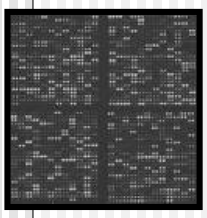
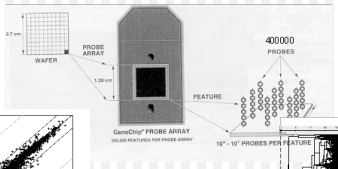


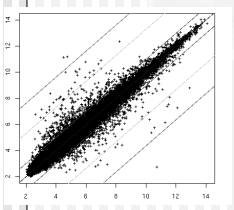
Microarray Informatics

Donald Dunbar

MSc Seminar

30th January 2008





Aims

- To give a biologist's view of microarray experiments
- To explain the technologies involved
- To describe typical microarray experiments
- To show how to get the most from and experiment
- To show where the field is going

Introduction

- Part 1
  - Microarrays in biological research
  - A typical microarray experiment
  - Experiment design, data pre-processing
- Part 2
  - Data analysis and mining
  - Microarray standards and resources
  - Recent advances

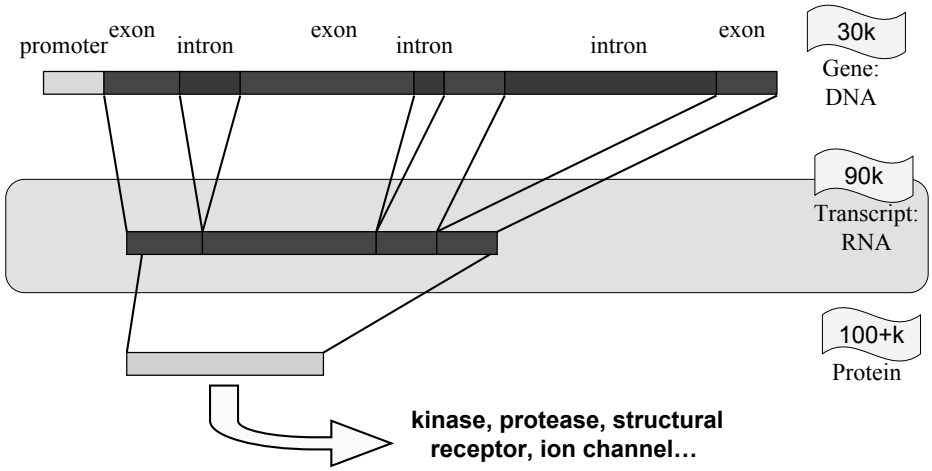
Microarray Informatics

Part 1

Biological research

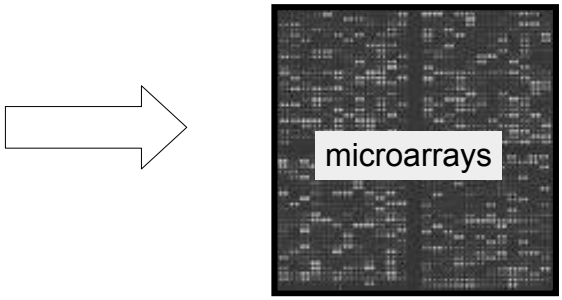
- Using a wide range of experimental and computational methods to answer biological questions
- Genetics, physiology, molecular biology…
- Biology and informatics → bioinformatics
- Genomic revolution
- What can we measure?

The central dogma



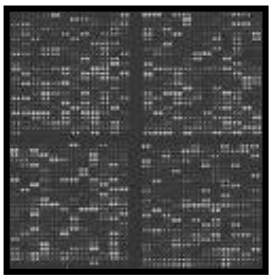
Measuring transcripts

- Genome level sequencing
- New miniaturisation technologies
- Better bioinformatics

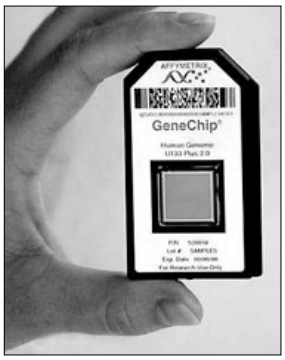


Microarrays: wish list

- Include all genes in the genome
- Include all splice variants
- Give reliable estimates of expression
- Easy to analyse
  - bioinformatics tools available
- Cost effective

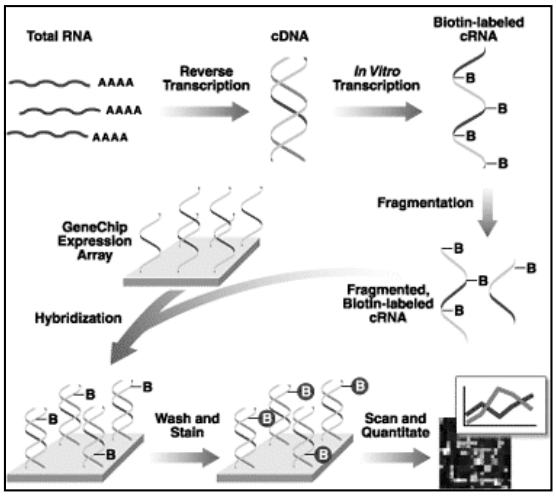


Microarray technologies - 1



- Oligonucleotides - Affymetrix
- One chip all genes
- Chips for many species
- Several oligos per transcript
- Use of control, mismatch sequences
- One sample per chip
  - absolute quatification
- Well established in research
- Expensive

