KTH Royal Institute of Technology
School of Biotechnology

Bioinformatics and Biostatistics BB2440 – lecture 14

Gene expression
RNA-seq

Olof Emanuelsson
olofem@kth.se

Lecture 14, 2015-10-01, 08:15-10:00 FA32

Enabler for Life Sciences

SciLifeLab

Gene expression. RNA-seq.

1. Why measure gene expression
2. History of gene expression analysis
3. Microarrays for gene expression
4. Microarray data analysis
5. RNA sequencing (RNA-seq)
6. RNA-seq data analysis
7. Summary

Reading instructions lecture 14 (today)

Z.B.:

C15: 599-604 (*not* SAGE)

C16: 625-630 (end after section 'Expression levels are often…')

651-657 (end after 'Nonparametric tests…'; *not* Box16.2)

Wang, Gerstein, and Snyder:

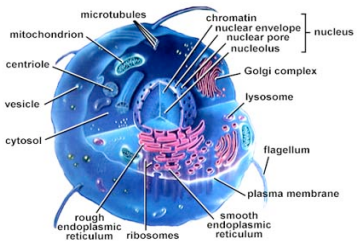
"RNA-seq: a revolutionary tool for transcriptomics". *Nature Rev Genet* vol. **10,** p. 57-63 (2009):

57-61, 63 (i.e., *not* 'New transcriptomic insights')

('Glossary' on page 62 is useful)

[1] Why measure gene expression

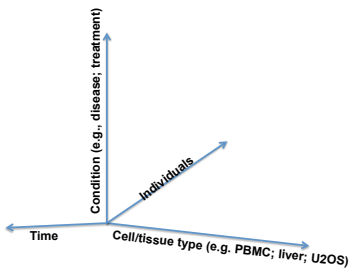
The eukaryotic cell



www.terrebonneonline.com

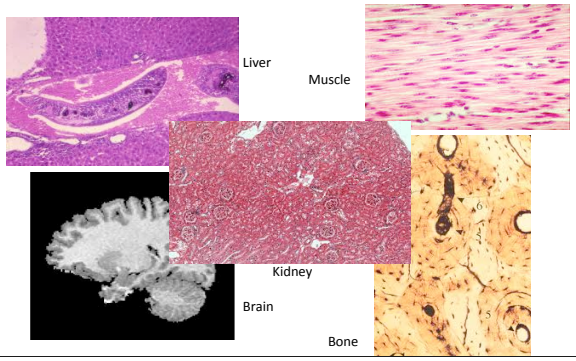
Cells are different

.



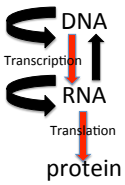
(A fifth axis could be different species)

Tissues are different



What are the effectors of diversity?

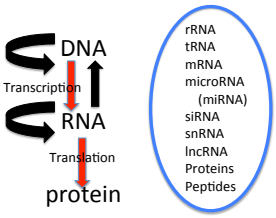
Transcribed and translated entities



rRNA
tRNA
mRNA
microRNA
(miRNA)
siRNA
snRNA
lncRNA
Proteins
Peptides

What are the effectors of diversity?

Transcribed and translated entities



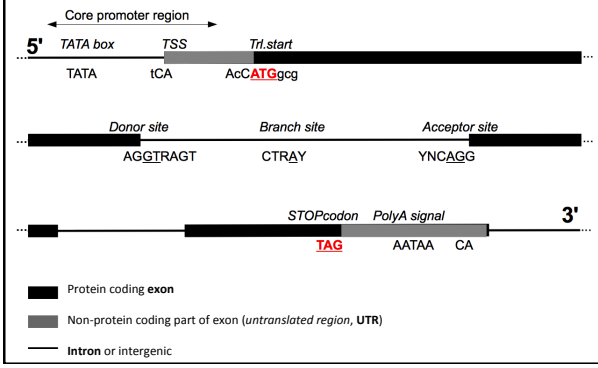
rRNA
tRNA
mRNA
microRNA
(miRNA)
siRNA
snRNA
lncRNA
Proteins
Peptides

The effectors are biopolymers:
DNA – 4 bases: A, C, G, T
RNA – 4 bases: A, C, G, U
Proteins – 20 amino acids: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y

