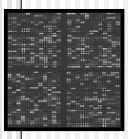
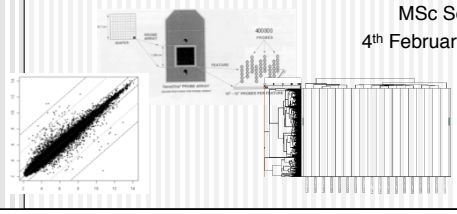


Microarray Informatics

Donald Dunbar
MSc Seminar
4th February 2009



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

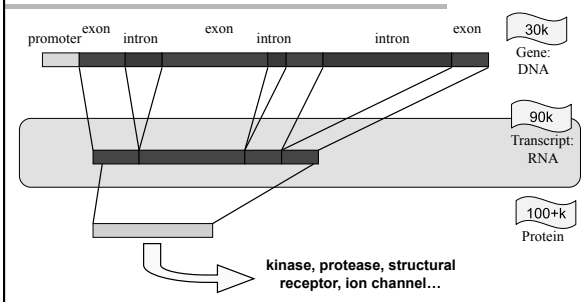
Microarray Informatics

Part 1

Biological research

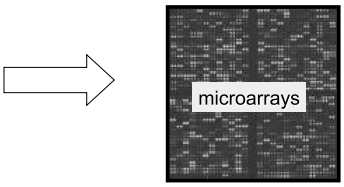
- Using a wide range of experimental and computational methods to answer biological questions
- Genetics, physiology, molecular biology…
- Biology and informatics → bioinformatics
- Genomic revolution
- What can we measure?

The central dogma



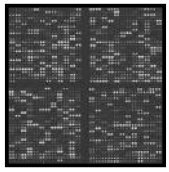
Measuring transcripts

- Genome level sequencing
- New miniaturisation technologies
- Better bioinformatics

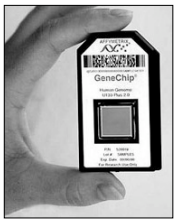


Microarrays: wish list

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

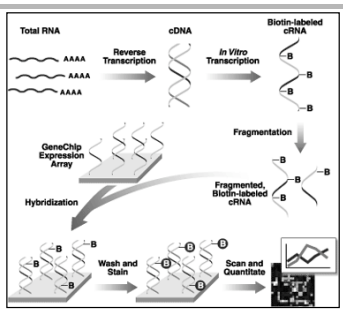


Microarray technologies - 1



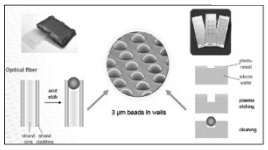
- Oligonucleotides - Affymetrix
- One chip all genes
- Chips for many species
- Several oligos per transcript
- Use of control, mismatch sequences
- One sample per chip
  - 'absolute quantification'
- Well established in research
- Expensive

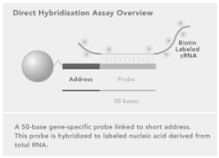
Microarray technologies - 1



Microarray technologies - 2

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





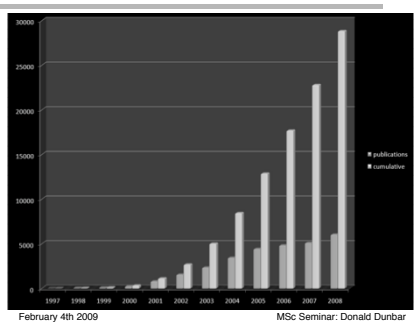
Problems with transcriptomics

- 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

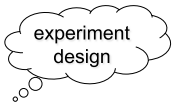
- Usually **control** v **test(s)**

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

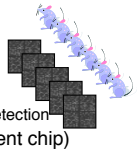


Experiment design: system

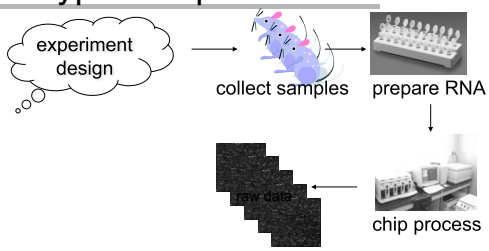
- What is your model?
  - animal, cell, tissue, drug, time…
- What comparison?
- What platform
  - microarray? oligo, cDNA?
- Record all information: see “standards”

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

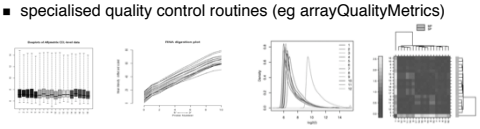
Bioconductor (BioC)