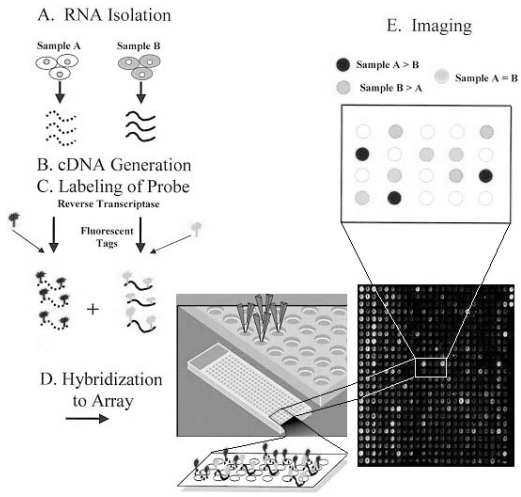
Microarray technologies - 1



Microarray technologies - 2

- cDNAs - Agilent
- One chip all genes
- Chips for many species
- One cDNA per transcript
- Two samples per chip
  - quantification of ratios
- Established in research
- Expensive

Microarray technologies - 2



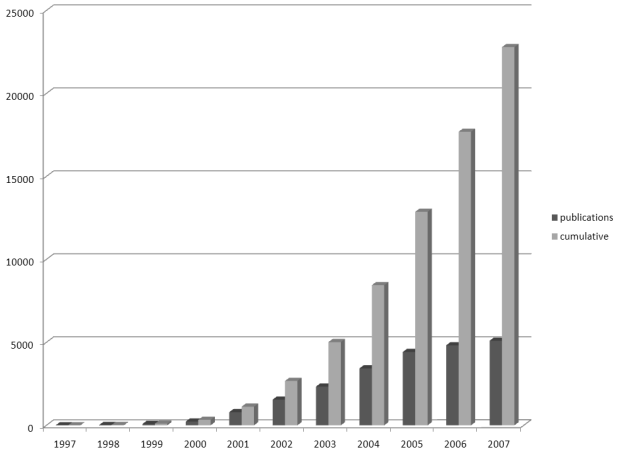
Problems with transcriptomics

- The gene might not be on the chip
- Can't differentiate splice variants
- The gene might be below detection limit
- Can't differentiate RNA synthesis and degradation
- Can't tell us about post translational events
- Bioinformatics can be difficult
- Relatively expensive

History of Microarrays

- Developed in early 1990s after larger macro-arrays (100-1000 genes)
- Microarrays were spotted on glass slides
- Labs spotted their own (Southern, Brown)
- Then companies started (Affymetrix, Agilent)
- Some early papers:
  - *Int J Immunopathol Pharmacol.* 1990 19(4):905-914. Raloxifene covalently bonded to titanium implants by interfacing with (3-aminopropyl)-triethoxysilane affects osteoblast-like cell gene expression. Bambini et al
  - *Nature* 1993 364(6437): 555-6 Multiplexed biochemical assays with biological chips. Fodor SP, et al
  - *Science* 1995 Oct 20;270(5235):467-70 Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Schena M, et al

Microarray publications



Types of experiment

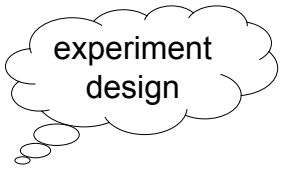
- Usually **control** v **test(s)**

Placebo	Drug treatment Drug 2…
Wild-type	Knockout
Healthy	Patient
Normal tissue	Cancerous tissue
Time = 0	Time = 1 Time = 2…

Types of experiment

- Usually **control v test(s)**
- But also **test v test(s)**
- Comparison:
  - placebo v drug treatment
  - drug 1 v drug 2
  - tissue 1 v tissue 2 v tissue 3 (pairwise)
  - time 0 v time 1, time 0 v time 2, time 0 v time 3
  - time 0 v time 1, time 1 v time 2, time 2 v time 3

A typical experiment

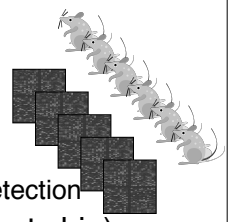


Experiment design: system

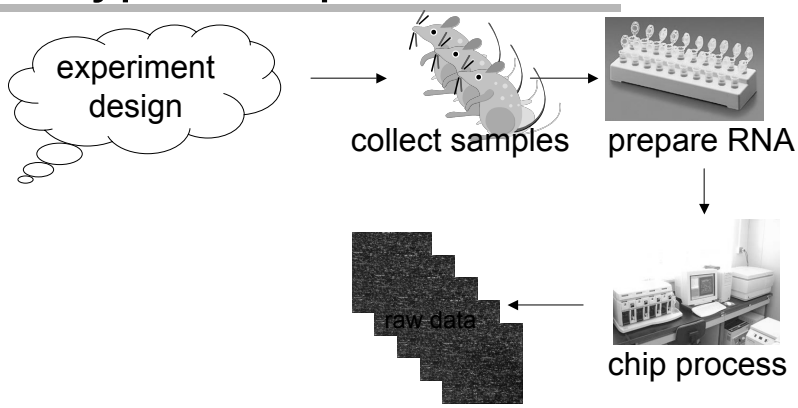
- What is your model?
  - animal, cell, tissue, drug, time…
- What comparison?
- What platform
  - microarray? oligo, cDNA?
- Record all information: see “standards”

