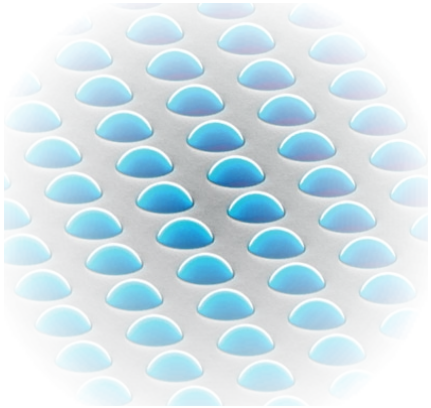
Analysis of Illumina Gene Expression Microarray Data



Asta Laiho, Msc. Tech.

Bioinformatics research engineer
The Finnish DNA Microarray Centre
Turku Centre for Biotechnology, Finland

The Finnish DNA Microarray Centre

- **National core facility** and part of the Turku Centre for Biotechnology (BTK)
- Provides state-of-the-art research technologies and services
- Services cover a full range of options in life science
- Customers: the Finnish and international scientific community
- FDMC has an active national and international collaborative network

illumina CSPro CERTIFIED SERVICE PROVIDER

FDMC Services - http://microarrays.btk.fi

- Gene expression analysis
  - Microarray analysis
  - Real-Time PCR
  - Exon array
  - miRNA

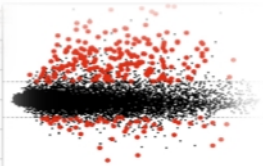
- ChIP-on-chip analysis

- Genotyping
  - Whole genome genotyping
  - SNP genotyping
  - aCGH

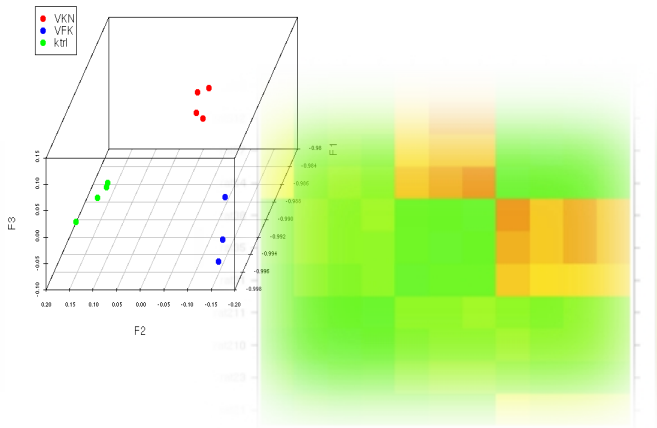
- Custom spotting

- Sequencing

- Bioinformatics



- Education and Training



Gene expression analysis service

What we offer for you:

- Experimental planning and selection of the most suitable technology platform
  (based on project size, organism, number of samples and genes, etc.)
- Fast and high quality service including full data analysis!

A few words on experimental design

- Typical experiment setup:
  disease samples vs. healthy controls

- **Replicates:**
  - Biological replicates: how many?
  - Technical replicates:
    not often used nowadays

Biological replicates: how many?

- No simple answer
- **General guide line:** at least 3 per condition group!
- Having more replicates increases sensitivity in detecting differential expression

- **Needed replicate number depends on:**
  - Strength of the studied effect
  - Within group variation
  - Level of technical noise

- When studying cell cultures or laboratory animals 3 replicates per group is often enough
- Patient studies generally require a lot more because biological variation between individuals is large

Example of a good biological replication

- Mouse experiment with wild type mice and knockout mice
- WT: 4 mice

- KO: 4 mice

What kind of samples can be compared?