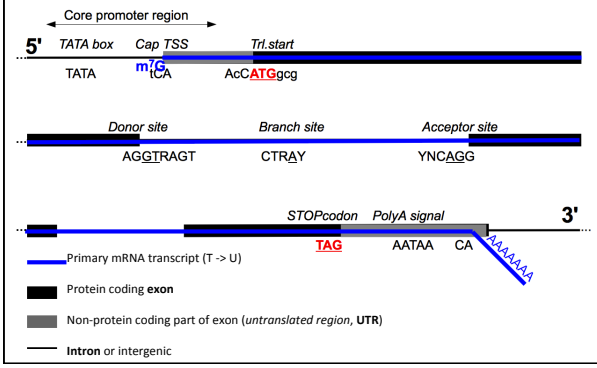
What are the effectors of diversity?

mRNA is the template for proteins
microRNA (miRNA) targets and regulates mRNA
tRNA are needed in protein synthesis
rRNA are part of the ribosome
lncRNA epigenetic activity
Proteins are enzymes, signaling molecules, building blocks, …

mRNA is transcribed from genes

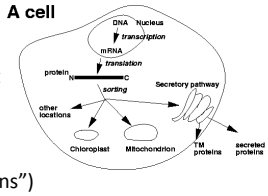


mRNA is transcribed from genes



Gene expression in the cell

The primary mRNA transcript is processed:
Exons are kept, introns spliced out
⇒ mature mRNA
⇒ which is translated to protein
⇒ which may be further modified
("post-translational modifications")



Measure the mRNA levels in a sample =>
Information about what proteins are active =>
Information about the biological processes in the sample

Gene expression differs between tissues

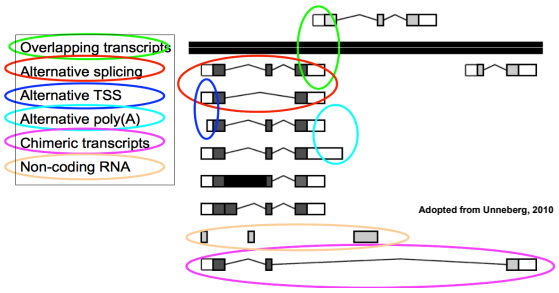
In a human cell: ca. 20,000 protein coding genes.

40-60% of these are expressed in a particular tissue type at a particular time point.

Different genes expressed in different human tissues.

Also differences between healthy and disease tissues, different developmental stages, different cell cycle stages, etc..

Gene vs. transcripts



One **gene** can produce many possible **transcripts**.

Gene – a genomic sequence encoding a functional product (or several functional products)

Transcript – an mRNA species transcribed from a *gene.* One gene may produce many different transcripts. Each transcript is typically represented by many identical mRNA molecules (dynamic range of transcription).

Transcriptome – the set of *transcripts* present in a cell/ tissue/organism (at a particular time point or integrated over many or all time points)

Transcriptomics – finding out everything about the *transcriptome*

Goals of gene expression studies

1. Detect what genes are expressed within a sample
2. Quantify expression of the genes within a sample
3. Differential expression of genes between samples

⇒ **Define a set of genes that play a role in the sample**

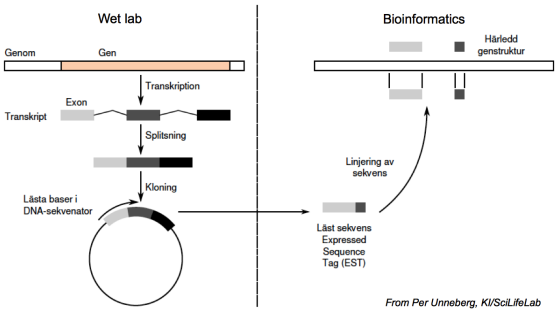
⇒ **Understand what's going on in the sample through further investigation of the interesting genes:**

A. Further bioinformatics investigation

B. Wet lab experiments targeted towards the interesting genes

[2] History of gene expression analysis

Expressed sequence tag – EST



From Per Unneberg, KI/SciLifeLab

The history of gene expression analysis

EST – expressed sequence tag

Other tag-based – e.g.
CAGE (Cap analysis of gene expression; 5', 20 nt)
MPSS – massively parallel signature sequencing (~20 nt)

nt = nucleotide

Microarray – cDNA or oligo arrays; up to 20 million features ("gene expression microarrays").