Experiment design: replicates

- Microarrays are noisy: need extra confidence in the measurements
- We usually don’t want to know about a specific individual
  - eg not an individual mouse, but the strain
  - although sometimes we do (eg people)
- Biological replicates needed
  - independent biological samples
  - number depends on variability and required detection
- Technical replicates (same sample, different chip) usually not needed



A typical experiment



Raw data

- Affymetrix GeneChip process generates:
  - DAT image file
  - CEL raw data file
  - CDF chip definition file
- Processing then involves CEL and CDF
- Will use Bioconductor

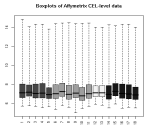
Bioconductor (BioC)

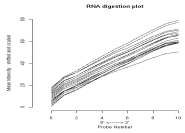
- http://www.bioconductor.org/
- “Bioconductor is an open source software project for the analysis and comprehension of genomic data”
- Started 2001, developed by expert volunteers
- Built on statistical programming environment “R”
- Provides a wide range of powerful statistical and graphical tools

- Use BioC for most microarray processing and analysis
- Make experiment design file and import data

Quality control (QC)

- Affymetrix gives data on QC
  - the microarray team will record these for you
  - scaling factor, % present, spiked probes, internal controls
- Bioconductor offers:
  - boxplots and histograms of raw and normalised data
  - RNA degradation plots
  - specialised quality control routines (eg in simpleaffy)





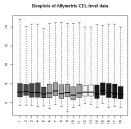
Pre-processing: background

- Signal corresponds to expression…
  - plus a non-specific component (noise)
- Non specific binding of labelled target
- Need to exclude this background
- Several methods exist
  - eg Affy: PM-MM but many complications
  - eg RMA PM=B+S (don't use MM)

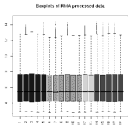
Pre-processing: normalisation

- In addition to background corrections
  - chip, probe, spatial, intra and inter- [illegible]ation
  - need to remove to get at real ex[illegible] differences
- Make use of
  - statistics

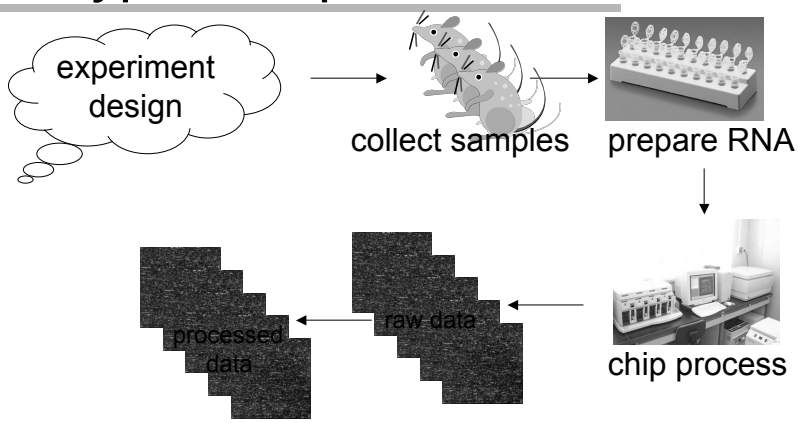
combined with [illegible]set summary:
get an expres[illegible]alue for the gene



- But seems to be no[illegible] dependency on intensity
  - additive and m[illegible]ive errors
- Quantile norm[illegible]n often used
- Normalisati[illegible]e complicated for 2-colour arrays
- Try to re[illegible]ost noise at lab stage (ie control things well statistical[illegible]



A typical experiment



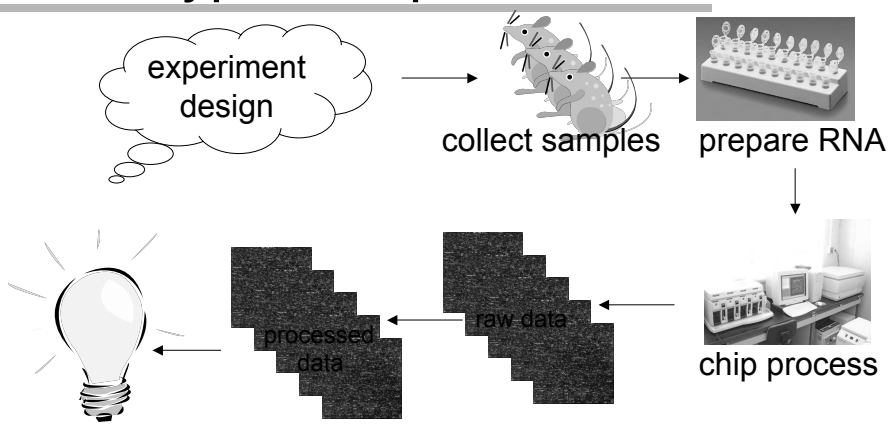
Part 1 Summary

- Microarrays in biological research
- Two types of microarray
- A typical microarray experiment
- Experiment design
- Data pre-processing

Microarray Informatics

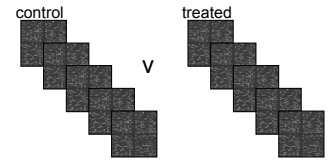
Part 2

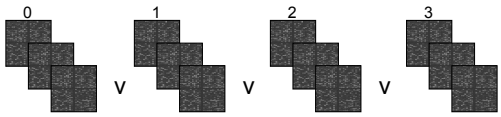
A typical experiment



Data analysis

- Identifying differential expression
- Compare control and test(s)
  - t-test
  - ANOVA
  - SAM (FDR)
  - Limma
- Time series





