Introduction to RNA sequencing

Bioinformatics perspective

Olga Dethlefsen

NBIS, National Bioinformatics Infrastructure Sweden

November 2017

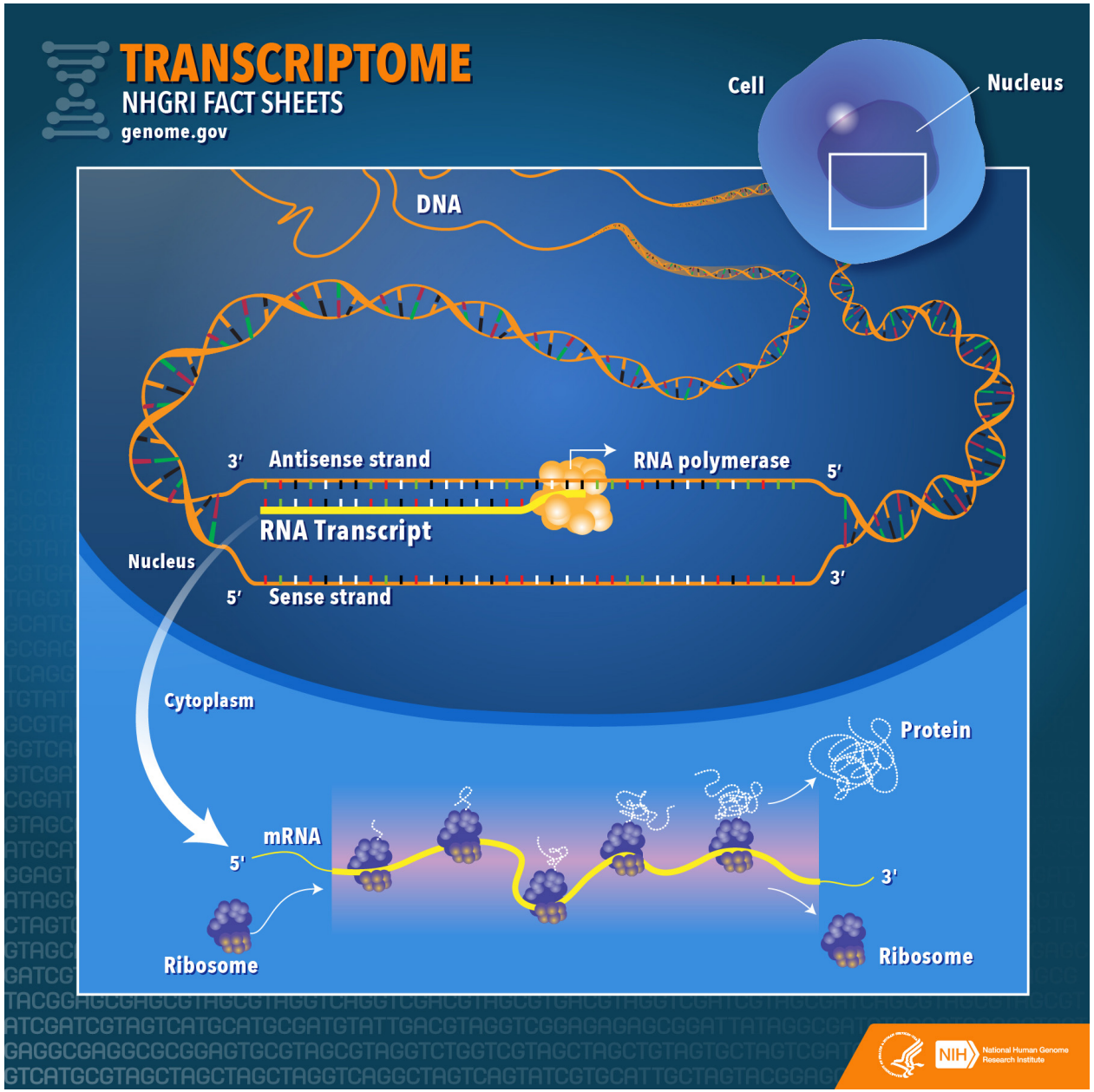
NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Outline

- Why sequence transcriptome?
- From RNA to sequence
- The most common way: reference based analysis pipeline
- What about de-novo assembly of transcriptomes?
- And what about scRNA-seq?
- Introduction to exercises

Why sequence transcriptome?


TRANSCRIPTOME
NHGRI FACT SHEETS
genome.gov
Cell
Nucleus
DNA
3′
Antisense strand
RNA polymerase
5′
RNA Transcript
Nucleus
5′
Sense strand
3′
Cytoplasm
Protein
mRNA
5'
3'
Ribosome
Ribosome
NIH
National Human Genome Research Institute

An RNA sequence mirrors the sequence of the DNA from which it was transcribed.

Consequently, by analyzing transcriptome we can determine when and where each gene is turned on or off in the cells and tissues of an organism.

What can a transcriptome tell us about?

- gene sequences in genomes
- gene functions
- gene activity / gene expression
- isoforms and allelic expression
- fusion transcripts and novel transcripts
- SNPs in genes
- co-expression of genes
- cell-to-cell heterogeneity (scRNA-seq)

Transcriptomes are:

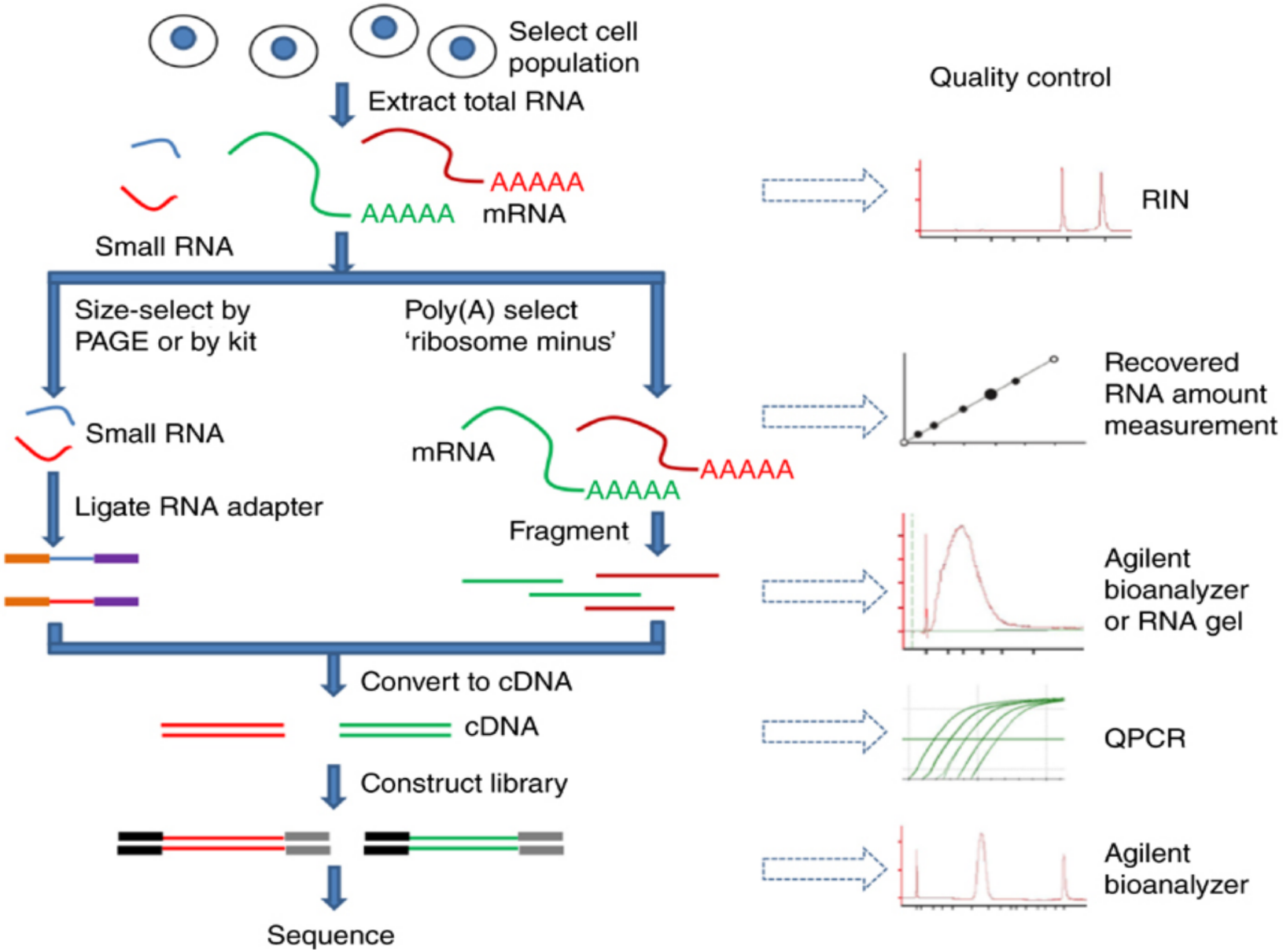
dynamic, that is not the same over tissues and time points

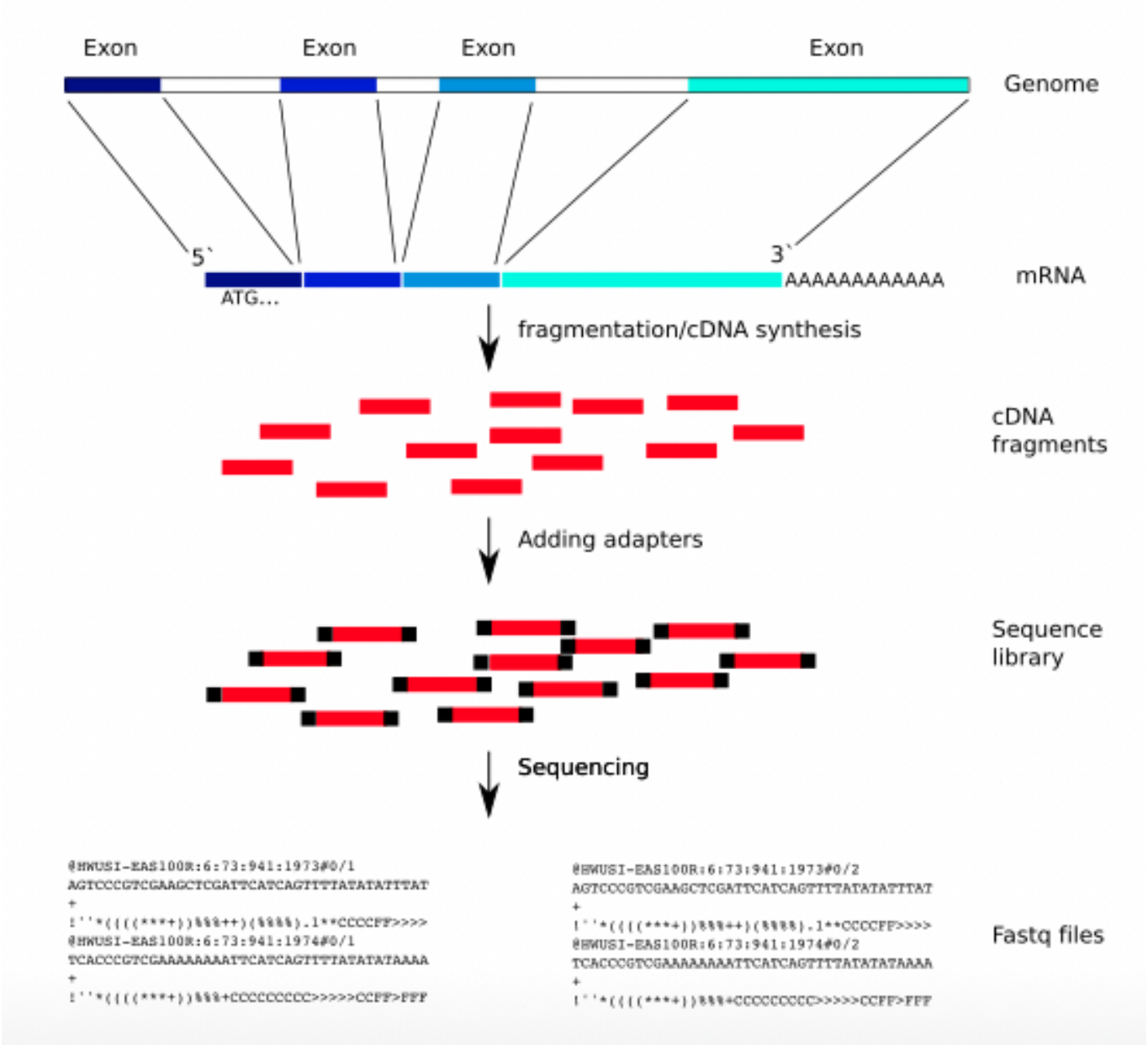
directly derived from functional genomics elements, that is mostly protein-coding genes, providing a useful functionally relevant subset of the genome, translating into smaller sequence space

