g. yeast two-hybrid)
- Co-expression (e.g. microarrays, RNAseq)
- Same operon
- Regulation of gene expression (protein to gene)
- Catalysis (e.g. metabolic networks)



Interaction databases

IntAct

Home
Advanced Search
Tools
Data Submission
Downloads
Documentation
Contact Us

Printer Friendly View

News
26/07/2011
UniProt collaboration on annotation of protein interactions to MINTx standard
UniProt curators will now curate protein interaction data to MINTx standards as part of their normal workflow, directly within the UniProt interface. These interactions will form a part of the IntAct dataset, which undergoes a MEx-level curation. A subset of binary interactions is extracted from this dataset.

EBI > Databases > Pathways & Networks > IntAct > View

Search: Search Clear Show Advanced Fields >

Home Search Interactions (4696) Browse Lists Interaction Details Molecule View Graph

Browse by taxonomy, gene ontology, ChEBI ontology

> 4,696 binary interactions were found in IntAct. 1,494 of them are originated from spoke expanded co-complexes and you may want to filter them. Your query also matches 3,155 interaction evidences from 10 other databases. Your query also matches 317 interaction evidences from 4 other IMEx databases.

Previous 1-30 of 4696 Next 30 Export to: Select format... Export Change Columns Displayed Previous

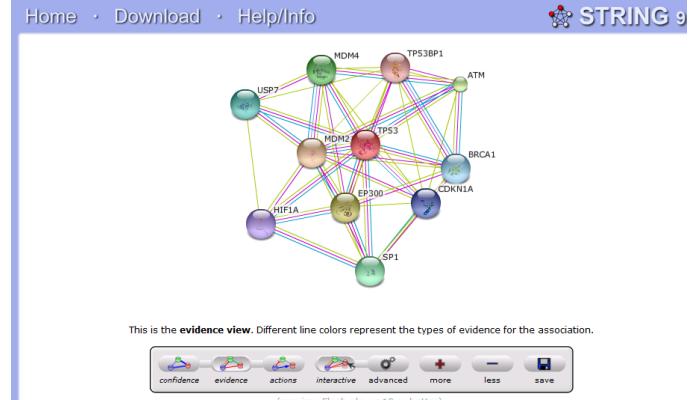
Name molecule A	Links	Aliases	Aliases	Species	Species	Publication	Interaction Detection Method
BABAM1	IPN UN	UnProtein	USP7	BA UN	UniProt	USP7	MINT
HTT	BA UN	UnProtein	HTT	BA UN	UniProt	HTT	MINT
BABAM1	BA UN	UnProtein	USP7	BA UN	UniProt	USP7	MINT
BBPBP6	BA UN	UnProtein		BA UN	UniProt		MINT

click on a protein name or the View interacting proteins link to see the list of protein it has been shown to interact with.

Protein view: Select proteins and connect them in the MINT Viewer: Include connecting proteins not in the list Only consider proteins in this list CONNECT

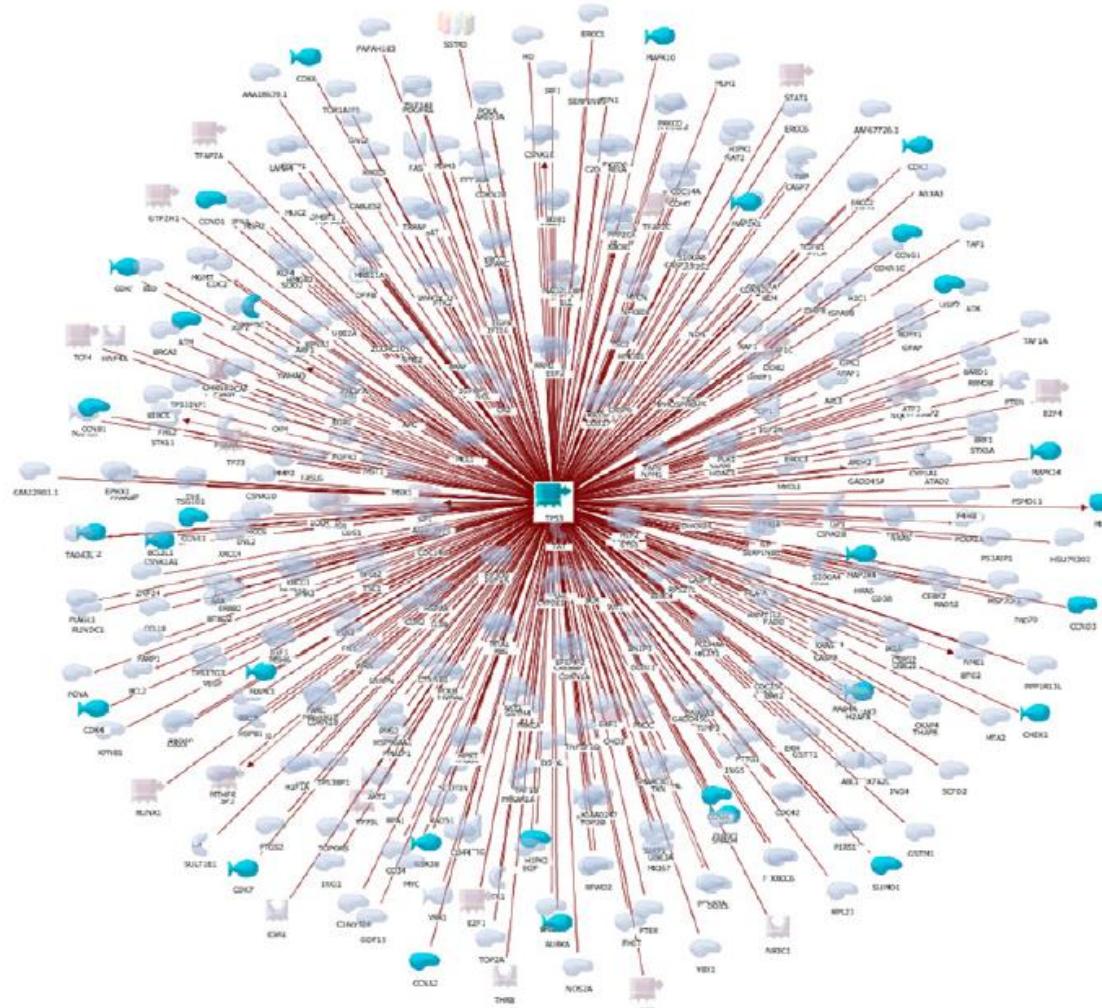
check all uncheck all

- p53 *Drosophila melanogaster* (7227) View interacting proteins
- MDM2 *Homo sapiens* (9606) E3 ubiquitin-protein ligase Mdm2 View interacting proteins
- TP53BP2 *Homo sapiens* (9606) Apoptosis-stimulating of p53 protein 2 View interacting proteins
- MDM4 *Homo sapiens* (9606) Protein Mdm4 View interacting proteins
- LGALS7 *Homo sapiens* (9606) Galectin-7 View interacting proteins
- names and synonyms: p53-induced gene 1 protein P17 HKL-14 PIG1 LGALS7B uniprotkb ac: P47929, Q8jB87, domains: ConA_like_subgrp IPR013320 [Galectin_b] IPR001079 |IPR008985|
- orthologs:
 - Lgals7: Galectin-7 *Rattus norvegicus* (P97590)
 - Lgals7: *Mus musculus* (Q9ML7)
 - Lgals7: *Mus musculus* (Q8CRB1)
 - Lgals7: Galectin-7 *Mus musculus* (O54974)
- BBPBP6 *Homo sapiens* (9606) Retinoblastoma-binding protein 6 View interacting proteins