- 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
- Most platforms now have BioC packages
- 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 arrayQualityMetrics)



Pre-processing: background

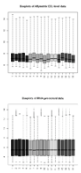
- Signal corresponds to expression…
  - plus a non-specific component (noise)
- Non specific binding of labelled target
- Need to exclude this background
- Several methods exist
  - eg Affy: PM-MM but many complications
  - eg RMA PM=B+S (don't use MM)

Pre-processing: normalisation

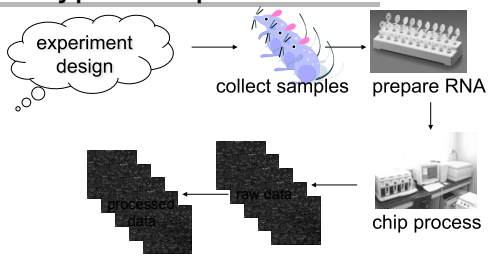
- In addition to background corrections
  - chip, probe, spatial, intra and inter ation
  - need to remove to get at real ex differences
- Make use of
  - statistics
- But seems to be n dependency on intensity
  - additive and m ive errors
- Quantile norm often used
- Normalisati e complicated for 2-colour arrays
- Try to re ost noise at lab stage (ie control things well statistica

combined with m set summary: get an expres alue for the gene

Now carry on with analysis!



A typical experiment



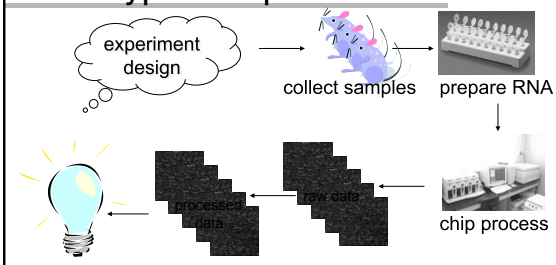
Part 1 Summary

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

Microarray Informatics

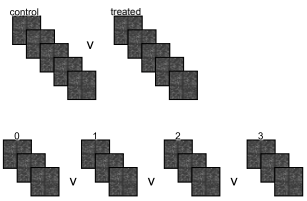
Part 2

A typical experiment



Data analysis

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



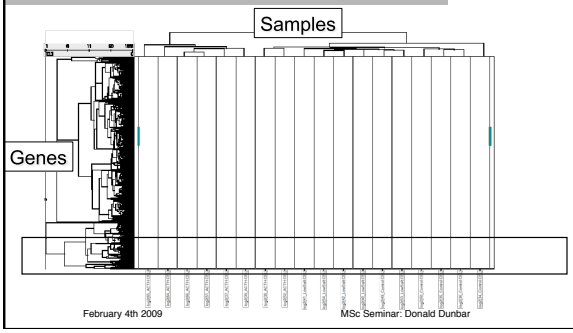
Multiple testing

- Problem:
  - statistical testing of 30,000 genes
  - at α = 0.05 → 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

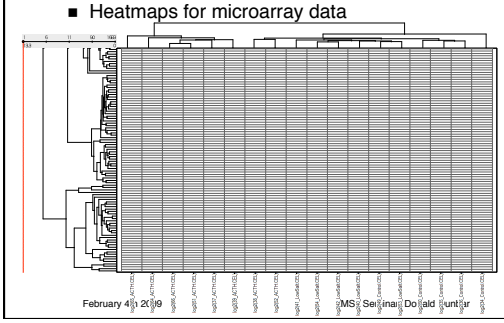
- Look for structure within dataset
  - similarities between genes
- Compare gene expression profiles
  - Euclidian distance
  - Correlation
  - Cosine correlation
- Calculate with distance matrix
- Combine closest, recalculate, combine closest… (or split!)
- Draw dendrogram and heatmap

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

Sample classification

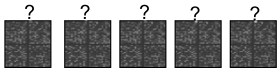
- Supervised or non-supervised
- Non-supervised
  - like hierarchical clustering of samples
- Supervised
  - have training (known) and test (unknown) datasets
  - use training sets to define robust classifier
  - apply to test set to classify new samples

