Introduction to RNA sequencing

Bioinformatics perspective

Olga Dethlefsen

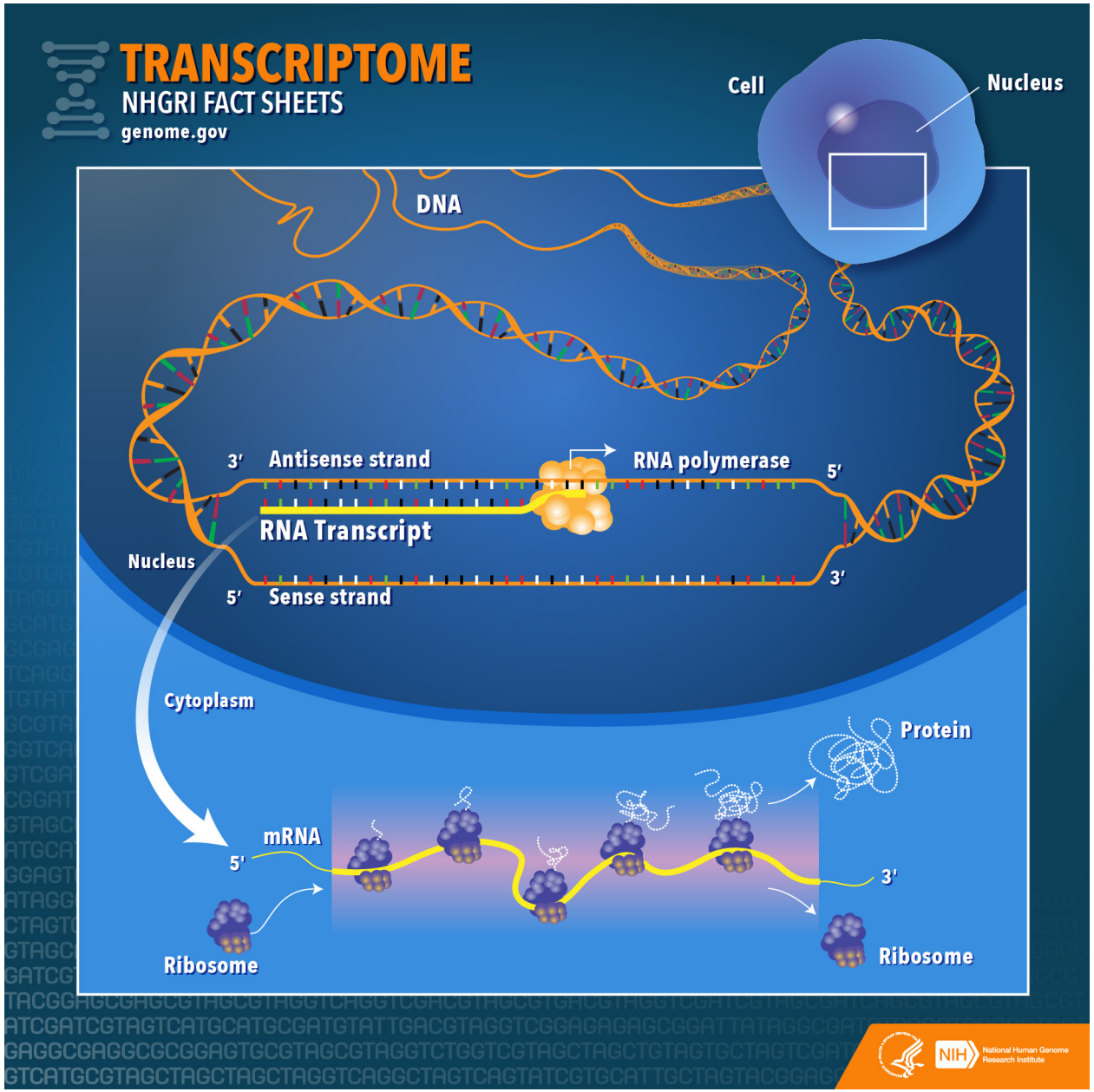
NBIS, National Bioinformatics Infrastructure Sweden

November 2017

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?
- Summary
- 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:

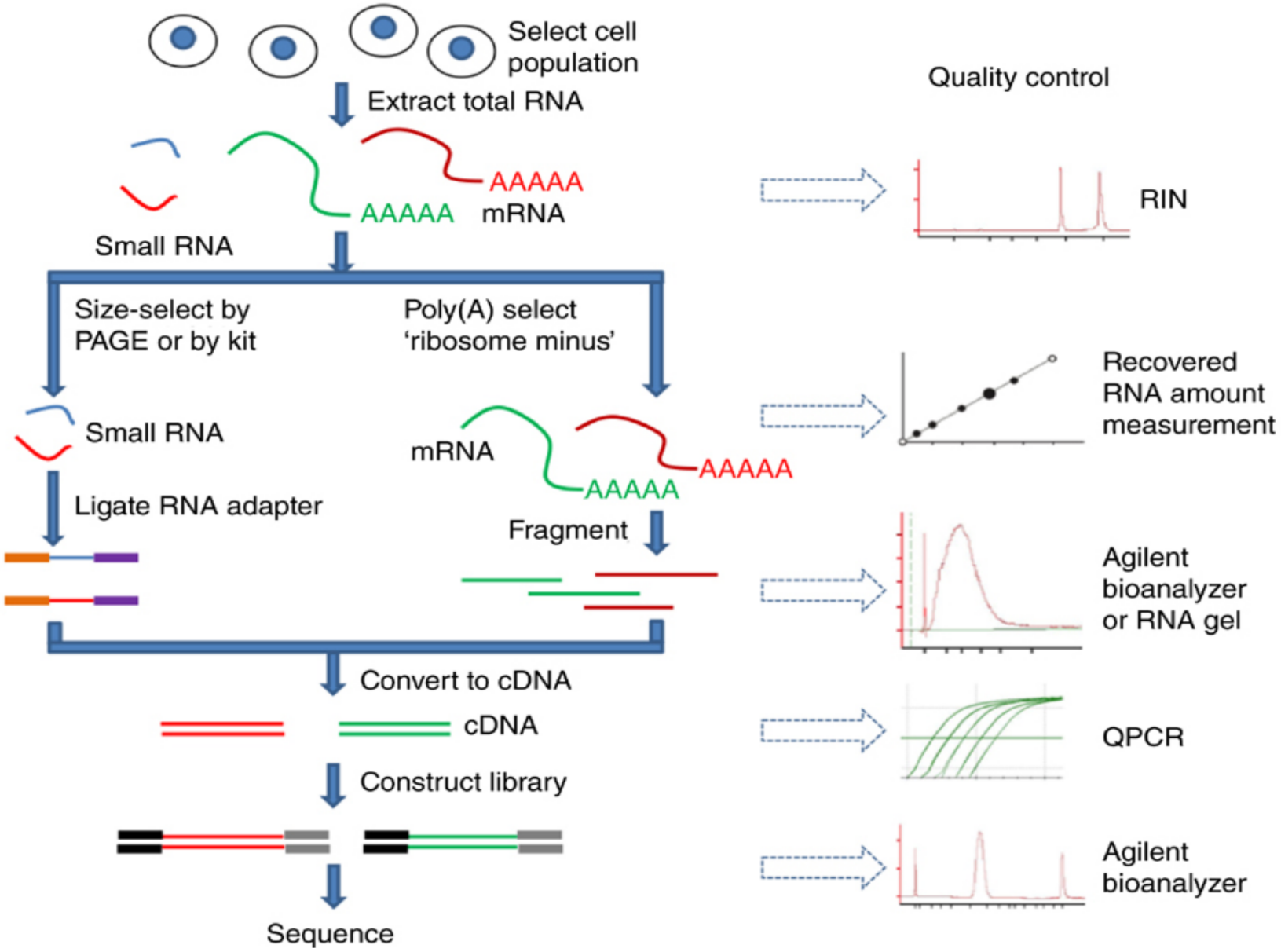
dynamic, that is not the same over tissues and time points

