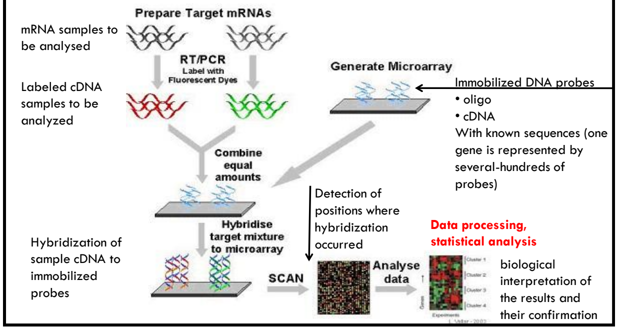
L7: Transcriptome challenge for bioinformatics

DNA microarrays
- Overview of a technique
- Microarray experiment design
- Data analysis – general considerations & databases
- Data analysis – steps and methods

SAGE
- Overview of a technique
- Data analysis – general considerations & databases

RNA sequencing
- Overview of next generation sequencing techniques
- Data analysis – general considerations & databases

DNA microarray experiment – the princple



mRNA samples to be analysed

Labeled cDNA samples to be analyzed

Hybridization of sample cDNA to immobilized probes

Immobilized DNA probes
- oligo
- cDNA

With known sequences (one gene is represented by several-hundreds of probes)

Detection of positions where hybridization occurred

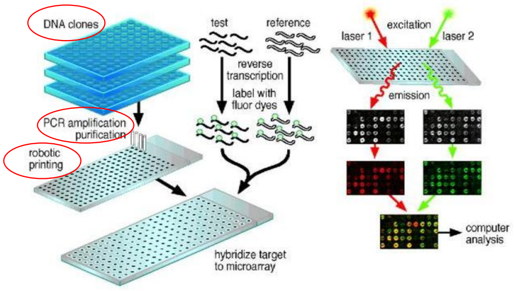
Data processing, statistical analysis

biological interpretation of the results and their confirmation

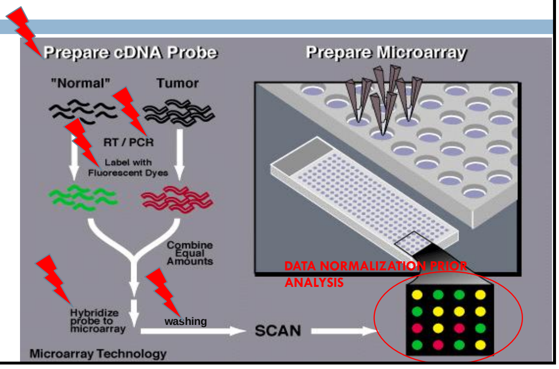
DNA microarray experiment variations

DNA probes immobilized on solid support	Sample and control RNA preparation	cDNA (cRNA) labeling	cDNA (cRNA) hybridization
oligonucleotides	Rewrite mRNA to cDNA	Fluorescently	Single sample hibridizations
cDNA	mRNA→ cDNA→biotin-labelled cRNA	Radioactively	Competitive hibridization (previously labelled with two dyes e.g. Cy3 & Cy5)

Types of DNA probes - cDNA



DNA microarray experiment error-inducing points



Stage 1: Experimental design

Stage 2: RNA and probe preparation

Stage 3: Hybridization to DNA arrays

Stage 4: Image analysis

Stage 5: Microarray data analysis

Stage 6: Biological confirmation

Stage 7: Microarray databases

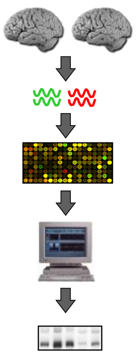


Fig. 8.17
Page 314

Stage 1: Experimental design

[1] Biological samples; technical and biological replicates;
determine the data analysis approach at the beginning!

[2] RNA extraction, conversion, labeling, hybridization
RNA procedures must be standarized (but still laboratory operator impact may be huge)

[3] Arrangement of DNA elements on a solid surface
randomization in cDNA printing can reduce spatially-based artifacts

Stage 2: RNA preparation

For Affymetrix chips, need about 5 µg total RNA

Confirm purity and integrity by running agarose gel

Measure A_{260}/A_{280} to confirm purity and quantity

One of the greatest sources of error in microarray experiments is artifacts associated with RNA isolation:
be sure to create an appropriately balanced, randomized experimental design (do not isolate samples for the same analysis on different days!).

Stage 3: Hybridization to DNA arrays

The array consists of cDNA or oligonucleotides (several 25-mers/gene).

Oligonucleotides can be deposited by photolithography (Affymetrix)

The sample is converted to cRNA or cDNA (most commonly)

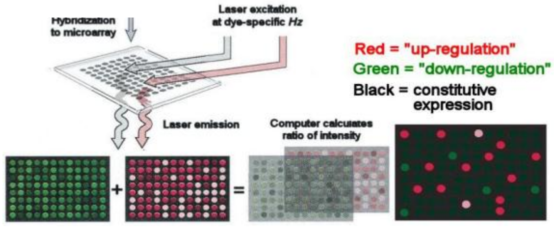
For competitive approaches equal amount of 2 samples is essential

Hybridization is done in specific apparatus stabilizing temperature.
Buffer conditions and step washes to be optimized.

Note that the terms "probe" and "target" may refer to the element immobilized on the surface of the microarray, or to the labeled biological sample; for clarity, it may be simplest to avoid both terms.

Stage 4: Image analysis

Fluorescence intensity is measured with a scanner
If competitive hibridization was used two readings are merged


Hybridization to microarray
Laser excitation at dye-specific *Hz*
Red = "up-regulation"
Green = "down-regulation"
Black = constitutive expression
Laser emission
Computer calculates ratio of intensity
+
=

Stage 4: Image analysis

Fluorescence intensity is measured with a scanner
If competitive hybridization was used two readings are merged

RNA transcript levels are quantified

Many experimental designs provide set of so-called „house keeping" genes which expression is unchanged in several cell types in broad range of conditions.
(Trascript level normalization as in RT-PCR and qRT-PCR).
Common examples:
Beta-actin, glyceraldehyde 3-phosphate dehydrogenase (GAPDH)

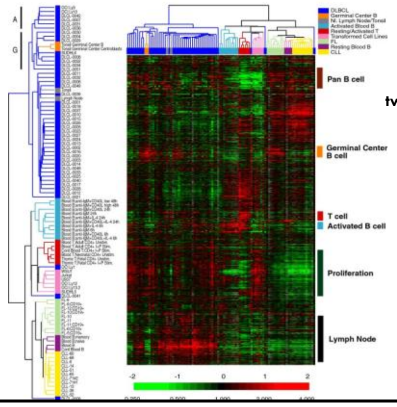
Stage 5: Microarray data analysis

Preprocessing

Hypothesis testing
- How can arrays be compared?
- Which RNA transcripts (genes) levels are changed?
- Are differences authentic?
- What are the criteria for statistical significance?