Tiling microarray – covering the entire non-repetitive part of a genome (not only genes)

RNA-seq – current state of the art

[3] Microarrays for gene expression

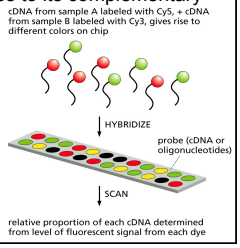
Gene expression microarray

A microarray contains a large number of *probes* (*spots* or *array features)*. Each feature contains part of the DNA sequence of a gene (or other genome feature).

The mRNA, turned into cDNA, hybridizes to its complementary sequence (if it is present).

Two-colour arrays: label 2 different samples with 2 different colours that compete for hybridization. (Also called two-channel arrays).

One-colour arrays: label one sample; one colour, no competition.



The workflow when using DNA microarrays

1. Collect the RNA from the samples

2. Convert the RNA to cDNA (by reverse transcription) and label it with a fluorescing compound (e.g. Cy5, Cy3).

3. Pour the labelled cDNA onto the microarray, where it hybridizes to complementary DNA strings attached to the surface of the microarray (these are called *probes*, and the labelled cDNA is called *target*).

4. Wash away unbound sample.

5. Scan the array with a laser to generate a picture of what has hybridized => you obtain for each probe an intensity (one intensity recording per fluorescent)

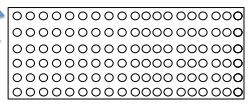
6. Analyze the intensities.

Two-colour microarray

Label cDNA from [a] normal (with Cy3) and [b] disease (with Cy5) tissue.

Let them hybridize competitively to complementary DNA strands attached to a glass slide (microarray).

normal
disease



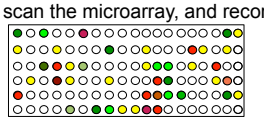
Each circle ("array feature") contains attached DNA (probes) representing a gene. The probes are PCR products of genes, or oligonucleotides (designed from the genomic sequence).

Two-colour microarray

Label cDNA from [a] normal (with Cy3) and [b] disease (with Cy5) tissue.

Let them hybridize competitively to complementary DNA strands attached to a glass slide (microarray).

Use a laser to scan the microarray, and record the fluorescence intensities

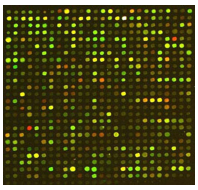


Some genes upregulated in disease tissue (here: red).
Some downregulated in disease tissue (here: green).
Some are unchanged (here: yellow).

Two-colour microarray

A real two-colour microarray picture (only small part of it):

(*C. elegans*)



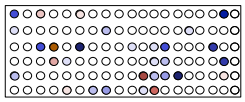
The colours are not visible to the eye, they are representations of the *intensities* that were obtained when scanning the microarray with the laser. The spots are very small, typically on the order of *µm*.

One-colour microarray

Label cDNA from [a] normal or [b] disease tissue.

Let them hybridize to complementary DNA strands attached to a glass slide (microarray) or magnetic beads (bead array).

Use a laser to scan the microarray, and record the fluorescence intensities



Each circle ("array feature") contains attached DNA (probes) representing a gene. The probes on one-colour microarrays are typically oligonucleotides of length 25-80 nt, designed from the genomic sequence. Affymetrix GeneChip, Illumina BeadChip

Output data from microarrays: light intensities

For each array feature on the array (gene ≈ array feature), a set of raw intensities are recorded: (2-colour array example)

gene ID	I_R	I_G
gene_001	165	106
gene_002	1329	224
gene_003	51	184

…

[4] Microarray data analysis

Data normalization and transformation

The intensities are supposed to reflect the expression level of the genes. However, there is a lot of **artefacts** and **noise** in microarray data:

1. uneven distribution of the sample on the microarray

2. uneven number of DNA molecules attached at each array feature

3. all probes have different sequences, hence they have different T_m, resulting in different ability to hybridize at the single fixed temperature at which the experiment is performed

4. if more than one array is involved (which should be the case since you want to assess the technical reproducibility), you need to adjust for different sample amounts on the microarrays.

… and **much** more => the data has to be *normalized*

Data normalization and transformation

Get the data in a form where it is comparable with the data from other arrays. To achieve this you have to:

1.adjust for all the factors described on the previous slide

2. get the data (if possible) into a form where standard parametric statistical methods can be used. To do this one usually performs the following operations:

- take the ratio of the two intensities I_R and I_G
- take the logarithm (use base 2) of the ratio

=>You end up with the signal $S_i = \log_2\left(\frac{I_R}{I_G}\right)$ for each feature *i*.

- adjust all S_i values with a constant so that the median *S* will be 0. (S_i = 0 means no difference between the samples)

Why log-transforming the data?