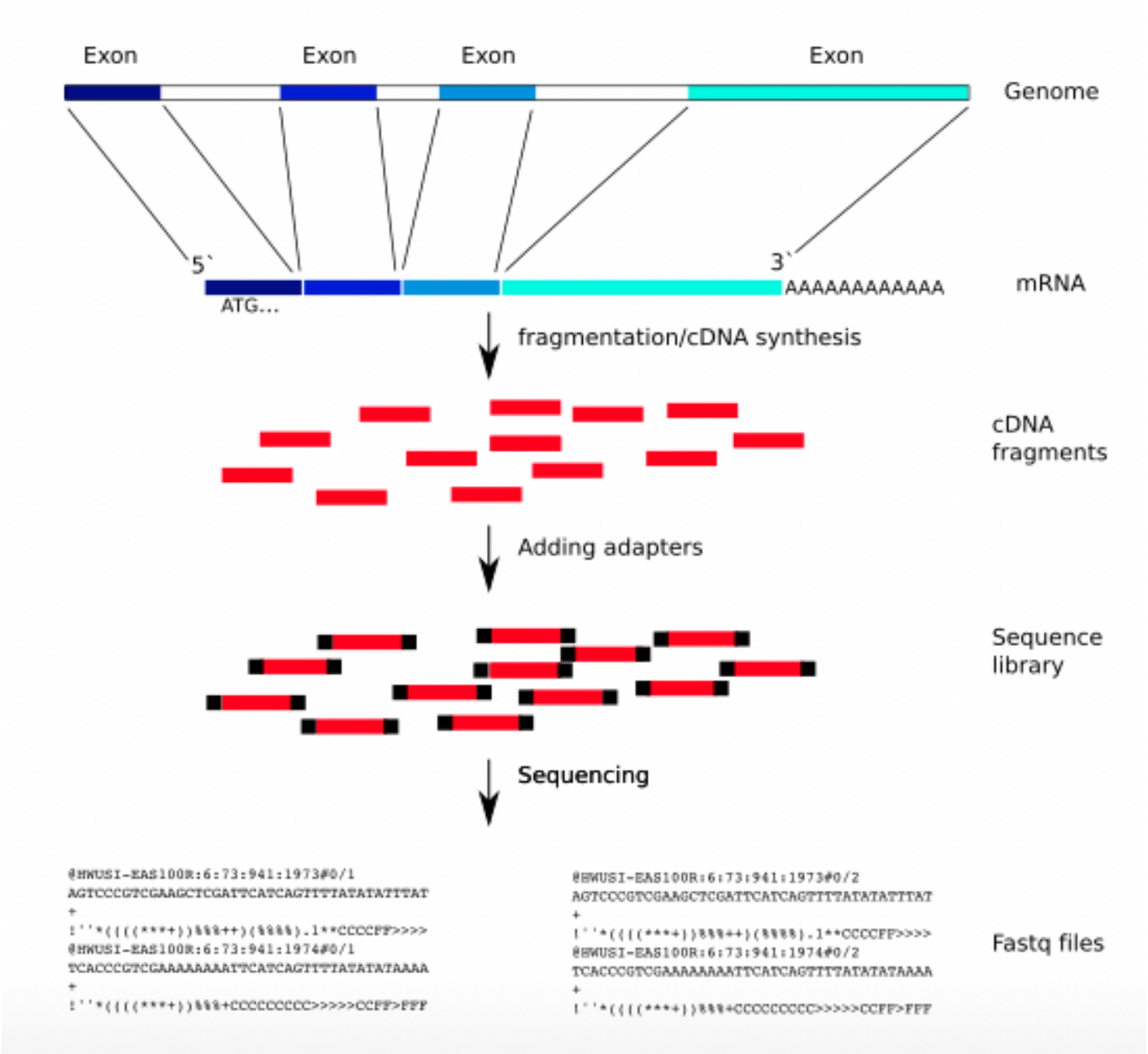
directly derived from functional genomics elements, that is mostly protein-coding genes, providing a useful functionally relevant subset of the genome, translating into smaller sequence space

Overview

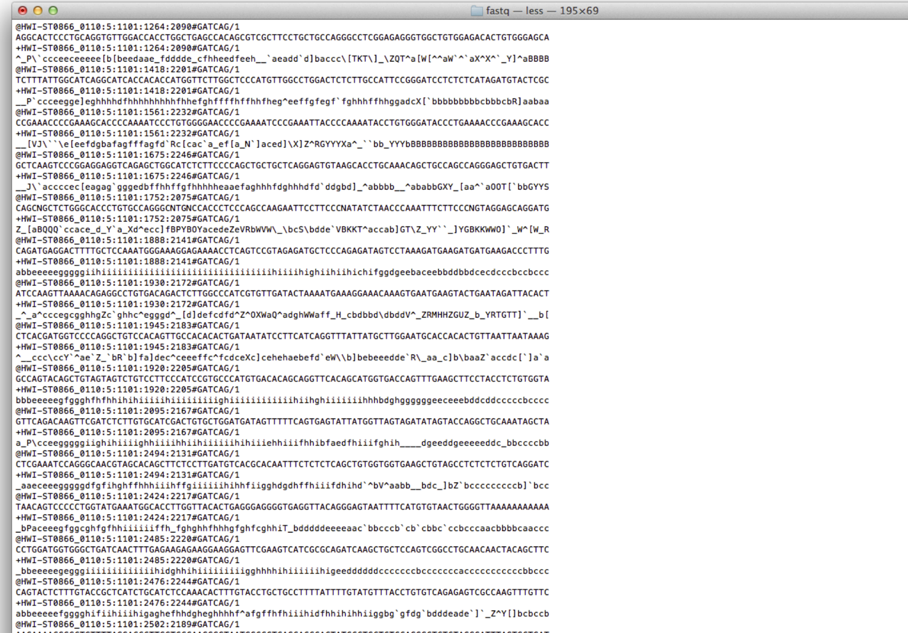
- Experimental design (biology, medicine, statistics)
- RNA extraction (biology, biotechnology)
- Library preparation (biology, biotechnology)
- High throughput sequencing (engineering, biology, chemistry, biotechnology, bioinformatics)
- Data processing (bioinformatics)
- Data analysis (bioinformatics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