- **Do not try to compare apples and pears:**
  If the samples are too different – all genes will be differentially expressed!
  -> no useful information can be gained

- Two different tissues (e.g., kidney vs. spleen) are usually too different to be compared directly

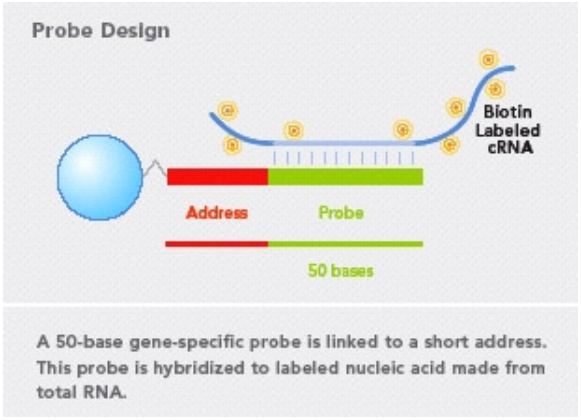
- If several tissue samples (meant to represent the same tissue) contain **varying amounts of different cell types** this can also be a problem

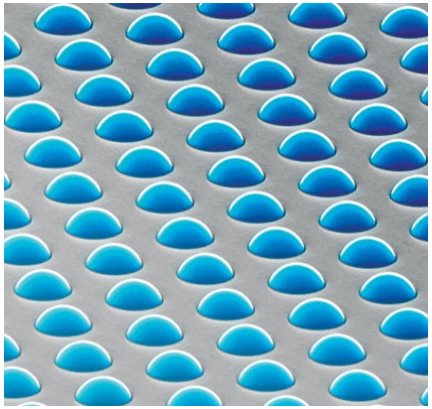
Other important issues

- **RNA sample quality!**
- **Standardise conditions** for all samples in the experiment set (e.g. age, sex, RNA extraction method...)
- Choose the correct **time point!**
- Only **pool samples** when sample material is scarce
- Be prepared to **validate your microarray results** with some other technique like **real-time pcr**
- **Data analysis issues** should always be considered when making experimental design
  - Experienced data analyst / bioinformatician should be consulted

Illumina Expression BeadChips

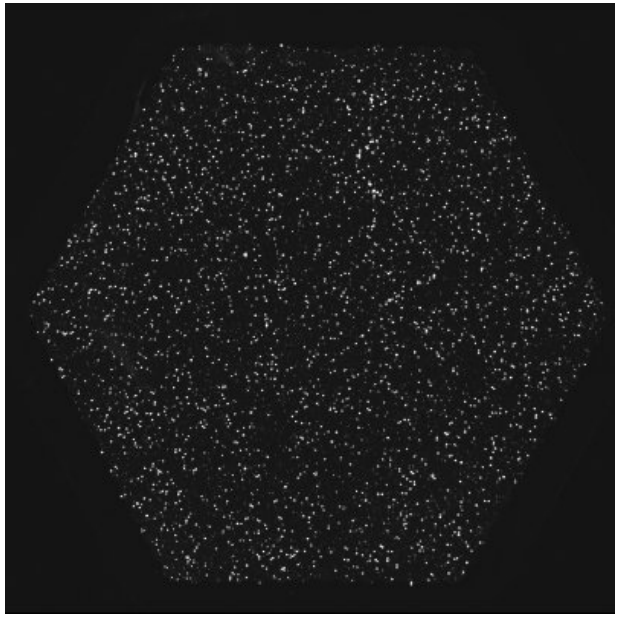
- 6 – 12 samples on one chip





- 15 – 30 replicate beads per array target on the average

Extracting information from the image



Raw data file

Feature identifiers

Sample columns

ProbeID	AVG_Signal	AVG_Signal	AVG_Signal	AVG_Signal
6450255	94.94183	89.89078	93.88631	81.89128
2570615	129.1059	132.6304	116.0114	106.537
6370619	93.40171	100.4024	115.1009	101.5393
2600039	97.51926	101.0345	106.5934	96.65336
2650615	107.3257	100.0602	101.7889	112.9131
5340672	97.75134	97.8282	97.43371	94.85838
1090041	94.68685	88.76044	100.1513	122.9476
6380561	103.3985	97.36023	92.73147	80.71864
7570255	91.33523	90.54805	113.7738	105.1146
4920477	100.0958	94.74992	89.17274	102.0431
2000519	124.8985	134.2148	111.1489	125.1393
3870044	82.87843	82.12807	94.96744	84.37419
7050209	112.6961	92.72503	83.39499	100.4304