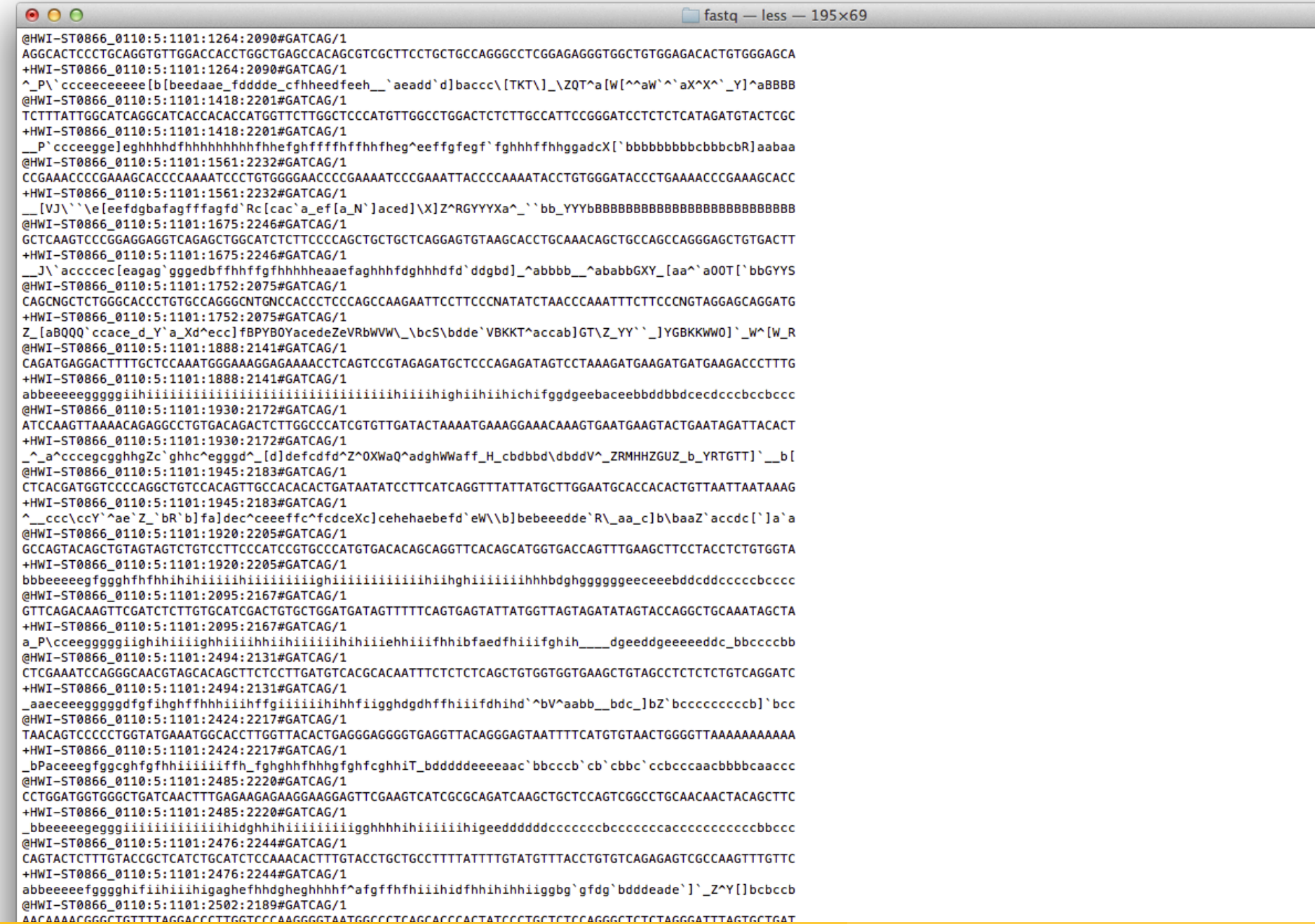
Overview

- Experimental design (biology, medicine, statistics)
- RNA extraction (biology, biotechnology)
- Library preparation (biology, biotechnology)
- High throughput sequencing (engineering, biology, chemistry, biotechnology, bioinformatics)
- Data processing (bioinformatics)
- Data analysis (bioinformatics & biostatistics)

From RNA to sequence


Select cell population
Extract total RNA
AAAAA
AAAAA
mRNA
Small RNA
Size-select by PAGE or by kit
Poly(A) select 'ribosome minus'
Small RNA
mRNA
AAAAA
AAAAA
Ligate RNA adapter
Fragment
Convert to cDNA
cDNA
Construct library
Sequence
Quality control
RIN
Recovered RNA amount measurement
Agilent bioanalyzer or RNA gel
QPCR
Agilent bioanalyzer


Exon
Exon
Exon
Exon
Genome
5`
3`
ATG...
AAAAAAAAAAAA
mRNA
fragmentation/cDNA synthesis
cDNA fragments
Adding adapters
Sequence library
Sequencing
@HWUSI-EAS100R:6:73:941:1973#0/1
AGTCCCGTCGAAGCTCGATTCATCAGTTTTATATATTTAT
+
!''*((((****+))%%%++)(%%%%).1**CCCCFF>>>>
@HWUSI-EAS100R:6:73:941:1974#0/1
TCACCCGTCGAAAAAAAATTCATCAGTTTTATATATAAAA
+
!''*((((****+))%%%+CCCCCCCCC>>>>>CCFF>FFF
@HWUSI-EAS100R:6:73:941:1973#0/2
AGTCCCGTCGAAGCTCGATTCATCAGTTTTATATATTTAT
+
!''*((((****+))%%%++)(%%%%).1**CCCCFF>>>>
@HWUSI-EAS100R:6:73:941:1974#0/2
TCACCCGTCGAAAAAAAATTCATCAGTTTTATATATAAAA
+
!''*((((****+))%%%+CCCCCCCCC>>>>>CCFF>FFF
Fastq files


fastq — less — 195×69
@HWI-ST0866_0110:5:1101:1264:2090#GATCAG/1
AGGCACTCCCTGCAGGTGTTGGACCACCTGGCTGAGCCACAGCGTCGCTTCCTGCTGCCAGGGCCTCGGAGAGGGTGGCTGTGGAGACACTGTGGGAGCA
+HWI-ST0866_0110:5:1101:1264:2090#GATCAG/1
^_P\`cceeeceeeee[b[beedaae_fdddde_cfhheedfeeh__`aeadd`d]baccc\[TKT\]_\ZQT^a[W[^^aW`^`aX^X^`_Y]^aBBBB
@HWI-ST0866_0110:5:1101:1418:2201#GATCAG/1
TCTTTATTGGCATCAGGCATCACCACACCATGGTTCTTGGCTCCCATGTTGGCCTGGACTCTCTTGCCATTCCGGGATCCTCTCTCATAGATGTACTCGC
+HWI-ST0866_0110:5:1101:1418:2201#GATCAG/1
__P`cceeggge]eghhhhdfhhhhhhhhhfhhefghffffhffhhfheg^eeffgfegf`fghhhffhhggadcX[`bbbbbbbbbcbbbcbR]aabaa
@HWI-ST0866_0110:5:1101:1561:2232#GATCAG/1
CCGAAACCCCGAAAGCACCCCAAAATCCCTGTGGGGAACCCCGAAAATCCCGAAATTACCCCAAAATACCTGTGGGATACCCTGAAAACCCGAAAGCACC
+HWI-ST0866_0110:5:1101:1561:2232#GATCAG/1
__[VJ\``\e[eefdgbafagfffagfd`Rc[cac`a_ef[a_N`]aced]\X]Z^RGYYYXa^_``bb_YYYbBBBBBBBBBBBBBBBBBBBBBBBBB
@HWI-ST0866_0110:5:1101:1675:2246#GATCAG/1
GCTCAAGTCCCGGAGGAGGTCAGAGCTGGCATCTCTTCCCCAGCTGCTGCTCAGGAGTGTAAGCACCTGCAAACAGCTGCCAGCCAGGGAGCTGTGACTT
+HWI-ST0866_0110:5:1101:1675:2246#GATCAG/1
__J\`accccec[eagag`gggedbffhhffgfhhhhheaaefaghhhfdghhhdfd`ddgbd]_^abbbb__^ababbGXY_[aa^`aOOT[`bbGYYS
@HWI-ST0866_0110:5:1101:1752:2075#GATCAG/1
CAGCNGCTCTGGGCACCCTGTGCCAGGGCNTGNCCACCCTCCCAGCCAAGAATTCCTTCCCNATATCTAACCCAAATTTCTTCCCNGTAGGAGCAGGATG
+HWI-ST0866_0110:5:1101:1752:2075#GATCAG/1
Z_[aBQQQ`ccace_d_Y`a_Xd^ecc]fBPYBOYacedeZeVRbWVW_\bcS\bdde`VBKKT^accab]GT\Z_YY``_]YGBKKWWO]`_W^[W_R
@HWI-ST0866_0110:5:1101:1888:2141#GATCAG/1
CAGATGAGGACTTTTGCTCCAAATGGGAAAGGAGAAAACCTCAGTCCGTAGAGATGCTCCCAGAGATAGTCCTAAAGATGAAGATGATGAAGACCCTTTG
+HWI-ST0866_0110:5:1101:1888:2141#GATCAG/1
abbeeeeeggggiihiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiihiiihighiihiihichifggdgeebaceebbddbbdcecdcccbccbccc
@HWI-ST0866_0110:5:1101:1930:2172#GATCAG/1
ATCCAAGTTAAAACAGAGGCCTGTGACAGACTCTTGGCCCATCGTGTTGATACTAAAATGAAAGGAAACAAAGTGAATGAAGTACTGAATAGATTACACT
+HWI-ST0866_0110:5:1101:1930:2172#GATCAG/1
_^_a^cccegcgghhgZc`ghhc^egggd^_[d]defcdfd^Z^OXWaQ^adghWWaff_H_cbdbbd\dbddV^_ZRMHHZGUZ_b_YRTGTT]`__b[
@HWI-ST0866_0110:5:1101:1945:2183#GATCAG/1
CTCACGATGGTCCCCAGGCTGTCCACAGTTGCCACACACTGATAATATCCTTCATCAGGTTTATTATGCTTGGAATGCACCACACTGTTAATTAATAAAG
+HWI-ST0866_0110:5:1101:1945:2183#GATCAG/1
^__ccc\ccY`^ae`Z_`bR`b]fa]dec^ceeeffc^fcdceXc]cehehaebefd`eW\\b]bebeeedde`R_aa_c]b\baaZ`accdc[`]a`a
@HWI-ST0866_0110:5:1101:1920:2205#GATCAG/1
GCCAGTACAGCTGTAGTAGTCTGTCCTTCCCATCCGTGCCCATGTGACACAGCAGGTTCACAGCATGGTGACCAGTTTGAAGCTTCCTACCTCTGTGGTA
+HWI-ST0866_0110:5:1101:1920:2205#GATCAG/1
bbbeeeeegfggghfhfhhihihiiiiihiiiiiiiiiighiiiiiiiiihhhbdghgggggeeceeebddcddcccccbcccc
@HWI-ST0866_0110:5:1101:2095:2167#GATCAG/1
GTTCAGACAAGTTCGATCTCTTGTGCATCGACTGTGCTGGATGATAGTTTTTCAGTGAGTATTATGGTTAGTAGATATAGTACCAGGCTGCAAATAGCTA
+HWI-ST0866_0110:5:1101:2095:2167#GATCAG/1
a_P\ccegggggiighihiiiighhiiihhiihiiiiiihihiiiehhiiifhhibfaedfhiiifghih____dgeeddgeeeeeddc_bbccccbb
@HWI-ST0866_0110:5:1101:2494:2131#GATCAG/1
CTCGAAATCCAGGGCAACGTAGCACAGCTTCTCCTTGATGTCACGCACAATTTCTCTCTCAGCTGTGGTGGTGAAGCTGTAGCCTCTCTCTGTCAGGATC
+HWI-ST0866_0110:5:1101:2494:2131#GATCAG/1
_aaeceeeggggdfgfihghffhhhiiihffgiiiiihihhfiigghdgdhffhiiifdhihd`^bV^aabb__bdc_]bZ`bccccccccb]`bcc
@HWI-ST0866_0110:5:1101:2424:2217#GATCAG/1
TAACAGTCCCCCTGGTATGAAATGGCACCTTGGTTACACTGAGGGAGGGGTGAGGTTACAGGGAGTAATTTTCATGTGTAACTGGGGTTAAAAAAAAAAA
+HWI-ST0866_0110:5:1101:2424:2217#GATCAG/1
_bPaceeegfggcghfgfhhiiiiiiffh_fghghhfhhhgfghfcghhiT_bddddeeeeaac`bbcccb`cb`cbbc`ccbcccaacbbbbcaaccc
@HWI-ST0866_0110:5:1101:2485:2220#GATCAG/1
CCTGGATGGTGGGCTGATCAACTTTGAGAAGAGAAGGAAGGAGTTCGAAGTCATCGCGCAGATCAAGCTGCTCCAGTCGGCCTGCAACAACTACAGCTTC
+HWI-ST0866_0110:5:1101:2485:2220#GATCAG/1
_bbeeeeegegggiiiiiiiiiiiiihidghhihiiiiiiiiiggghhhhhihiiiiiihigeeddddddccccccccbccccccccacccccccccccbbccc
@HWI-ST0866_0110:5:1101:2476:2244#GATCAG/1
CAGTACTCTTTGTACCGCTCATCTGCATCTCCAAACACTTTGTACCTGCTGCCTTTTATTTTGTATGTTTACCTGTGTCAGAGAGTCGCCAAGTTTGTTC
+HWI-ST0866_0110:5:1101:2476:2244#GATCAG/1
abbeeeeefgggghifiihiiihigaghefhhdgheghhhhf^afgffhfhiiihidfhhihihhiiggbg`gfdg`bdddeade`]`_Z^Y[]bcbccb
@HWI-ST0866_0110:5:1101:2502:2189#GATCAG/1

