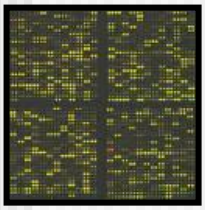
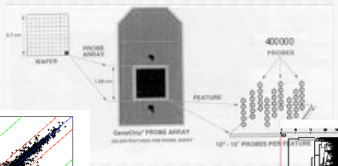


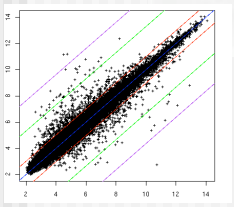
Microarray Informatics

Donald Dunbar

MSc Seminar

3rd February 2010





Aims

- To give a biologist's view of microarray experiments
- To explain some 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

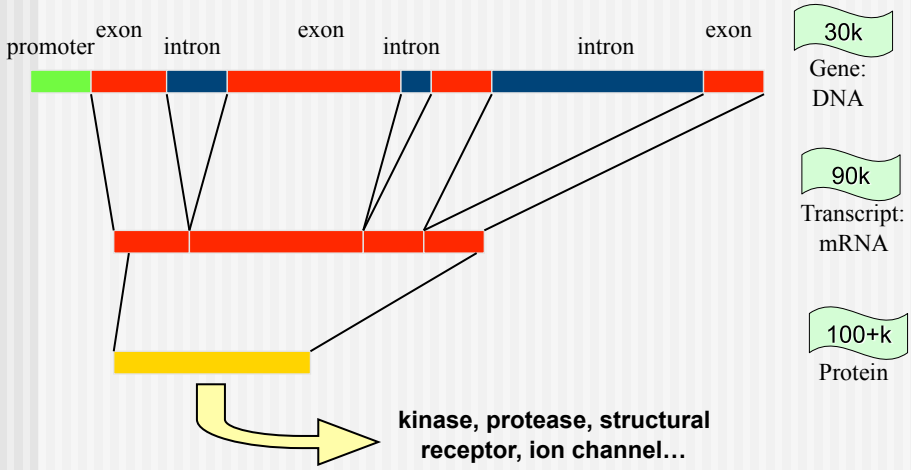
Microarray Informatics

Part 1

Biological research

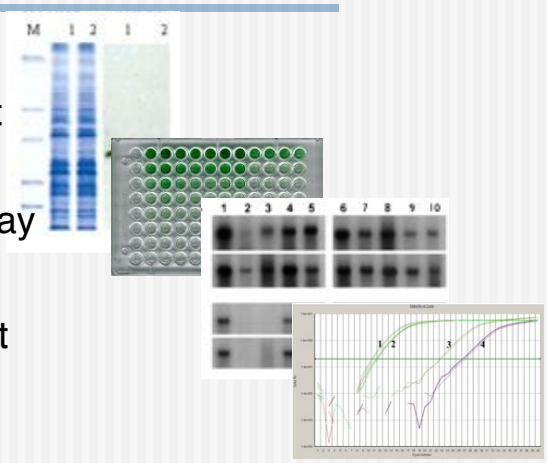
- Using a wide range of experimental and computational methods to answer biological questions
- Genetics, physiology, molecular biology…
- Biology and informatics → bioinformatics
- Genomic revolution
- What can we measure?

The central dogma



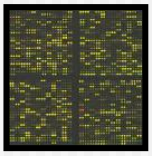
Measuring RNA and proteins

- Proteins
  - Western blot
  - ELISA
  - Enzyme assay
- mRNA
  - Northern blot
  - RT-PCR



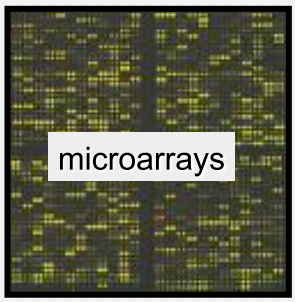
Measuring RNA and proteins

- Protein levels/activities would be best
  - no real high throughput method
- mRNA levels will have to do
  - genome-wide physical microarrays
  - other ‘array-like’ technologies
  - sequencing (see later)



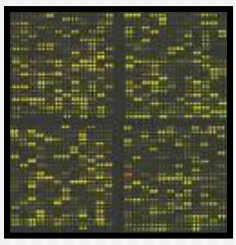
Measuring transcripts

- Genome level sequencing
- New miniaturisation technologies
- Better bioinformatics



Microarrays: wish list

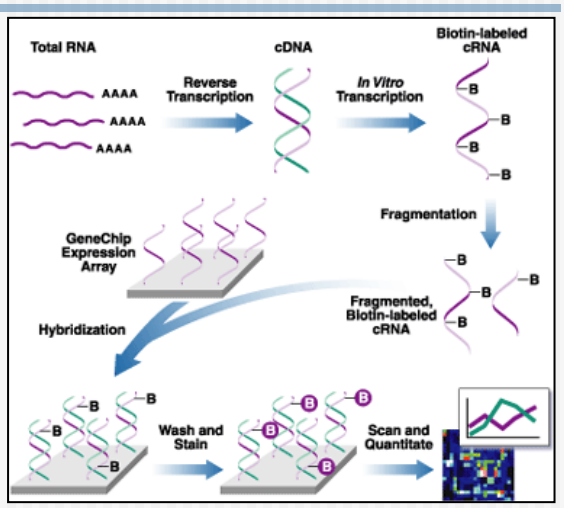
- Include all genes in the genome
- Include all splice variants
- Give reliable estimates of expression
- Easy to analyse
  - bioinformatics tools available
- Cost effective