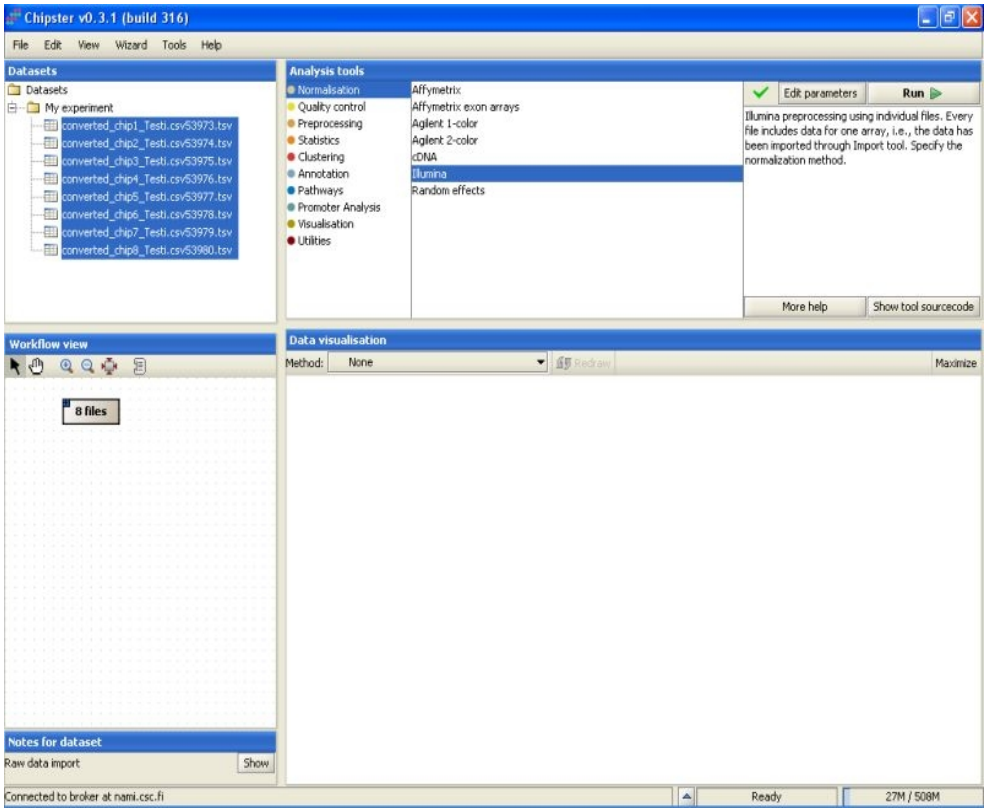
Intensity measurements

Analysis software

- R/Bioconductor (free)
- GeneSpring (commercial)
- Chipster

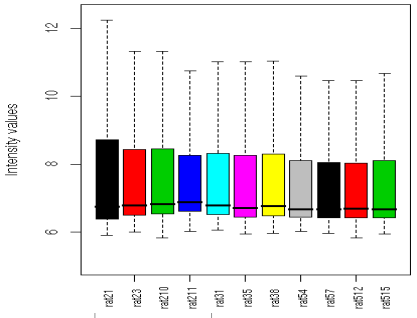
  (http://chipster.csc.fi/)
  (free for academic use)

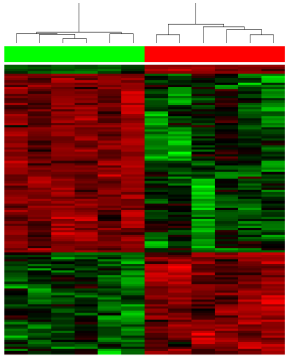
- Lots of other free & commercial tools exist