Clustering
- Are there meaningful patterns in the data (e.g. groups)?

Classification
- Do RNA transcripts predict predefined groups, such as disease subtypes?



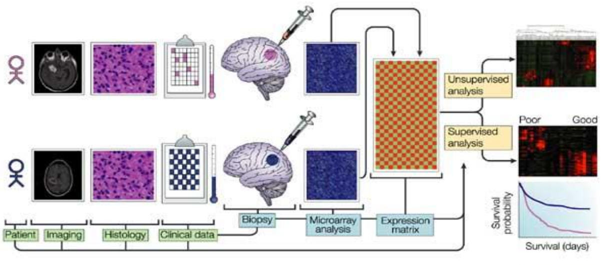
An example of two-way clustering of

genes (y-axis) and cell lines (x-axis)

(Alizadeh et al., 2000)

Fig. 9.18
Page 362

Classification: predict a specific pattern & conclude about the sample origin (DIAGNOSIS)



Nature Reviews | Neuroscience

Stage 6: Biological confirmation

Microarray experiments can be thought of as "hypothesis-generating" experiments.

The differential up- or down-regulation of specific RNA transcripts can be measured using independent assays such as:

-- **Northern blots**
-- polymerase chain reaction (RT-PCR, **qRT-PCR**)
-- ***in situ* hybridization** (confirming localization as well)

Page 320

MIAME (http://www.mged.org)

MIAME = Minimum Information About a Microarray Experiment was established in an effort to standardize microarray data presentation and analysis.

The MIAME framework standardizes six areas of information:

- experimental design
- microarray design
- sample preparation
- hybridization procedures
- image analysis
- controls for normalization

Page 319

Stage 5: Microarray data analysis

Preprocessing

Hypothesis testing (statistics)
- How can arrays be compared?
- Which RNA transcripts (genes) are regulated?
- Are differences authentic?
- What are the criteria for statistical significance?

Clustering
- Are there meaningful patterns in the data (e.g. groups)?

Classification
- Do RNA transcripts predict predefined groups, such as disease subtypes?

Page 320

Microarray data analysis: preprocessing

Observed differences in gene expression (fluorescence) could be due to transcriptional changes, or they could be caused by artifacts such as:

- different labeling efficiencies of Cy3, Cy5
- uneven spotting of DNA onto an array surface
- variations in RNA purity or quantity
- variations in washing efficiency
- variations in scanning efficiency

Page 337

Microarray data analysis: preprocessing

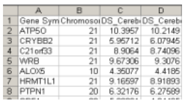
The main goal of data preprocessing is to remove the systematic bias in the data as completely as possible, while preserving the variation in gene expression that occurs because of biologically relevant changes in transcription.

A basic assumption of most normalization procedures is that the average gene expression level does not change in an experiment.

Page 337

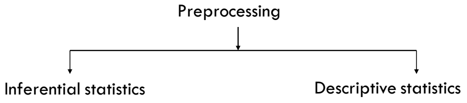
Microarray data analysis - preprocessing

We begin with a data matrix (gene expression values versus samples)



	A	B	C	D
1	Gene Sym	Chromosoı	DS_Cerebı	DS_Cereb
2	ATP5O	21	10.3957	10.2149
3	CRYBB2	21	5.95712	6.07945
4	C21orf33	21	8.9064	8.74096
5	WRB	21	9.67306	9.3076
6	ALOX5	10	4.35077	4.4185
7	HRMT1L1	21	9.16597	8.91893
8	PTPN1	20	6.32176	6.27589

Typically, there are many genes (>> 20,000) and few samples (< 10)



Preprocessing

Inferential statistics

Descriptive statistics

Fig. 9.1
Page 333

Data analysis: global normalization

Global normalization is used to correct two or more data sets.

Sample RNA is labeled with **Cy3** (cyanine3 - green dye) and control RNA with **Cy5** (red dye). After hybridization and washing, probes are excited with a laser and detected with a scanning confocal microscope.

Example: total fluorescence read in
Cy3 channel = 4 million fluoresc. units
Cy 5 channel = 2 million fluoresc. units

Then the **uncorrected ratio for a gene** could show 2,000 units versus 1,000 units. This would artifactually appear to show **2-fold regulation**!

Page 343

Data analysis: global normalization

Global normalization procedure

Step 1: subtract background intensity values
(use a blank (with no RNA added) region of the array)

Step 2: globally normalize so that the average ratio = 1
(apply this to 1-channel or 2-channel data sets)

Page 343

Scatter plots

Are useful to represent and compare gene expression values from two microarray experiments (e.g. from control & tested conditions, two different cell types, normal and diseased samples etc.)

Each dot corresponds to a gene expression value

Most dots fall along a line

Outliers represent up-regulated or down-regulated genes

Page 338

Scatter plots

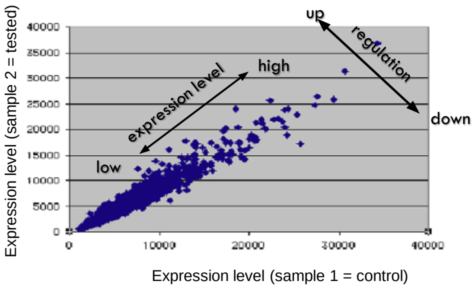


Fig. 9.2
Page 338

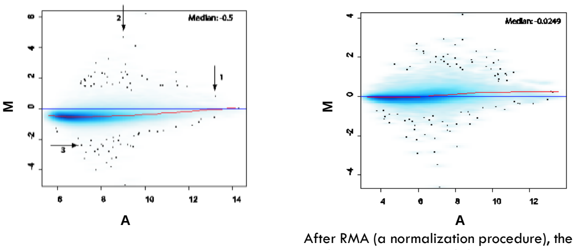
Scatter plots

Typically, data are plotted on log-log coordinates (in stead of row data)