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 …

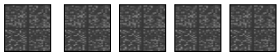


Sample classification

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…

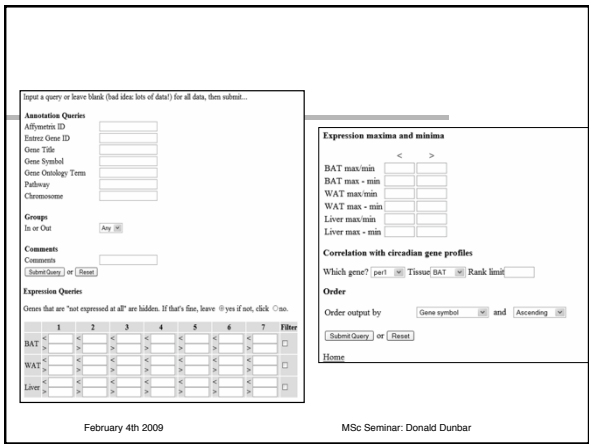
Data-mining

Microarrays are a waste of time

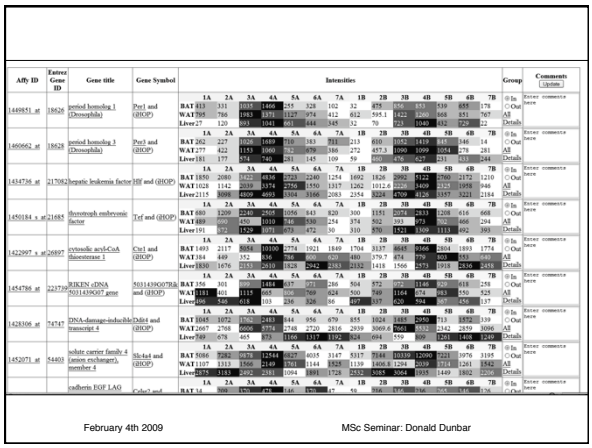
…unless you do something with the data

Data-mining

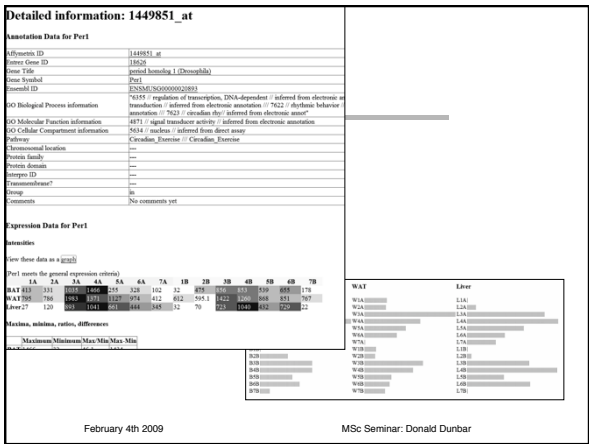
- Once data are statistically analysed:
  - pull out genes of interest
  - pull out pathways of interest
  - mine data based on annotation
    - what are the expression patterns of these genes
    - what are the expression patterns in this pathway
  - mine genes based on expression pattern
    - what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



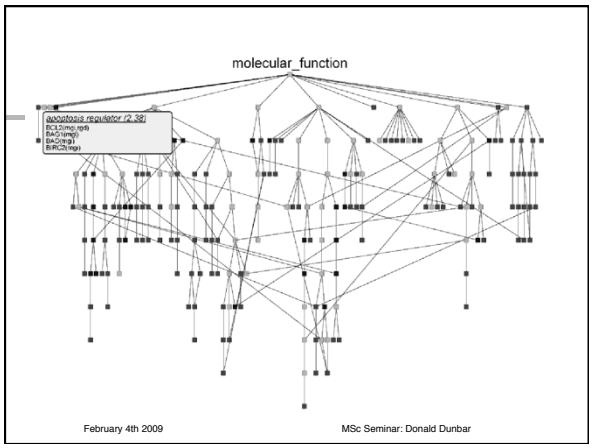
February 4th 2009 MSc Seminar: Donald Dunbar



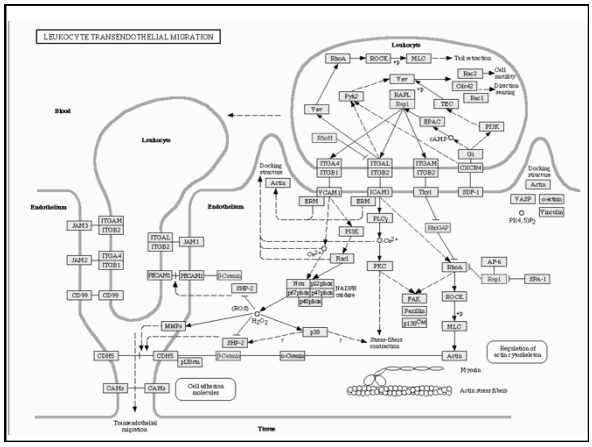
February 4th 2009 MSc Seminar: Donald Dunbar