Microarray technologies - 1

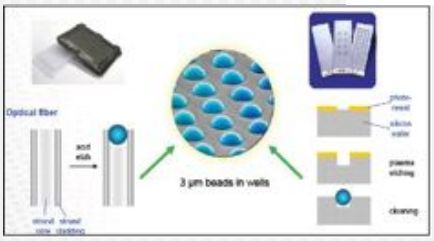
- Oligonucleotides - Affymetrix
- One chip all genes
- Chips for many species
- Several oligos per transcript
- Use of control, mismatch sequences
- One sample per chip
  - ‘absolute quantification’
- Well established in research
- Expensive

Microarray technologies - 1



Microarray technologies - 2

- Illumina BeadChip
- Oligos on beads
- Hybridise in wells
- Compared to Affy
  - Higher throughput
  - Less RNA needed
  - Cheaper



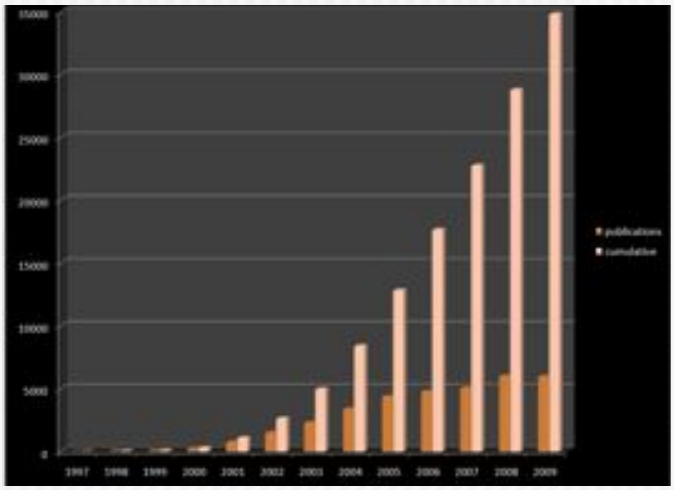
Problems with microarrays

- 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:
  - *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

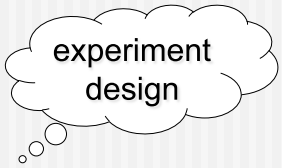
- Usually **control** v **test(s)**

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

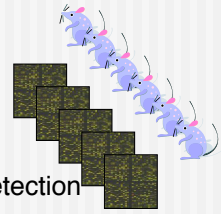


Experiment design: system

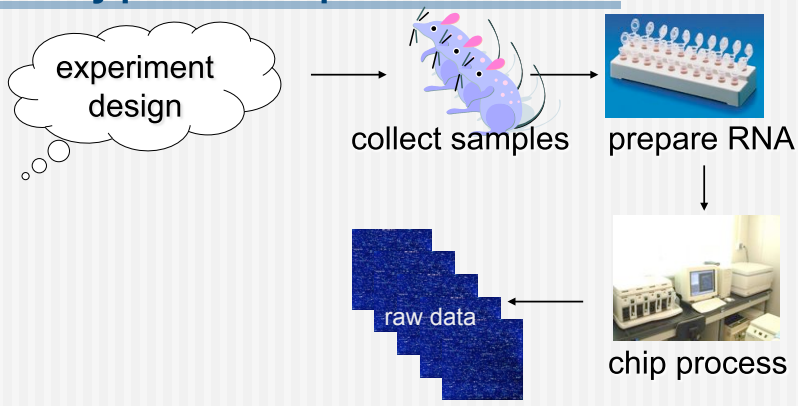
- What is your model?
  - animal, cell, tissue, drug, time…
- What comparison?
- What platform
  - microarray? oligo, cDNA?
- Record all information: see “standards”

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

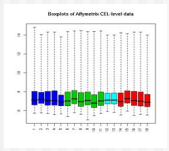
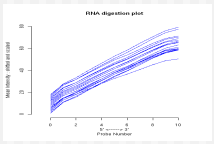
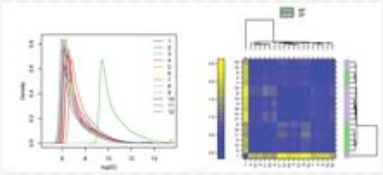
- Affymetrix GeneChip process generates:
  - DAT image file
  - CEL raw data file
  - CDF chip definition file
- Processing then involves CEL and CDF
- Will use Bioconductor

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
- Most platforms now have BioC packages
- Tutorial: manuals.bioinformatics.ucr.edu/home/R_BioCondManual