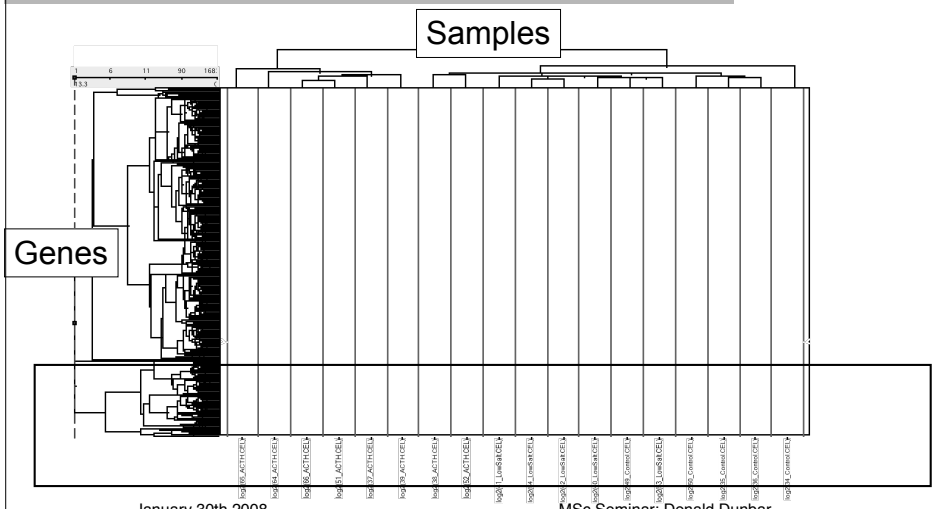
Multiple testing

- Problem:
  - statistical testing of 30,000 genes
  - at $\alpha = 0.05 \rightarrow 1500$ genes
- Need to correct this
  - Multiply p-value by number of observations
    - Bonferroni, too conservative
  - False discovery
    - defines a q value: expected false positive rate
    - Less conservative, but higher chance of type I error
    - Benjamini and Hochberg
- Then regard genes as differentally expressed

Hierarchical clustering

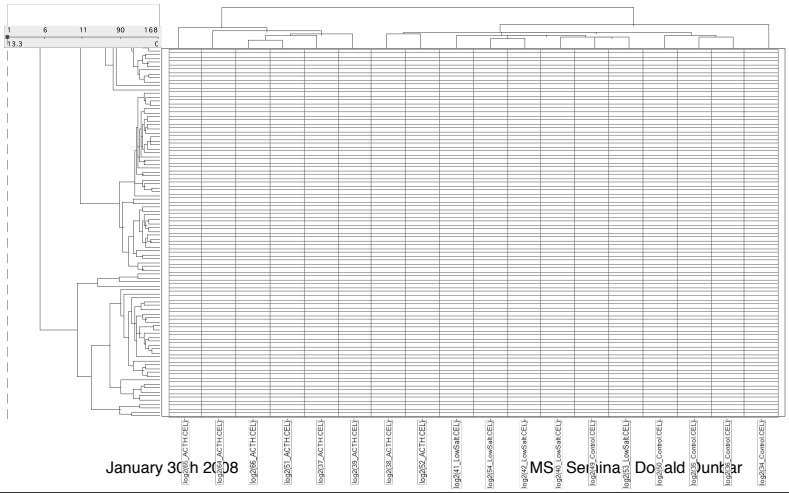
- Look for structure within dataset
  - similarities between genes
- Compare gene expression profiles
  - Euclidian distance
  - Correlation
  - Cosine correlation
- Calculate with distance matrix
- Combine closest, recalculate, combine closest… (or split!)
- Draw dendrogram and heatmap

Hierarchical clustering



Hierarchical clustering

- Heatmaps for microarray data



Hierarchical clustering

- Predicting association of known and novel genes
- Class discovery in samples: new subtypes
- Visualising structure in data (sample outliers)
- Classifying groups of genes
- Identifying trends and rhythms in gene expression
- Caveat: you will always see clusters, even when they are not particularly meaningful

Sample classification

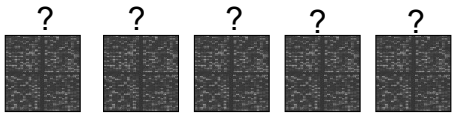
- Supervised or non-supervised
- Non-supervised
  - like hierarchical clustering of samples
- Supervised
  - have training (known) and test (unknown) datasets
  - use training sets to define robust classifier
  - apply to test set to classify new samples

Sample classification

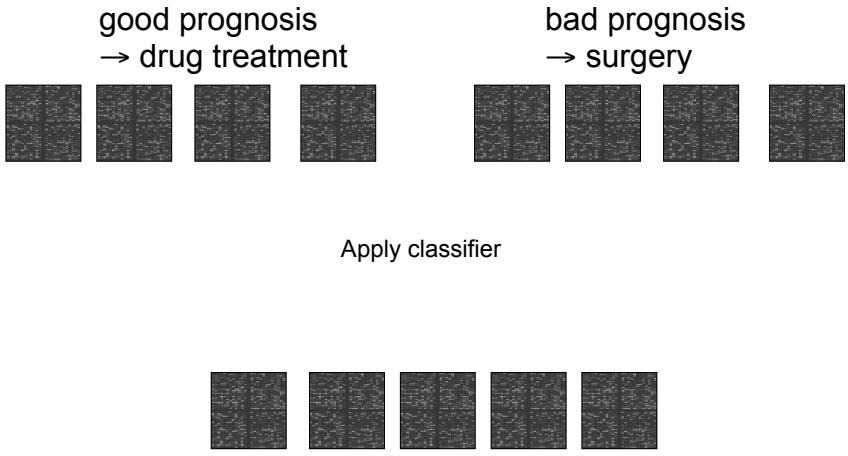
good prognosis
→ drug treatment

bad prognosis
→ surgery

Gene selection, training, cross validation →
classifier: gene x * 0.5 gene y * 0.25 gene z …