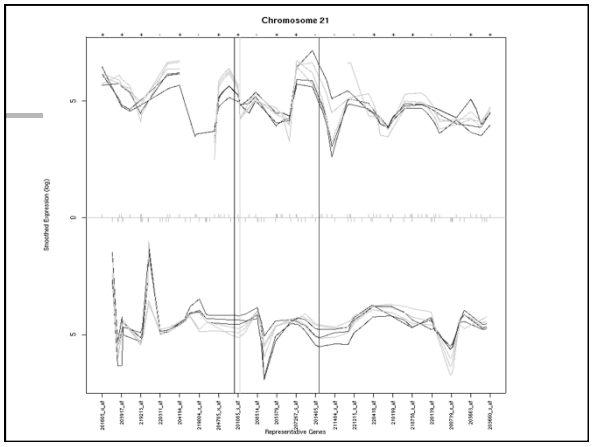


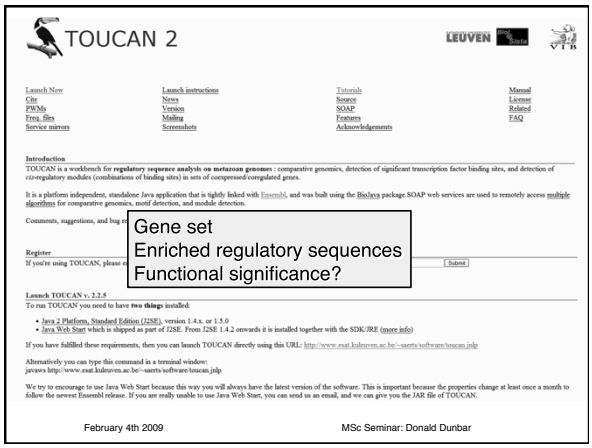
February 4th 2009 MSc Seminar: Donald Dunbar