From RNA to sequence



```
@HWI-ST0866_0110:5:1101:1264:2090#GATCAG/1
AGGCACTCCCTGCAGGTGTTGGACCACCTGGCTGAGCCACAGCGTCGCTTCCTGCTGCCAGGGCCTCGGAGAGGGTGGCTGTGGAGACACTGTGGGAGCA
+HWI-ST0866_0110:5:1101:1264:2090#GATCAG/1
^_P\`ccceeceeeee[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`ccceegge]eghhhhdfhhhhhhhhhfhhefghffffhffhhfheg^eeffgfegf`fghhhffhhggadcX[`bbbbbbbbcbbbcbR]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_YYYbBBBBBBBBBBBBBBBBBBBBBBBBBB
@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^`a0OT[`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
abbeeeeeggggggiihiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiihiiiihighiihiihichifggdgeebaceebbddbbdcecdcccbccbccc
@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
bbbeeeeegfggghfhfhhihihiiiiihiiiiiiiiighiiiiiiiiiiiihiihghiiiiiiihhhbdghggggggeeceeebddcddcccccbcccc
@HWI-ST0866_0110:5:1101:2095:2167#GATCAG/1
GTTCAGACAAGTTCGATCTCTTGTGCATCGACTGTGCTGGATGATAGTTTTTCAGTGAGTATTATGGTTAGTAGATATAGTACCAGGCTGCAAATAGCTA
+HWI-ST0866_0110:5:1101:2095:2167#GATCAG/1
a_P\cceeggggggiighihiiiighhiiiihhiihiiiiiihihiiiehhiiifhhibfaedfhiiifghih____dgeeddgeeeeeddc_bbccccbb
@HWI-ST0866_0110:5:1101:2494:2131#GATCAG/1
CTCGAAATCCAGGGCAACGTAGCACAGCTTCTCCTTGATGTCACGCACAATTTCTCTCTCAGCTGTGGTGGTGAAGCTGTAGCCTCTCTCTGTCAGGATC
+HWI-ST0866_0110:5:1101:2494:2131#GATCAG/1
_aaeceeeggggggdfgfihghffhhhiiihffgiiiiiihihhfiigghdgdhffhiiifdhihd`^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_bdddddeeeeaac`bbcccb`cb`cbbc`ccbcccaacbbbbcaaccc
@HWI-ST0866_0110:5:1101:2485:2220#GATCAG/1
CCTGGATGGTGGGCTGATCAACTTTGAGAAGAGAAGGAAGGAGTTCGAAGTCATCGCGCAGATCAAGCTGCTCCAGTCGGCCTGCAACAACTACAGCTTC
+HWI-ST0866_0110:5:1101:2485:2220#GATCAG/1
_bbeeeeegegggiiiiiiiiiiiiiihidghhihiiiiiiiiiigghhhhihiiiiiihigeeddddddcccccccbccccccaccccccccccccbbccc
@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
AACAAAACGGGCTGTTTTAGGACCCTTGGTCCCAAGGGGTAATGGCCCTCAGCACCCACTATCCCTGCTCTCCAGGGCTCTCTAGGGATTTAGTGCTGAT
```

.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:

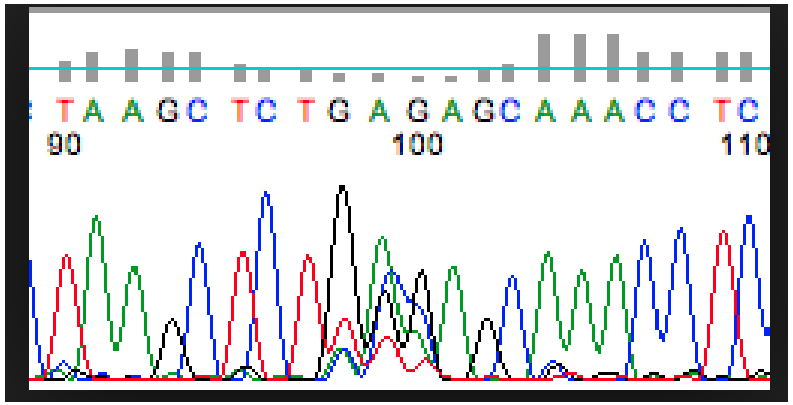
.fastq

```
@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

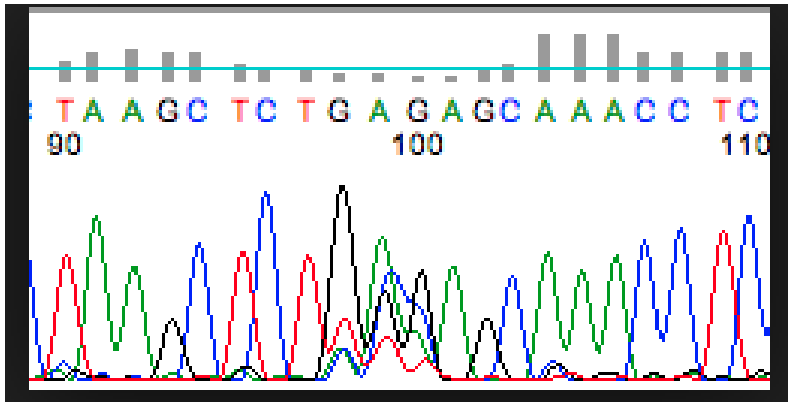
- Q = -10 x log P
- where:
  - P, probability of base calling being incorrect
  - High Q = high probability of the base being correct



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

Phred Quality Score

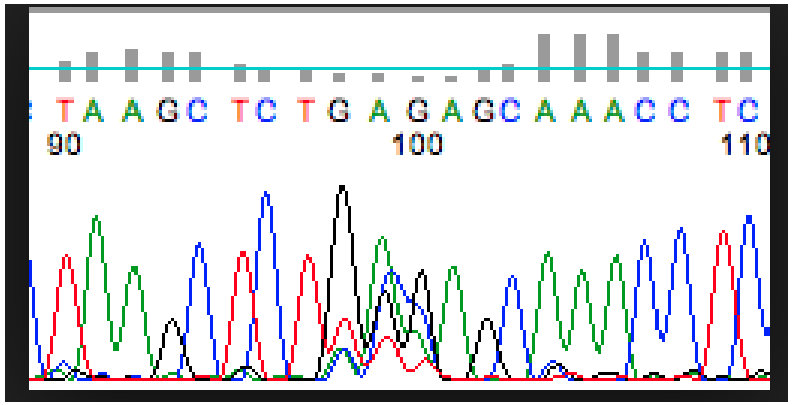
- Q = -10 x log P
- where:
  - P, probability of base calling being incorrect
  - High Q = high probability of the base being correct



- 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

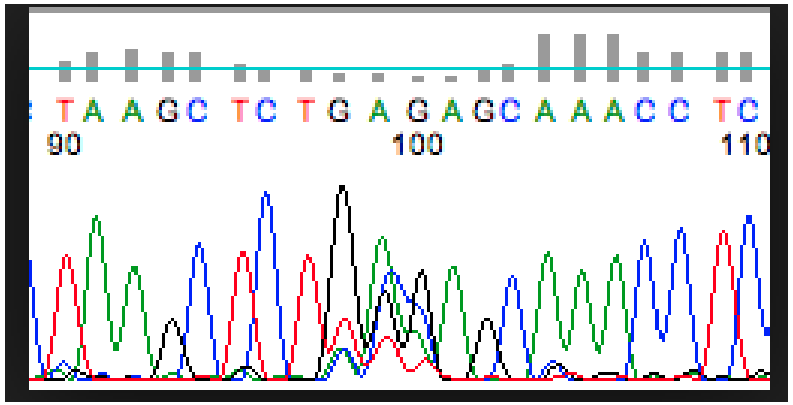
- Q = -10 x log P
- where:
  - P, probability of base calling being incorrect
  - High Q = high probability of the base being correct



- 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
```

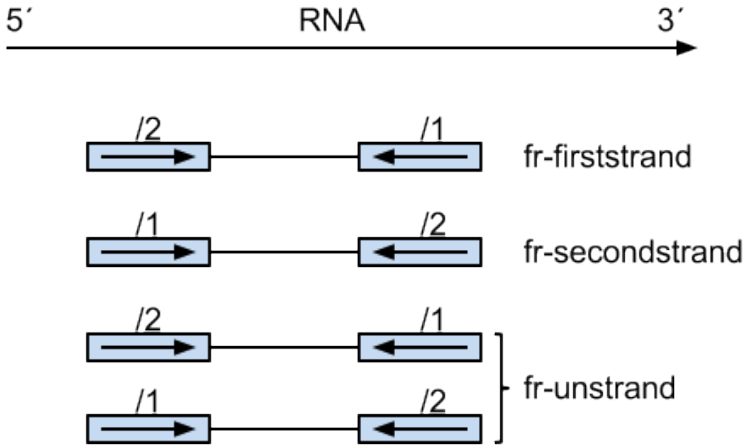
```
@61DFRAAXX100204:1:100:10494:3070/2
ATCCAAGTTAAAACAGAGGCCTGTGACAGACTCTTGGCCCATCGTGTTGATA
+
_^_a^cccegcgghhgZc`ghhc^egggd^_[d]defcdfd^Z^OXWaQ^ad
```