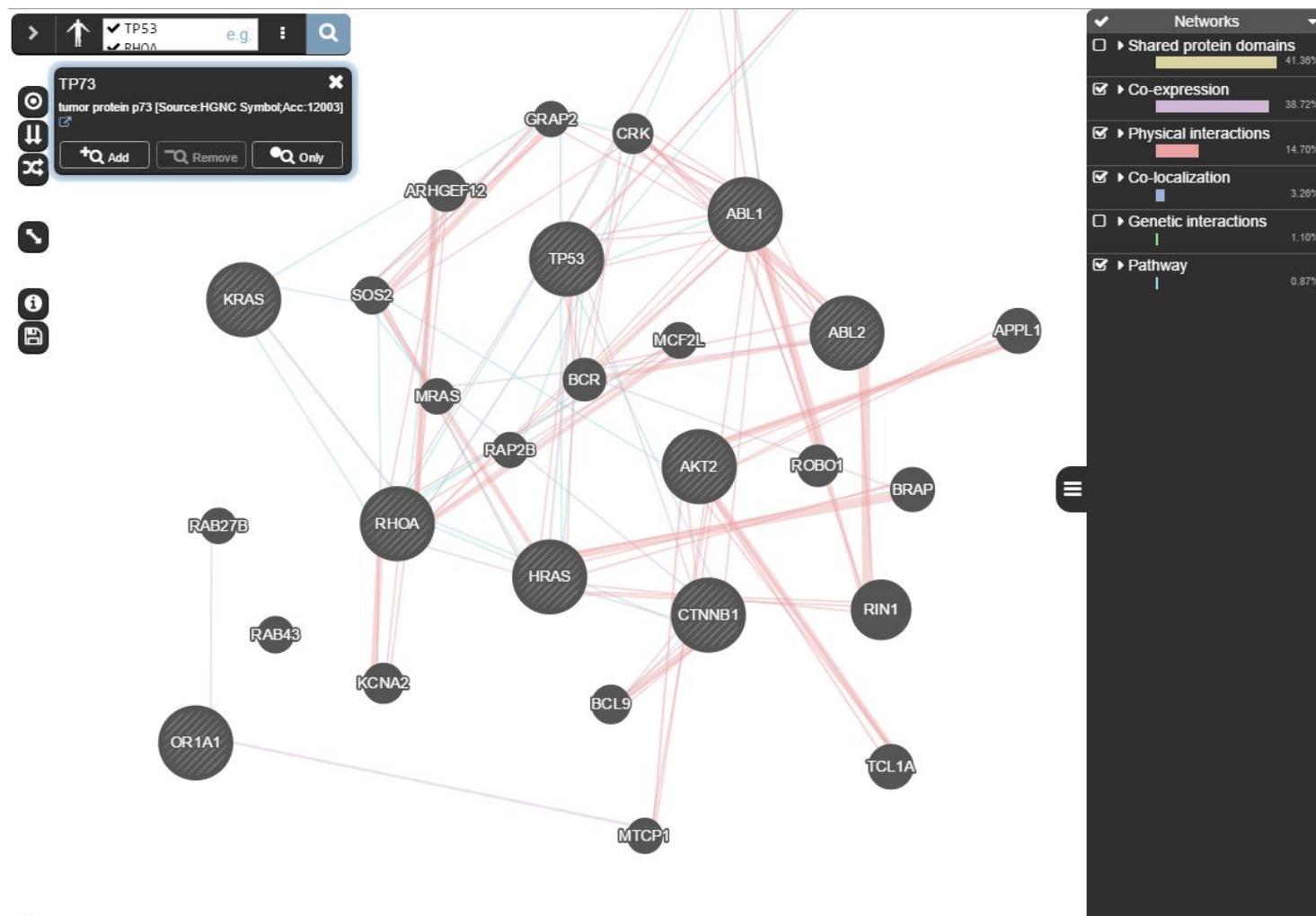


- Resources are very different in appearance and content
- Efforts are underway to make a unified search/view, but not complete
- Thus one needs currently to look at several sites to check if an interaction is known
- Some are purely content (e.g. IntAct, MINT) others are processed and augmented (e.g. STRING, GeneMANIA) with additional predicted/inferred interactions

p53 – the promiscuous transcription factor

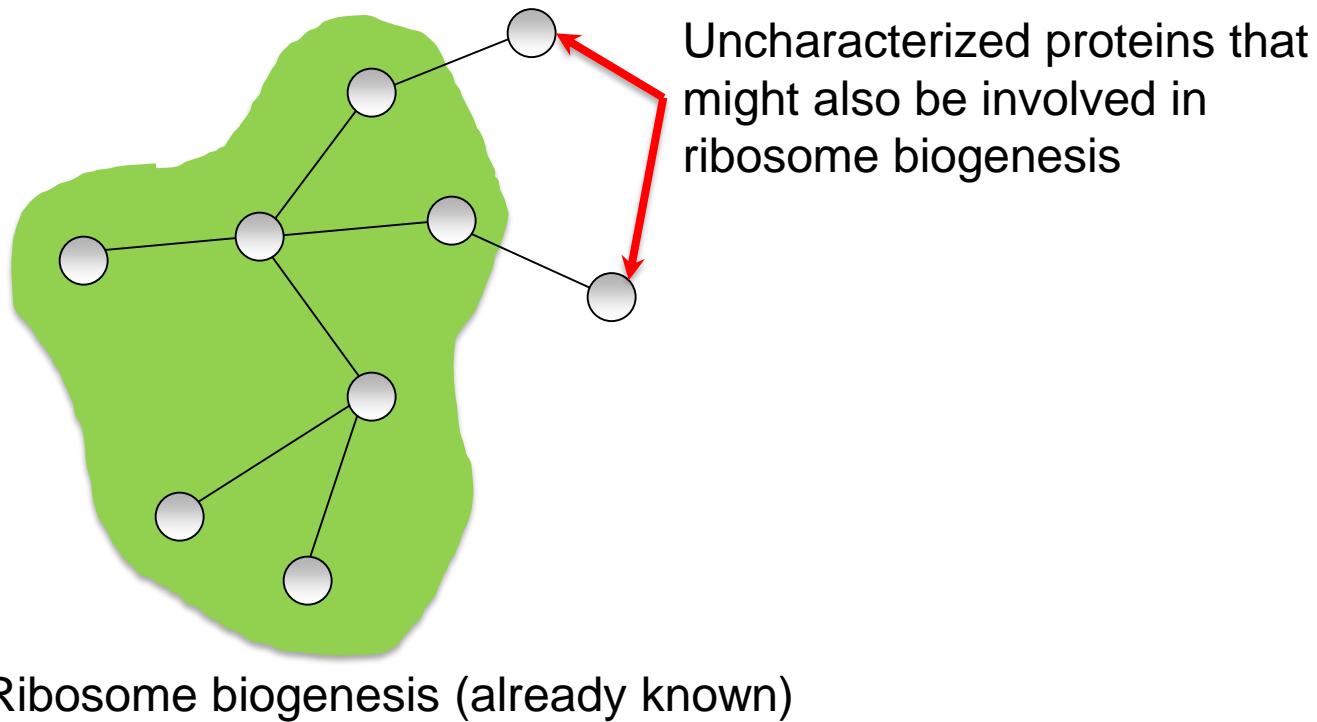


GeneMANIA: a tool for building networks from a set of genes and known interactions



What are networks good for? Concept of “guilt by association”

- This flawed legal concept is very useful in biology.
- If a protein that is poorly understood is somehow connected (e.g. physical interaction, same expression patterns, same phenotype) to one or more well-understood proteins, then it can (to some extent) inherit a similar function.