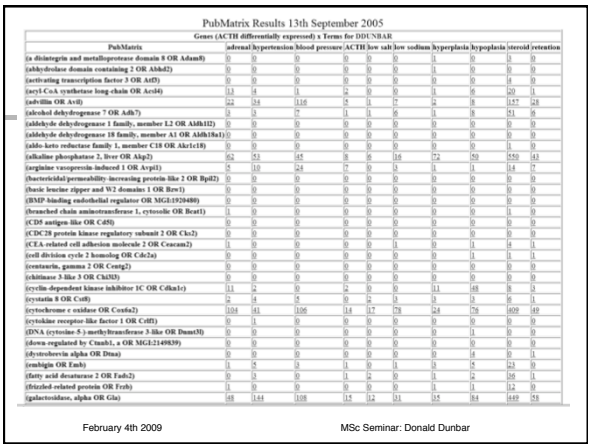


February 4th 2009 MSc Seminar: Donald Dunbar



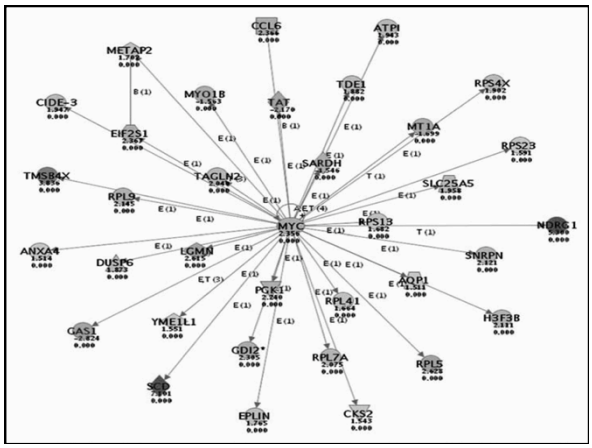






PubMatrix Results 13th September 2005

Genes (ACTH differentially expressed) x Terms for DDUNBAR										
PubMatrix	adrenal	hypertension	blood pressure	ACTH	low salt	low sodium	hyperplasia	hypoplasia	steroid	retention
(a disintegrin and metalloprotease domain 8 OR Adam8)	0	0	0	0	0	0	1	0	1	0
(abhydrolase domain containing 2 OR Abhd2)	0	0	0	0	0	0	1	0	0	0
(activating transcription factor 3 OR Atf3)	0	0	0	0	0	0	0	0	4	0
(acyl-CoA synthetase long-chain OR Acsl4)	13	4	1	2	0	0	1	6	20	1
(advillin OR Avil)	22	34	116	5	1	7	2	8	157	28
(alcohol dehydrogenase 7 OR Adh7)	3	3	7	1	1	6	1	8	51	6
(aldehyde dehydrogenase 1 family, member L2 OR Aldh1l2)	0	0	0	0	0	0	0	0	0	0
(aldehyde dehydrogenase 18 family, member A1 OR Aldh18a1)	0	0	0	0	0	0	0	0	0	0
(aldo-keto reductase family 1, member C18 OR Akr1c18)	0	0	0	0	0	0	0	0	1	0
(alkaline phosphatase 2, liver OR Alp2)	62	53	45	8	6	16	72	50	550	43
(arginine vasopressin-induced 1 OR Avpi1)	5	10	24	7	0	3	1	1	14	7
(bactericidal/permeability-increasing protein-like 2 OR Bpil2)	0	0	0	0	0	0	0	0	0	0
(basic leucine zipper and W2 domains 1 OR Bzw1)	0	0	0	0	0	0	0	0	0	0
(BMP-binding endothelial regulator OR MGI:1920480)	0	0	0	0	0	0	0	0	0	0
(branched chain aminotransferase 1, cytosolic OR Bcat1)	1	0	0	0	0	0	0	0	1	0
(CD5 antigen-like OR Cd5l)	0	0	0	0	0	0	0	0	0	0
(CDC28 protein kinase regulatory subunit 2 OR Cks2)	0	0	0	0	0	0	0	0	0	0
(CEA related cell adhesion molecule 2 OR Ceacam2)	1	0	0	0	0	1	0	1	4	1
(cell division cycle 2 homolog OR Cdc2a)	0	0	0	0	0	0	0	1	1	1
(centaurin, gamma 2 OR Centg2)	0	0	0	0	0	0	0	0	0	0
(chitinase 3-like 3 OR Chi3l3)	0	0	0	0	0	0	0	0	0	0
(cyclin-dependent kinase inhibitor 1C OR Cdkn1c)	11	2	0	2	0	0	11	48	8	1
(cystatin 8 OR Cst8)	2	4	5	0	2	1	1	3	6	1
(cytochrome c oxidase OR Cox6a2)	104	41	106	14	17	78	24	76	409	49
(cytokine receptor-like factor 1 OR Crlf1)	0	1	0	0	0	0	0	0	0	0
(DNA (cytosine-5)-methyltransferase 3-like OR Dnmt3l)	0	0	0	0	0	0	0	1	0	0
(down-regulated by Ctnnb1, a OR MGI:2149839)	0	0	0	0	0	0	0	0	0	0
(dystrobrevin alpha OR Dtna)	0	0	0	0	0	0	0	4	0	1
(embigin OR Emb)	1	5	3	1	0	1	3	5	23	0
(fatty acid desaturase 2 OR Fads2)	0	3	0	1	2	0	1	2	26	1
(frizzled-related protein OR Frzb)	1	0	0	0	0	0	1	1	12	0
(galactosidase, alpha OR Gla)	48	184	108	15	12	31	35	84	449	58



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

Microarray Resources

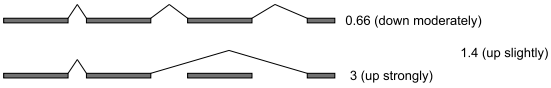
- Microarray data repositories
  - Array express (EBI, UK)
  - Gene Expression Omnibus (NCBI, USA)
  - CIBEX (Japan)
- Annotation
  - NetAffx, Ensembl, TIGR, Stanford…

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

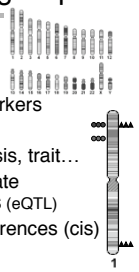
Recent advances: Exon chips

- 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 5 years: the end of microarrays?

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

PhD opportunity

- Centre for Cardiovascular Science (Edinburgh)
- The Cellular and Molecular Basis of Cardiovascular Disease
- BHF funded PhDs
  - biologists (x4)
  - physical scientists (informatics, physics, maths….)
- Details on web:
  - http://www.cvs.med.ed.ac.uk/Training/content.asp?SubCatID=44