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 arrayQualityMetrics)

Pre-processing: background

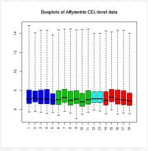
- Signal corresponds to expression…
  - plus a non-specific component (noise)
- Non specific binding of labelled target
- Need to exclude this background
- Several methods exist
  - eg Affy: PM-MM but many complications
  - eg RMA PM=B+S (don't use MM)

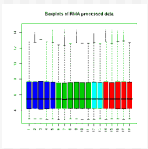
Pre-processing: normalisation

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

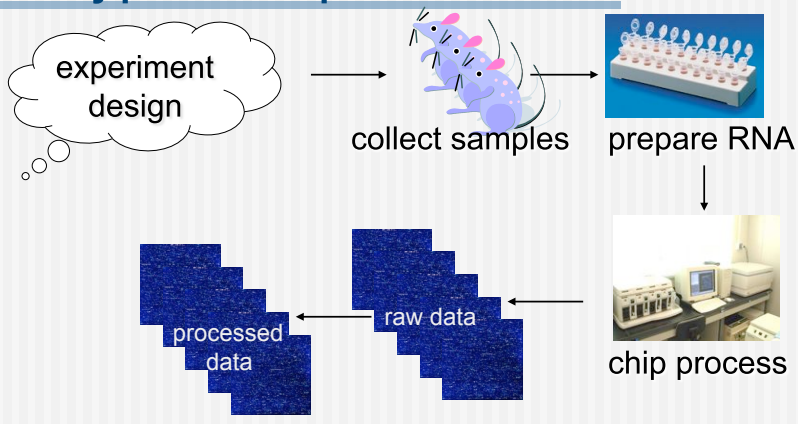
Now carry on with analysis!

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





A typical experiment



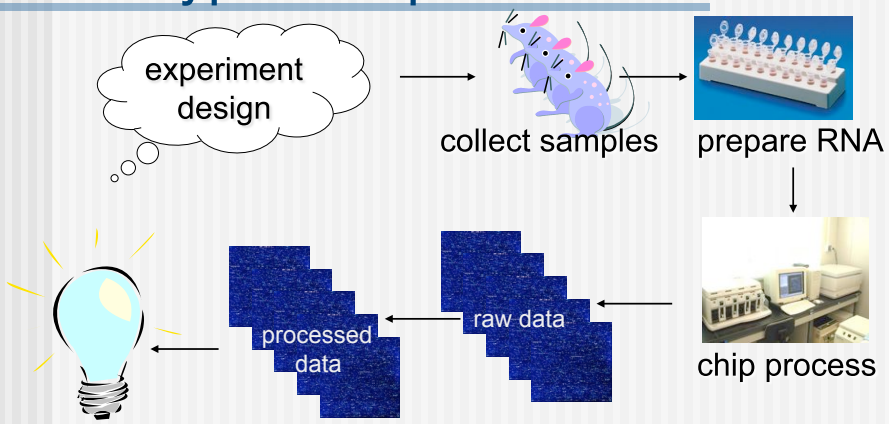
Part 1 Summary

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

Microarray Informatics

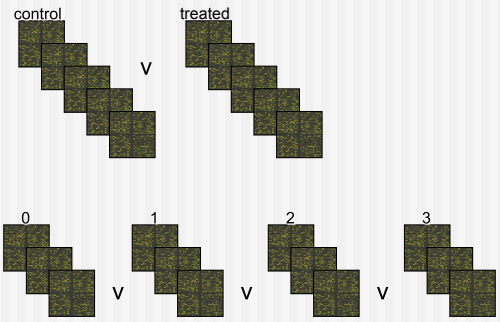
Part 2

A typical experiment



Data analysis

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



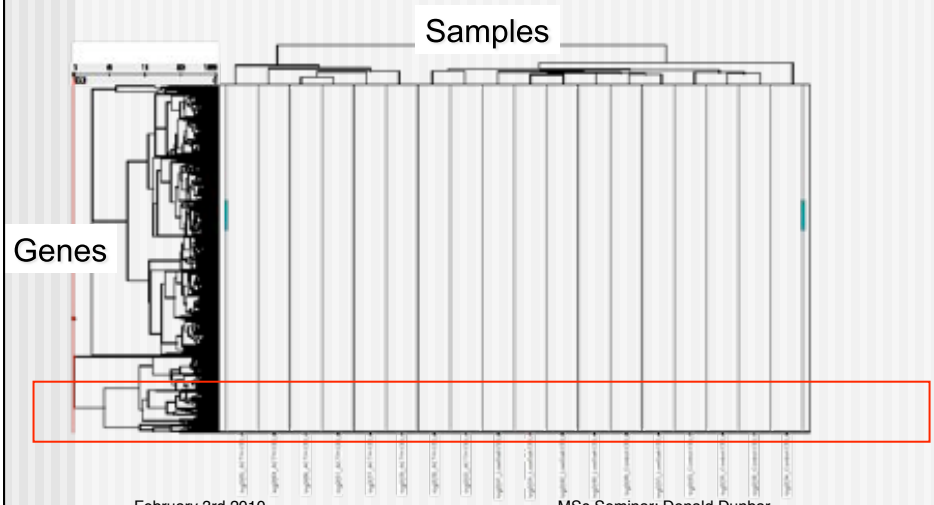
Multiple testing

- 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 differentially expressed
- Depends on follow-up procedure!