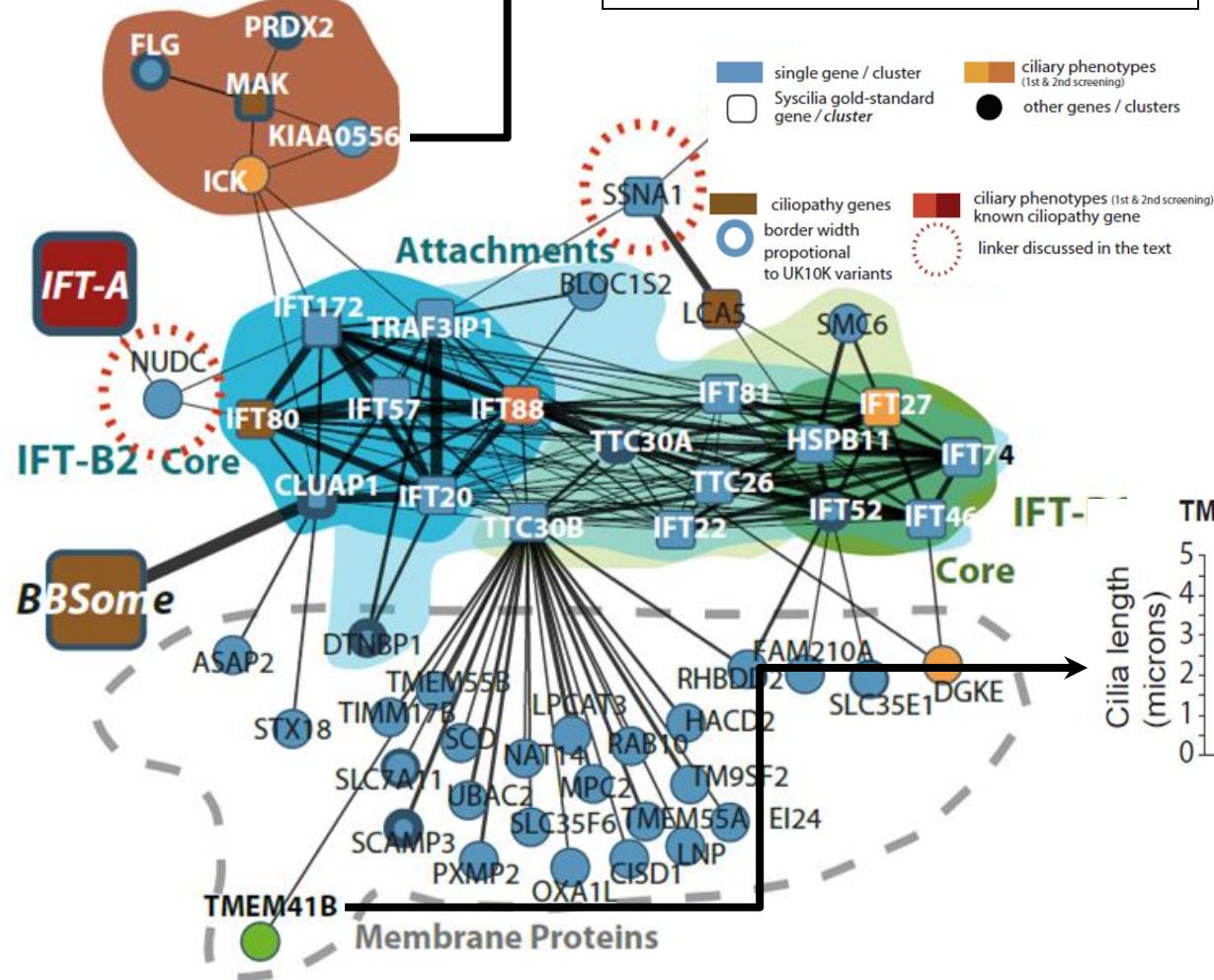




EXZELLENZCLUSTER

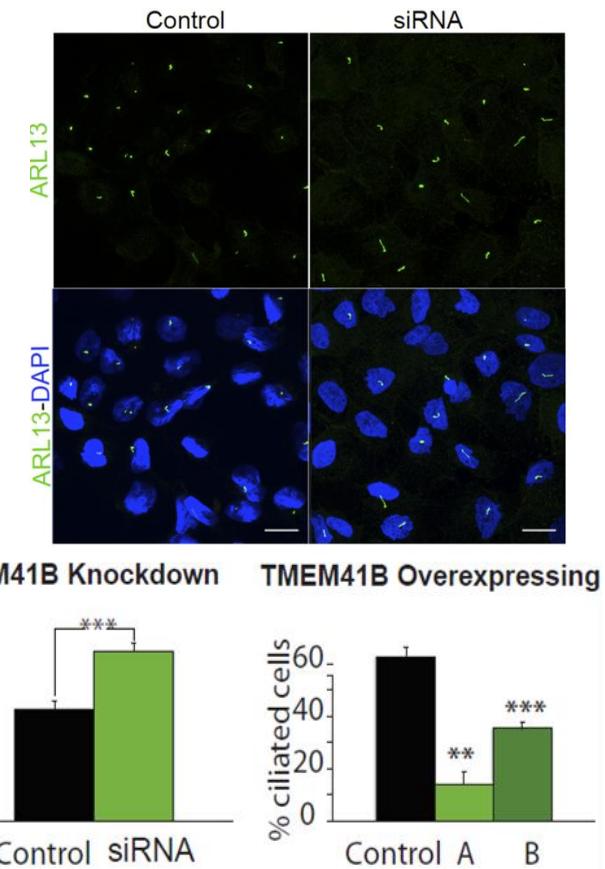
CellNetworks

MAK/ICK/KIAA0556



Interaction proteomics and guilt-by-association – recent example

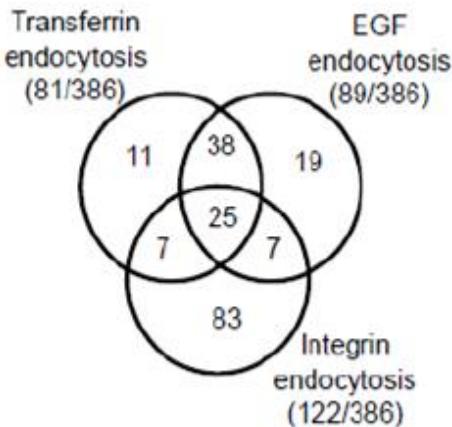
Sanders et al, Genome Biol, 2015



The need to combine or *integrate* datasets

- One dataset is *usually* not enough.
- Trivially, just identifying genes means nothing unless you know what the genes do (e.g. one is in principle always integrating omics data with gene databases, or textbook knowledge)
- More significantly, a single technique can have many problems:
 - Lack of complete coverage – e.g. low expressed genes or proteins might be missed, or a whole class of proteins/genes might be missed (e.g. extracellular proteins in the two-hybrid system)
 - A potential disconnect with reality: e.g. gene expression can differ significantly from protein expression for many reasons
- By *integrating* datasets one can clean the data up (rule out spurious observations & strengthen real ones), put the dataset in context, allow more elaborate queries (e.g. proteins that are most often over-expressed in RNAseq also tend to make more interactions).