.fastq

@MISEQ:233:000000000-AGJP2:1:1101:15260:1358
CTGTAAATTGCCTGACTTGCTAATTGTGATTAACTTAGTTT
+
BBBBBFFFFFFFGGGGGGGGGGHFFFHGHHGFFHHHHHAG

- Line1:

.fastq

```
@MISEQ:233:000000000-AGJP2:1:1101:15260:1358
CTGTAAATTGCCTGACTTGCTAATTGTGATTAACTTAGTTT
+
BBBBBFFFFFFFGGGGGGGGGGHFFFHGHHGFFHHHHHAG
```

- Line1: begins with a '@' character and is followed by a sequence identifier and an optional description
- Line2:

.fastq

@MISEQ:233:000000000-AGJP2:1:1101:15260:1358
CTGTAAATTGCCTGACTTGCTAATTGTGATTAACTTAGTTT
+
BBBBBFFFFFFFGGGGGGGGGGHFFFHGHHGFFHHHHHAG

- Line1: begins with a '@' character and is followed by a sequence identifier and an optional description
- Line2: is the raw sequence letters
- Line3:

.fastq

```
@MISEQ:233:000000000-AGJP2:1:1101:15260:1358
CTGTAAATTGCCTGACTTGCTAATTGTGATTAACTTAGTTT
+
BBBBBFFFFFFFGGGGGGGGGGHFFFHGHHGFFHHHHHAG
```

- Line1: begins with a '@' character and is followed by a sequence identifier and an optional description
- Line2: is the raw sequence letters
- Line3: begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again
- Line4:

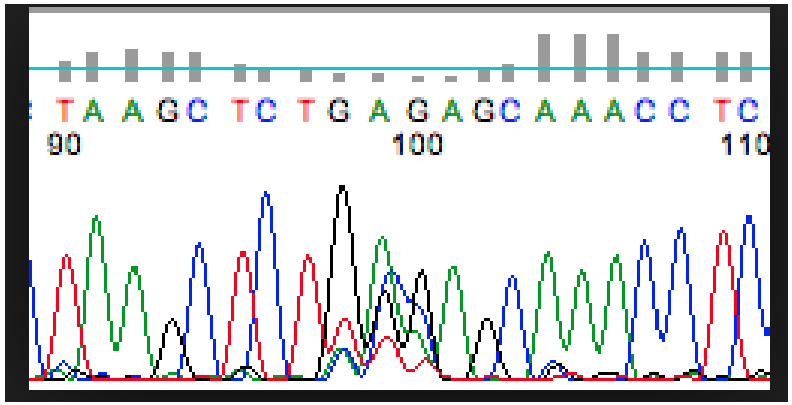
.fastq

```
@MISEQ:233:000000000-AGJP2:1:1101:15260:1358
CTGTAAATTGCCTGACTTGCTAATTGTGATTAACTTAGTTT
+
BBBBBFFFFFFFGGGGGGGGGGHFFFHGHHGFFHHHHHAG
```

- Line1: begins with a '@' character and is followed by a sequence identifier and an optional description
- Line2: is the raw sequence letters
- Line3: begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again
- Line4: encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence

Phred Quality Score

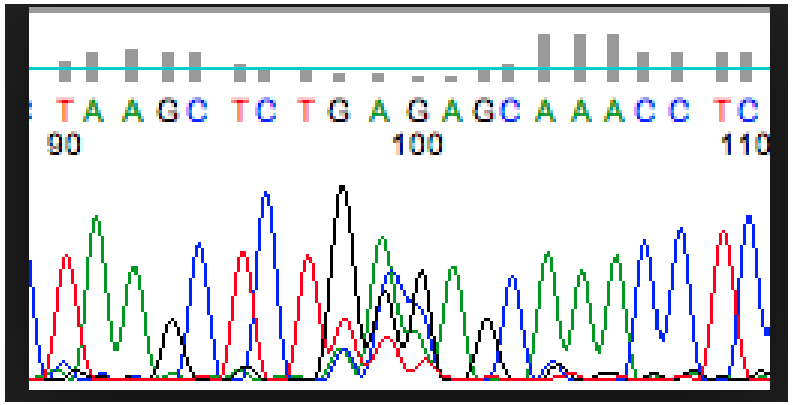
- Q = -10 x log P
- where:
  - P, probability of base calling being incorrect
  - High Q = high probability of the base being correct



- A Phred quality score of 10 to a base means that the base is called incorrectly in 1 out of...

Phred Quality Score

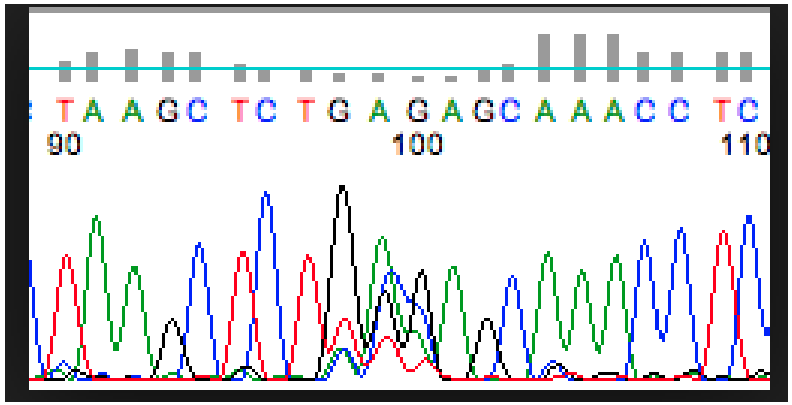
- Q = -10 x log P
- where:
  - P, probability of base calling being incorrect
  - High Q = high probability of the base being correct



- A Phred quality score of 10 to a base means that the base is called incorrectly in 1 out of...10 times.
- A Phred quality score of 20 to a base, means that the base is called incorrectly in 1 out of...

Phred Quality Score

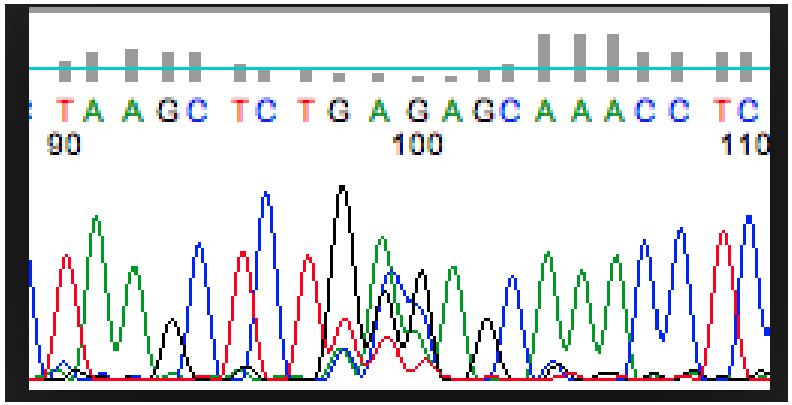
- Q = -10 x log P
- where:
  - P, probability of base calling being incorrect
  - High Q = high probability of the base being correct



- A Phred quality score of 10 to a base means that the base is called incorrectly in 1 out of...10 times.
- A Phred quality score of 20 to a base, means that the base is called incorrectly in 1 out of...100 times.
- A Phred quality score of 30 to a base, means that the base is called incorrectly in 1 out of...