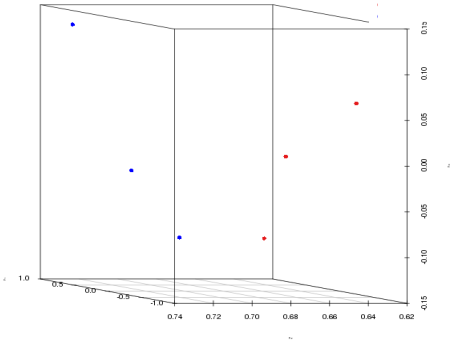


General outline of expression data analysis

- Normalisation
- Quality inspection (for raw + normalised data)
- Prefiltering (optional)
- Running statistical tests
- Filtering **differentially expressed (DE)** genes
- Visualising the results
- Carrying out gene functional analysis

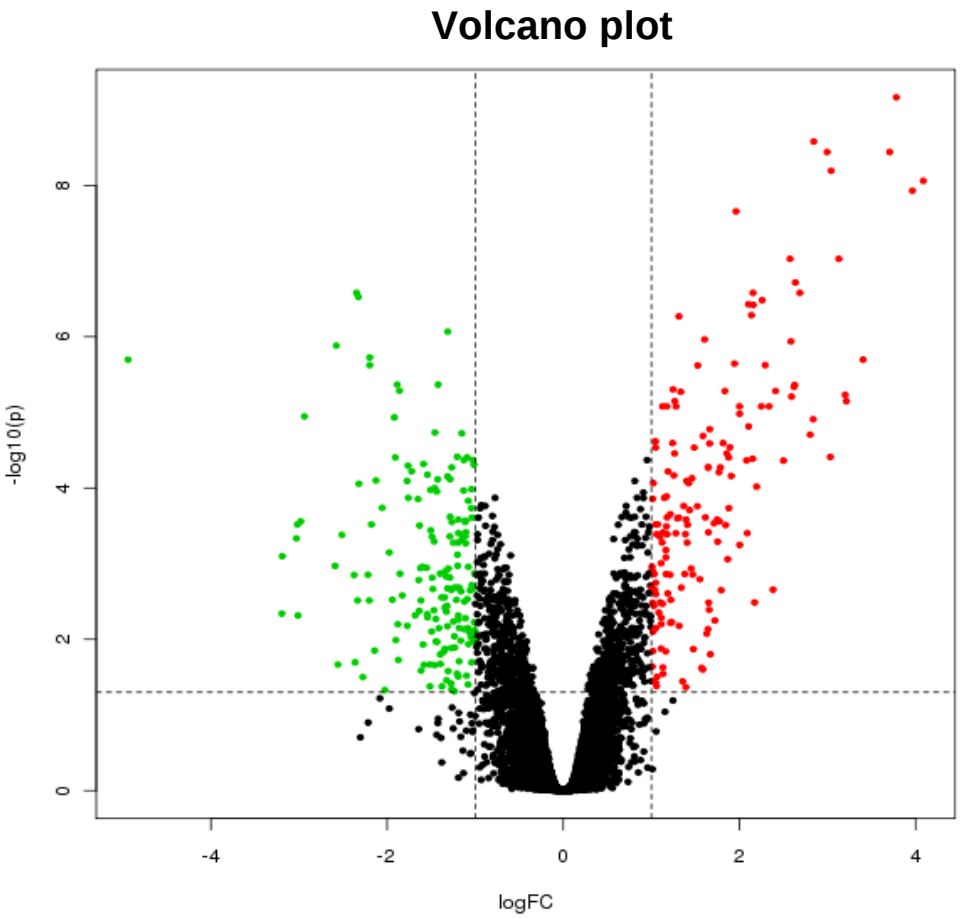






Filtering for differential expression

- Genes that have **similar behavior within each sample group** but the **group means clearly differ** from each other
- Fold-change (FC) for the **size** of the change in gene expression
- P-values and false discovery rates for the **reliability** of the change
- **Goal:**
  To produce a reasonable sized list of the most differentially expressed genes



Functional analysis

- Focus in **pathways** or other **functional categorisations** rather than individual genes