Hierarchical clustering

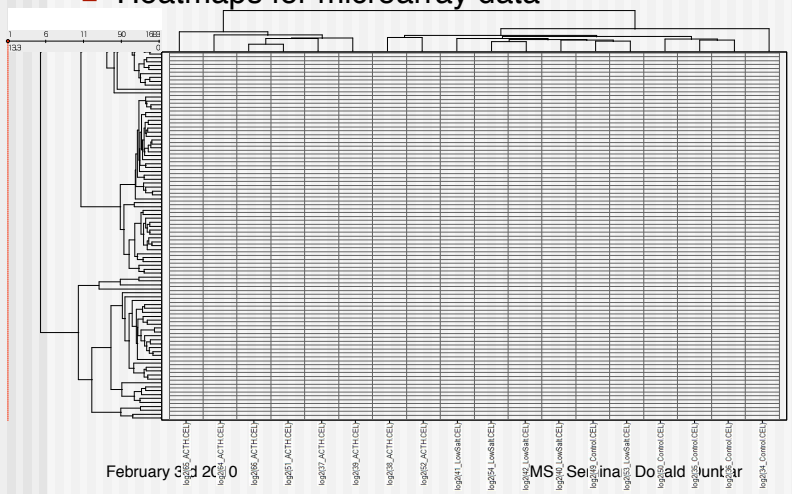
- Look for structure within dataset
  - similarities between genes
- Compare gene expression profiles
  - Euclidian distance
  - Correlation
  - Cosine correlation
- Calculate with distance matrix
- Combine closest, recalculate, combine closest… (or split!)
- Draw dendrogram and heatmap

Hierarchical clustering



Hierarchical clustering

- Heatmaps for microarray data



Hierarchical clustering

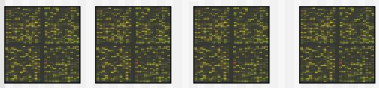
- Predicting association of known and novel genes
- Class discovery in samples: new subtypes
- Visualising structure in data (sample outliers)
- Classifying groups of genes
- Identifying trends and rhythms in gene expression
- Caveat: you will always see clusters, even when they are not particularly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

Sample classification

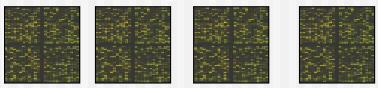
- Supervised or non-supervised
- Non-supervised
  - like hierarchical clustering of samples
- Supervised
  - have training (known) and test (unknown) datasets
  - use training sets to define robust classifier
  - apply to test set to classify new samples

Sample classification

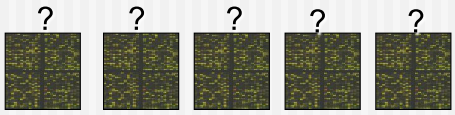
good prognosis
→ drug treatment



bad prognosis
→ surgery

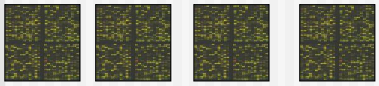


Gene selection, training, cross validation →
classifier: gene x * 0.5 gene y * 0.25 gene z …

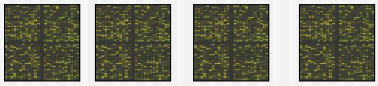


Sample classification

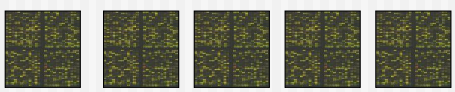
good prognosis
→ drug treatment



bad prognosis
→ surgery



Apply classifier



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
  - classifier loses specificity