Sample classification



Sample classification

- Class prediction for new samples
  - cancer prognosis
  - pharmacogenomics (predict drug efficacy)
- Need to watch for overfitting
  - using too much of the data to classify

Annotation

- Big problem for microarrays
- Genome-wide chips need genome-wide annotation
- Good bioinformatics essential
  - use several resources (Affymetrix, Ensembl)
  - keep up to date (as annotation changes)
  - genes have many attributes
    - name, symbol, gene ontology, pathway…

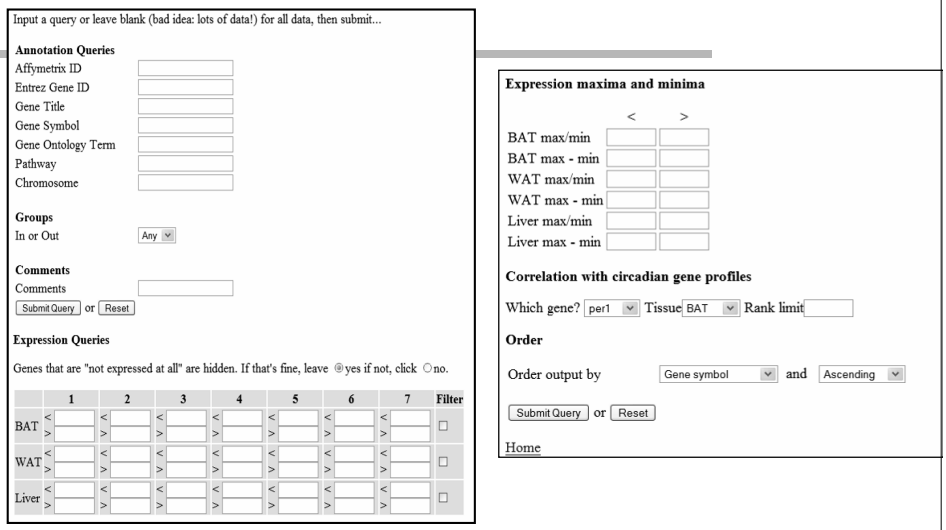
Data-mining

Microarrays are a waste of time

…unless you do something with the data

Data-mining

- Once data are statistically analysed:
  - pull out genes of interest
  - pull out pathways of interest
  - mine data based on annotation
    - what are the expression patterns of these genes
    - what are the expression patterns in this pathway
  - mine genes based on expression pattern
    - what types of genes are up-regulated …
    - fold change, p-value, expression level, correlation



Affy ID	Entrez Gene ID	Gene title	Gene Symbol	Intensities														Group	Comments Update
					1A	2A	3A	4A	5A	6A	7A	1B	2B	3B	4B	5B	6B	7B	
1449851_at	18626	period homolog 1 (Drosophila)	Per1 and (iHOP)	BAT	413	331	1035	1466	255	328	102	32	475	856	853	539	655	178	In / Out
				WAT	795	786	1983	1371	1127	974	412	612	595.1	1422	1260	868	851	767	All
				Liver	27	120	893	1041	661	444	345	32	70	723	1040	432	729	22	Details
1460662_at	18628	period homolog 3 (Drosophila)	Per3 and (iHOP)	BAT	262	227	1026	1689	710	383	711	213	610	1052	1419	845	346	14	In / Out
				WAT	277	422	1153	1060	782	679	386	272	457.3	1090	1099	1054	278	281	All
				Liver	181	177	574	740	281	145	109	59	460	476	627	231	433	244	Details
1434736_at	217082	hepatic leukemia factor	Hlf and (iHOP)	BAT	1850	2080	3422	4836	2723	2240	1254	1692	1826	2992	5122	2760	2172	1210	In / Out
				WAT	1028	1142	2039	3374	2756	1550	1317	1262	1012.6	2226	3409	2325	1958	946	All
				Liver	2115	3098	4809	4693	3304	3166	2083	2354	3224	4709	4126	3357	3221	2184	Details
1450184_s_at	21685	thyrotroph embryonic factor	Tef and (iHOP)	BAT	680	1209	2240	2505	1056	843	820	300	1151	2074	2833	1208	616	668	In / Out
				WAT	489	690	450	1010	746	530	254	374	502	393	973	702	466	294	All
				Liver	191	872	1529	1071	673	472	30	310	570	1521	1309	1113	492	393	Details
1422997_s_at	26897	cytosolic acyl-CoA thioesterase 1	Cte1 and (iHOP)	BAT	1493	2117	5054	10100	2774	1921	1849	1704	3137	4645	9366	2804	1893	1774	In / Out
				WAT	384	449	352	836	786	600	620	480	379.7	474	779	803	553	640	All
				Liver	1830	1676	2153	2610	1828	2942	2383	2132	1418	1566	2573	1918	2836	2458	Details
1454786_at	223739	RIKEN cDNA 5031439G07 gene	5031439G07Rik and (iHOP)	BAT	356	301	899	1484	637	971	286	504	572	972	1146	929	618	258	In / Out
				WAT	1181	401	1115	665	806	769	624	500	749	1164	674	983	550	525	All
				Liver	496	546	618	103	236	326	86	497	337	620	594	367	456	137	Details
1428306_at	74747	DNA-damage-inducible transcript 4	Ddit4 and (iHOP)	BAT	1045	1072	1762	2483	844	956	679	855	1024	1485	2950	713	1572	339	In / Out
				WAT	2667	2768	6606	5774	2748	2720	2816	2939	3069.6	7661	8532	2342	2859	3096	All
				Liver	749	678	465	873	1166	1317	1192	824	694	559	809	1261	1408	1249	Details
1452071_at	54403	solute carrier family 4 (anion exchanger), member 4	Slc4a4 and (iHOP)	BAT	5086	7282	9878	12544	6827	4035	3147	5317	7144	10339	12090	7221	3976	3195	In / Out
				WAT	1107	1313	1566	2149	1761	1144	1525	1139	1406.8	1294	2039	1714	1261	1542	All
				Liver	2875	3183	2492	2381	1094	1891	1728	2532	3085	3064	1935	1449	1802	2206	Details
		cadherin EGF LAG	Celsr2 and	BAT	34	209	870	478	146	870	47	59	216	585	236	265	346	126	In / Out