Integrating datasets



> 30 % overlap

Often one wants to compare similar datasets, using the same technology, that have been differently applied.

Comparison of the Three siRNA Knockout Screens

Host Cell Line	Time of siRNA Treatment	Virus Challenge	Time of Scoring Post-infection	Readout	# of Filtered Hits	
Brass et al.	HeLa (CD4 ⁺ , β-gal reporter)	72 hr	Live HIV-1 (III B)	48 hr; 48 hr in new cells	p24 (CA); reporter activation	273
König et al.	293T	48 hr	HIV-1 luc vector, VSV-G pseudotyped	24 hr	Luc reporter	295
Zhou et al.	HeLa (CD4 ⁺ , β-gal reporter)	24 hr	Live HIV-1 (HXB2 isolate)	48 hr; 96 hr	β-gal reporter activation	224

< 4 % overlap

An example of a large systems biology project: SEURAT-1 (Safety Evaluation Ultimately Replacing Animal Testing)

SEURAT-1 is developing knowledge and technology building blocks required for the development of solutions for the replacement of current repeated dose systemic toxicity testing *in vivo* used for the assessment of human safety.



- Funding 50 Milion EURs
- Funded by the EC and CE (formerly named Colipa)
- 2011-2016
- www.seurat-1.eu



EUROPEAN COMMISSION
Research & Innovation



Cosmetics Europe
the personal care association

Integrated data analysis in NoTox (SEURAT)

- SEURAT-1 is designed as a cluster of seven projects
- The project develops infrastructure and service functions to create a sustainable predictive toxicology support resource going beyond the lifetime of the program
- Providing tools for long-term toxicity prediction using repeat dose

Valproic Acid

Valproic Acid

Executive Summary Information

Compound	Valproic Acid
Toxicities	Steatosis, cytotoxicity
Mechanisms	As a fatty acid analogue, the compound is a competitive inhibitor of fatty acid metabolism, which accounts for steatosis. The parent compound is also hepatotoxic by a mechanism that has not been resolved; however, this hydrophobic compound is used at very high concentrations and its promiscuous activity at these concentrations is likely due to disruption of membrane integrity. P450 ω -oxidation produces a reactive alkylating and free radical-propagating agent that adds to the toxicity profile.
Comments	This compound was selected as a reference standard for steatosis via inhibition of β -oxidation.



EXZELLENZCLUSTER

CellNetworks

Example input (expression data)

Analysis of Valproic acid (VPA) expression data within SEURAT-1 Project

	A	B	C	D	E	F
1	symbol	FCd2-0.5	FCd8-0.15	FCd8-0.5	FCd14-0.15	FCd14-0.5
2	GNMT	-3.4200512	-1.5056477	-2.0550076	-1.4837006	-2.43474587
3	CCL20	-3.0082003	-1.7569313	-2.0010222	1.27835876	-1.28902101
4	IGSF23	-3.007573	-1.4323363	-2.3809579	-1.4008533	-2.85662217
5	TAT	-2.6461464	-1.1485478	-1.1461061	1.09276156	-1.13326837
6	AHSG	-2.3255954	-1.3480667	-2.2125185	-1.6927436	-3.34501294
7	ABCB11	-2.210286	-1.2933891	-2.0223599	-1.2689865	-2.5839639
8	CHAC1	-2.1852995	-1.3072288	-2.3030681	-1.2740291	-1.42586597
9	INHBE	-2.149874	-1.3668306	-2.0658698	-1.3681003	-2.07231396
10	AKR1D1	-2.1396615	-1.1987528	-1.3082941	-1.1836023	-1.3347069
11	UBD	-2.1057916	-1.4296438	-1.8328095	-1.4606478	-2.61794341
12	FETUB	-2.0378259	-1.2604623	-1.93777909	-1.4817262	-3.11805694
13	UBD	-2.0153373	-1.3874231	-1.744188	-1.4774916	-2.5001375
14	UBD	-2.0153373	-1.3874231	-1.744188	-1.4774916	-2.5001375
15	UBD	-2.0153373	-1.3874231	-1.744188	-1.4774916	-2.5001375
16	CYP7A1	-1.9917766	-1.0285462	-2.0432567	-1.1549422	-2.66780126
17	UBD	-1.964659	-1.364476	-1.7437102	-1.4207546	-2.29640835
18	PSAT1	-1.9619694	-1.5143786	-1.5348269	1.06968099	1.04682857
19	UBD	-1.9543502	-1.47444535	-1.9544974	-1.544179	-2.77409556
20	CRP	-1.9532213	-1.6989478	-3.9198775	-2.4439166	-4.62963202
21	GLYATL1	-1.8873292	-1.3431879	-1.878024	-1.3145902	-2.30649382
22	PFKFB1	-1.887076	-1.1417868	-1.5341713	-1.2051444	-2.11780676
23	GBP7	-1.8821852	-1.2685911	-1.6025085	-1.3950364	-2.22057136
24	IL8	-1.8662524	-1.2505138	-1.7161424	1.69668459	1.20143748
25	UBD	-1.8638358	-1.3660403	-1.732867	-1.3953078	-2.24111663

Figure legend:

FC – Fold Change

d – day (2, 8 and 15)