Phred Quality Score

- Q = -10 x log P
- where:
  - P, probability of base calling being incorrect
  - High Q = high probability of the base being correct



- A Phred quality score of 10 to a base means that the base is called incorrectly in 1 out of...10 times.
- A Phred quality score of 20 to a base, means that the base is called incorrectly in 1 out of...100 times.
- A Phred quality score of 30 to a base, means that the base is called incorrectly in 1 out of...1000 times etc...

PE, paired-end

- Two .fastq files are created per sequenced library
- The order of reads in files is identical and naming of reads is the same with the exception of the end information
- The way of naming reads are changing over time so the read names depend on software version

```
@61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCCC@@CACCCCCA
```

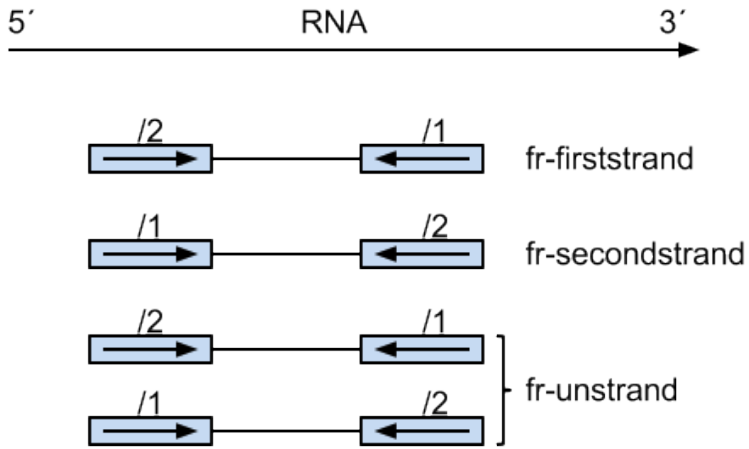
```
@61DFRAAXX100204:1:100:10494:3070/2
ATCCAAGTTAAAACAGAGGCCTGTGACAGACTCTTGGCCCATCGTGTTGATA
+
_^_a^cccegcgghhgZc`ghhc^egggd^_[d]defcdfd^Z^OXWaQ^ad
```

SE

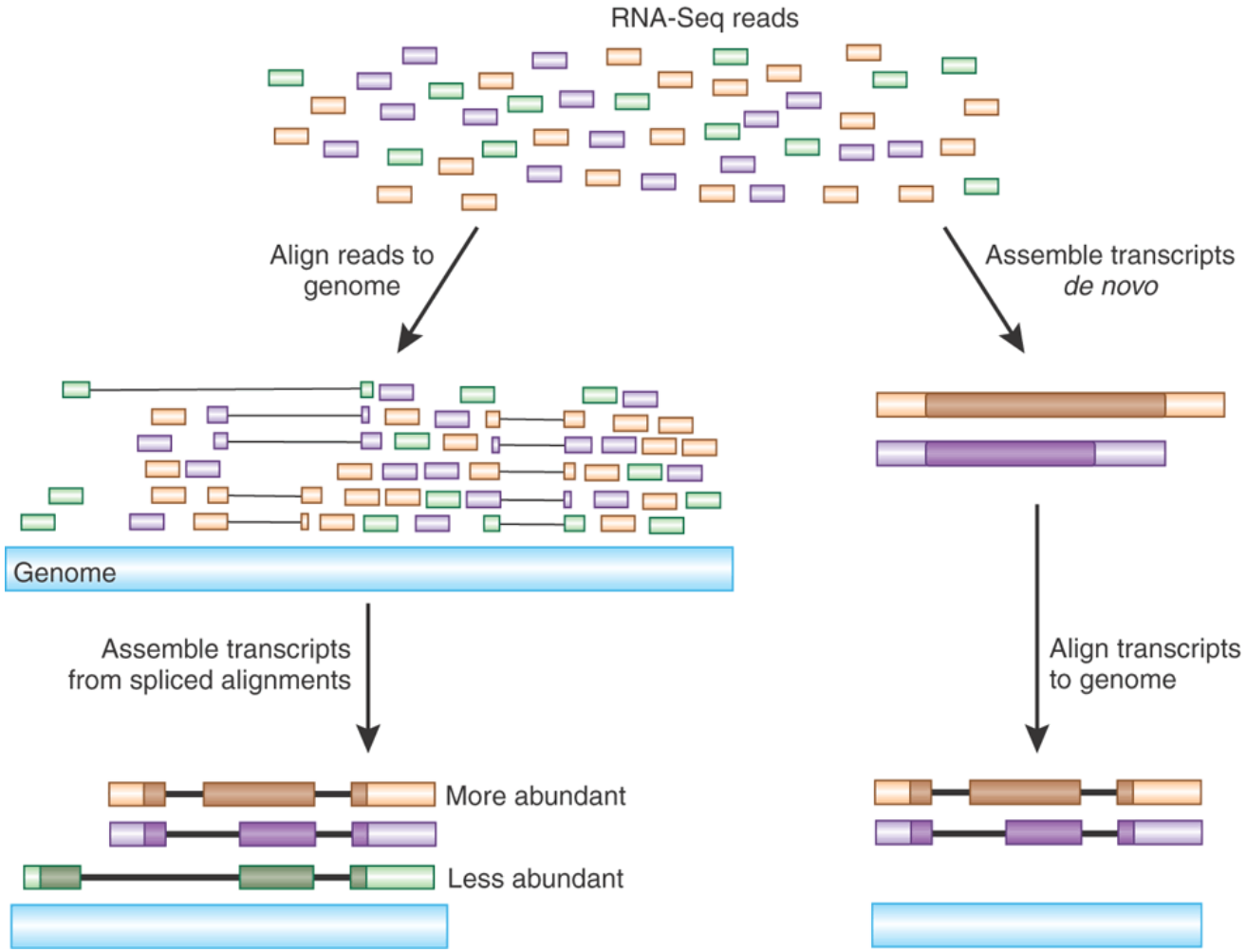
- F: the single read is in the sense (F, forward) orientation
- R: the single read is in the antisense (R, reverse) orientation

PE

- RF: first read (/1) is sequenced as anti-sense (R) & second read (/2) is in the sense strand (F)
- FR: first read (/1) is sequenced as sense (F) & second read (/2) is in the antisense strand (R)



Reference based data analysis pipeline

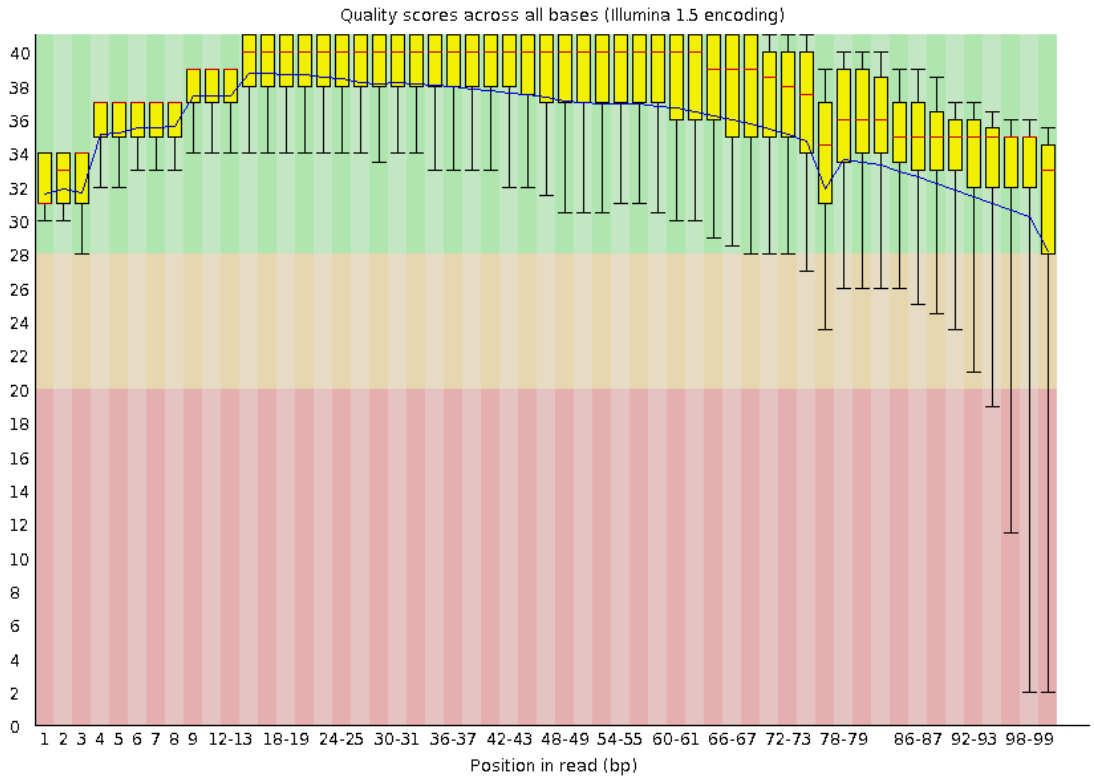

RNA-Seq reads
Align reads to genome
Assemble transcripts de novo
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

Main steps

- Initial processing incl. QC
- Aligning reads to reference genome
- Counting reads
- Differential gene expression
- Further analysis

Initial processing incl. QC

- Demultiplex by index or barcode
- Remove adapter sequences
- Trim reads by quality
- Discard reads by quality/ambiguity