Visually, this spreads out the data and offers **symmetry:**

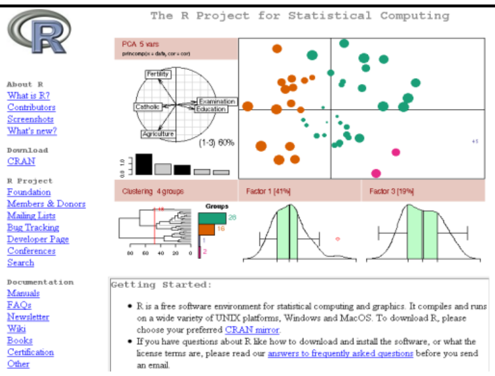
Sam- ple	time point	transcript behavior	raw fluorescence ratio value	$\log_2$ ratio value
C	t=0	basal	1.0	0.0
1	t=1h	no change	1.0	0.0
2	t=2h	2-fold up	2.0	1.0
3	t=3h	2-fold down	0.5	-1.0

Scatter plots of log-log coordinates



After RMA (a normalization procedure), the median is near zero, and skewing is corrected.

Scatterplots above display the effects of normalization.



The R Project for Statistical Computing

About R
What is R?
Contributors
Screenshots
What's new?

Download
CRAN

R Project
Foundation
Members & Donors
Mailing Lists
Bug Tracking
Developer Page
Conferences
Search

Documentation
Manuals
FAQs
Newsletter
Wiki
Books
Certification
Other

Getting Started:

- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.
- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

http://www.r-project.org

Robust multi-array analysis (RMA)

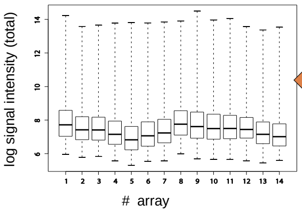
- Developed by Rafael Irizarry, Terry Speed and others
- Available at www.bioconductor.org as an R package

There are three RMA steps:

[1] Background adjustment
based on a normal plus exponential model (no mismatch data are used)

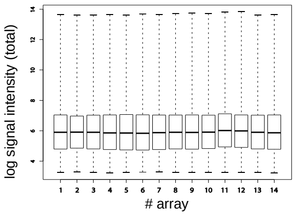
[2] Quantile normalization
(nonparametric fitting of signal intensity data to normalize their distribution)

[3] Fitting a log scale additive model robustly.
The model is additive: probe effect + sample effect



Histograms of raw intensity values for 14 arrays (plotted in R) **before RMA was applied.**

Histograms of raw intensity values for 14 arrays (plotted in R) **after RMA was applied.**



Inferential statistics

Inferential statistics are used to make conclusions about a population from a sample.

Hypothesis testing is a common form of inferential statistics.
A **null hypothesis** is stated, such as:
"There is no difference in signal intensity for the gene expression measurements in normal and diseased samples."
The alternative hypothesis is that there is a difference.

We use a test statistic to decide whether to accept or reject the null hypothesis.
For many applications, we set the significance level to $p < 0.05$.

Inferential statistics

A **t-test** is a commonly used test statistic to assess the difference in mean values between two groups.

$$t = \frac{x_1 - x_2}{SE} = \frac{\textbf{difference between mean values}}{\textbf{variability (standard error of the difference)}}$$

Questions:
Is the sample size (n) adequate?
Are the data normally distributed?
Is the variance of the data known?
Is the variance the same in the two groups?
Is it appropriate to set the significance level to $p < 0.05$?

Inferential statistics methods

Paradigm	Parametric test	Nonparametric
Compare two unpaired groups	Unpaired t-test (independent samples)	Mann-Whitney test
Compare two paired groups	Paired t-test (e.g. repetitions of measurement)	Wilcoxon test
Compare 3 or more groups	ANOVA	

Paired t-test is more powerful as paired units are similar with respect to noise factors

Inferential statistics

Is it appropriate to set the significance level to $p < 0.05$?
If you hypothesize that a specific gene is up-regulated, you can set the probability value to 0.05.

You might measure the expression of 10,000 genes and hope that *any* of them are up- or down-regulated. But you can expect to see 5% (500 genes) regulated at the $p < 0.05$ level by chance alone.

To account for the thousands of repeated measurements you are making, ***some researchers apply*** a Bonferroni correction. The level for statistical significance is divided by the number of measurements, e.g. the criterion becomes:

$p < (0.05)/10{,}000$ or $p < 5 \times 10^{-6}$

But the Bonferroni correction is generally considered to be <u>*too*</u> *conservative…*

Inferential statistics: false discovery rate

The **false discovery rate** (FDR) is a popular multiple corrections correction.

A **false discovery** = false positive (also called a type I error)

The FDR equals the p value of the t-test multiplied by the number of genes measured (e.g. for 10,000 genes and a p value of 0.01, FDR=100 means that there are 100 expected false positives).

You can adjust the false discovery rate. For example:

FDR	# regulated transcripts	# false discoveries
0.1	100	10
0.05	45	3
0.01	20	1

Descriptive statistics

Microarray data are highly dimensional: there are many thousands of measurements made from a small number of samples.

Descriptive (exploratory) statistics help you to find meaningful patterns in the data.

A first step is to arrange the data in a matrix.
Next, use a **distance metric** to define the relatedness of the different data points. Two commonly used distance metrics are:
-- Euclidean distance
-- Pearson coefficient of correlation

Principal components analysis (PCA)

Principal components analysis is an exploratory technique used to reduce the dimensionality of the data set to 2D or 3D.

For a matrix of ***m* genes x *n* samples,**
create a new, covariance matrix of size ***n* x *n***

Thus transform some large number of variables into a smaller number of uncorrelated variables called **principal components** (PCs).

Principal components analysis (PCA): objectives

- to reduce dimensionality
- to determine the linear combination of variables
- to choose the most useful variables (features)
- to visualize multidimensional data
- to identify groups of objects (e.g. genes/samples)
- to identify outliers

Page 364

Stage 5: Microarray data analysis

Preprocessing

Hypothesis testing
- How can arrays be compared?
- Which RNA transcripts (genes) are regulated?
- Are differences authentic?
- What are the criteria for statistical significance?

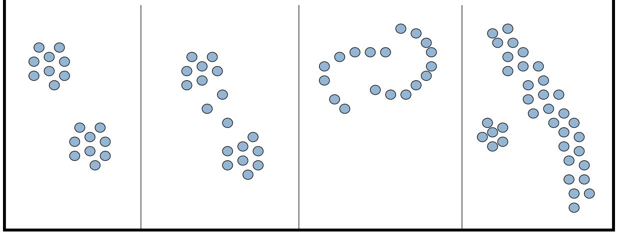
Clustering
- Are there meaningful patterns in the data (e.g. groups)?