Tested two concentrations: 0.5 and 0.15

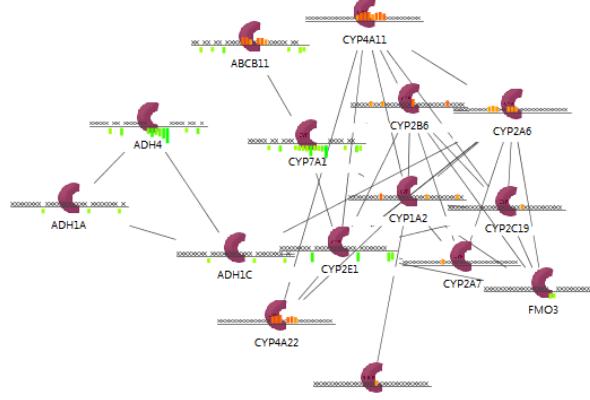


EXZELLENZCLUSTER

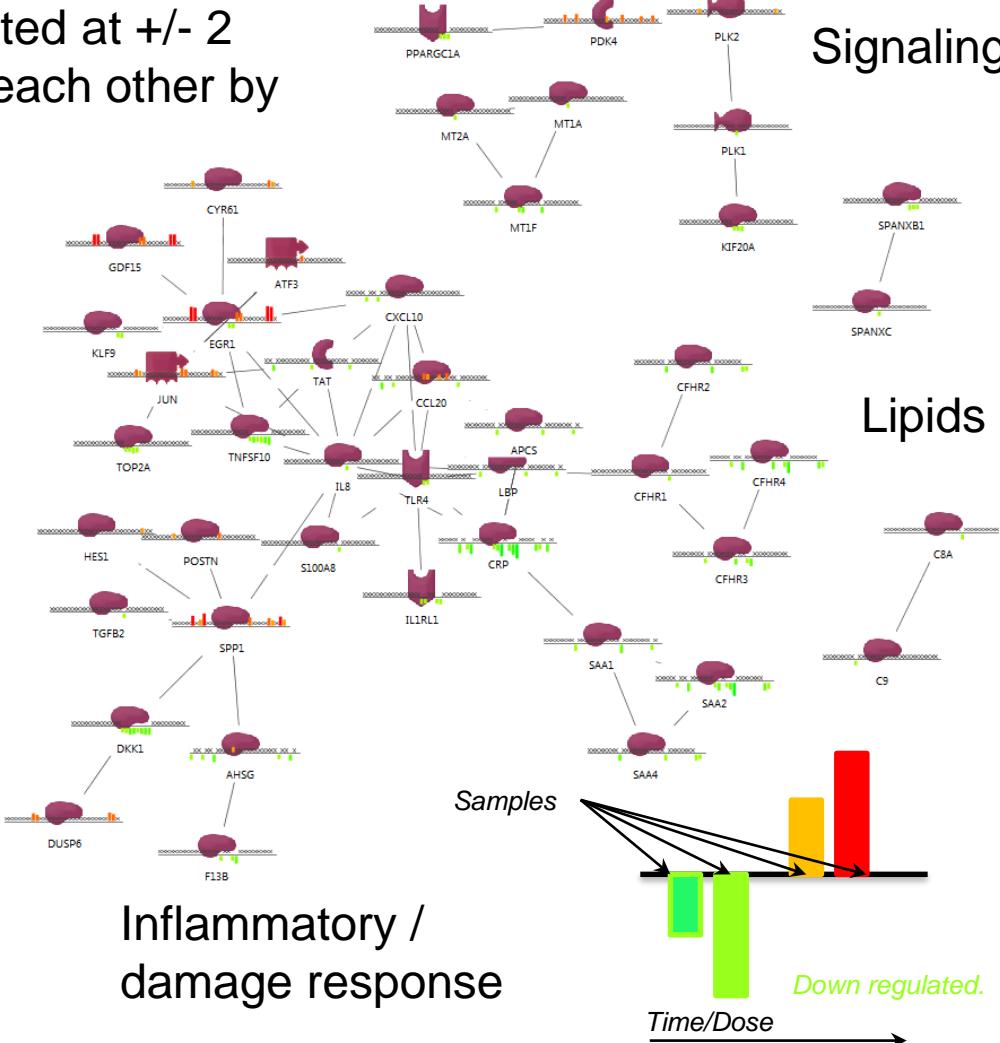
CellNetworks

Merging gene expression & proteomics data

These are all genes dysregulated at +/- 2 fold that can be connected to each other by known interactions.



Drug metabolism





EXZELLENZCLUSTER

CellNetworks

One can now probe for what is new and what is not (e.g. new biomarkers?)

These are all genes dysregulated at +/- 2 fold that can be connected to each other by

↳ [Epilepsy Res. 2010 Oct;91\(2-3\):187-92. doi: 10.1016/j.epilepsyres.2010.07.011. Epub 2010 Aug 14.](#)

Effects of AEDs on biomarkers in people with epilepsy: CRP, HbA1c and eGFR.

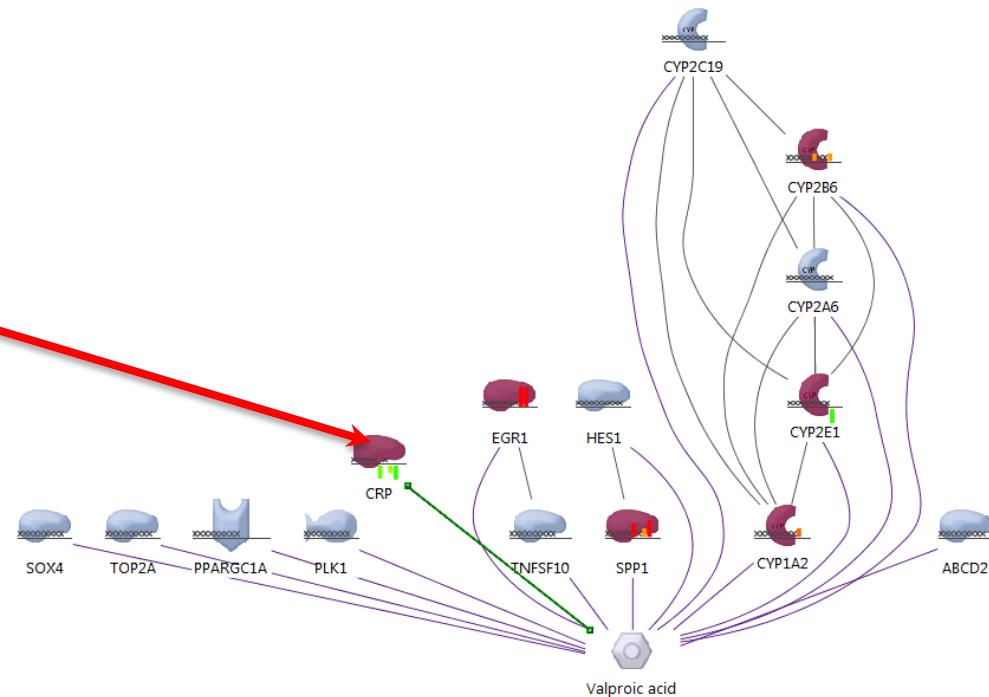
[Yuen AW¹, Bell GS, Peacock JL, Koepp MM, Patsalos PN, Sander JW.](#)

⊕ Author information

Abstract

The standardised mortality ratio in people with epilepsy is raised to population. Some biomarker levels, including higher C-reactive protein (HbA1c) and lower estimated glomerular filtration rate (eGFR), are associated with premature mortality. These biomarkers were measured in 125 people with epilepsy to assess the potential effect of antiepileptic drug (AED) use on these markers. Monotherapy with valproate (N=50) use was associated with 55% lower mean CRP compared to carbamazepine (N=30) and phenytoin (N=32) use with 4% lower mean HbA1c values. These improvements may be explained by AEDs. On the other hand, lamotrigine use (N=19) was associated with 10% higher mean eGFR. This may represent a negative effect on a health marker. These preliminary results suggest that controlled studies are needed ideally in people on AED monotherapy.

Copyright © 2010 Elsevier B.V. All rights reserved.

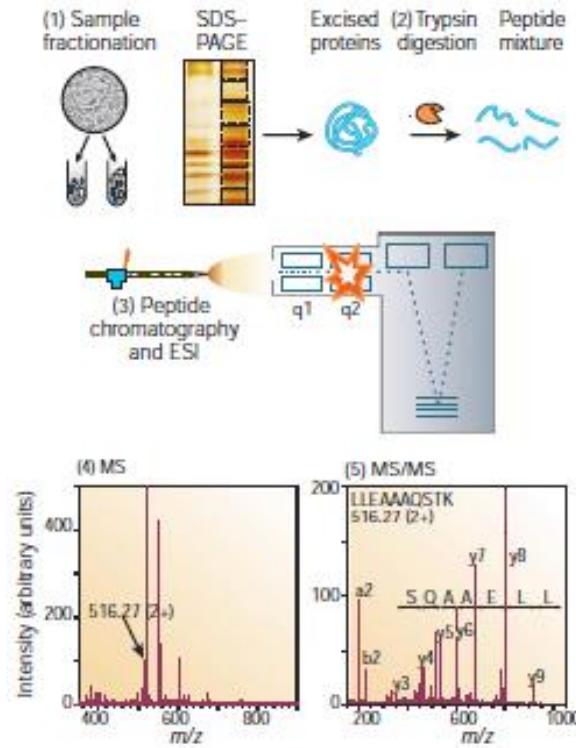
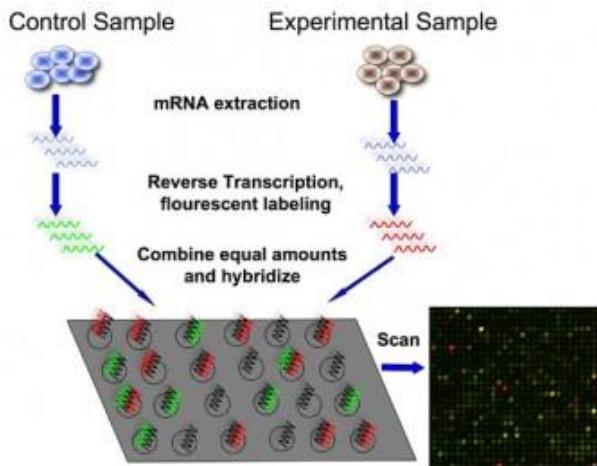




EXZELLENZCLUSTER

CellNetworks

How do I get a gene/protein list?