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…

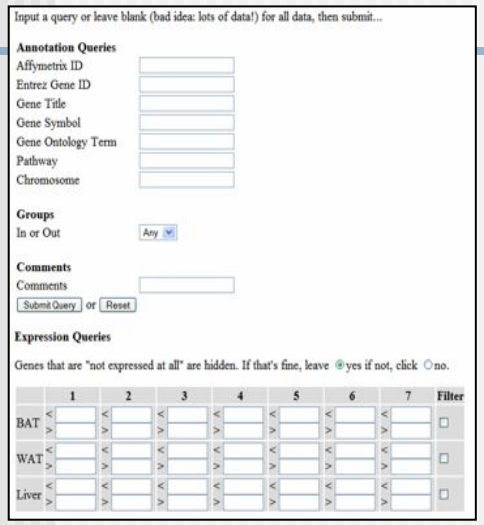
Data-mining

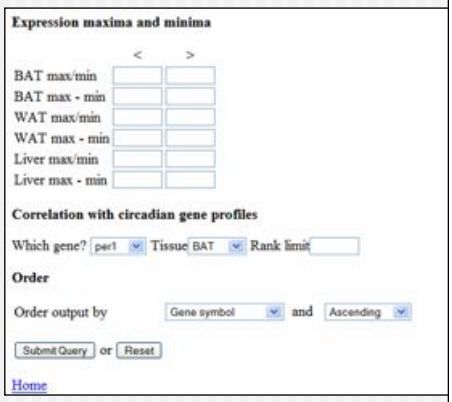
Microarrays are a waste of time

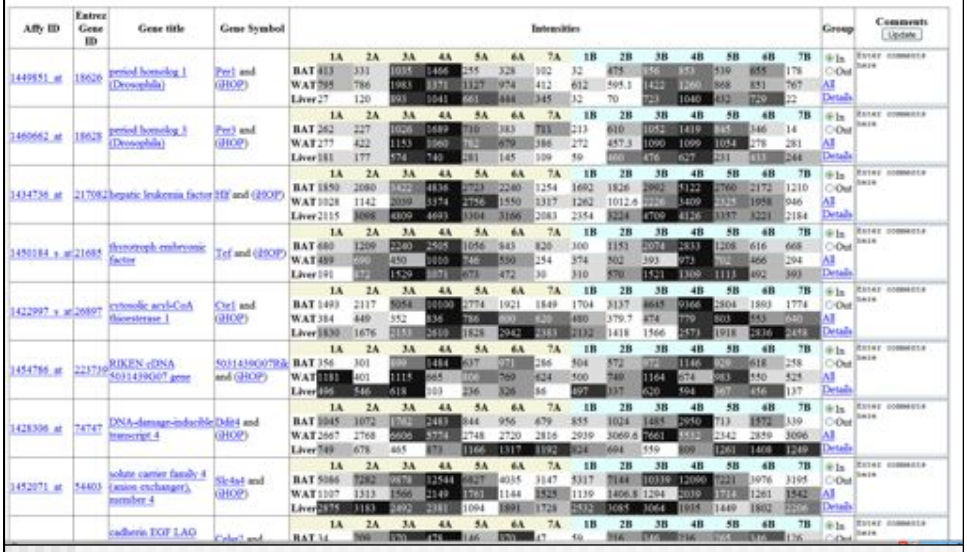
…unless you do something with the data

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
- Should be driven by the biological question


Input a query or leave blank (bad idea: lots of data!) for all data, then submit...
Annotation Queries
Affymetrix ID
Entrez Gene ID
Gene Title
Gene Symbol
Gene Ontology Term
Pathway
Chromosome
Groups
In or Out
Any
Comments
Comments
Submit Query or Reset
Expression Queries
Genes that are "not expressed at all" are hidden. If that's fine, leave yes if not, click no.
1 2 3 4 5 6 7 Filter
BAT
WAT
Liver


Expression maxima and minima
< >
BAT max/min
BAT max - min
WAT max/min
WAT max - min
Liver max/min
Liver max - min
Correlation with circadian gene profiles
Which gene? per1 Tissue BAT Rank limit
Order
Order output by Gene symbol and Ascending
Submit Query or Reset
Home


Affy ID
Entrez Gene ID
Gene title
Gene Symbol
Intensities
1A 2A 3A 4A 5A 6A 7A 1B 2B 3B 4B 5B 6B 7B
BAT
WAT
Liver
Group
Comments
Update
In
Out
All
Details

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 ar 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

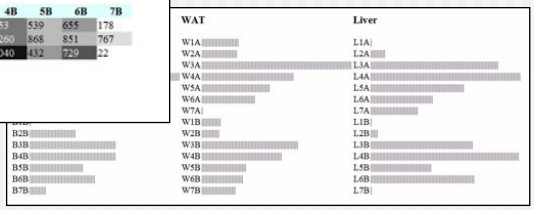
View these data as a graph

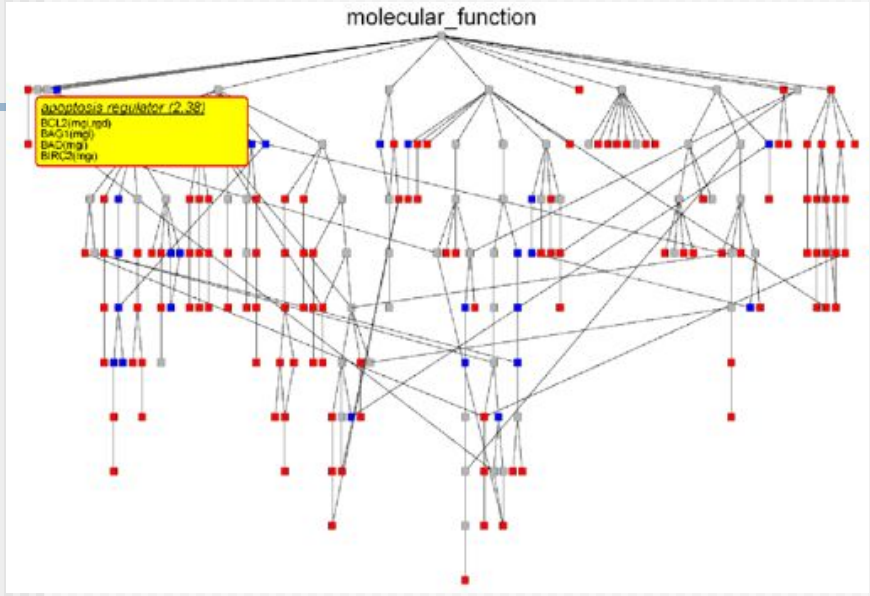
(Per1 meets the general expression criteria)