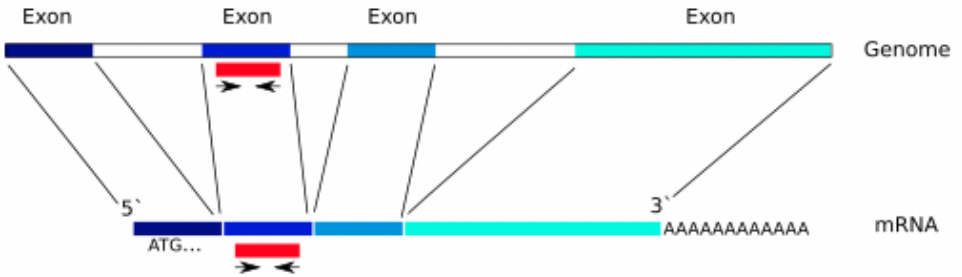
Available tools

FastQC, PRINSEQ, TRIMMOMATIC, TrimGalore, FastX, Cutadapt

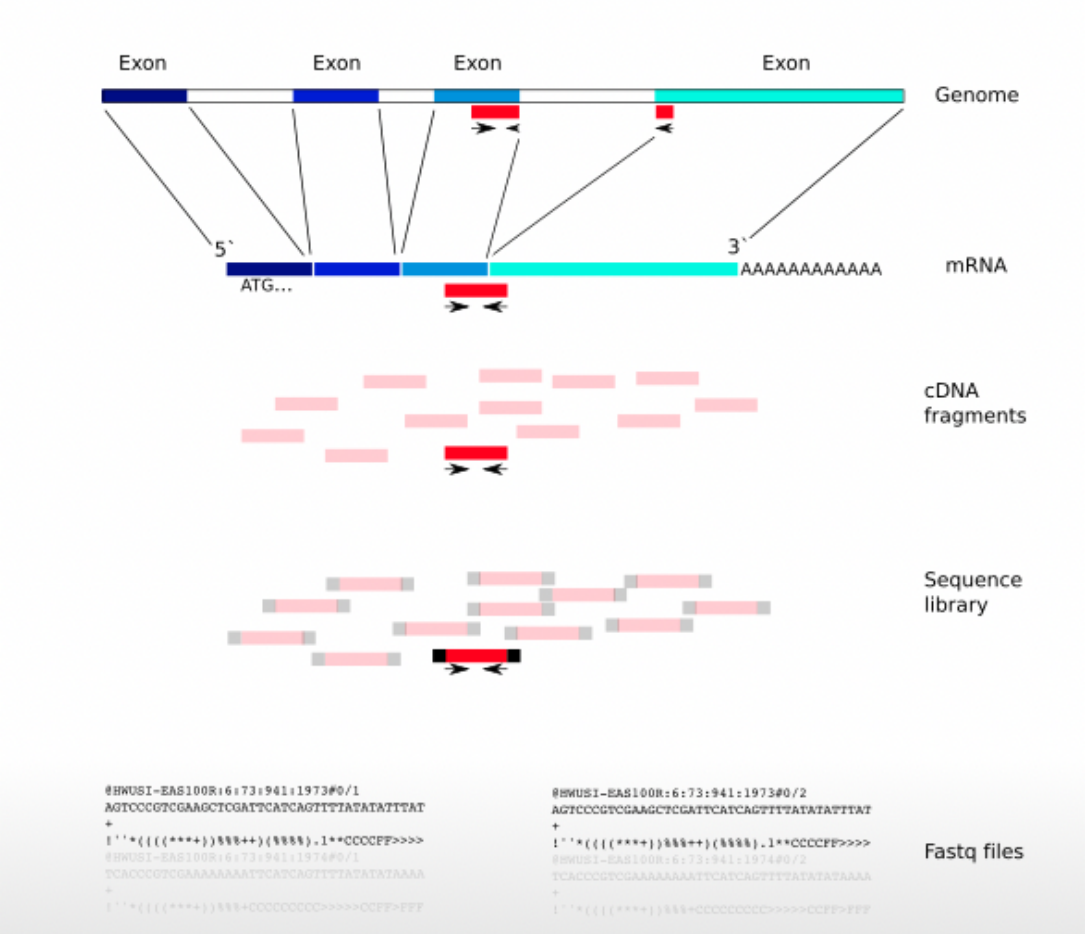
Initial processing incl. QC

- filtering reads for quality score, e.g. with avg. quality below 20 defined within 4-base wide sliding window
- filtering reads for read length, e.g. reads shorter than 36 bases
- removing artificial sequences, e.g. adapters

Aligning reads



Aligning reads



Aligning reads: mappers

- important to use mappers allowing for a read to be "split" between distant regions of the reference in the event that the read spans two exons
- lots of different aligners exists based on various algorithms e.g. brute force comparison, Burrows-Wheeler Transform, Smith-Waterman, Suffix tree
- usually there is a trade-off between speed versus accuracy and sensitivity
- usually the "biggest difference" is with default settings, most mappers will allow to optimize settings
- performance vary by genome complexity

A good read: Barruzo et. al. Nature Methods 14, (2017)
https://www.nature.com/articles/nmeth.4106