- **Different approaches exist for this:**

  - **Detect functional enrichment in the DE target list**

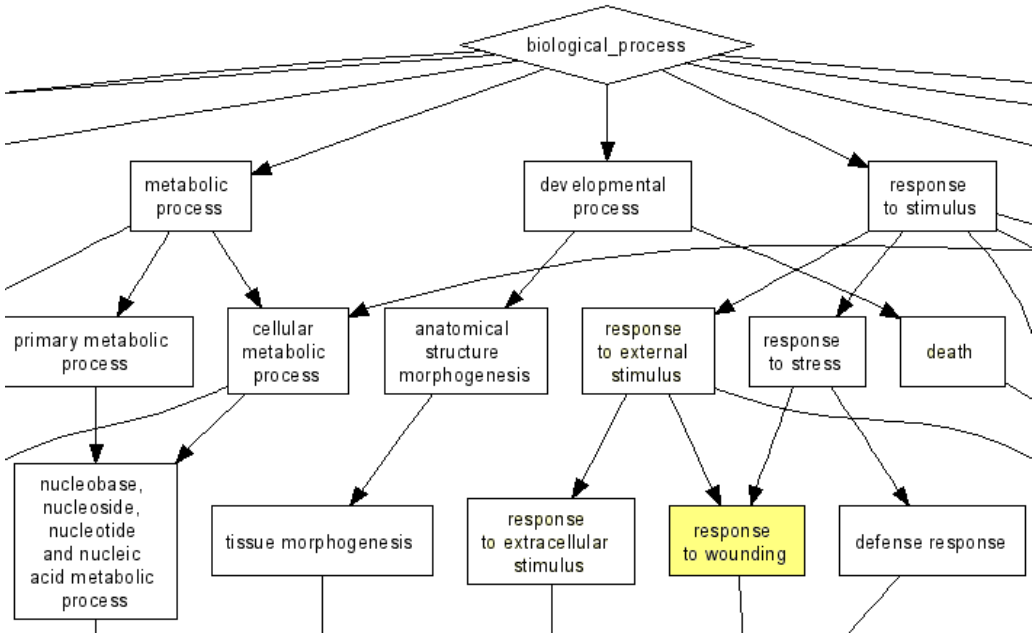
  - Detect functional enrichment towards the top of the list when all array targets have been ranked according to the evidence for being differentially expressed

  - Make the statistical test between sample groups not assuming **independence** between array targets (as usually) but taking the **dependence** between **genes belonging to same functional categorisations** into account

GO (Gene Ontology)

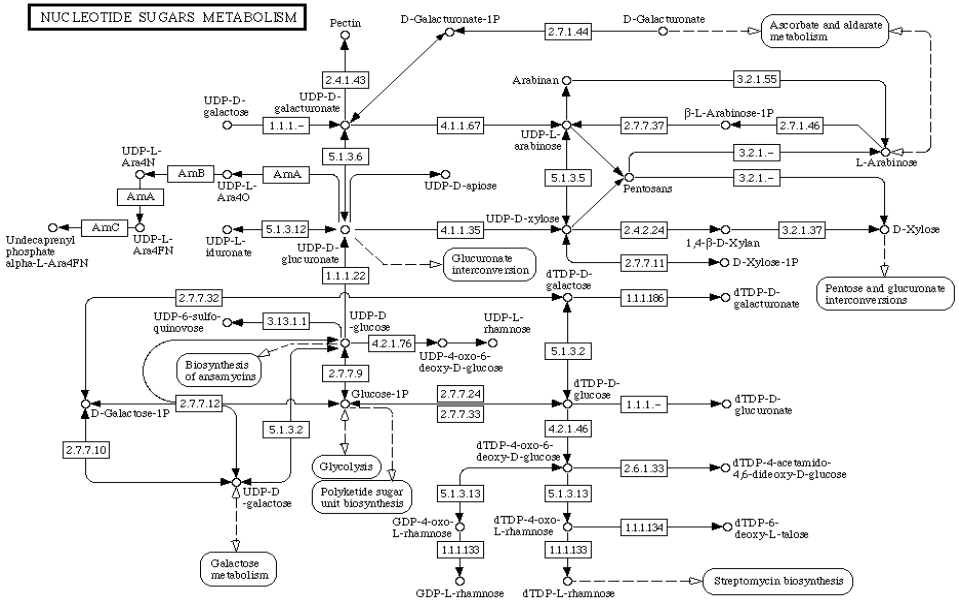
- http://www.geneontology.org
- Classifies genes into a hierarchy, placing gene products with similar functions together