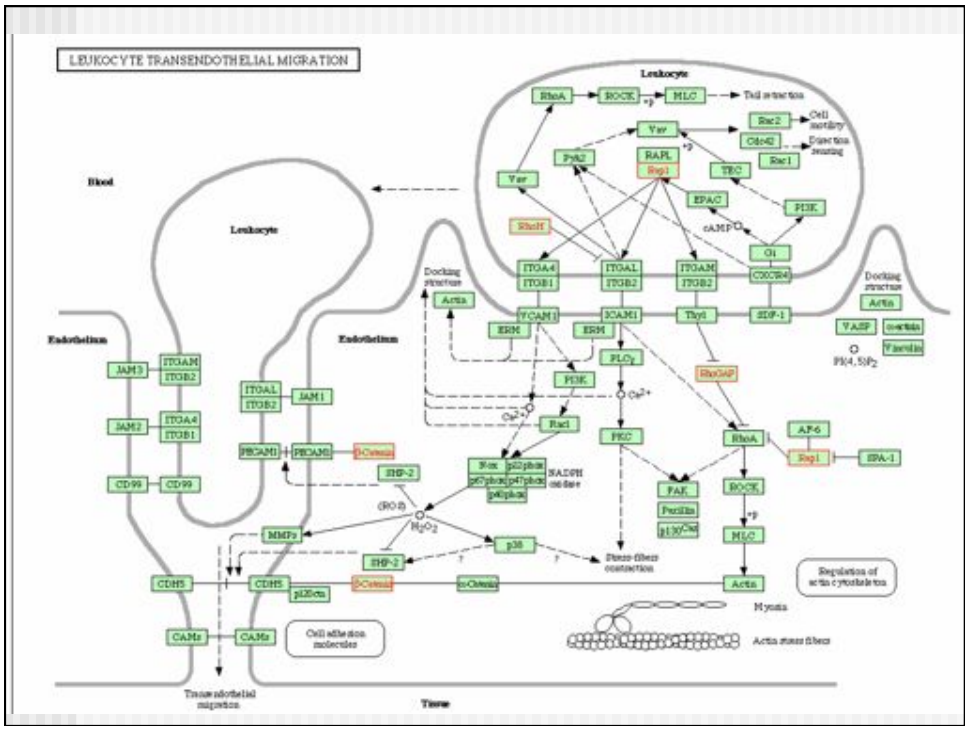
	1A	2A	3A	4A	5A	6A	7A	1B	2B	3B	4B	5B	6B	7B
BAT	413	331	1035	1466	255	328	102	32	475	856	853	539	655	178
WAT	795	786	1983	1371	1127	974	412	612	595.1	1422	1260	868	851	767
Liver	27	120	893	1041	661	444	345	32	70	723	1040	432	729	22