SE

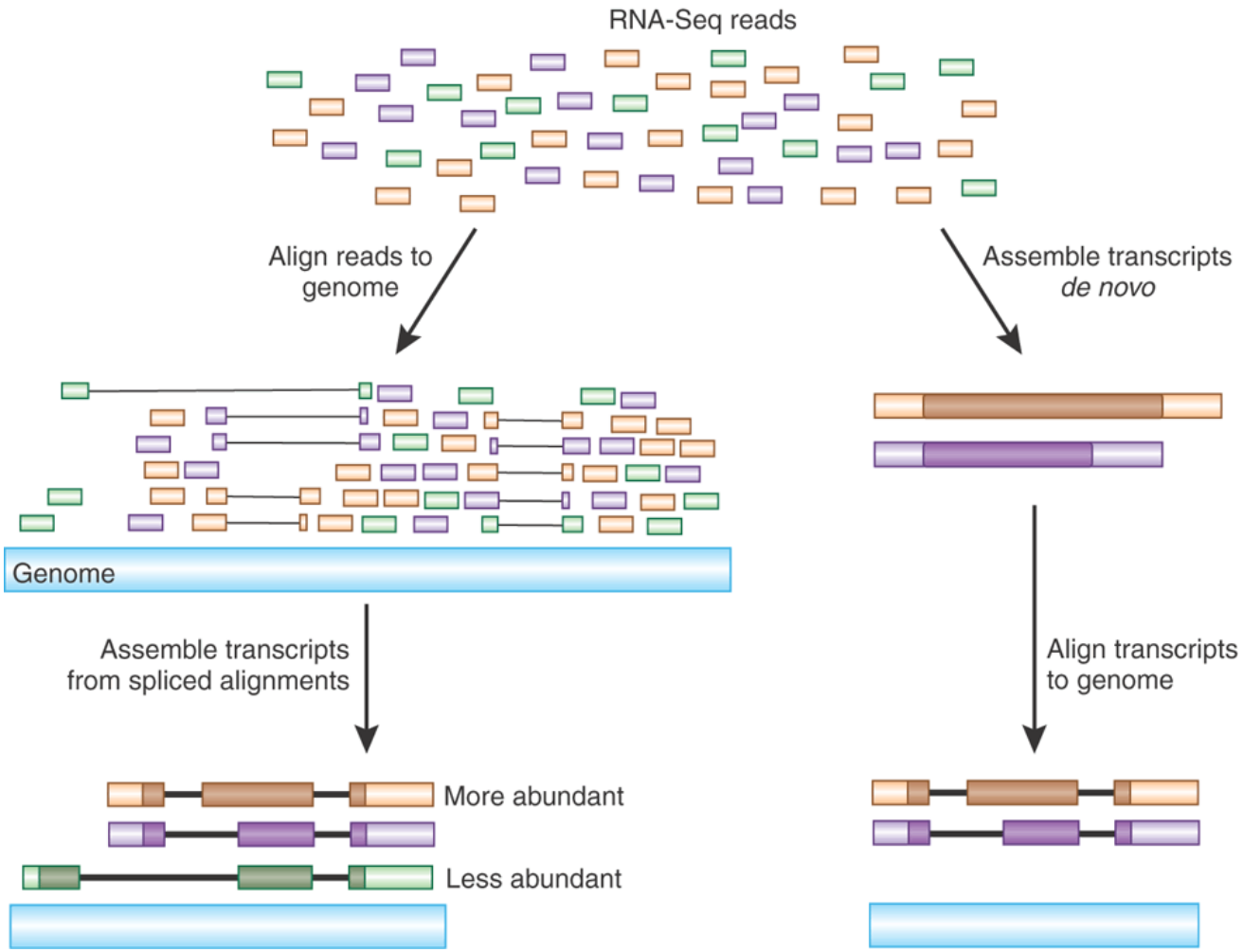
- F: the single read is in the sense (F, forward) orientation
- R: the single read is in the antisense (R, reverse) orientation