DNA microarray

Mass spectrometry

RNAseq

“Raw” data

Working with Raw data

- Hard to generalize, but...
- ... in the 1990s (the bioinformatics “cowboy” days) one would have to write programs for everything oneself
- In the civilized 2010s, now one exploits the work of large communities of thousands of scientists:
 - Data standardizations (e.g. “vcf” format, etc.)
 - Common code bases that are relatively userfriendly:
 - BioPERL
 - BioPython
 - Whole language systems dedicated to bioinformatics (e.g. BioConductor in R)

An example of software for genomics statistical analysis and graphics – R & Bioconductor

- **What is R?**

- R is a free statistical package designed for easy analysis of data
- R is also a *language* for programming
- Brief History: S and Splus were commercial packages that Robert Gentleman (hence R) mimicked to create R, which is open source and heavily supported by a large user and developer community worldwide
- R statistical package should be downloaded from <http://www.r-project.org>

- **What is Bioconductor?**

- Bioconductor is a suite of R routines and libraries for use within the life sciences
- Also developed by a large community of developers and users worldwide
- It contains many extra packages required for microarray analysis

R/Bioconductor & Microarrays

- There are dozens of different packages in R for processing microarray data of various kinds (and for that matter, most other data types)
- Can be daunting, but googling usually finds out what you need (e.g. platform specific objects).
- For example, we can use a simple routine for analysing Affymetrix data:
`library(simpleaffy)`



[Home](#) » [Help](#) » [Workflows](#) » Oligonucleotide Arrays

Using Bioconductor for Microarray Analysis

Bioconductor has advanced facilities for analysis of microarray platforms including Affymetrix, Illumina, Nimblegen, Agilent, and other one- and two-color technologies.

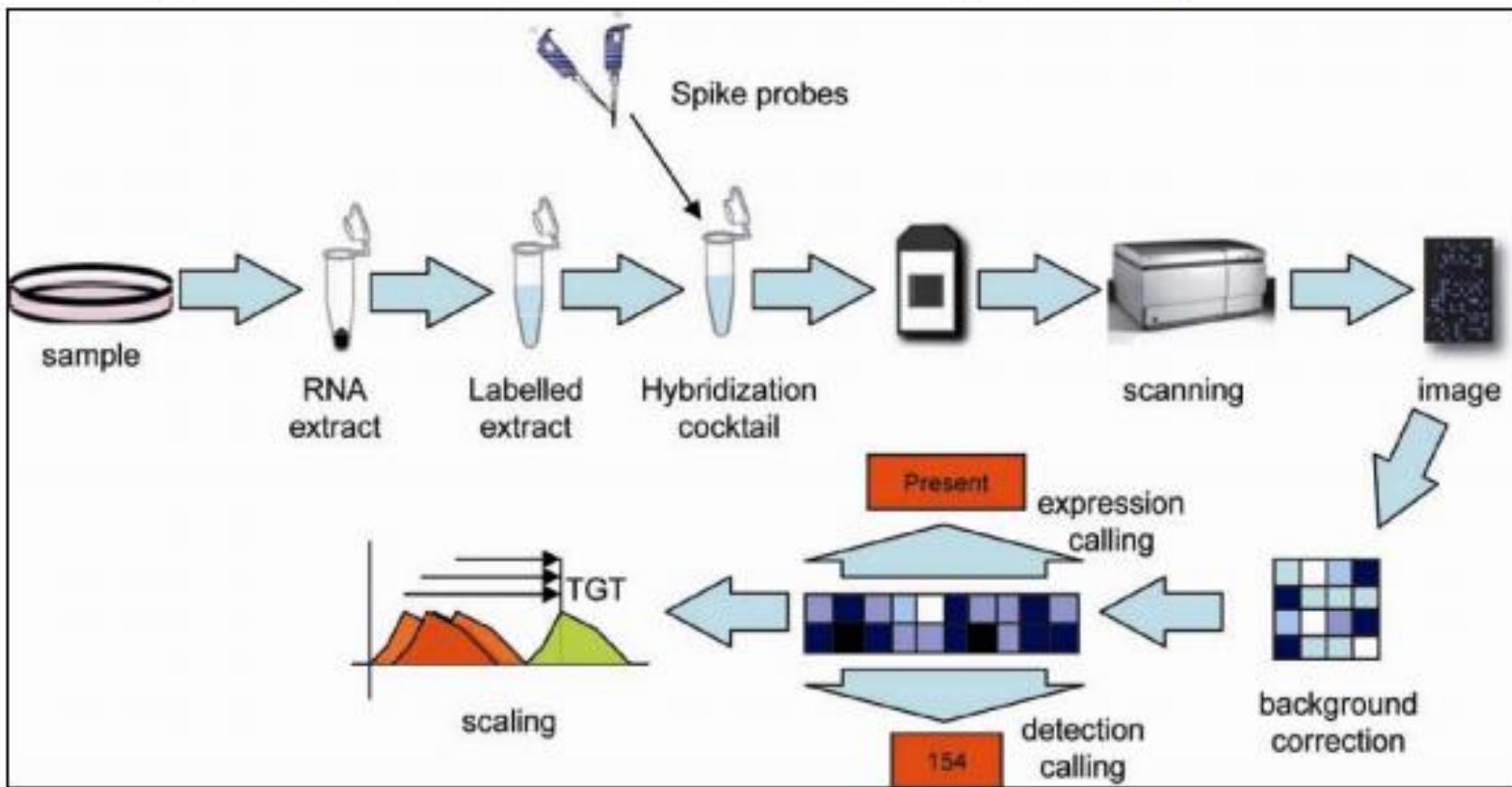
Bioconductor includes extensive support for analysis of expression arrays, and well-developed support for exon, copy number, SNP, methylation, and other assays.

Major workflows in Bioconductor include pre-processing, quality assessment, differential expression, clustering and classification, gene set enrichment analysis, and genetical genomics.

Bioconductor offers extensive interfaces to community resources, including GEO, ArrayExpress, Biomart, genome browsers, GO, KEGG, and diverse annotation sources.