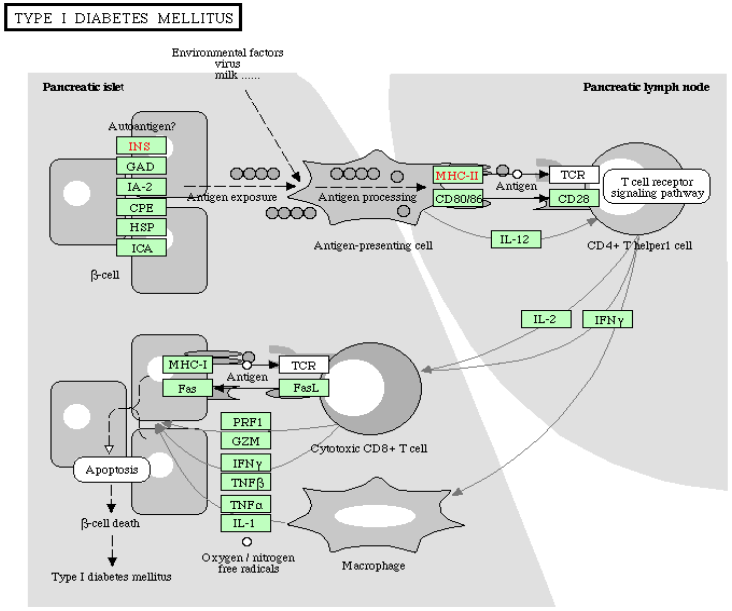
- **Three main categories:**
  - Biological process (BP)
  - Molecular function (MF)
  - Cellular component (CC)



KEGG

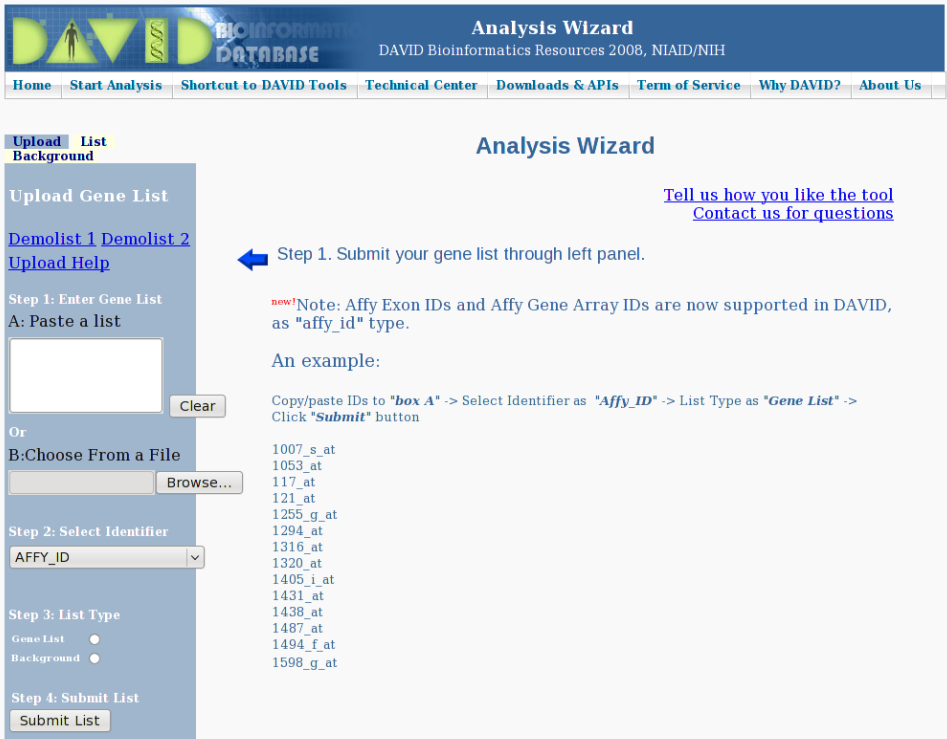
- The Kyoto Encyclopaedia of Genes and Genomes
- http://www.genome.jp/kegg/
- Provides searchable pathways for **molecular interaction and reaction networks for metabolism**, **various cellular processes** and **human diseases**
- Manually entered from published materials