Detailed information: 1449851_at

Annotation Data for Per1

Affymetrix ID	1449851_at
Entrez Gene ID	18626
Gene Title	period homolog 1 (Drosophila)
Gene Symbol	Per1
Ensembl ID	ENSMUSG00000020893
GO Biological Process information	"6355 // regulation of transcription, DNA-dependent // inferred from electronic an transduction // inferred from electronic annotation /// 7622 // rhythmic behavior // annotation /// 7623 // circadian rhy// inferred from electronic annot"
GO Molecular Function information	4871 // signal transducer activity // inferred from electronic annotation
GO Cellular Compartment information	5634 // nucleus // inferred from direct assay
Pathway	Circadian_Exercise /// Circadian_Exercise
Chromosomal location	---
Protein family	---
Protein domain	---
Interpro ID	---
Transmembrane?	---
Group	in
Comments	No comments yet

Expression Data for Per1

Intensities

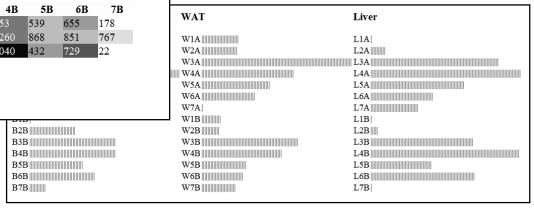
View these data as a graph

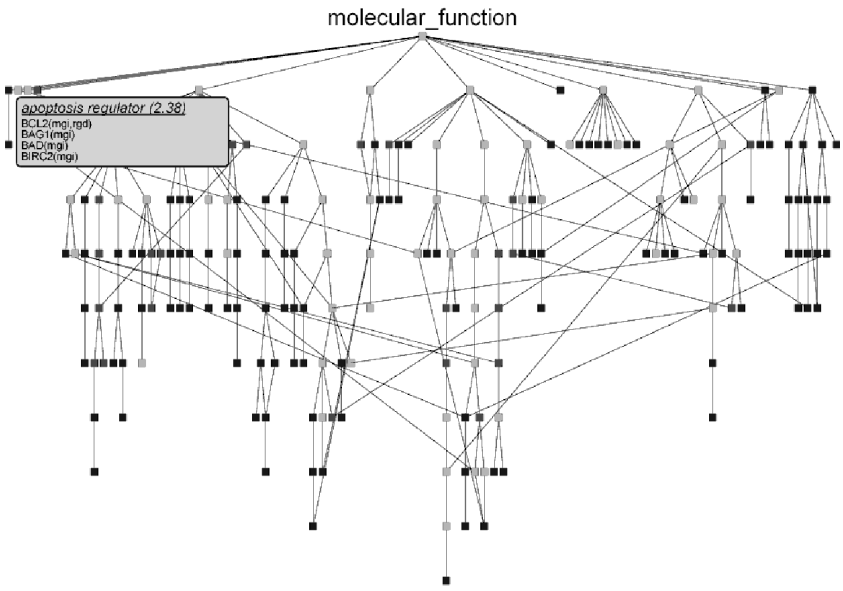
(Per1 meets the general expression criteria)