Available tools

STAR, HISAT, MapSlice2, Subread, TopHat

Aligning reads: reference files

.fasta (download reference genome FASTA file)

```
>1 dna:chromosome chromosome:GRCm38:1:1:195471971:1 REF
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

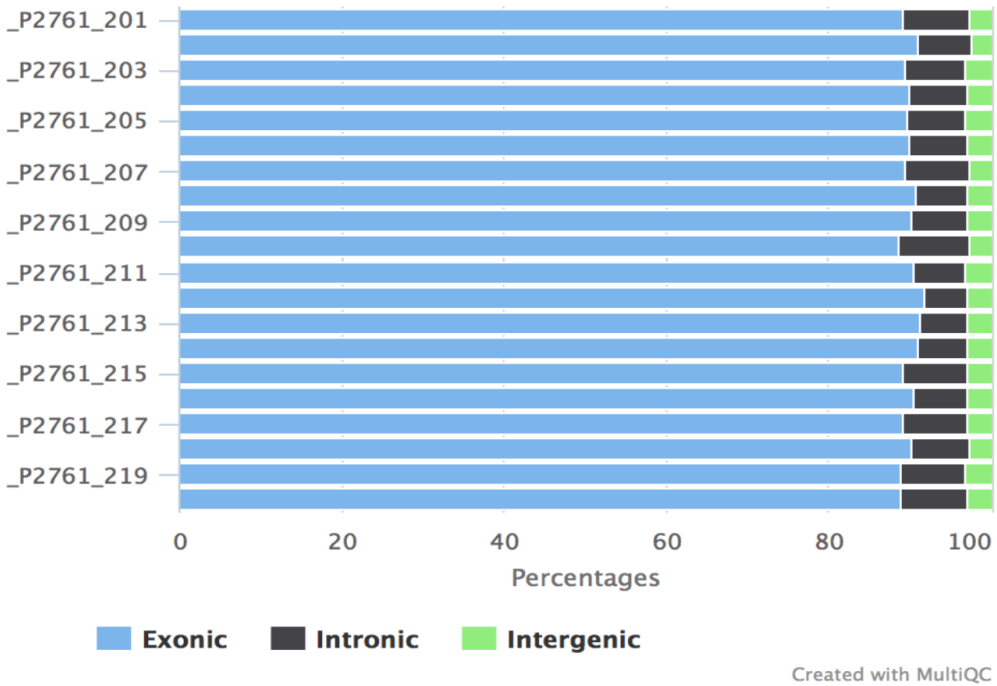
.gtf (download the corresponding genome annotation in GTF or GFF)

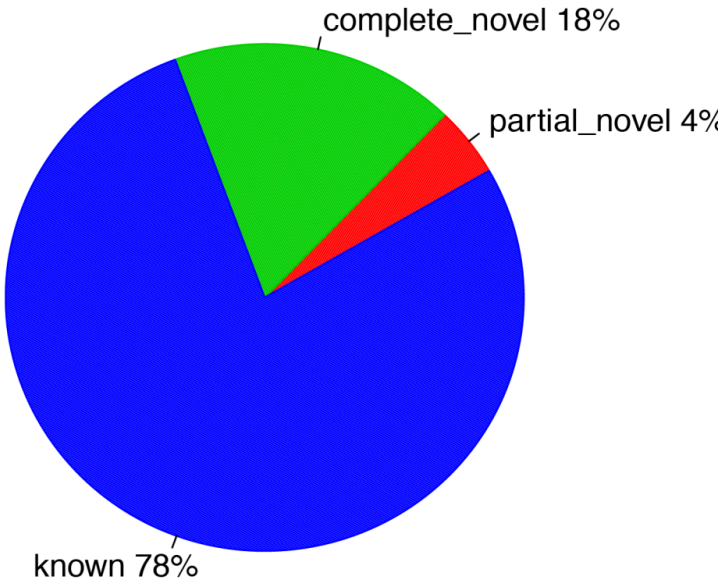
```
#!genome-build GRCm38.p4
#!genome-version GRCm38
#!genome-date 2012-01
#!genome-build-accession NCBI:GCA_000001635.6
#!genebuild-last-updated 2015-07
1       havana  gene    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693"; gene_version "1"; gene_name "493340
1J01Rik"; gene_source "havana"; gene_biotype "TEC"; havana_gene "OTTMUSG00000049935"; havana_gene_version "1";
1       havana  transcript      3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693"; gene_version "1"; transcrip
t_id "ENSMUST00000193812"; transcript_version "1"; gene_name "4933401J01Rik"; gene_source "havana"; gene_biotype "TEC"; havana_ge
ne "OTTMUSG00000049935"; havana_gene_version "1"; transcript_name "4933401J01Rik-001"; transcript_source "havana"; transcript_bio
type "TEC"; havana_transcript "OTTMUST00000127109"; havana_transcript_version "1"; tag "basic"; transcript_support_level "NA";
```

Source

ENSEMBL, NCBI

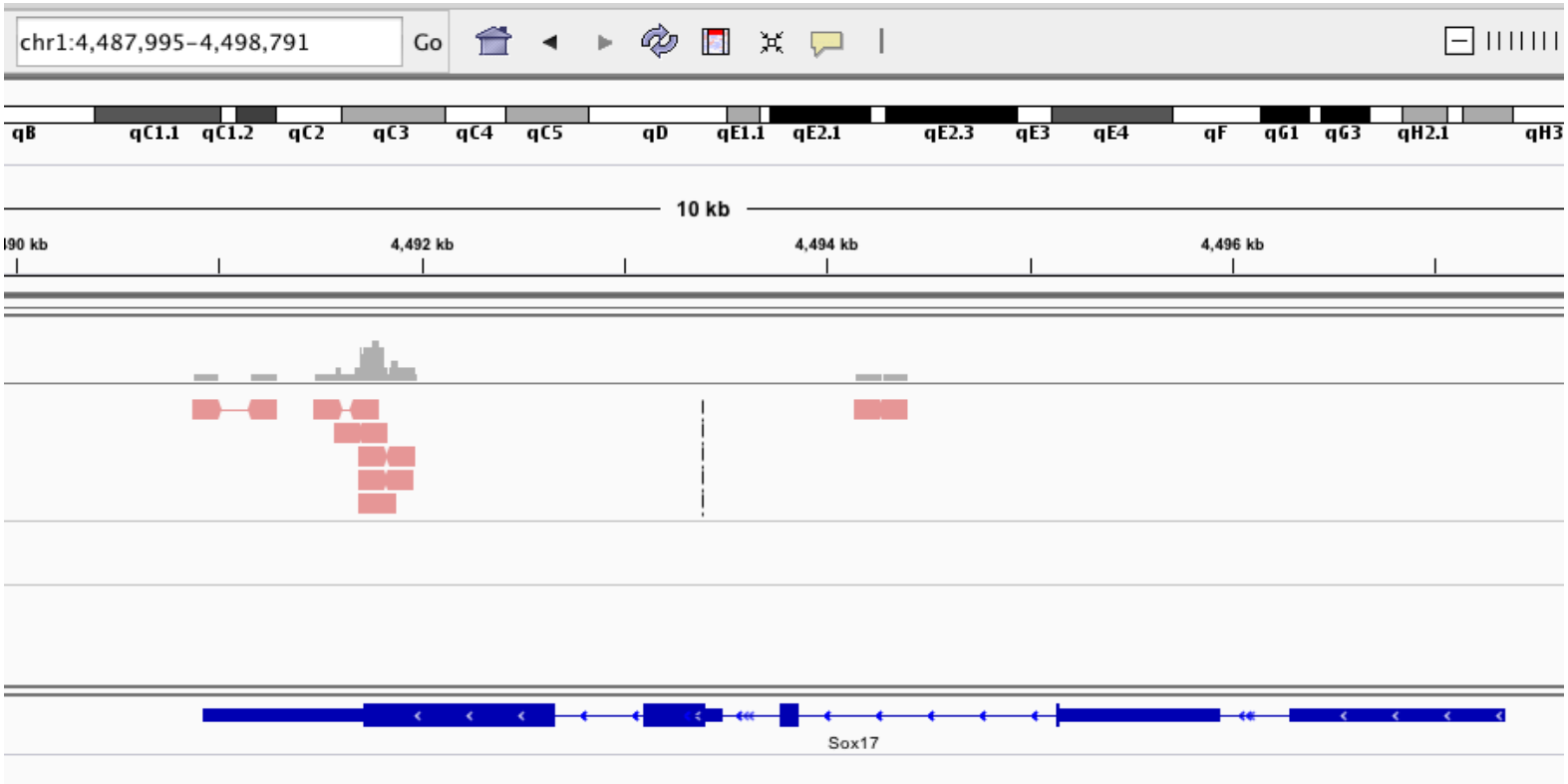
Aligning reads: QC





Post mapping QC, e.g. reads should mostly map to known genes, most splice event should be known and canonical (GU-AG)

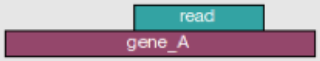
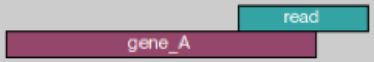
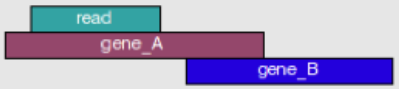
Counting reads



Available tools

HTSeq, featureCounts, R

Counting reads

	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

from: http://www-huber.embl.de/users/anders/HTSeq/doc/count.html

Counting reads

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Transcript	P1822_1	P1822_2	P1822_3	P1822_4	P1822_5	P1822_6	P1822_7	P1822_8	P1822_9	P1822_10	P1822_11	P1822_12	P1822_13	P1822_14	P1822_15	P1822_16
2	ENSMUSG00000102693	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	ENSMUSG00000088000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	ENSMUSG00000103265	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
32	ENSMUSG00000103922	7	7	7	4	1	12	3	6	14	3	9	3	9	7	9	7
33	ENSMUSG00000033845	972	860	878	1085	1058	1009	992	1143	947	1059	970	1147	801	837	1042	927
34	ENSMUSG00000102275	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	ENSMUSG00000025903	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	ENSMUSG00000104217	16	13	17	16	22	17	12	27	11	5	12	15	8	9	9	12
37	ENSMUSG00000033813	2560	2581	2937	3904	2975	3100	3027	3417	2272	2801	2266	3294	2491	2578	2554	2806
38	ENSMUSG00000062588	3	1	1	1	0	1	0	3	3	0	4	0	2	1	0	0
39	ENSMUSG00000103280	1	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0
40	ENSMUSG00000002459	7	10	5	7	4	6	3	8	2	5	7	8	1	5	4	1
41	ENSMUSG00000091305	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	ENSMUSG00000102653	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	ENSMUSG00000085623	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
44	ENSMUSG00000091665	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	ENSMUSG00000033793	3682	3757	4414	5978	3774	4102	3815	4250	4193	4962	4240	5694	3565	3757	3849	4094
46	ENSMUSG00000104352	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	ENSMUSG00000104046	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
48	ENSMUSG00000102907	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
49	ENSMUSG00000025905	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
50	ENSMUSG00000103936	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
51	ENSMUSG00000093015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
52	ENSMUSG00000103519	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
53	ENSMUSG00000033774	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
54	ENSMUSG00000103090	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
55	ENSMUSG00000025907	1816	2087	2088	2820	2012	2236	2065	2727	2586	2931	2813	3667	2410	2739	2479	2745
56	ENSMUSG00000090031	43	58	55	73	38	38	57	96	89	107	98	123	76	93	66	69

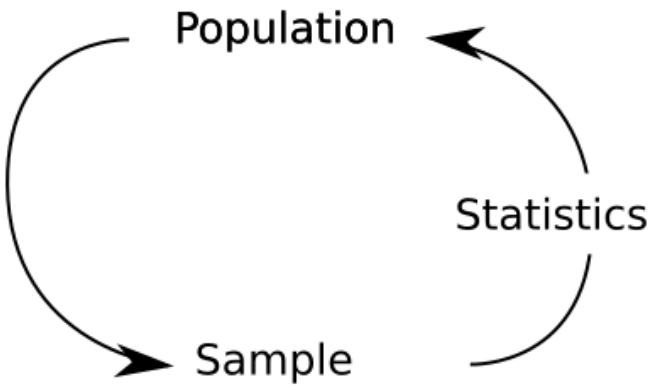
Normalizing counts

Gene counts depend e.g. on sequencing depth of a sample and on the sequence length of the gene/transcript. Raw read counts cannot be used to compare gene expression across libraries.

Normalization methods

- CPM, counts per million, accounts for sequencing depth
- RPKM/FPKM, Reads/Fragments Per Kilobase Per Milion accounts for sequencing depth and transcript length
- TMM, Trimmed Mean of M-values, accounts for sequencing depth and transcript length and composition of the RNA population
- and few other using scaling factors methods...

Differential gene expression