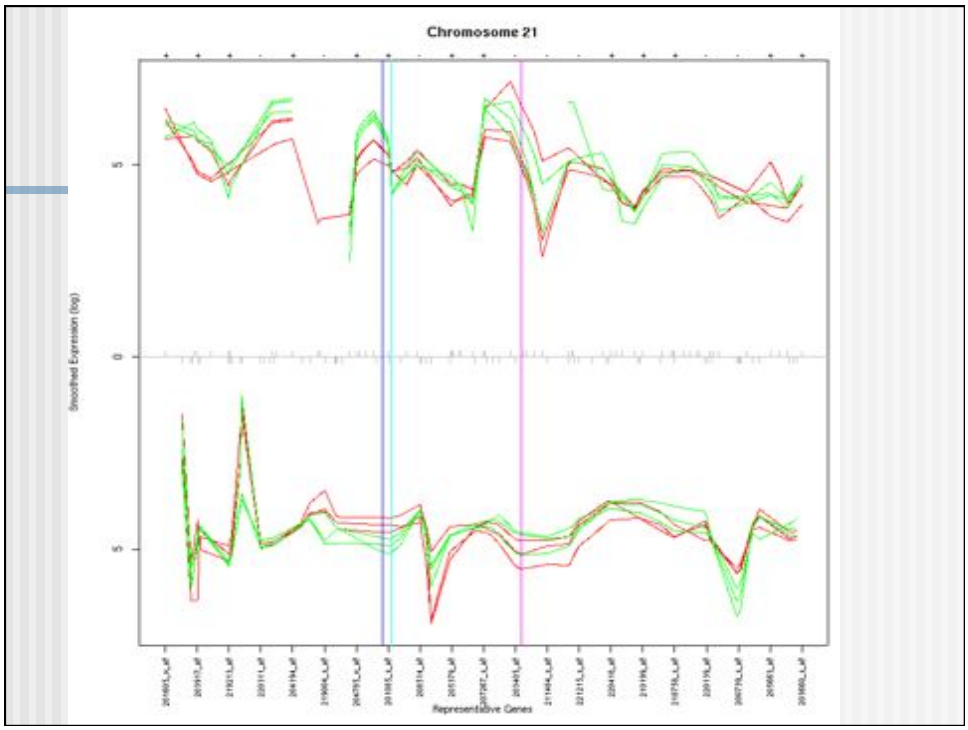
Maxima, minima, ratios, differences

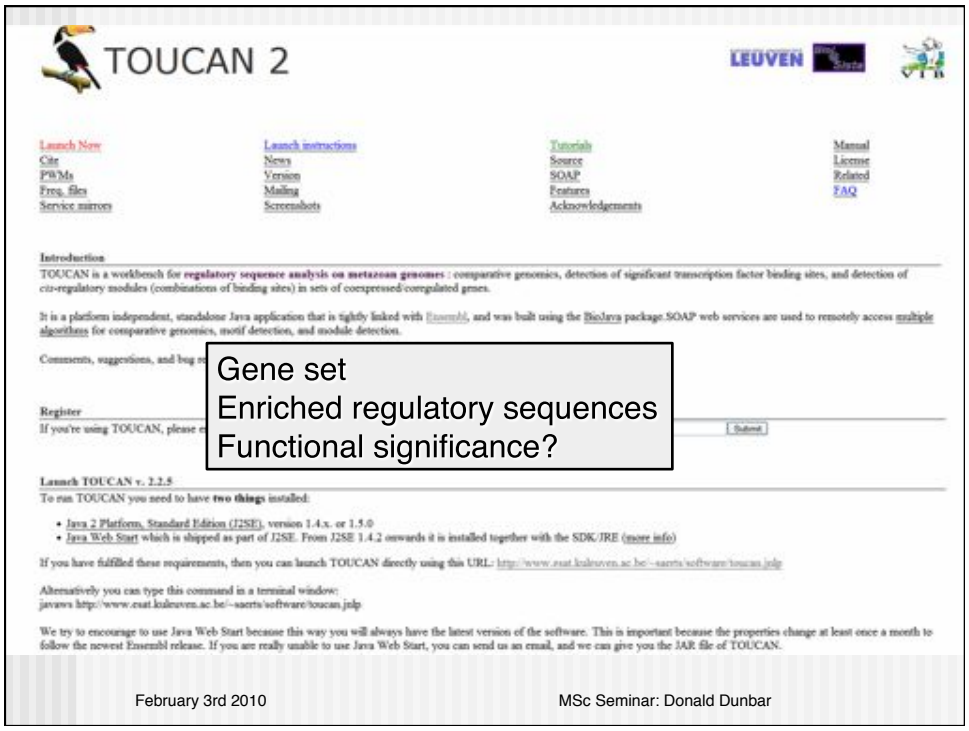
	Maximum	Minimum	Max/Min	Max-Min

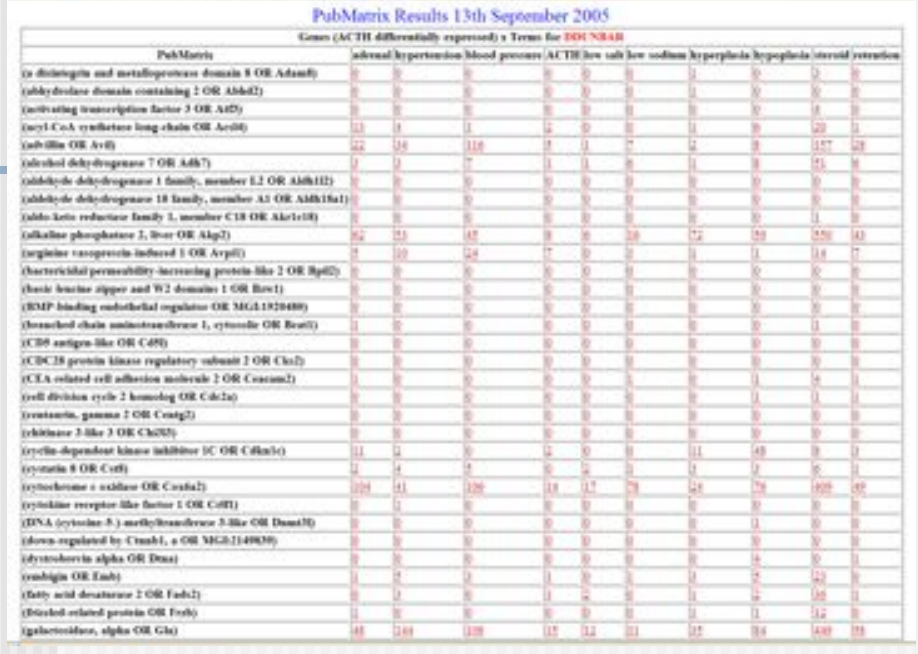


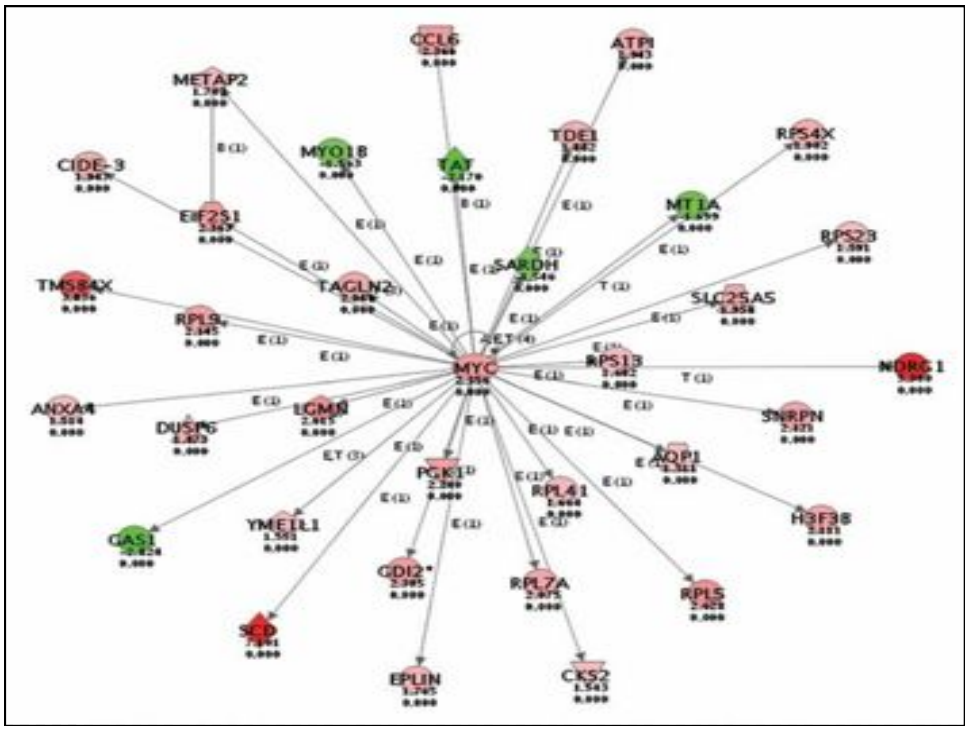



LEUKOCYTE TRANSENDOTHELIAL MIGRATION
Leukocyte
Blood
Endothelium
Tissue
Docking structure
Actin
Cell adhesion molecules
Regulation of actin cytoskeleton
Myosin
Actin stress fibers
Transendothelial migration


Chromosome 21
Smoothed Expression (log)
Representative Genes


TOUCAN 2
LEUVEN
Launch Now
Cite
PWMs
Freq. files
Service mirrors
Launch instructions
News
Version
Mailing
Screenshots
Tutorials
Source
SOAP
Features
Acknowledgements
Manual
License
Related
FAQ
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 r
Gene set
Enriched regulatory sequences
Functional significance?
Register
If you're using TOUCAN, please r
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.


PubMatrix Results 13th September 2005
Genes (ACTH differentially expressed) x Terms for DOCSRAR
PubMatrix
adrenal
hypertension
blood pressure
ACTH
low salt
low sodium
hyperplasia
hypoplasia
steroid
retention