Classification
- Do RNA transcripts predict predefined groups, such as disease subtypes?

Page 320

What is a cluster?

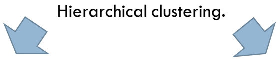
A cluster is a group that has **homogeneity** (internal cohesion) and **separation** (external isolation). The relationships between objects being studied are assessed by similarity or dissimilarity measures.



Descriptive statistics: clustering

Clustering algorithms offer useful visual descriptions of microarray data.

We may wish to claster: genes, samples or both



Hierarchical clustering.

Agglomerative
(beginning with the two most closely related objects (*like UPGMA*)

Divisive
(beginning by finding the most dissimilar objects first).

In each case, we end up with a tree having branches and nodes.

Page 355

Results of clustering: finding a relationship between gene expression profiles and phenotypes

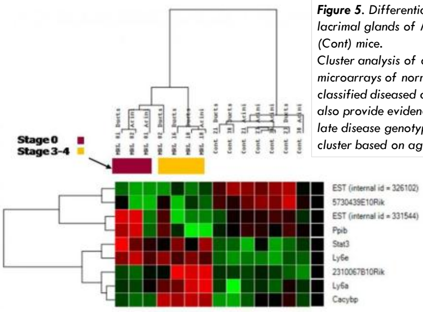


Figure 5. *Differential gene expression between lacrimal glands of MRL/lpr (MRL) and normal (Cont) mice.*
Cluster analysis of combined duct and acinar microarrays of normal and MRL/lpr mice correctly classified diseased and healthy glands. The data also provide evidence for differences in early and late disease genotype […] Control animals did not cluster based on age.

Hope E. Shaffer, Journal of Young Investigators, 19: (5) 31 October 2008; http://www.jyi.org/research/re.php?id=1664

Problems with microarray experiments

Cost ■ hard to afford to do appropriate numbers of internal controls, technical replicates (min 5) & biological replicates (min 3)

Knowledge limitations
■ Available only for known genomes!
■ Noncoding (regulatory) RNAs not yet fully represented

Quality control
■ Artifacts with image analysis
■ Artifacts with data analysis
■ Attention to experimental design needed
■ Tight collaboration with statisticians exceptionally important

SAGE is not only a herb…

- **SAGE = serial analysis of gene expression** (V. Velculescu,1995)

The principle:

- Calculate the amount of mRNA molecules using its unique, short representatives – so called „**tags**"
- Gene expression level measure is **quantitative**
- It is possible to analyze **unknown** sequences (you don't need to know genome sequence *a priori*)

Assumptions:

- It is possible to prepare tags representing every trapscript in the cell
- 18bp-long tag is unique in transcriptome (classic version of SAGE produce 14bp-long tags which are shared by some mRNAs)
- All transcripts posses appropriate restriction sites

SAGE

[1] mRNA isolation and cDNA synthesis & streptavidin binding

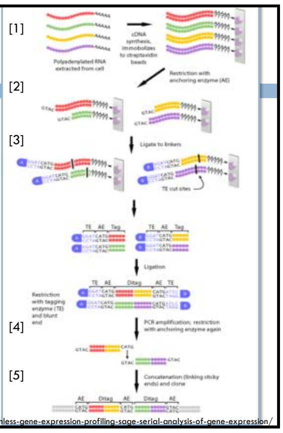
[2] tags preparation: restriction digestion with AE (frequently cutting restrictase with 4bp recognition site),

[3] ditags preparation: ligation with two types of linkers (A,B) possessing type IIS restriction sites; restriction digestion with TE (20bp away from its recognition site) – bead release, blunt ends ligation

[4] PCR amplification of ditags with primers complimentary to the two linkers' sequences (A&B), *enrichment of ditags in a mixture*

[5] AE digestion of amplified ditags (linker release) and their ligation to create a construct ready for vector cloning

Sequencing


[1]
Polyadenylated RNA extracted from cell
cDNA synthesis, immobolizes to streptavidin beads
Restriction with anchoring enzyme (AE)
[2]
GTAC
[3]
Ligate to linkers
TE cut sites
TE AE Tag
Ligation
TE AE Ditag AE TE
Restriction with tagging enzyme (TE) and blunt end
[4]
PCR amplification; restriction with anchoring enzyme again
CATG
[5]
Concatenation (linking sticky ends) and clone
AE Ditag AE Ditag AE

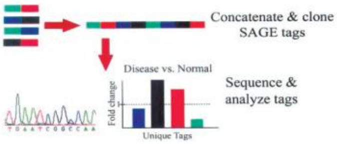
http://www.scq.ubc.ca/painless-gene-expression-profiling-sage-serial-analysis-of-gene-expression/

SAGE data analysis – general considerations

The output of SAGE:

a list of tags and the number of times they were observed

- Statistical methods are applied to count lists of tags from different samples



SAGE data analysis – general considerations

SAGE results analysis:

[1] decipher the SAGE tags from the sequence data files

[2] download a sequence database from the NCBI

[3] associate the tags to the expressed gene database

The relative transcript abundance can then be calculated by dividing the **unique tag count** by the total tags sequenced, and the **fold change** can be determined by the **ratio of tags between libraries (samples).**

Patino et al., *Circ Res. 2002;91:565-569*

SAGE data analysis – [1] from ditags to tags

- Locate the punctuation – restriction site of AE: "CATG"
- Extract ditags of length 20-26 between the punctuation
- Discard duplicate ditags (including in reverse direction) -- probably PCR artifacts
- Take extreme 10 bases as the two tags, reversing right-hand tag
- Discard linker sequences
- Count occurrences of each tag

SAGE data analysis – [3] from tags to genes

- Collect sequence records from GenBank (UniGene collection of ESTs)
- Assign sequence orientation (by finding poly-A tail or poly-A signal or from annotations)
- Extract 10-bases 3'-adjacent to 3'-most CATG
- Assign UniGene identifier to each sequence with a SAGE tag
- Record (for each tag-gene pair)
  - #sequences with this tag
  - #sequences in gene cluster with this tag

Ideal situation:
one gene = one tag

True situation:
one gene = many tags (alternative splicing & polyadenylation)
one tag = many genes (conserved 3' regions)