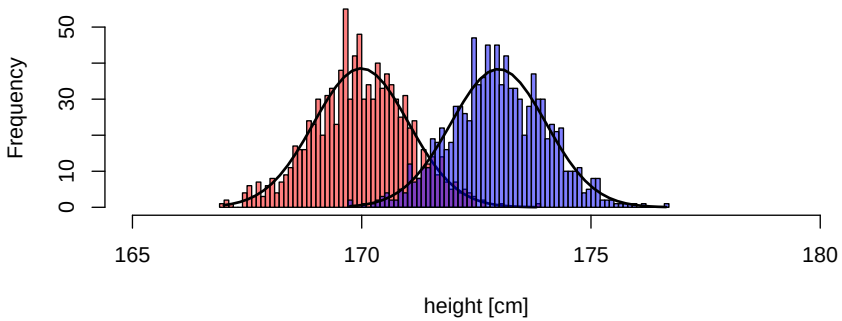


$$Outcome_i = (Model_i) + error_i$$

- we collect data on a *sample* from a much larger *population*. *Statistics* lets us to make inferences about the population from which it was derived
- we try to predict the outcome given a model fitted to the data

Differential gene expression

$$t = \frac{x_1 - x_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$



in RNA-seq case:

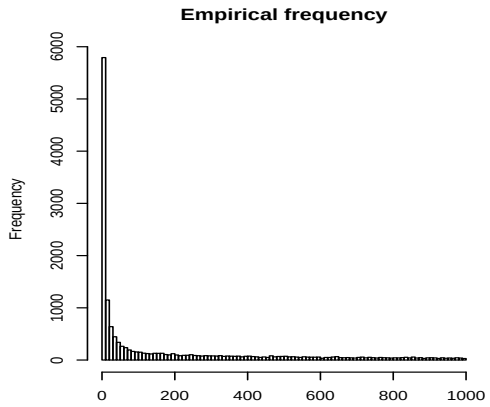
- we take the normalized read counts
- and we perform statistical analysis to discover quantitative changes in expression levels between experimental groups
- e.g. to decide whether, for a given gene, an observed difference in read counts is significant, that is, whether it is greater than what would be expected just due to natural random variation.

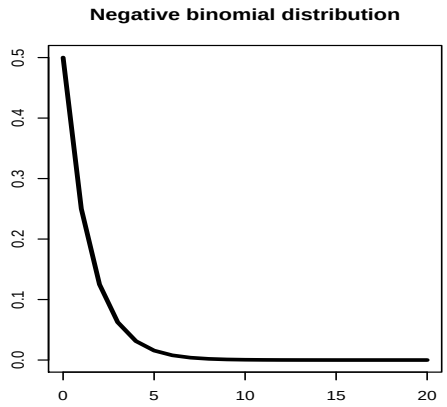
Differential expression

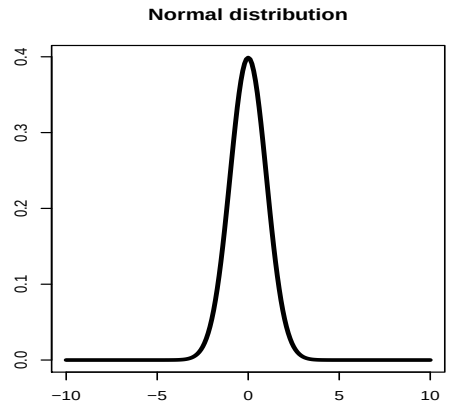
Usually, reads counts do not follow normal distribution & we work with low number of biological replicates

DE methods

- Discrete distribution models, e.g. edgeR, DESeq2
- Continuous discrete models, e.g. t-test
- Non-parametric model, e.g. SAMseq





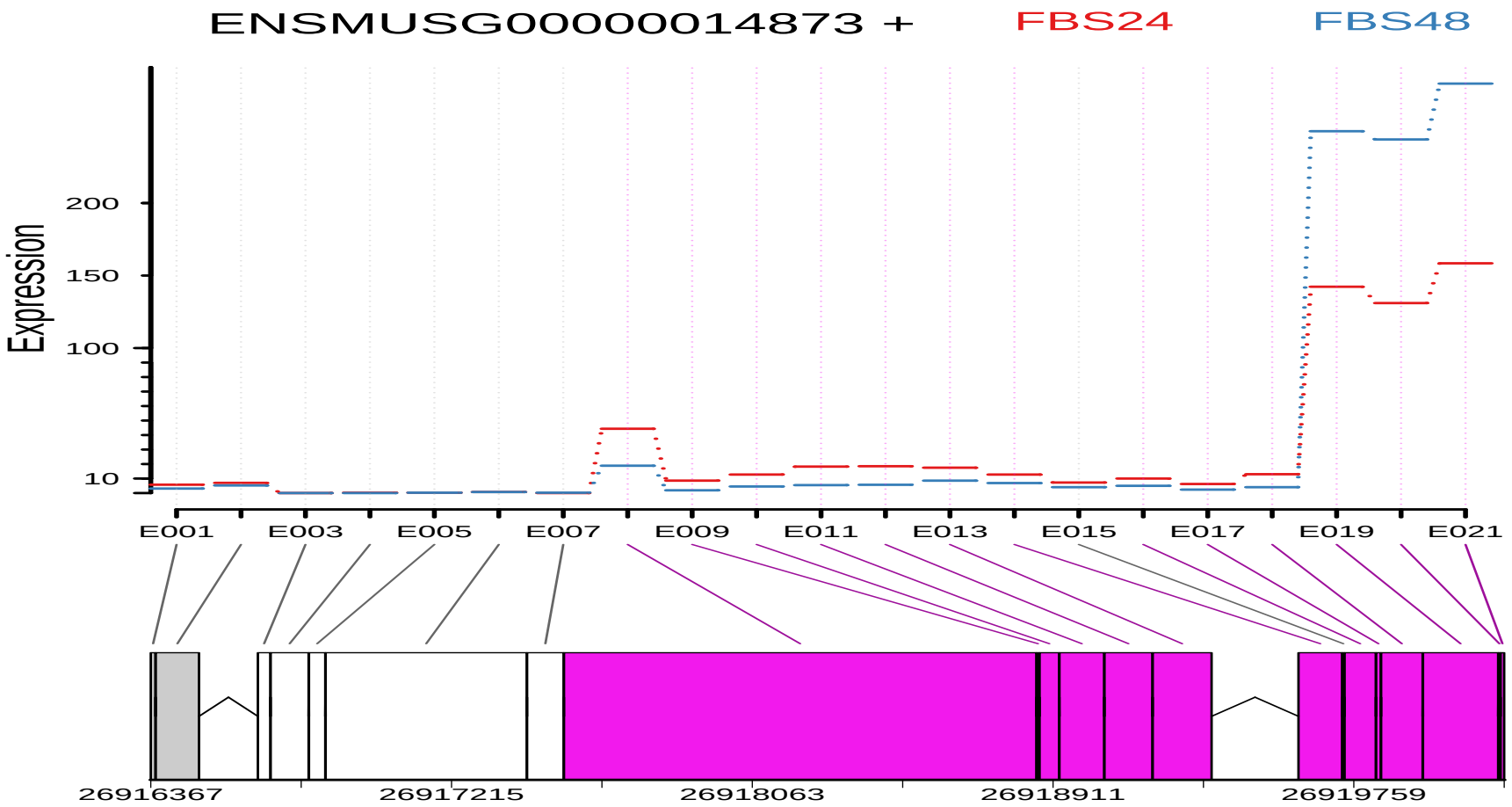


Differential expression

	A	B	C	D	E	F	G	H	I	J
1	ensembl_gene_id	ensembl_transcript_id	chromosome_name	mgi_symbol	description	logFC	logCPM	LR	PValue	FDR
2	ENSMUSG00000028328	ENSMUST00000107773	4	Tmod1	tropomodulin 1 [Source:MGI Symbol;Acc:MGI:98775]	1.971089	5.958225	581.2916	1.96E-128	2.79E-124
3	ENSMUSG00000066705	ENSMUST00000085939	9	Fxyd6	FXYD domain-containing ion transport regulator 6 [Source:MGI Symbol;Acc	3.18062	5.916499	553.8787	1.80E-122	1.28E-118
4	ENSMUSG00000049112	ENSMUST00000053306	6	Oxtr	oxytocin receptor [Source:MGI Symbol;Acc:MGI:109147]	3.820952	3.423774	375.1689	1.40E-83	6.65E-80
5	ENSMUSG00000017446	ENSMUST00000124861	11	C1qtnf1	C1q and tumor necrosis factor related protein 1 [Source:MGI Symbol;Acc:M	1.484213	7.145099	345.7577	3.56E-77	1.26E-73
6	ENSMUSG00000029123	ENSMUST00000094836	5	Stk32b	serine/threonine kinase 32B [Source:MGI Symbol;Acc:MGI:1927552]	3.453001	2.321613	338.7155	1.22E-75	3.46E-72
7	ENSMUSG00000009378	ENSMUST00000009522	19	Slc16a12	solute carrier family 16 (monocarboxylic acid transporters), member 12 [Sc	4.173029	3.89466	335.706	5.50E-75	1.30E-71
8	ENSMUSG00000025355	ENSMUST00000026411	10	Mmp19	matrix metallopeptidase 19 [Source:MGI Symbol;Acc:MGI:1927899]	1.940915	8.973932	328.4969	2.04E-73	4.15E-70
9	ENSMUSG00000029671	ENSMUST00000128245	6	Wnt16	wingless-type MMTV integration site family, member 16 [Source:MGI Syml	2.339149	5.673738	315.6779	1.27E-70	2.25E-67
10	ENSMUSG00000042190	ENSMUST00000047936	5	Cmklr1	chemokine-like receptor 1 [Source:MGI Symbol;Acc:MGI:109603]	2.518748	3.540638	305.0157	2.66E-68	4.20E-65
11	ENSMUSG00000028035	ENSMUST00000134701	3	Dnajb4	DnaJ (Hsp40) homolog, subfamily B, member 4 [Source:MGI Symbol;Acc:M	1.417856	7.292192	297.1316	1.39E-66	1.98E-63
12	ENSMUSG00000048960	ENSMUST00000027056	1	Prex2	phosphatidylinositol-3,4,5-trisphosphate-dependent Rac exchange factor 2	1.706461	6.676335	283.7963	1.12E-63	1.44E-60
13	ENSMUSG00000002289	ENSMUST00000002360	17	Angptl4	angiopoietin-like 4 [Source:MGI Symbol;Acc:MGI:1888999]	-1.73049	7.972378	282.7705	1.87E-63	2.22E-60

The likelihood of observing a significant p-value increases as we do more tests, i.e. testing more than one gene. Modern FDR adjustment techniques take into account of background expectation of a uniformly distributed p-values and adjust their values accordingly to how significantly different things are, so the p-values from multiple testing can be interpreted more accurately.

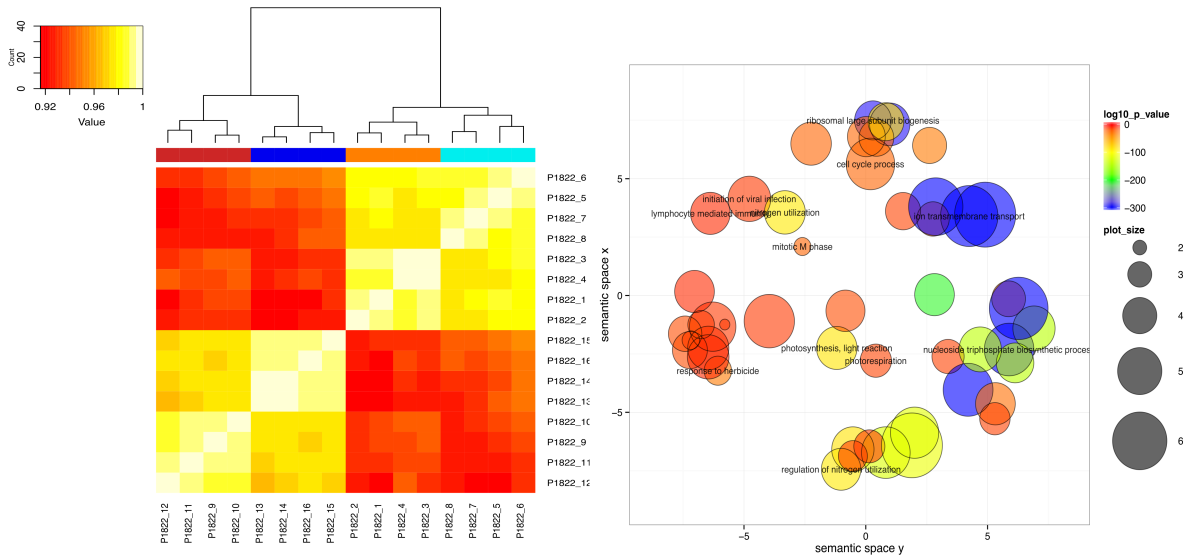
Differential expression



Available tools

edgeR, DEXSeq

Further analysis

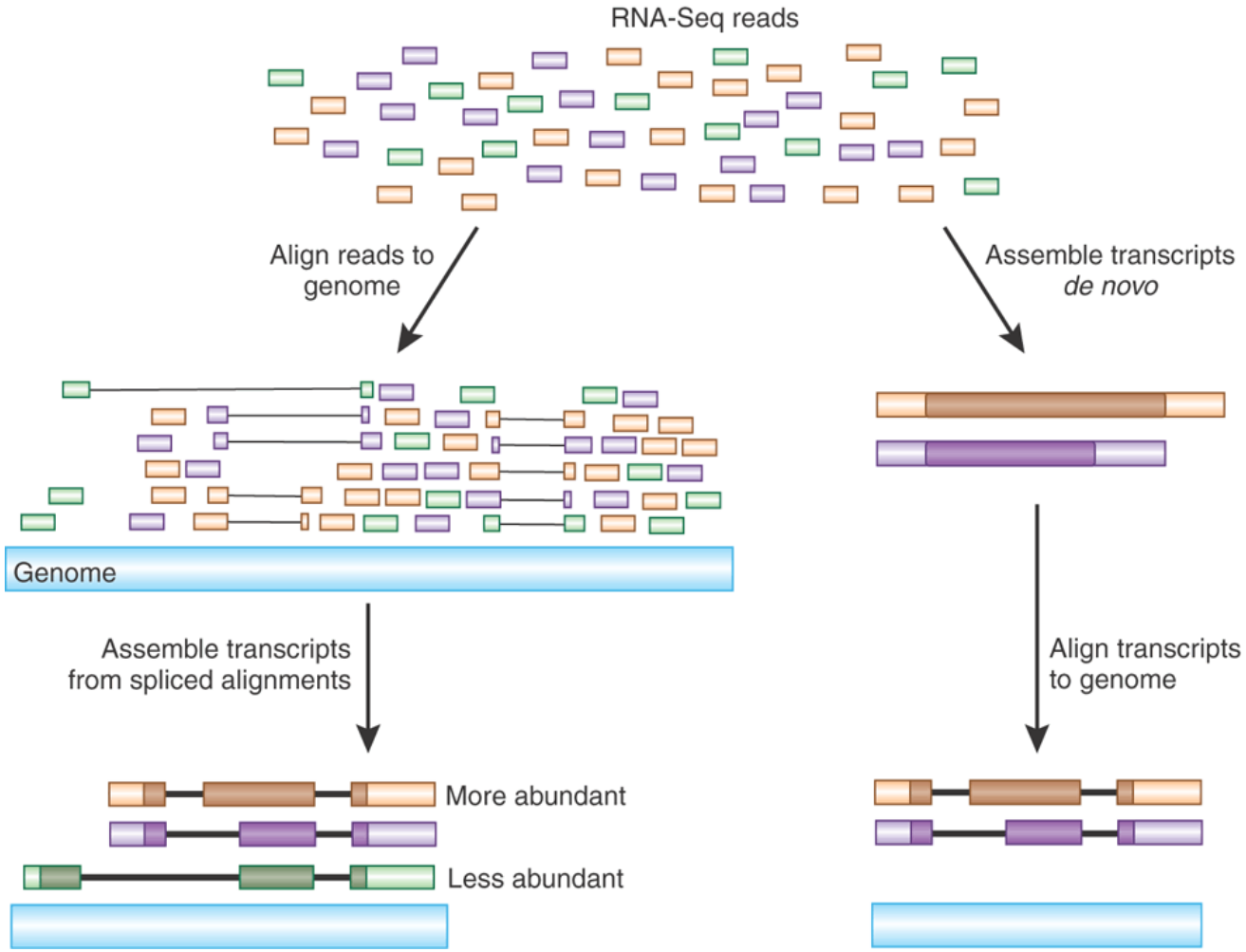


- Annotating the results e.g. with gene symbols, GO terms
- Visualizing the results, e.g. Volcano plots
- Gene set analysis etc...

Available tools

bioMart (R), DAVID, GOrilla, REVIGO, ClustVis...

What about de-novo assembly of transcriptomes?


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

Building a reference transcriptome

- alternative strategy when well-assembled reference genome from a relatively recently diverged organism is not available
- primary goal: assembling a transcriptome *de novo* to reconstruct a set of contigous sequences (contigs) presumed to reflect accurately a large portion of the RNAs actually transcribed in the cells

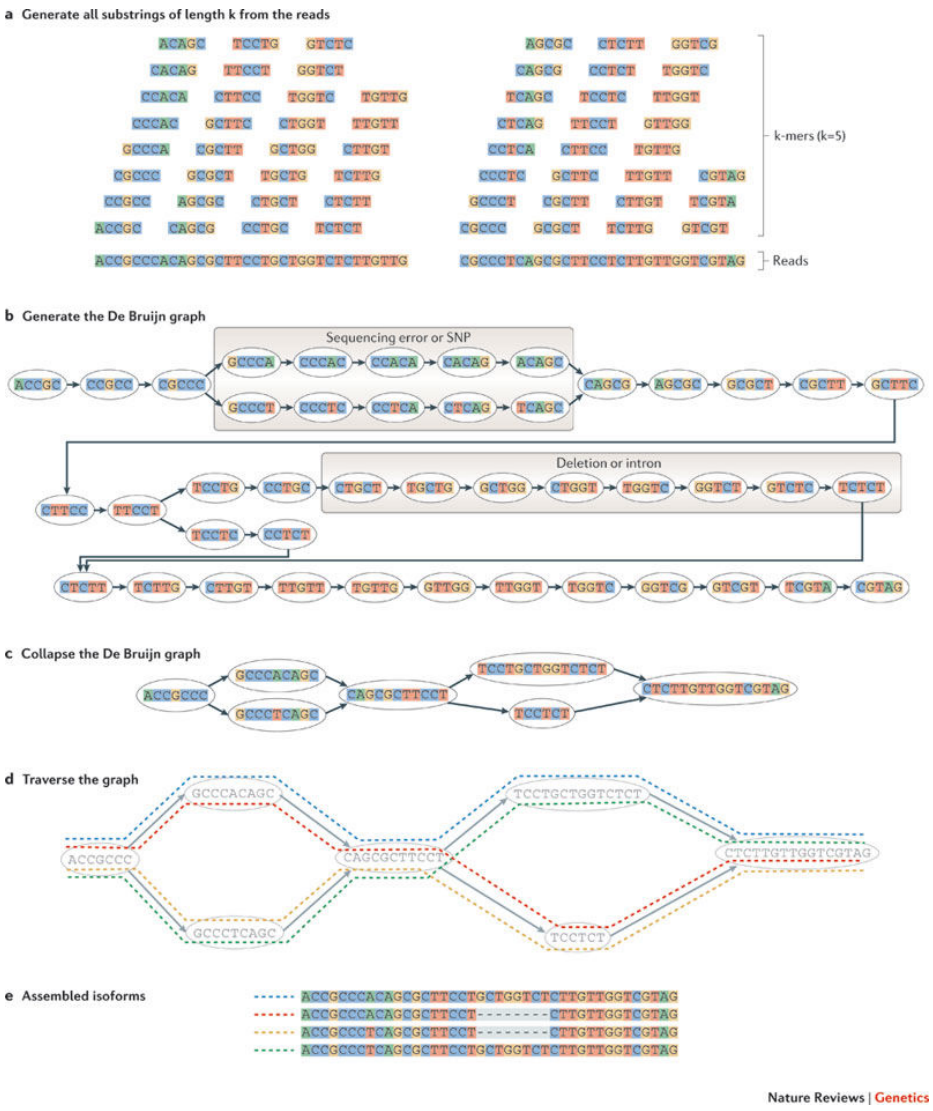
not a trivial task, because

- a limited amount of information about the original gene transcripts is retained in the short reads produced by a sequencer
- genes show different levels of gene expression (uneven coverage)