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

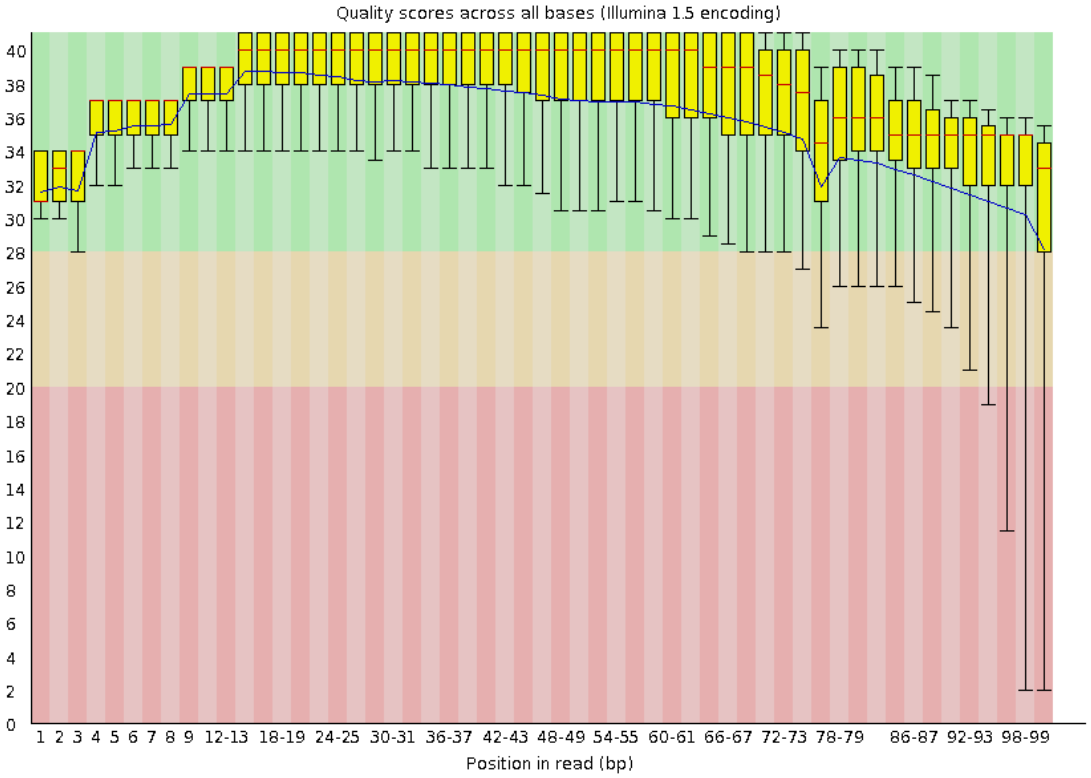

RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

Main steps

- Initial processing incl. QC
- Aligning reads to reference genome
- Counting reads
- Differential gene expression
- Annotations of transcripts