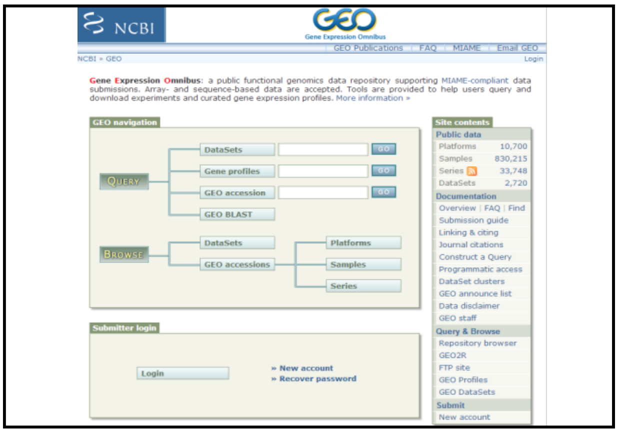
SAGE v. microarray

- SAGE generates absolute, rather than relative, measurements of RNA abundance levels
- SAGE data analysis are far easier, preprocessing less complicated; statistical methods less challenging
- It is possible to reliably compare your SAGE data to those produced by other laboratories
- SAGE may be used with unsequenced genomes (SAGE experiments may be the source of new genes discoveries!)
- With longer tags the method is highly more specific than microarray with its cross-hybridisation false positive errors risk

SAGE & microarray experimental databases

There are two main repositories:

- Gene Expression Omnibus (GEO) at NCBI
- ArrayExpress & Gene Expression Atlas at EBI


NCBI GEO Gene Expression Omnibus

GEO Publications | FAQ | MIAME | Email GEO

NCBI » GEO Login

Gene Expression Omnibus: a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles. More information »

GEO navigation

Submitter login

Site contents

Public data

Platforms	10,700
Samples	830,215
Series	33,748
DataSets	2,720

Documentation

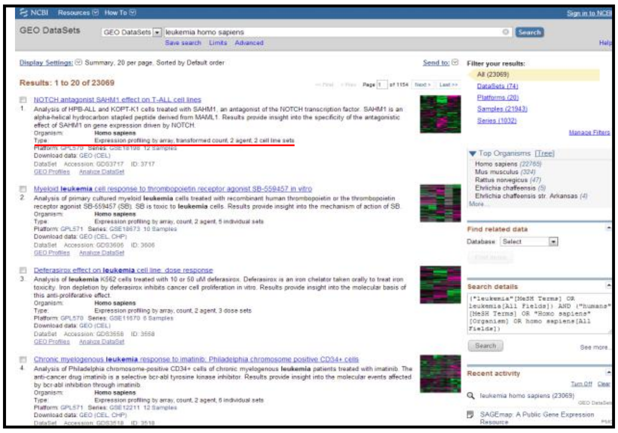
Overview | FAQ | Find
Submission guide
Linking & citing
Journal citations
Construct a Query
Programmatic access
DataSet clusters
GEO announce list
Data disclaimer
GEO staff

Query & Browse

Repository browser
GEO2R
FTP site
GEO Profiles
GEO DataSets

Submit

New account


NCBI Resources How To Sign in to NCBI
GEO DataSets
GEO DataSets
Search
Save search Limits Advanced
Help

leukemia homo sapiens

Display Settings: Summary, 20 per page, Sorted by Default order

Send to:

Results: 1 to 20 of 23069

<< First < Prev Page 1 of 1154 Next > Last >>

1. NOTCH antagonist SAHM1 effect on T-ALL cell lines
Analysis of HPB-ALL and KOPT-K1 cells treated with SAHM1, an antagonist of the NOTCH transcription factor. SAHM1 is an alpha-helical hydrocarbon stapled peptide derived from MAML1. Results provide insight into the specificity of the antagonistic effect of SAHM1 on gene expression driven by NOTCH.
Organism: **Homo sapiens**
Type: Expression profiling by array, transformed count, 2 agent, 2 cell line sets
Platform: GPL570 Series: GSE18198 12 Samples
Download data: GEO (CEL)
DataSet Accession: GDS3717 ID: 3717
GEO Profiles Analyze DataSet

2. Myeloid **leukemia** cell response to thrombopoietin receptor agonist SB-559457 in vitro
Analysis of primary cultured myeloid **leukemia** cells treated with recombinant human thrombopoietin or the thrombopoietin receptor agonist SB-559457 (SB). SB is toxic to **leukemia** cells. Results provide insight into the mechanism of action of SB.
Organism: **Homo sapiens**
Type: Expression profiling by array, count, 2 agent, 5 individual sets
Platform: GPL571 Series: GSE18673 10 Samples
Download data: GEO (CEL, CHP)
DataSet Accession: GDS3606 ID: 3606
GEO Profiles Analyze DataSet

3. Deferasirox effect on **leukemia** cell line: dose response
Analysis of **leukemia** K562 cells treated with 10 or 50 uM deferasirox. Deferasirox is an iron chelator taken orally to treat iron toxicity. Iron depletion by deferasirox inhibits cancer cell proliferation in vitro. Results provide insight into the molecular basis of this anti-proliferative effect.
Organism: **Homo sapiens**
Type: Expression profiling by array, count, 2 agent, 3 dose sets
Platform: GPL570 Series: GSE11670 6 Samples
Download data: GEO (CEL)
DataSet Accession: GDS3558 ID: 3558
GEO Profiles Analyze DataSet

4. Chronic myelogenous **leukemia** response to imatinib: Philadelphia chromosome positive CD34+ cells
Analysis of Philadelphia chromosome-positive CD34+ cells of chronic myelogenous **leukemia** patients treated with imatinib. The anti-cancer drug imatinib is a selective bcr-abl tyrosine kinase inhibitor. Results provide insight into the molecular events affected by bcr-abl inhibition through imatinib.
Organism: **Homo sapiens**
Type: Expression profiling by array, count, 2 agent, 6 individual sets
Platform: GPL571 Series: GSE12211 12 Samples
Download data: GEO (CEL, CHP)
DataSet Accession: GDS3518 ID: 3518

Filter your results:
All (23069)
DataSets (74)
Platforms (20)
Samples (21943)
Series (1032)
Manage Filters

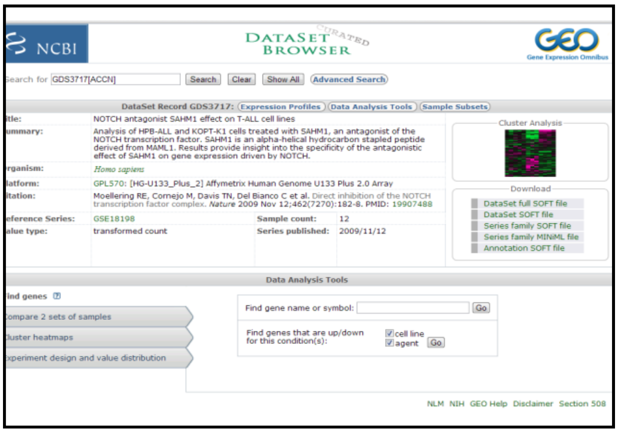
Top Organisms [Tree]
Homo sapiens (22765)
Mus musculus (324)
Rattus norvegicus (47)
Ehrlichia chaffeensis (5)
Ehrlichia chaffeensis str. Arkansas (4)
More...