- more sequencing depth is needed to represent less abundant genes and rare events
- reads from the same transcript must be placed together in the face of variants introduced by polymorphism and sequencing errors
- and the process must assemble reads from different but often similar, paralogous transcripts as separate contigs

Solutions to sequence assembly arose from the field of mathematics known as graph theory. These approaches were designed with genome assembly in mind but have been adapted for transcriptome assembly as necessary. Most of them are based on de Brujin graphs.

Available tools

- Velvet/Oases: Velvet constructs de Bruijn graphs, simplifies the graphs, and corrects the graphs for errors and repeats. Oases post-processes Velvet assemblies with different k-mer sizes
- Trans-ABySS: much like the Velvet/Oases model, Trans-ABySS (Robertson et al. 2010) takes multiple ABySS assemblies (Simpson et al. 2009) produced from a range of k-mer sizes to optimize transcriptome assemblies in the face of varying coverage across transcripts
- Trinity: "Inchworm" builds initial contigs by finding paths through k-mer graphs. "Chrysalis" groups these contigs together and builds de Bruijn graphs for these groups, in which the overlaps are nodes and the k-mers connecting edges. "Butterfly" simplifies the graphs when possible, then reconciles the graphs with original reads to output individual contigs representative of unique splice variants and paralogous transcripts



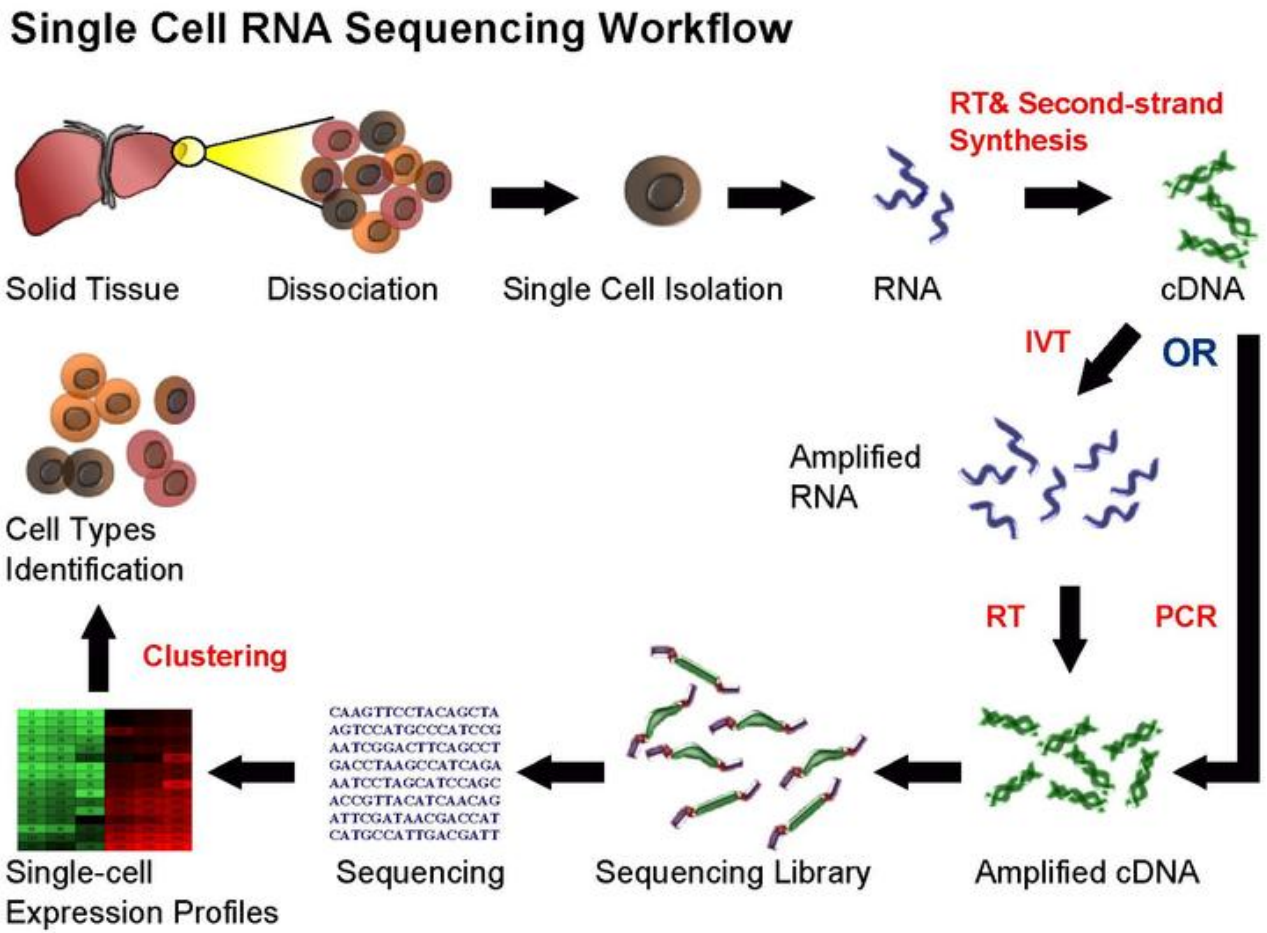
- a) all substrings of length k (k-mers) are generated from each read
- b) each unique k-mer is used to represent a node in the De Bruijn graph,pairs of nodes are connected if shifting a k-mer by one character creates an exact k???1 overlap between the two k-mers.
- The example (5-mers) illustrates a SNP or sequencing error and an example of an intron or a deletion.
- Single-nucleotide differences cause 'bubbles' of length k in the De Brujin graph, whereas introns or deletions introduce a shorter path in the graph
- c,d) chains of adjacent nodes in the graph are collapsed into a single node when the first node has an out degree of one and the second node has an in degree of one
- e) the isoforms are then assembled. See more http://rdcu.be/zSpz

If a reference genome is available, annotation is relatively straightforward: genomic coordinates from the reference genome are normally associated with various forms of annotation information through databases. A transcriptome assembled de novo, on the other hand, is often annotated from scratch

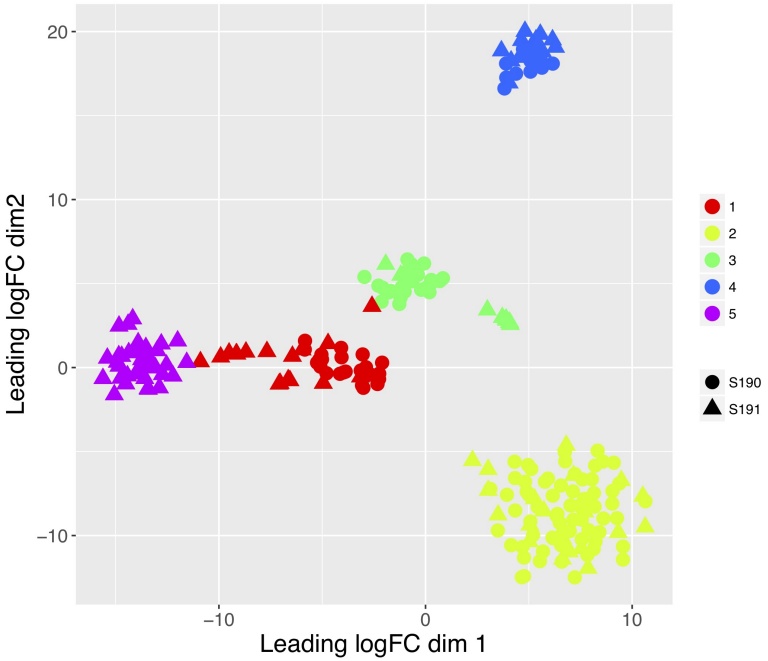
NCBI-supported BLAST

- "match" query sequences to one or more databases of curated, annotated sequences, using an efficient local sequence alignment approach.
- it may be adequate to blast against a database of known or predicted transcripts from the reference genome of a closely-related organism
- it may be desirable to blast contigs against all nucleotide sequences in an inclusive database
- if the annotation emphasis is on protein-coding transcripts, BLASTx, which translates each query sequence (in all six reading frames) to amino acid sequences and uses these to query a protein database, may be an appropriate tool

And what about scRNA-seq?



- scRNA-seq are affected by higher noise (amplification biases, dropout event, 3'bias, partical coverage, uneven depth, stochatic nature of transcription, multimodality in gene expression)
- read processing steps to generate count matrix are largely the same as for bulk RNA-seq, but the spike-in normalization is a must



Common steps

- Quality control on the cells
- Classification of the cell-cycle phase
- Normalization of cell-specific biases
- Checking for important technical factors
- Modelling and removing technical noise
- Data exploration with dimensionality reduction
- Clustering cells into putative subpopulations
- Detecting marker genes between subpopulations
- see more: Bioconductor simpleSingleCell workflow

Exercises

Main exercise

- checking the quality of the raw reads with FastQC
- mapping the reads to the reference genome using STAR
- converting between SAM and BAM files format using Samtools
- assessing the post-alignment reads quality using QualiMap
- counting reads overlapping with genes regions using featureCounts
- building statistical model to find DE genes using edgeR called from a prepared R script

Bonus exercises

- functional annotation, putting DE genes in the biological context
- exon usage, studying the alternative splicing
- data visualisation and graphics
- de novo transcriptome assembly

Thank you for attention
Questions?

Enjoy the rest of the course

Read more

- RNA-seqlopedia
- RNA-Seq blog
- Conesa et al. Genome Biology, 2016, A survey of best practices for RNA-seq data analysis