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, Ingenuity…)
- … to make sense of the data
- Links at: www.bioinf.mvm.ed.ac.uk/projects/analysis_tools.html

Microarray Resources

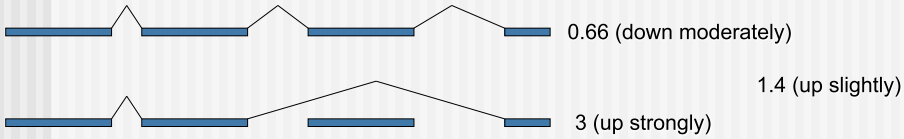
- Microarray data repositories
  - Array express (EBI, UK)
  - Gene Expression Omnibus (NCBI, USA)
  - CIBEX (Japan)
- Annotation
  - NetAffx, Ensembl, TIGR, Stanford…

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

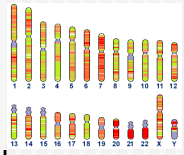
Recent advances: Exon chips

- 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 TRAIT LOCUS (eQTL)
  - Overlap of expression and genetic differences (cis)
  - Correlation at different locus (trans)

Next Generation Sequencing

- Sequence rather than hybridisation
- Gene expression, genotyping, epigenetics
- New technologies: much cheaper than before
- Gene expression, genotyping, epigenetics
- Open ended (no previous knowledge required)
- Will take over in 2 years: the end of microarrays?

Part 2 Summary

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

Seminar Summary

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

Contact

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