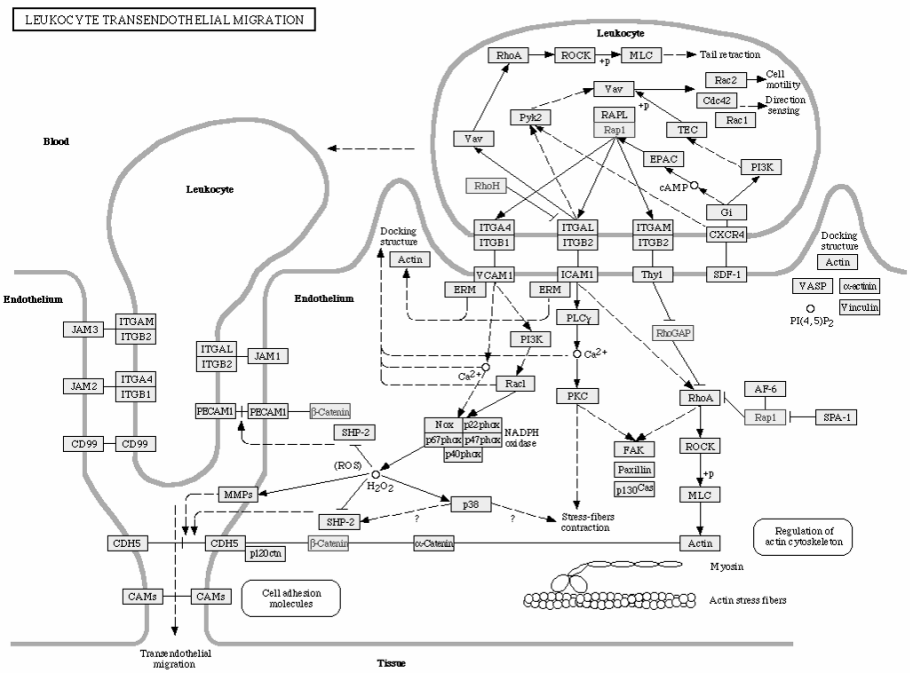
	1A	2A	3A	4A	5A	6A	7A	1B	2B	3B	4B	5B	6B	7B
BAT	413	331	1035	1466	255	328	102	32	475	856	853	539	655	178
WAT	795	786	1983	1371	1127	974	412	612	595.1	1422	1260	868	851	767
Liver	27	120	893	1041	661	444	345	32	70	723	1040	432	729	22

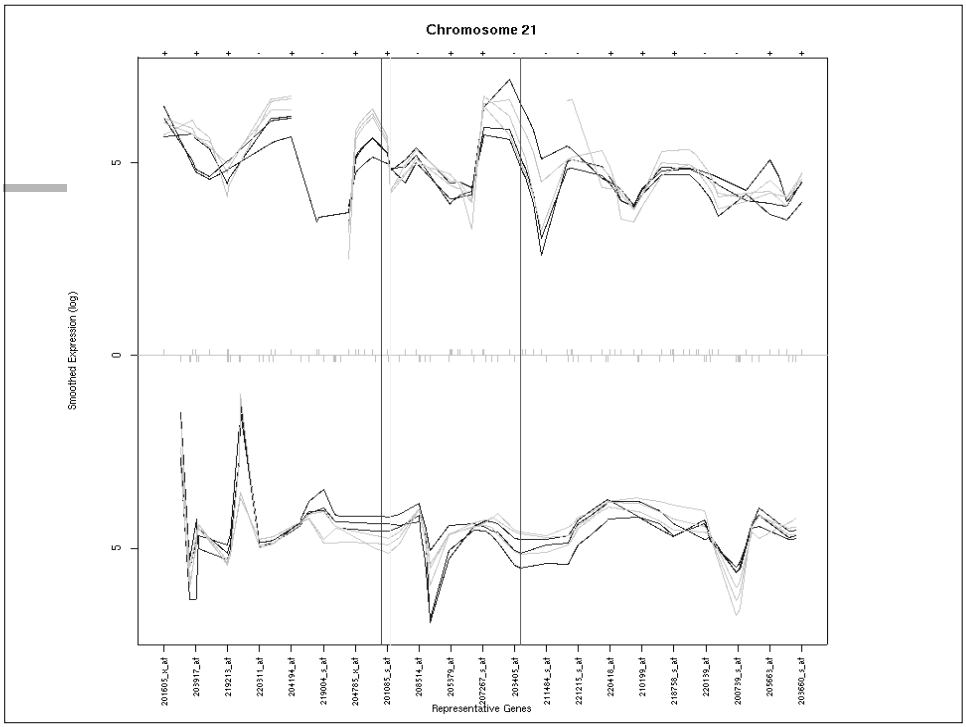
Maxima, minima, ratios, differences

	Maximum	Minimum	Max/Min	Max-Min




molecular_function
apoptosis regulator (2,38)
BCL2(mgi,rgd)
BAG1(mgi)
BAD(mgi)
BIRC2(mgi)


LEUKOCYTE TRANSENDOTHELIAL MIGRATION
Leukocyte
Blood
Endothelium
Tissue
Docking structure
Actin
Cell adhesion molecules
Regulation of actin cytoskeleton
Stress-fibers contraction
Actin stress fibers
Myosin
Trans endothelial migration
NADPH oxidase
Cell motility
Direction sensing



TOUCAN 2

Launch Now	Launch instructions	Tutorials	Manual
Cite	News	Source	License
PWMs	Version	SOAP	Related
Freq. files	Mailing	Features	FAQ
Service mirrors	Screenshots	Acknowledgements	

Introduction

TOUCAN is a workbench for **regulatory sequence analysis on metazoan genomes** : comparative genomics, detection of significant transcription factor binding sites, and detection of *cis*-regulatory modules (combinations of binding sites) in sets of coexpressed/coregulated genes.

It is a platform independent, standalone Java application that is tightly linked with Ensembl, and was built using the BioJava package.SOAP web services are used to remotely access multiple algorithms for comparative genomics, motif detection, and module detection.