If I_R= 5 and I_G= 1, then I_R / I_G = 5 but...

If I_R=1 and I_G=5, then I_R / I_G = 0.2

---> *limes* 0 when I_G ---> infinity

All ratios where the I_G sample has the higher intensity will be squished between 0 and 1.

Taking the log of the data will give a distribution that is more symmetric especially in the tails.

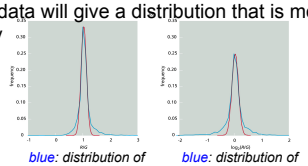


Figure 16.4
red: a normal distribution
blue: distribution of I_R/I_G ratios
blue: distribution of $log_2(I_R/I_G)$ ratios

The intensities

For each gene on the array (gene ≈ array feature), a set of normalized intensities and their log ratios:

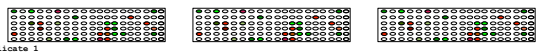
gene ID	I_R	I_G	I_R/I_G	$log_2(I_R/I_G)$
gene_001	145.3	106.2	1.368	0.452
gene_002	1029.0	150.8	6.824	2.771
gene_003	66.4	180.7	0.367	-1.446
…				

Replicates

Replicates needed: (1) assess technical variation
(2) enable the use of statistical methods

The same sample probed on three different microarrays:

Normal vs. disease — Replicate 1 | Normal vs. disease — Replicate 2 | Normal vs. disease — Replicate 3



Replicate 1

gene ID	I_R	I_G	I_R/I_G	$log_2(I_R/I_G)$
gene_001	145.3	106.2	1.368	0.452
gene_002	1029.0	150.8	6.824	2.771
gene_003	66.4	180.7	0.367	-1.446
. . .				

Replicate 2

gene ID	I_R	I_G	I_R/I_G	$log_2(I_R/I_G)$
gene_001	156.3	148.0	1.056	0.079
gene_002	2227.6	250.1	8.907	3.155
gene_003	47.9	152.4	0.314	-1.670
. . .				

Replicate 3

gene ID	I_R	I_G	I_R/I_G	$log_2(I_R/I_G)$
gene_001	122.0	151.8	0.804	-0.315
gene_002	3939.1	292.5	13.467	3.751
gene_003	100.4	152.6	0.658	-0.604
. . .				

The intensities – replicates

For each gene, normalized log ratios for each technical replicate (normal vs. disease) hybridization will be obtained:

gene ID	$log_2(I_R/I_G)_{Rep1}$	$log_2(I_R/I_G)_{Rep2}$	$log_2(I_R/I_G)_{Rep3}$
gene_001	0.452	0.079	-0.315
gene_002	2.771	3.155	3.751
gene_003	-1.446	-1.670	-0.604
…			

Assessing differential expression

What genes are differentially expressed between 2 samples? How much must the expression levels differ between samples?
Very often a 2-fold change is required (*ad hoc* threshold…).
Corresponding to a log_2 change >= 1.
The log_2 value is called the **effect size**.

Microarray intensities

gene ID	I_R	I_G	I_R/I_G	$log_2(I_R/I_G)$
gene_001	145.3	106.2	1.368	0.452
gene_002	1029.0	150.8	6.824	2.771
gene_003	66.4	180.7	0.367	-1.446
. . .				

[Note that "effect size" is a generic name for something that shows how much a measurement of something differs from another measurement of the same thing; it is not always defined as a fold change or log(fold change)].

Assessing differential expression with replicates

Use replicates to calculate averages and use these to determine differential expression.

Replicate 1

gene ID	I_R	I_G	I_R/I_G	$\log_2(I_R/I_G)$
gene_001	145.3	106.2	1.368	0.452
gene_002	1029.0	150.8	6.824	2.771
gene_003	66.4	180.7	0.367	-1.446
. . .				

Replicate 2

gene ID	I_R	I_G	I_R/I_G	$\log_2(I_R/I_G)$
gene_001	156.3	148.0	1.056	0.079
gene_002	2227.6	250.1	8.907	3.155
gene_003	47.9	152.4	0.314	-1.670
. . .				

Replicate 3

gene ID	I_R	I_G	I_R/I_G	$\log_2(I_R/I_G)$
gene_001	122.0	151.8	0.804	-0.315
gene_002	3939.1	292.5	13.467	3.751
gene_003	100.4	152.6	0.658	-0.604
. . .				