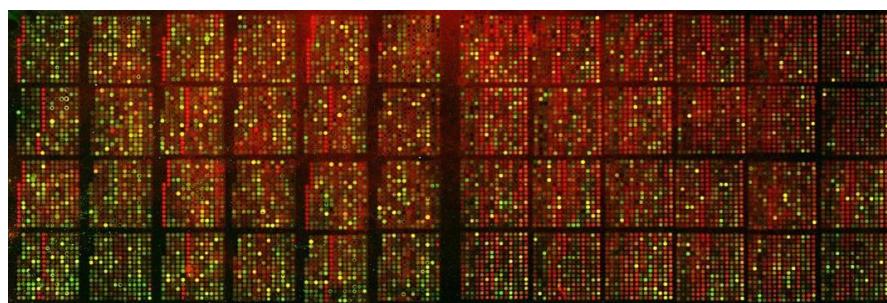
Simpleaffy a very simple to use high level analysis of Affymetrix data in Bioconductor

The QC metrics implemented in simpleaffy

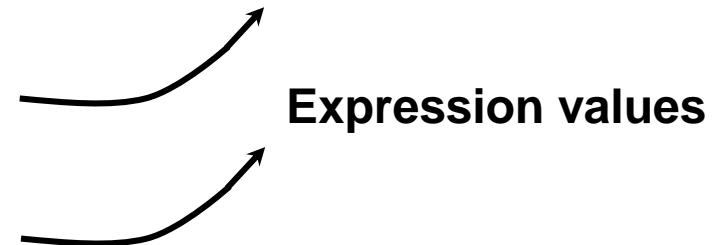


Raw data from microarray experiments

- CEL files are essentially the raw images from the microarray reader
- You'll need to tell R/Bioconductor what kind of microarray datasets you are dealing with (protocols below presume Affymetrix chips)
- You need to know what each CEL file (i.e. image) corresponds to: is it a control experiment or one corresponding to a test condition?

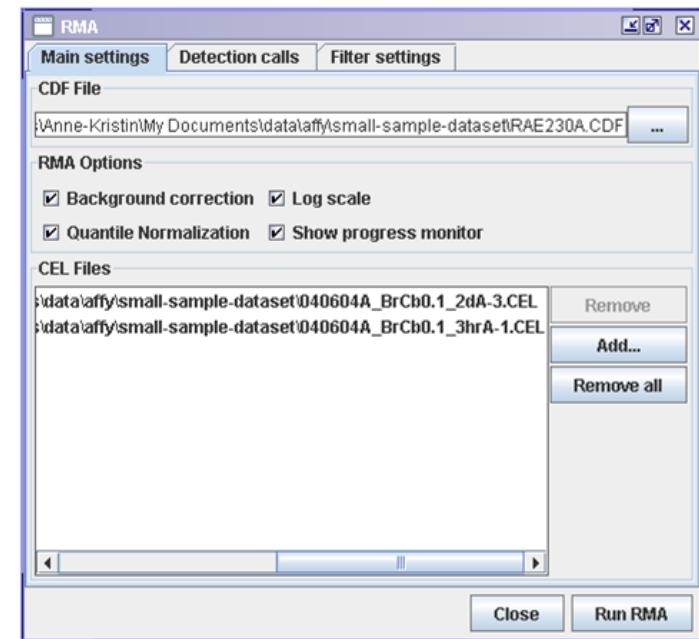


**Fluorescent
images**



Averaging values, eg. RMA

- RMA - "Robust Multiarray Averaging" – combines all the repeats for each condition to define a single value for the expression level of each gene. And corrects for non-specific binding
- This process will give each gene on each treatment a number and also has some statistics that are used later for various purposes related to confidence of the observation etc.
- These numbers aren't in any way fold changes (we are dealing here with "single channel" microarrays), they are just a measure of how much of mRNA for each gene is present in each sample.

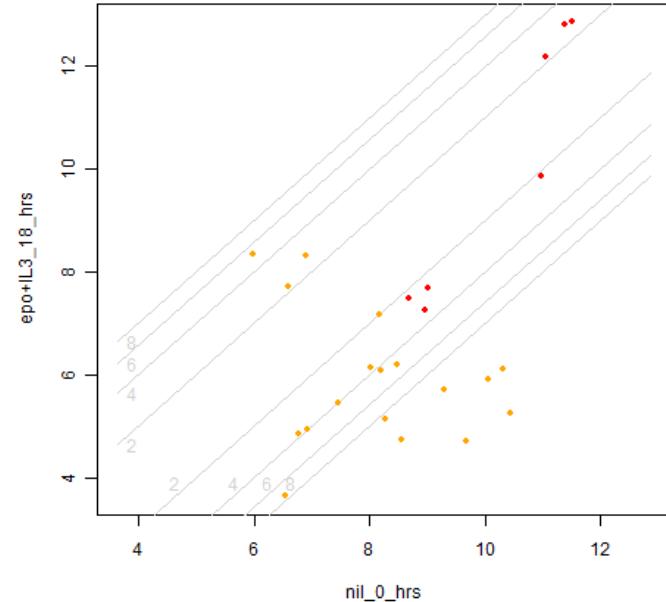
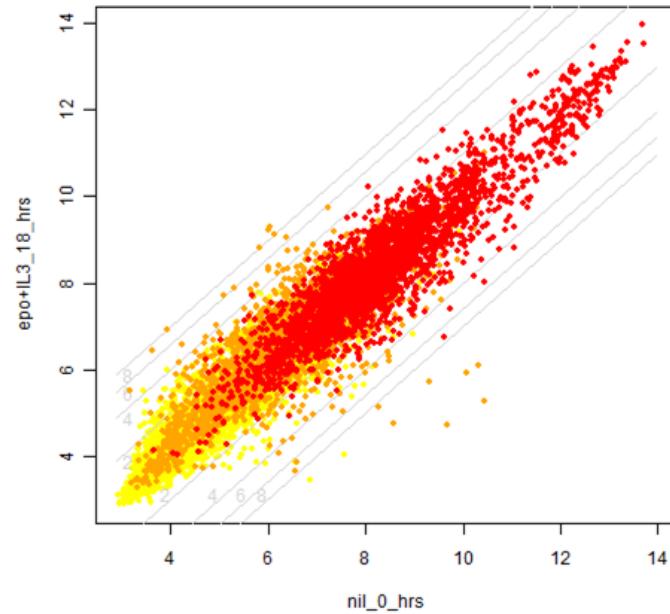




EXZELLENZCLUSTER

CellNetworks

Comparing treatment to control



- In the significant dataset we have compared treatment to control groups and created a set of data that is somehow significant ($fc \geq 2$ or ≤ -2 $p\text{-value} \leq 0.001$) and then writes it to a file and plots a scatter plot to display the results
- The files "**sig.rma.txt**" and "**sig_rma.png**" in the working folder will contain these two things