Find related data
Database: Select
Find items

Search details

```
("leukemia"[MeSH Terms] OR
leukemia[All Fields]) AND ("humans"
[MeSH Terms] OR "Homo sapiens"
[Organism] OR homo sapiens[All
Fields])
```

Recent activity
Turn Off Clear
leukemia homo sapiens (23069) GEO DataSets
SAGEmap: A Public Gene Expression Resource



Search for GDS3717[ACCN]

DataSet Record GDS3717: Expression Profiles | Data Analysis Tools | Sample Subsets

itle:	NOTCH antagonist SAHM1 effect on T-ALL cell lines		
ummary:	Analysis of HPB-ALL and KOPT-K1 cells treated with SAHM1, an antagonist of the NOTCH transcription factor. SAHM1 is an alpha-helical hydrocarbon stapled peptide derived from MAML1. Results provide insight into the specificity of the antagonistic effect of SAHM1 on gene expression driven by NOTCH.		
rganism:	*Homo sapiens*		
latform:	GPL570: [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array		
itation:	Moellering RE, Cornejo M, Davis TN, Del Bianco C et al. Direct inhibition of the NOTCH transcription factor complex. *Nature* 2009 Nov 12;462(7270):182-8. PMID: 19907488		
eference Series:	GSE18198	Sample count:	12
alue type:	transformed count	Series published:	2009/11/12

Cluster Analysis

Download

DataSet full SOFT file
DataSet SOFT file
Series family SOFT file
Series family MINiML file
Annotation SOFT file

Data Analysis Tools

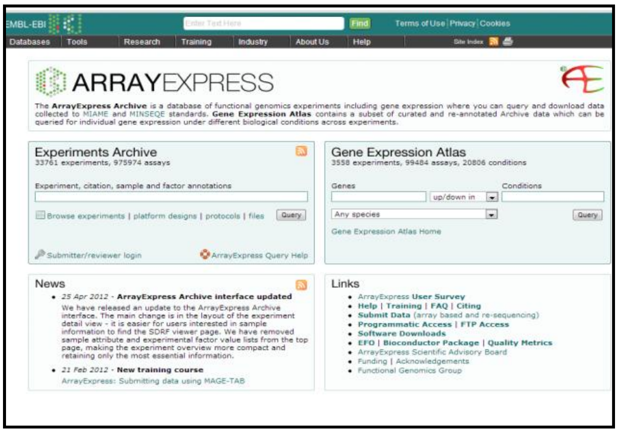
ind genes

ompare 2 sets of samples
luster heatmaps
xperiment design and value distribution

Find gene name or symbol: Go

Find genes that are up/down for this condition(s): ☑ cell line ☑ agent Go

NLM NIH GEO Help Disclaimer Section 508



The **ArrayExpress Archive** is a database of functional genomics experiments including gene expression where you can query and download data collected to MIAME and MINSEQE standards. **Gene Expression Atlas** contains a subset of curated and re-annotated Archive data which can be queried for individual gene expression under different biological conditions across experiments.

Experiments Archive
33761 experiments, 975974 assays

Experiment, citation, sample and factor annotations

Browse experiments | platform designs | protocols | files Query

Submitter/reviewer login ArrayExpress Query Help

Gene Expression Atlas
3558 experiments, 99484 assays, 20806 conditions

Genes up/down in Conditions

Any species Query

Gene Expression Atlas Home

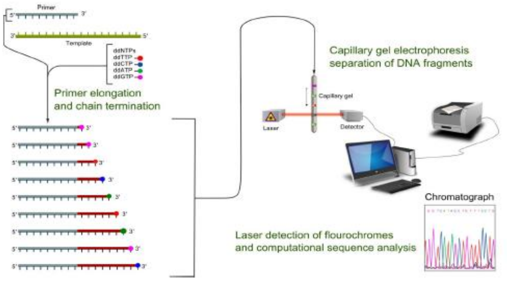
News

- *25 Apr 2012* - **ArrayExpress Archive interface updated**
We have released an update to the ArrayExpress Archive interface. The main change is in the layout of the experiment detail view - it is easier for users interested in sample information to find the SDRF viewer page. We have removed sample attribute and experimental factor value lists from the top page, making the experiment overview more compact and retaining only the most essential information.
- *21 Feb 2012* - **New training course**
ArrayExpress: Submitting data using MAGE-TAB

Links

- ArrayExpress **User Survey**
- **Help | Training | FAQ | Citing**
- **Submit Data** (array based and re-sequencing)
- **Programmatic Access | FTP Access**
- **Software Downloads**
- **EFO | Bioconductor Package | Quality Metrics**
- ArrayExpress Scientific Advisory Board
- Funding | Acknowledgements
- Functional Genomics Group

Ist generation sequencing



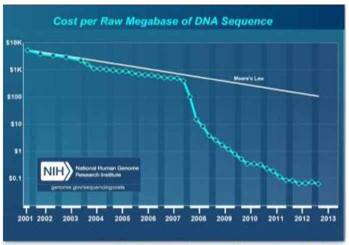
Conventional DNA sequencing limitations

- □ The rate-limiting step: the need to separate randomly terminated DNA fragments by gel electrophoresis
- □ Relatively low number of samples could be analyzed in parallel
- □ Total automation of the sample preparation methods is difficult
- □ DNA fragments need to be cloned into bacteria for larger sequences
- □ High cost of sequencing
- □ Sequencing errors. Level of sensitivity (generally estimated at 10-20%) insufficient for detection of clinically relevant low-level mutant alleles or organisms.
- □ *cis or trans orientation of* heterozygous positions may be difficult to resolve during data analysis.
- □ Not readily scalable to achieve a throughput capable of efficiently analyzing complex diploid genomes at low cost.
- □ de novo genome assembly is difficult