Comments, suggestions, and bug reports can be send to stein.aerts<at>med.kuleuven.ac.be or toucan@listserv.cc.kuleuven.ac.be

Register

If you're using TOUCAN, please enter your email adress: Submit

Launch TOUCAN v. 2.2.5

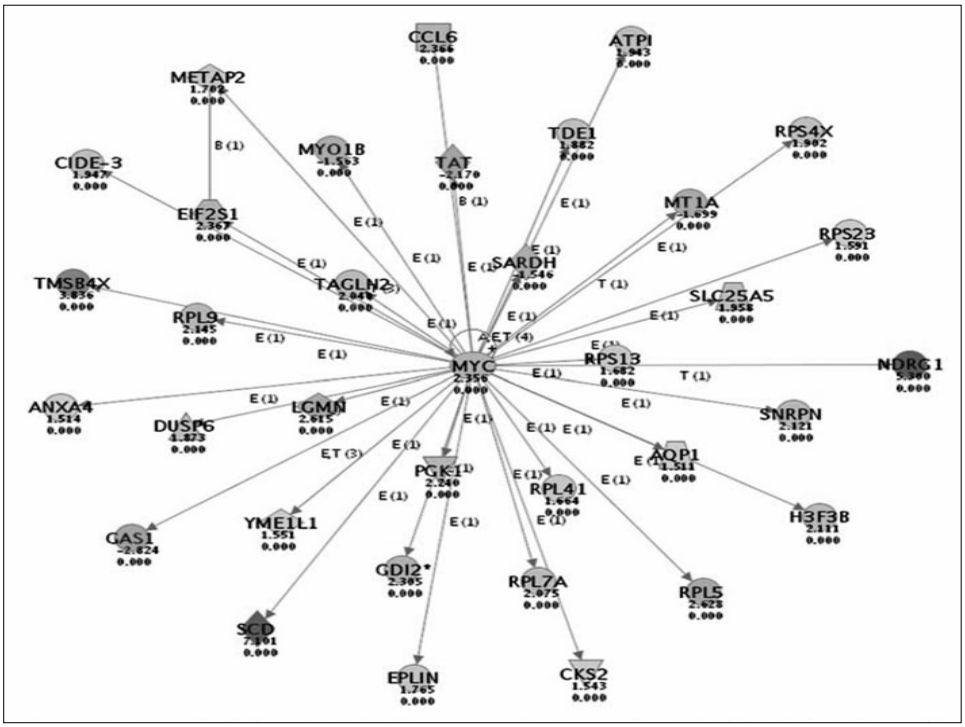
To run TOUCAN you need to have **two things** installed:

- Java 2 Platform, Standard Edition (J2SE), version 1.4.x. or 1.5.0
- Java Web Start which is shipped as part of J2SE. From J2SE 1.4.2 onwards it is installed together with the SDK/JRE (more info)

If you have fulfilled these requirements, then you can launch TOUCAN directly using this URL: http://www.esat.kuleuven.ac.be/~saerts/software/toucan.jnlp

Alternatively you can type this command in a terminal window:
javaws http://www.esat.kuleuven.ac.be/~saerts/software/toucan.jnlp

We try to encourage to use Java Web Start because this way you will always have the latest version of the software. This is important because the properties change at least once a month to follow the newest Ensembl release. If you are really unable to use Java Web Start, you can send us an email, and we can give you the JAR file of TOUCAN.



Further data-mining

- Other tools available using
  - gene ontology (GO)
  - biological pathways (eg KEGG)
  - genomic localisation (Ensembl)
  - regulatory sequence data (Toucan, BioProspector)
  - literature (eg Pubmatrix or our text mining tool)
- … to make sense of the data

Microarray Resources

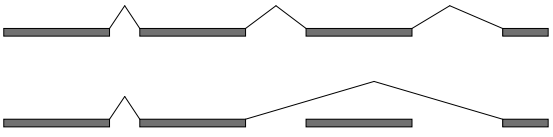
- Microarray data repositories
  - Array express (EBI, UK) 
  - Gene Expression Omnibus (NCBI, USA) 
  - CIBEX (Japan) 
- Annotation
  - NetAffx, Ensembl, TIGR, Stanford…

Microarray Standards

- MIAME
  - Minimum annotation about a microarray experiment
  - Comprehensive description of experiment
  - Models experiments well, and allows replication
    - chips, samples, treatments, settings, comparisons
  - Required for most publications now
- MAGE-ML
  - Microarray gene expression markup language
  - Describes experiment (MIAME) and data
  - Tools available for processing

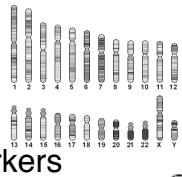
Recent advances: Exon chips

- Affymetrix now have chips that allow us to measure expression of splice variants



New chips will give us much more information

Recent advances: Genotyping chips



- All discussion on EXPRESSION chips
- Also can get chips looking at genotype
- Tell us the sequence for genome-wide markers
- Test 300,000 markers with one chip
- Look for association with disease, prognosis, trait…
- Combined with expression chips to generate
  - EXPRESSION QUANTITATIVE TRAITS LOCI (eQTL)
  - Overlap of expression and genetic differences (cis)
  - Correlation at different locus (trans)

Part 2 Summary

- Data analysis
- Data Mining
- Microarray Resources
- Microarray Standards
- Recent advances

Seminar Summary

- Part 1
  - Microarrays in biological research
  - A typical microarray experiment
- Part 2
  - Data analysis and mining
  - Recent advances

Contact

- Donald Dunbar
- CVS and CIR Bioinformatics
- donald.dunbar@ed.ac.uk
- 0131 242 6700
- Room W3.01, QMRI, Little France
- www.bioinf.mvm.ed.ac.uk