Averages and variances:

	Mean $\log_2(I_R/I_G)$	**var** $\log_2(I_R/I_G)$
gene_001	0.072	0.147
gene_002	3.226	0.244
gene_003	-1.240	0.316

Assessing statistical significance of differential expression with replicates

Need to assess the statistical significance of the observed differences in expression.
Replicates needed to enable the use of statistical methods.

Microarray intensities

gene ID	$\log_2(I_R/I_G)_{Rep1}$	$\log_2(I_R/I_G)_{Rep2}$	$\log_2(I_R/I_G)_{Rep3}$
gene_001	0.452	0.079	-0.315
gene_002	2.771	3.155	3.751
gene_003	-1.446	-1.670	-0.604
. . .			

Given our great normalization efforts mentioned earlier, we can (almost) assume that the $\log_2$ ratios are normally distributed and with mean=0 for the unchanged genes.
We can then use the *t-test* to test whether our measurements (i.e. the three $\log_2$ ratios for each gene) indicate that the $\log_2$ ratios truly are different from 0.
H0: true mean is 0 (i.e., null hypothesis is that the two measurements are the same)
H1: true mean is not 0

Assessing statistical significance of differential expression with replicates

Microarray intensities

gene ID	$\log_2(I_R/I_G)_{Rep1}$	$\log_2(I_R/I_G)_{Rep2}$	$\log_2(I_R/I_G)_{Rep3}$
gene_001	0.452	0.079	-0.315
gene_002	2.771	3.155	3.751
gene_003	-1.446	-1.670	-0.604
. . .			

H0: true mean is 0
H1: true mean is not 0
Significance level typically set to 0.05 = α = false positive rate (fp/(fp+tn))

Use R:

```
> gene_001 <-c(0.452,0.079,-0.315)
> t.test(gene_001,mu=0)
t = 0.3251, df = 2, p-value = 0.776
..
> gene_002 <-c(2.771,3.155,3.751)
> t.test(gene_002,mu=0)
t = 11.3142, df = 2, p-value = 0.007721
..
> gene_003 <-c(-1.446,-1.670,-0.604)
> t.test(gene_003,mu=0)
t = -3.812, df = 2, p-value = 0.06217
```

Thus only `gene_002` is significantly differentially expressed at α = 0.05.

Assessing statistical significance of differential expression with more replicates

Microarray intensities

gene ID	$\log_2(I_R/I_G)_{Rep1}$	$\log_2(I_R/I_G)_{Rep2}$	$\log_2(I_R/I_G)_{Rep3}$	$\log_2(I_R/I_G)_{Rep4}$	$\log_2(I_R/I_G)_{Rep5}$	$\log_2(I_R/I_G)_{Rep6}$
gene_001	0.452	0.079	-0.315	0.083	-0.221	-0.188
gene_002	2.771	3.155	3.751	3.167	0.659	2.694
gene_003	-1.446	-1.670	-0.604	-0.976	-1.592	-1.152
. . .						

What if having more replicates?
Increased statistical power?

H0: true mean is 0
H1: true mean is not 0
Significance level typically set to 0.05 = α = false positive rate (fp/(fp+tn))

Use R:

```
> gene_003 <-c(-1.446,-1.670,-0.604,-0.976,-1.592,-1.152)
> t.test(gene_003,mu=0)
t = -7.4407, df = 5, p-value = 0.0006913
```

Thus now also `gene_003` is significantly differentially expressed at α = 0.05. Despite the fact that the mean value of the $\log_2(I_R/I_G)$ for `gene_003` is the same as in the previous example (= -1.24)

[5] RNA sequencing (RNA-seq)

RNA-sequencing: **RNA-seq**

Microarrays have several problems:
- Limited dynamic range
- Built-in uncertainty: probes have different T_m – impossible to optimize experimental conditions
- Must define beforehand what you are looking for (probe design)

The recent advances within the sequencing technology has enabled fast and cheap sequencing of DNA. This can also be used to sequence RNA.
RNA-seq has the advantages:
- No need to pre-define what you are looking for
- Better detection of lowly expressed genes
- Possible to detect alternative splicing
- Possible to detect variation (*i.e.*, mutations in an individual)

RNA-seq

The state-of-the-art approach for analysis of gene expression
The most widely used sequencing based molecular biology assay

Desired measurement: RNA abundance → **Prepare for sequencing:** make cDNA libraries → **Sequencing** → **Bioinformatics:** Quality control; Quantification → **Bioinformatics:** Differential expression