Fakruddin and Abhijit Chowdhury, Pyrosequencing-An Alternative to Traditional Sanger Sequencing, American Journal of Biochemistry and Biotechnology 8 (1): 14-20, 2012

New (next) generation sequencing

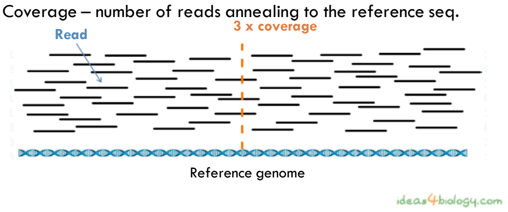
- □ Fragment DNA, multiply them and sequence in parralel huge number of short fragments



http://www.genome.gov/sequencingcosts/

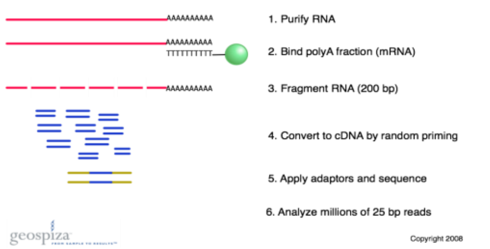
A few new words…

- □ Library – collection of DNA/RNA fragments with appropriate adaptors (sequences added for amplification and sequencing)
- □ Read – result of the sequencing experiment (adaptor seq. Most ofer deleted by sequencing machine software)
- □ Coverage – number of reads annealing to the reference seq.



RNA Sequencing

Steps in Preparing an RNA-Seq Library



1. Purify RNA
2. Bind polyA fraction (mRNA)
3. Fragment RNA (200 bp)
4. Convert to cDNA by random priming
5. Apply adaptors and sequence
6. Analyze millions of 25 bp reads

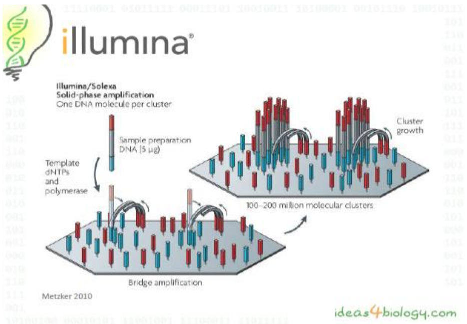
Copyright 2008

http://finchtalk.geospiza.com/2008_09_01_archive.html

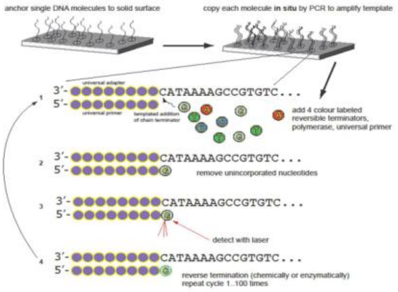
NGS platforms

- □ **Illumina** Genome Analyzer Iix, HiSeq HiSeq 2000, MiSeq
  (*in situ* synthesis approach; short reads: 50-250 nt; most frequently cited platform)
- □ Life Sciences/**Roche 454**
  (pyrosequencing; longer reads – up to 700 nt)
- □ **ABI Solid** Sequencing System
  (2-mer ligation strategy; short reads: 50 nt; the lowest cost per site)
- □ **Ion Torrent** (pH measurement, the lowest cost of the instrument)
- □ **Pacific Biosciences** (immobilised single polimerase molecule)
- □ Nanopore sequencing – research stage
- □ your idea?

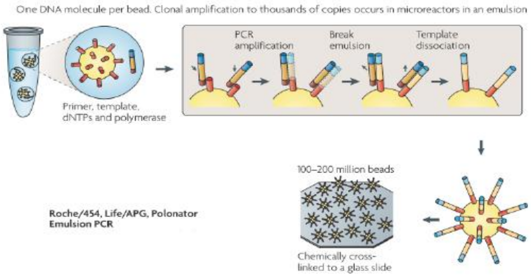
Bridge-PCR based cluster formation



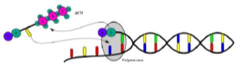
In situ synthesis approach


anchor single DNA molecules to solid surface
copy each molecule in situ by PCR to amplify template
universal adapter
1
3'-CATAAAAGCCGTGTC...
5'-
universal primer
templated addition of chain terminator
add 4 colour labeled reversible terminators, polymerase, universal primer
2
3'-CATAAAAGCCGTGTC...
5'-
remove unincorporated nucleotides
3
3'-CATAAAAGCCGTGTC...
5'-
detect with laser
4
3'-CATAAAAGCCGTGTC...
5'-
reverse termination (chemically or enzymatically)
repeat cycle 1..100 times

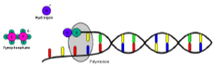
Roche 454 – emultion PCR



ion torrent by life technologies™

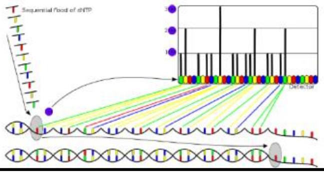


Polymerase integrates a nucleotide.



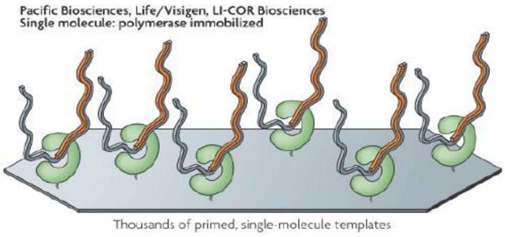
Hydrogen and pyrophosphate are released.