EXZELLENZCLUSTER

CellNetworks

Looking at the data

The final output be imported into excel and read by many other programs

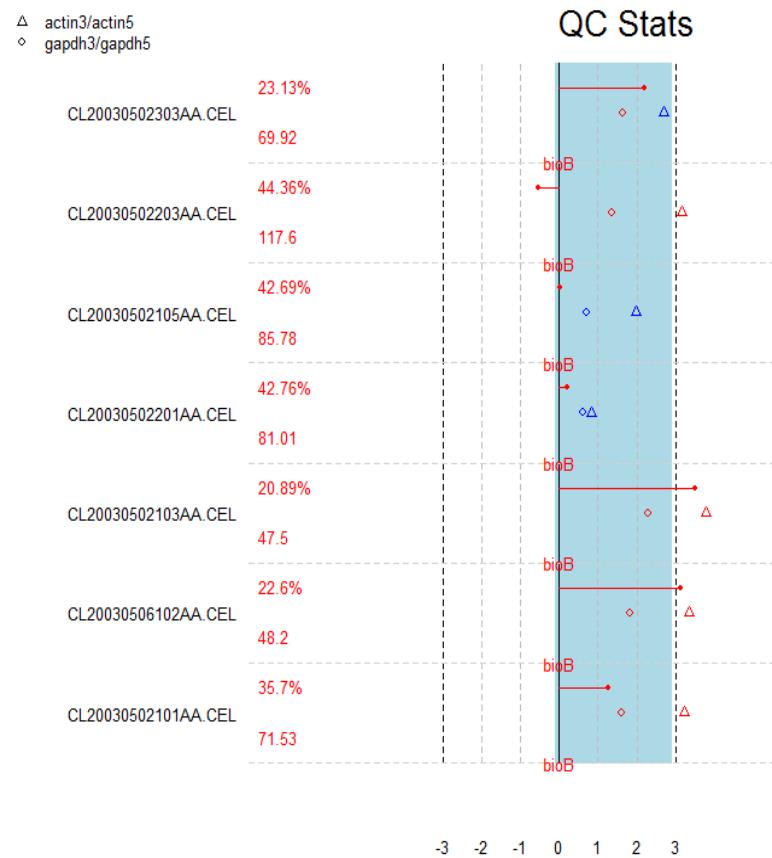
The screenshot shows a Microsoft Excel window titled "Microsoft Excel - sig.rma". The table contains 33 rows of data, each representing a different sample or condition. The columns are labeled A through L. Column A contains sample identifiers like "nil_0 hrs", "200862_at", etc. Columns B through D contain numerical values. Columns E through L contain categorical labels such as "P", "A", "M", and "AA". The data appears to be a list of genes or proteins with their corresponding values and annotations.

A	B	C	D	E	F	G	H	I	J	K	L
1	nil_0 hrs	lepo_IL3_dex_18_hfc.significant.	tt.significant.	CL20030502101	CL20030506	CL2003050	CL2003050	CL2003050	CL20030502306AA.CEL	present	
2	200862_at	5.55915562	7.352864521	-1.793708901	0.000806203	P	A	A	P	P	P
3	200904_at	8.308101509	6.397514865	1.910586644	0.000314517	P	P	P	P	P	
4	200905_x_at	9.930015548	8.669094091	1.260921457	0.000305992	P	P	P	P	P	
5	200971_s_at	9.743958768	8.593903868	1.1500549	0.000418943	P	P	P	P	P	
6	201012_at	9.788299698	8.424886035	1.363413664	0.000952912	P	P	P	P	P	
7	201041_s_at	10.01174531	5.849235804	4.162509511	4.00E-05	P	P	P	P	P	
8	201360_at	7.524044075	5.917660659	1.606383415	0.000855113	P	P	P	P	P	
9	201812_s_at	9.321968111	7.289202925	2.032765186	0.000185194	P	P	P	P	P	
10	202021_x_at	11.00820138	9.140023871	1.868177511	3.04E-05	P	P	P	P	P	
11	202388_at	8.25546873	4.643970298	3.611498432	0.000317081	P	P	P	A	A	
12	202768_at	10.43976304	5.279895504	5.159887533	0.000115297	P	P	P	A	A	
13	204070_at	6.940277819	5.237179721	1.703098099	0.000289908	P	P	P	P	A	
14	204805_s_at	8.574225043	7.187072639	1.387152404	0.000436175	P	P	P	P	P	
15	206310_at	8.693160047	6.388201694	2.304958353	0.000528228	P	P	P	P	A	
16	207132_x_at	8.739402951	7.310947824	1.428455127	0.000803924	P	P	P	P	P	
17	207341_at	10.22849889	7.783356406	2.445142482	0.000347142	P	P	P	P	P	
18	208746_x_at	10.113322	8.620145428	1.493176773	0.000766292	P	P	P	P	P	
19	209189_at	9.176001553	4.36956538	4.806436173	0.000162173	P	P	P	A	A	
20	209608_s_at	5.689552595	7.279811183	-1.590258588	0.000417486	P	P	M	P	P	
21	210140_at	9.059510256	7.386927403	1.672582853	0.000925094	P	P	P	P	P	
22	211031_s_at	6.17661994	5.149998503	1.026621436	0.000223973	P	P	P	P	A	
23	212130_x_at	11.02404725	9.22017048	1.803876766	0.000127835	P	P	P	P	P	
24	212227_x_at	11.07228872	9.197209948	1.87507877	0.000153485	P	P	P	P	P	
25	213549_at	6.37653649	4.754540679	1.621995811	0.000520199	P	P	P	P	P	
26	215946_x_at	7.516089964	5.681806561	1.834283402	0.000327864	P	P	P	P	P	
27	217801_at	9.361634704	8.01160848	1.350026224	0.000895374	P	P	P	P	P	
28	218213_s_at	9.222334173	7.568130103	1.65420407	0.00087139	P	P	P	P	P	
29	218225_at	7.284430491	5.829672418	1.454758073	0.000836052	P	P	P	P	A	
30	218723_s_at	8.746841877	4.729958269	4.016883608	0.000137355	P	P	P	M	A	
31	221841_s_at	7.901052277	3.866107556	4.03494472	0.000263778	P	P	P	A	A	
32	221875_x_at	9.250644009	8.220811928	1.029832081	0.00024265	P	P	P	P	P	
33	59375_at	7.288026846	6.026737786	1.26128906	0.000205622	P	P	P	P	P	

Array quality control

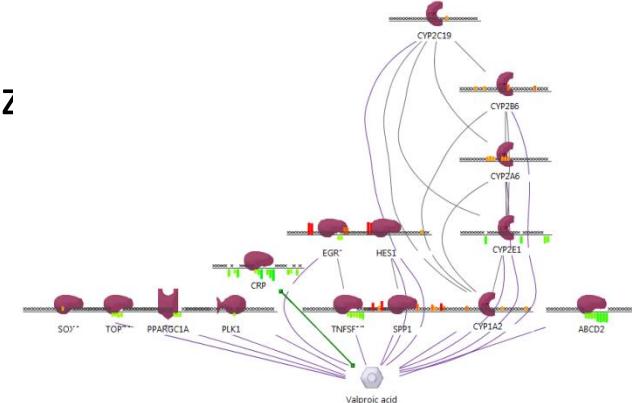
- There are many, many, many publications about this, which leaves you with the feeling that one will never get it right without a lot of work
- Fortunately, however, there are a number of established standards for correcting spot intensities (e.g. MAS5, RMA) and most array companies provide internal standards
- Simpleaffy has a simple scheme for judging array quality

```
x.mas5 <-  
call.exprs(x,"mas5");qcs <- qc(x,x.mas5);  
plot(qcs)
```



In the end, having used Bioconductor, we get a gene list

- You start with 12 CEL files (4 different biological samples in triplicate) with circa 18,000 individual gene and circa 70,000 individual expression (probe) measurements
- This is reduced by statistical measures down to 126 genes with significantly altered expression, and with limited experimental noise.
- These genes are then what one can use to (say) study pathways, integrate with other datasets, etc.
- Broadly speaking, this is how all these approaches work – one needs to process and combine raw data into easier to understand, significant changes in one condition relative to another.
- And we aren't expecting you to have memorized this was just an illustration



And if you want to learn programming...

- DO NOT SIT AND READ A TEXTBOOK.
- Find a real problem (e.g. processing some raw data) and try to do that.
- Learn on the job.

(A lot of students ask this question, so I'm pre-empting it here)

What I hope you have got from this lecture

- How to process gene-lists, the basics
- Gene ontology
- Pathways
- Networks (e.g. physical associations, functional associations)
- The concept of raw data and some basic ideas of how one processes it (e.g. what is Bioconductor)

Thank you.