Published papers with "RNA-seq" or "RNA-sequencing" in title or abstract (from Pubmed):

Year	Papers
2014	2456
2013	1583
2012	859
2011	478
2010	219
2009	59
2008	16

RNA-seq

What application(s) are you interested in:

- mRNA abundance
- differential expression
- novel transcription
- antisense transcription
- transcriptome reconstruction
- allele-specific expression
- non-coding RNAs

RNA-seq

What application(s) are you interested in:

- mRNA abundance – *what genes are expressed*
- differential expression – *difference in expression between two samples*
- novel transcription
- antisense transcription
- transcriptome reconstruction
- allele-specific expression
- non-coding RNAs

RNA-seq

Several different technologies available for sequencing.

They differ in: sequencing chemistry, amplification strategy, read length, number of reads, base calling accuracy, sequencing errors – rates and types, …

The output is called "a **read**" – a stretch of DNA sequence.

Sequencing platforms

ABI 3730xl Sanger Sequencing

454 Life Sciences pyrosequencing

SOLiD, Illumina

Pacific Biosciences, Oxford Nanopore etc Single-molecule sequencing

	ABI 3730xl Sanger Sequencing	454 Life Sciences pyrosequencing	SOLiD, Illumina	Pacific Biosciences, Oxford Nanopore etc Single-molecule sequencing
Length/read	800 bp	400 bp	150 bp	20 000+ bp
Reads/run	96	1 million	2 billion	5 million
Bases/run	60 kbp	400 Mbp	500 Gbp	100 Gbp
Speed (HG=human genome, *i.e.* 3 Gbases)	10 years/HG	1 month/HG	1 day/HG	10 min/HG
	"old school"	"2nd gen"		"3rd gen"

Illumina **HiSeq**

The most popular choice for RNA-sequencing
Read length up to 2x150 bases (paired-end reads)

Typically >10^9 reads corresponding to >200 Gbases per run

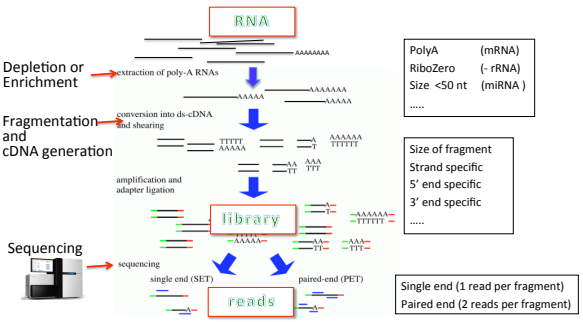
RNA-seq: Library preparation

From RNA to sequencing-ready DNA molecules

1. Sample extraction and quality checks on samples
2. rRNA depletion or mRNA enrichment
   [rRNA=ribosomal RNA, ca. **90%** of total RNA content]
3. Fragmentation
4. Random priming
5. cDNA generation (reverse transcription)
6. Adapter ligation, cluster generation

=> RNAs converted to DNA and ready for sequencing

RNA to reads



Note: it is not a linear relationship between the reads and the original RNA molecules

RNA-seq: output data

Fastq: *de facto* standard for output files

(1) DNA sequence for each read

(2) Quality for each base in each read

```
@SEQ_ID_1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
@SEQ_ID_2
TCCTA…
```

~25,000,000 reads per RNA-seq sample (typically)

>1,000,000,000 reads per machine run

RNA-seq: base quality

Base quality: $Q = -10 \log_{10} P$

Probability of wrong base: $P = 10^{\frac{-Q}{10}}$

Quality scores:

Quality score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

Base quality encoded using ASCII characters in fastq file.

[6] RNA-seq data analysis

RNA-seq bioinformatics

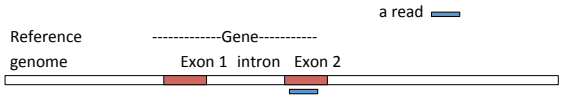
Reads have been generated. Then, typically:

1. Map (align) reads to reference genome or transcriptome, *or*
   Reconstruct the transcriptome from the reads without a reference
2. Count the reads in an entity of interest (gene, transcript, exon, …)
3. Quantify the abundance for the entity of interest
4. Differential expression between samples

Map reads to a reference

Find the place (=map or align) on the genome (=reference) from which the read originated (=was transcribed).