Tools for functional analysis

- Lots of free third-party tools exist
  - David
    http://david.abcc.ncifcrf.gov/home.jsp
  - Pathway-Express
    http://vortex.cs.wayne.edu/projects.htm#Pathway-Express
  - GSEA
    http://www.broad.mit.edu/gsea/
  - GOrilla
    http://cbl-gorilla.cs.technion.ac.il/
  - GenMapp
    http://www.genmapp.org/
  - Cytoscape
    http://www.cytoscape.org/



David

Functional Annotation Chart

Help and Manual

Current Gene List: Uploaded List_1

Current Background: RATTUS NORVEGICUS

171 DAVID IDs

⊞ Options

Rerun Using Options | Create Sublist

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
☐	GOTERM_BP_4	organic acid metabolism	RT		26	15,2	4,2E-12	2,8E-9
☐	GOTERM_BP_4	amino acid and derivative metabolism	RT		19	11,1	8,8E-10	2,9E-7
☐	GOTERM_BP_4	amine metabolism	RT		20	11,7	2,4E-9	5,2E-7
☐	GOTERM_BP_4	nitrogen compound catabolism	RT		7	4,1	4,3E-5	7,0E-3
☐	GOTERM_BP_4	sulfur metabolism	RT		7	4,1	6,1E-5	8,1E-3
☐	GOTERM_BP_4	nitrogen compound biosynthesis	RT		7	4,1	8,6E-5	9,4E-3
☐	GOTERM_BP_4	transport	RT		40	23,4	1,3E-4	1,2E-2
☐	GOTERM_BP_4	cellular catabolism	RT		13	7,6	1,5E-3	1,2E-1
☐	GOTERM_BP_4	urea metabolism	RT		3	1,8	2,3E-3	1,5E-1

Publishing microarray data

- **GEO** (Gene Expression Omnibus)
  www.ncbi.nlm.nih.gov/geo/
- **ArrayExpress**
  http://www.ebi.ac.uk/microarray-as/ae/

- Most journals require the expression data to be submitted to a public repository

  – some even before they will send the manuscript to referees for evaluation

- The data can be hidden from others than the authors and the referees before the official publication of the article

illumina CSPro CERTIFIED SERVICE PROVIDER

FDMC data analysis service for gene expression data

- **Project start meeting:**
  experimental design consultation
- **Analysis with R/Bioconductor**
  - Normalisation & quality inspection
  - Statistical testing
  - Filtering for differentially expressed genes
  - Producing annotated gene lists
  - Functional analysis
- **Sending the results to customer electronically**
  - Results include a comprehensive report
- **Result meeting**

- **Contact us:** bioinfo@btk.fi

Take home message:

- Data analysis issues should be considered during the **planning** and **experimental design** of the experiments

  - Who will analyse the data (what will it cost)?
  - What tools will be used for data analysis?
  - What questions do we want to answer by these experiments?

Asta Laiho
asta.laiho@btk.fi
bioinfo@btk.fi
http://microarrays.btk.fi

Contact us:
The Finnish DNA Microarray Centre
BioCity, Tykistökatu 6, Biocity, 5th floor
FIN-20521 Turku, Finland
Phone: +358 2 333 8603
Fax: +358 2 333 8000
General inquiries: microarrays@btk.fi