Hydrogen ions level (pH) is measured

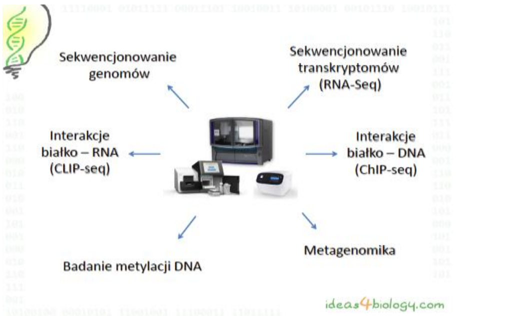


www.wikipedia.com

Immobilised polymerase instead of DNA


Pacific Biosciences, Life/Visigen, LI-COR Biosciences
Single molecule: polymerase immobilized
Thousands of primed, single-molecule templates

NGS applications



Sekwencjonowanie genomów

Sekwencjonowanie transkryptomów (RNA-Seq)

Interakcje białko – RNA (CLIP-seq)

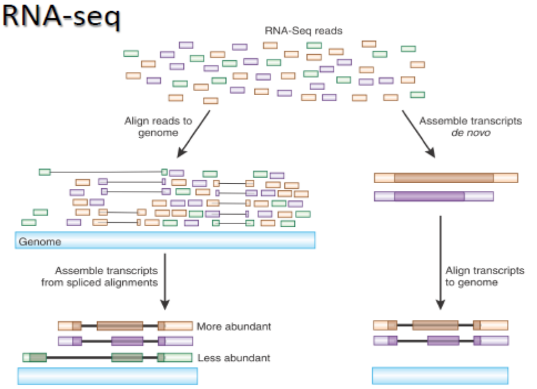
Interakcje białko – DNA (ChIP-seq)

Badanie metylacji DNA

Metagenomika

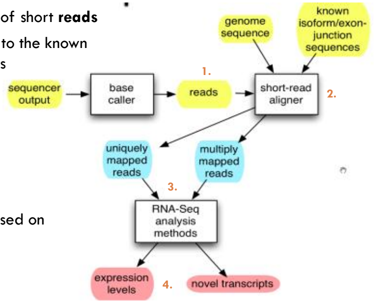
ideas4biology.com

RNA-seq


RNA-Seq reads
Align reads to genome
Assemble transcripts de novo
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

RNA Sequencing data analysis - the prnciples

1. You obtain: huge amount of short **reads**
2. They have to be **aligned** to the known genes/trascript sequences
3. You get mapped **read counts**
4. As a result you get info about the level of expression/ the presence of newly discovered transcripts based on read counts



http://www.docstoc.com/docs/48949299/Transcriptome-analysis-methods-for-RNA-Seq-data

Raw and Aligned Reads

- Raw data is a (large) set of sequences
- Typical read file format is FASTQ

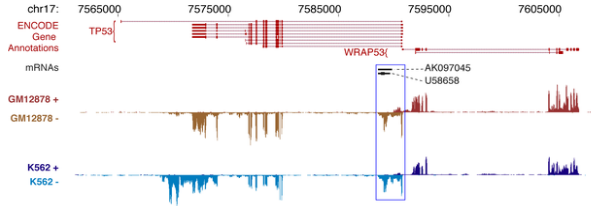
```
@HWI-EAS255_4_FC2010Y_1_43_110_790      Read identifier
TTAATCTACAGAATAGATAGCTAGCATATATTT       Bases called
+
hhhhhhhhhhhhhhhhdhhhhhhhhhhhhdRehdh     Base quality codes
```

- Alignment to genome is done by efficient indexing of seed sequences
- Aligned reads typically are in SAM format:

```
@HWI-… 163 chr19 9900 10000 16M2I25M
```

Read identifier — Where this read matched — Start and end positions — Codes for match: 16 matches, 2 extra,…

Finally RNA-Seq data is often represented by 'pile-up' diagrams:



From: The ENCODE Project Consortium (2011) A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol 9(4): e1001046.

This and previous slide from: Mark Reimers, Virginia Institute for Psychiatric and Behavioral Genetics, Transcriptome Analysis by High-Throughput Sequencing (RNA-Seq) – .ppt avaiable from: http://www.rna-seqblog.com/tag/transcriptome-analysis/

NGS results depositories – SRA (@NCBI)

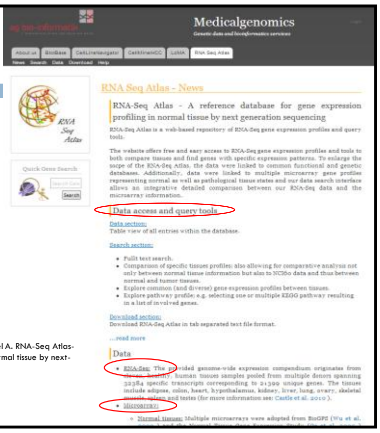
- SRA = Sequence Read Archive
- Data in SRA come from the following platforms: 454, IonTorrent, Illumina, SOLiD, Helicos, Complete Genomics (data are cataloged by the method used)

http://www.ncbi.nlm.nih.gov/Traces/sra/

- ENA – European Nucleotide Archive (@EBI)
- GEO – Gene expression Omnibus (@NCBI)

RNA Sequencing resources:

www.rna-seqblog.com
www.seqanswers.com
www.blueseq.com



Medicalgenomics
Genetic data and bioinformatics services

RNA Seq Atlas - News

RNA-Seq Atlas - A reference database for gene expression profiling in normal tissue by next generation sequencing

RNA-Seq Atlas is a web-based repository of RNA-Seq gene expression profiles and query tools.

The website offers free and easy access to RNA-Seq gene expression profiles and tools to both compare tissues and find genes with specific expression patterns. To enlarge the socpe of the RNA-Seq Atlas, the data were linked to common functional and genetic databases. Additionally, data were linked to multiple microarray gene profiles representing normal as well as pathological tissue states and our data search interface allows an integrative detailed comparison between our RNA-Seq data and the microarray information.

Data access and query tools

Data section:
Table view of all entries within the database.

Search section:

- Fullt text search.
- Comparison of specific tissues profiles: also allowing for comparative analysis not only between normal tissue information but also to NCI60 data and thus between normal and tumor tissues.
- Explore common (and diverse) gene expression profiles between tissues.
- Explore pathway profile: e.g. selecting one or multiple KEGG pathway resulting in a list of involved genes.

Download section:
Download RNA-Seq Atlas in tab separated text file format.

...read more

Data

- RNA-Seq: The provided genome-wide expression compendium originates from eleven healthy human tissues samples pooled from multiple donors spanning 32384 specific transcripts corresponding to 21399 unique genes. The tissues include adipose, colon, heart, hypothalamus, kidney, liver, lung, ovary, skeletal muscle, spleen and testes (for more information see: Castle et al. 2010).
- Microarray:
  - Normal tissues: Multiple microarrays were adopted from BioGPS (Wu et al.

Krupp M, Marquardt JU, Sahin U, Galle PR, Castle J, Teufel A. RNA-Seq Atlas--a reference database for gene expression profiling in normal tissue by next-generation sequencing. Bioinformatics. 2012 28:1184-5 ; http://medicalgenomics.org/rna_seq_atlas