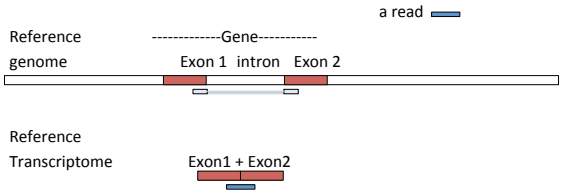
The reference genome is the genomic DNA sequence of the organism.



Map reads to a reference

Find the place (=map or align) on the *transcriptome* (=reference) from which the read originated.

The reference transcriptome is the total set of RNAs of the organism.



Tools for this: BWA, Bowtie, Maq, SOAP, Gnumap, ...

Map reads to a reference

Find the place (=map or align) on the genome (=reference) from which the read originated (=was transcribed). *Allow spliced reads*.

The reference genome is the genomic DNA sequence of the organism.

Tools for this: TopHat, GSNAP, STAR, ...

Map reads to a reference

Reads from the sequencer:

CAATCAGATGATCCCACTGTGG

AGACGATCCCACTGTGGGGTG

GTGAAGTGTCCGTAGATGTGTG

GCAAATGCAATCAGACGATCCC

Gene(or transcript) sequence (reference sequence):

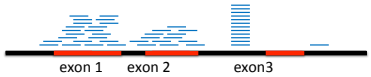
...GCAAATGCAATCAGACGGTCCCACTGTGGGGTGTGAAGTGTCCGTAGAGCTGTGAGAGAA...

Map reads to a reference

Adapted from Mikael Huss, SciLifeLab

Count the reads

Mapping result:

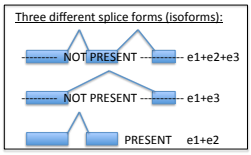


Exon 1: 21

Exon 2: 13

Exon 3: 0

Total for the gene: 34



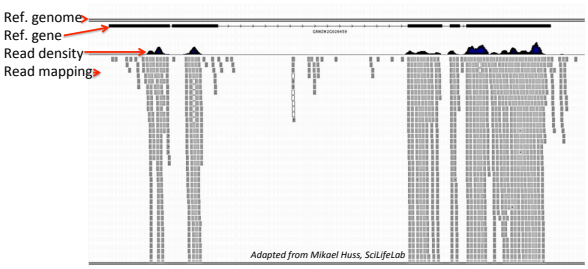
I.e., it seems as if only exon1 and exon2 of the gene is actually transcribed.

Question: Would this splicing pattern be detected if a microarray had been used?

Count the reads

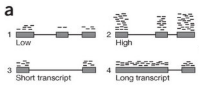
Microarrays give a **continuous** (floating-point) expression value for each gene

RNA-seq gives an **integer** value for each gene ("digital expression"): read counts



Estimate abundance

Read counts can be misleading:



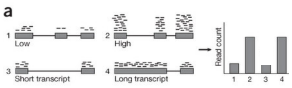
Transcript #3: 12 reads

Transcript #4: 31 reads

But… are #3 and #4 unequally expressed?

Estimate abundance

Read counts can be misleading:



Transcript #3: 12 reads

Transcript #4: 31 reads

But… are #3 and #4 unequally expressed?

Estimate abundance

=> Longer genes/transcripts are expected to generate more reads
=> The more you sequence, the more reads you get from each gene

RPKM – reads per kilobase of transcript per million mapped reads

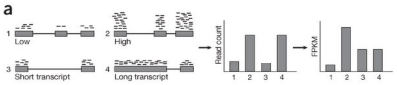
$$R = \frac{10^9 C}{LN}$$

- *R*=*RPKM* value (or *FPKM* for paired end reads)
- *C*= number of reads mapped to the transcript
- *L*= gene length
- *N*= number of million mappable reads

RPKM normalizes for: (i) transcript length and (ii) number of reads.
(FPKM: *fragments* per kilobase of transcript per million mapped reads)

Estimate abundance

Use *RPKM* instead of read counts:



Transcript #3: 12 reads

Transcript #4: 31 reads

But their RPKM is the same.

Tools for this: ERANGE, Myrna, eXpress

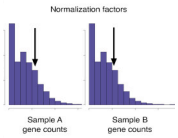
Differential expression (DE)

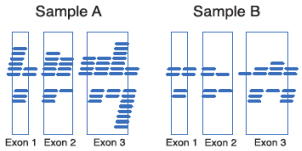
When is a difference in read count also statistically significant?
=> Model the variability in read count for each gene across replicates.

The read count variability has been modelled using
Poisson distribution (e.g., DEGseq, Myrna)
  models the variance between technical replicates
Negative binomial (e.g. edgeR, DEseq, CuffDiff)
  also models the overdispersion of read counts between biological replicates (biological replicates are less similar than technical replicates)

<u>Output is, for each gene</u>
⇒ a ***P*-value** describing the probability that the difference in counts is due to chance. (Should be corrected for multiple hypothesis testing).
⇒ **Effect size** (fold change)

Differential expression (DE)





P-value = 2.68e-4

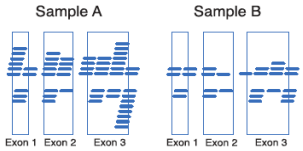
Some DE-tools calculate between-sample scaling factors (instead of using RPKMs).

Normalized counts or RPKM obtained for each gene in the two samples

[1] *P*-value is calculated

[2] Effect size is calculated (fold change) based on read counts *or* RPKM (approach differs between different

Differential expression (DE)



RPKM definition:

$$R = \frac{10^9 C}{LN}$$

P-value = 2.68e-4

<u>Effect size calculation, raw counts (*C*):</u>

Counts(Sample A): 54

Counts(Sample B): 27

Fold change (*C*(A)/*C*(B)): 54/27=2

Log_2 (Fold change): $log_2(54/27)=1$

<u>Effect size calculation, RPKM (*R*):</u>

RPKM(Sample A):

RPKM(Sample B):

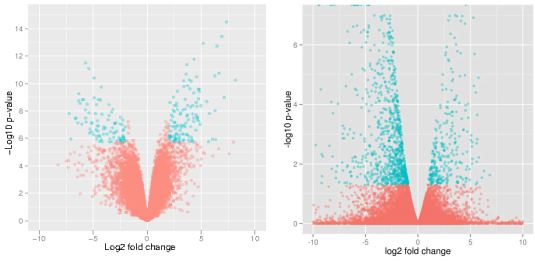
Fold change (*R*(A)/*R*(B)):

Log_2 (Fold change):

Do you have enough information to complete this calculation?

Volcano plot – effect size *vs.* p-value

Plot $-log_{10}$*(p-value) vs.* log_2*(fold change)*. Showing two different experiments in these two plots. Each dot represents one gene.



Red: **not reported** as differentially expressed gene
Cyan: **reported** as differentially expressed gene

Multiple testing correction

20000 genes in human genome and using an α = 0.05

⇒ We would expect 0.05 * 20000 = 1000 genes to be considered differentially expressed by random chance

⇒ Need to correct the p-values for the fact that we perform many independent tests on the same data set.

Simplest: Bonferroni correction: multiply each p-value with the number of tests performed. Then compare with α.

p-value = $2*10^{-4}$ => corrected p-value is 4 which is > α

p-value = $3*10^{-7}$ => corrected p-value is 0.006 which is < α

Better: Benjamini-Hochberg correction.

Multiple testing correction is included in (most) RNA-seq differential expression analysis tools.

[7] Summary

- Measure mRNA level to find out the biological processes in a sample
- Gene expression differs between tissues, between individuals, between different treatments, etc.
- One gene may produce many different transcripts
- Microarrays use hybridization and the output data are intensities
- RNA-seq uses cDNA sequencing and the output data are sequence reads
- Library preparation, e.g., depletion vs. enrichment, fragmentation
- The raw output data needs to be normalized ("pre-processed")
- Map RNA-seq reads to reference genome (or transcriptome)
- RPKM (or FPKM for paired-end reads) for quantification of reads in RNA-seq experiment
- Differential expression and statistical testing (microarrays and RNA-seq)
- The use of replicates to measure differential expression
- P-value and effect size
- RNA-seq advantages: dynamic range; no need to predefine what to look for; higher sensitivity; variant (mutation) detection; fairly cheap

Lab 6

RNA sequencing, differential gene expression
Thursday 2015-10-15
08:00-12:00 in ”4V2Röd”

Galaxy (galaxy.org): NOTE: need to register (see lab instructions)
- FastQC
- Cuffdiff
- SAMtools
- Picardtools

R (Volcano plot)
GTF files
SAM/BAM files
FPKM

Reading instructions lecture 15 (2015-01-02)

Z.B.:

C15: 606-611 (start at section *“The simplest method…”*)

C16: 631-646 (end after section about SOMs). But *not*:

‘The Mahalanobis distance…’