Initial processing incl. QC

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



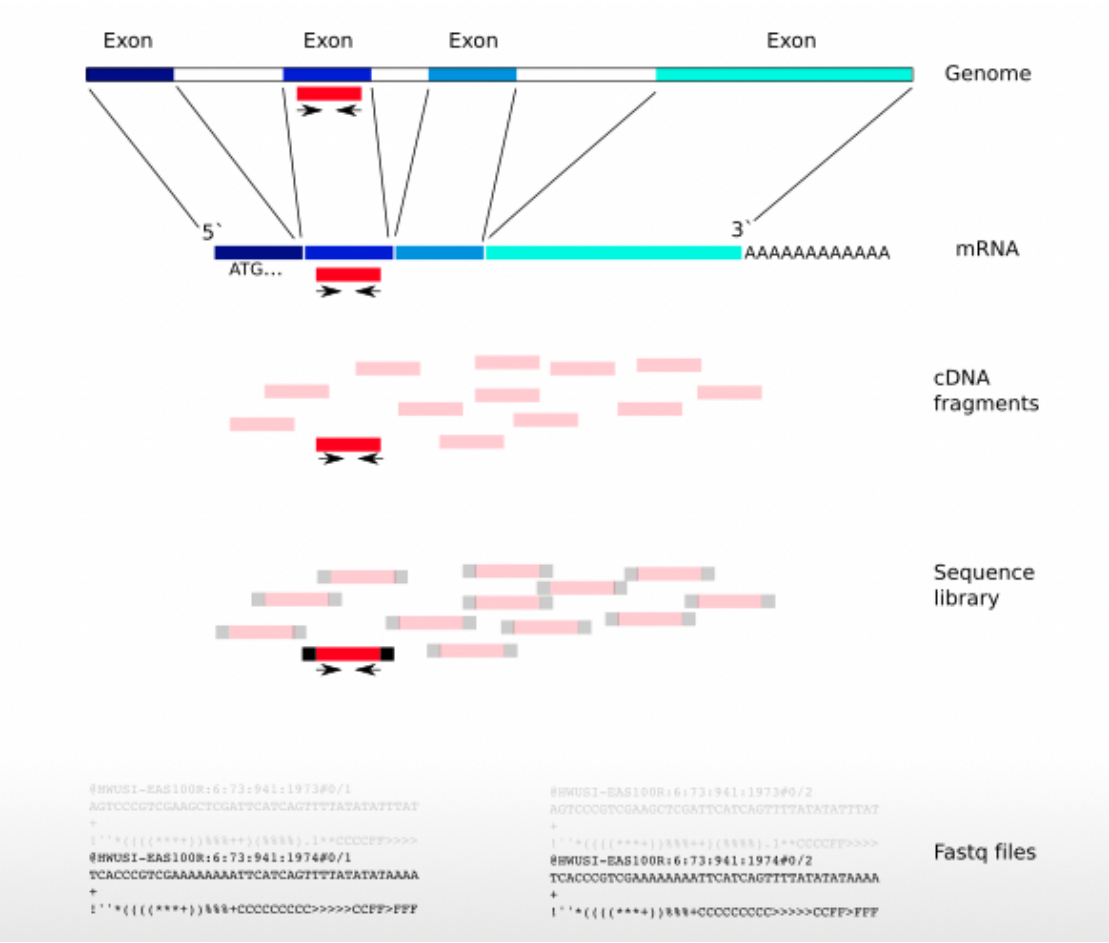
Available tools

FastQC, PRINSEQ, TRIMMOMATIC, TrimGalore, FastX, Cutadapt

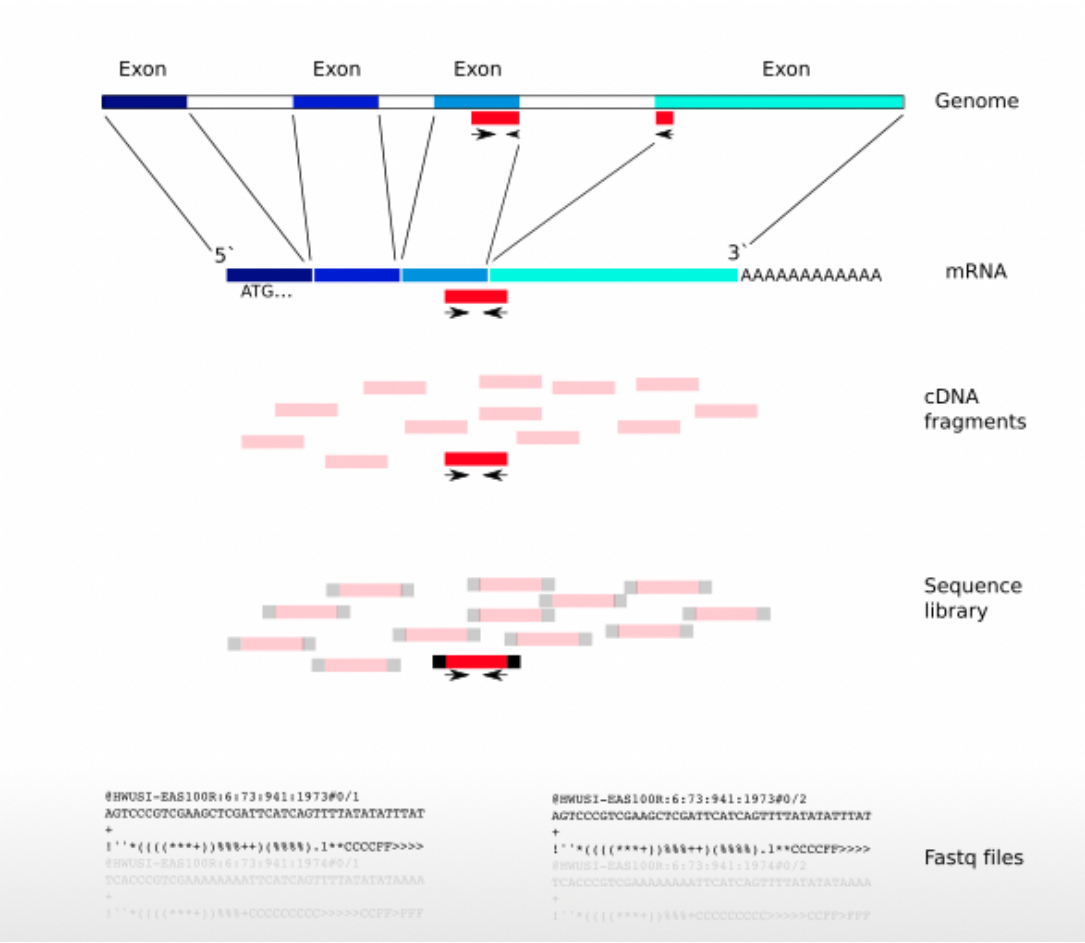
Initial processing incl. QC

- 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 differece" is with default settings, most mappers will allow to optimise settings
- perfomance vary by genome complexity

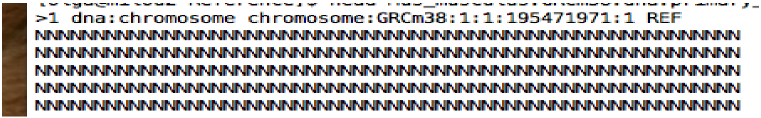
A good read: Barruzo et. al. Nature Methods 14, (2017)
https://www.nature.com/articles/nmeth.4106

Available tools

STAR, HISAT, MapSlice2, Subread, TopHat

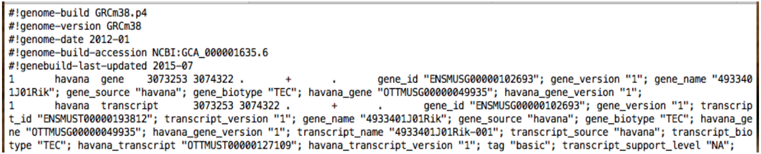
Aligning reads: reference files

.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)

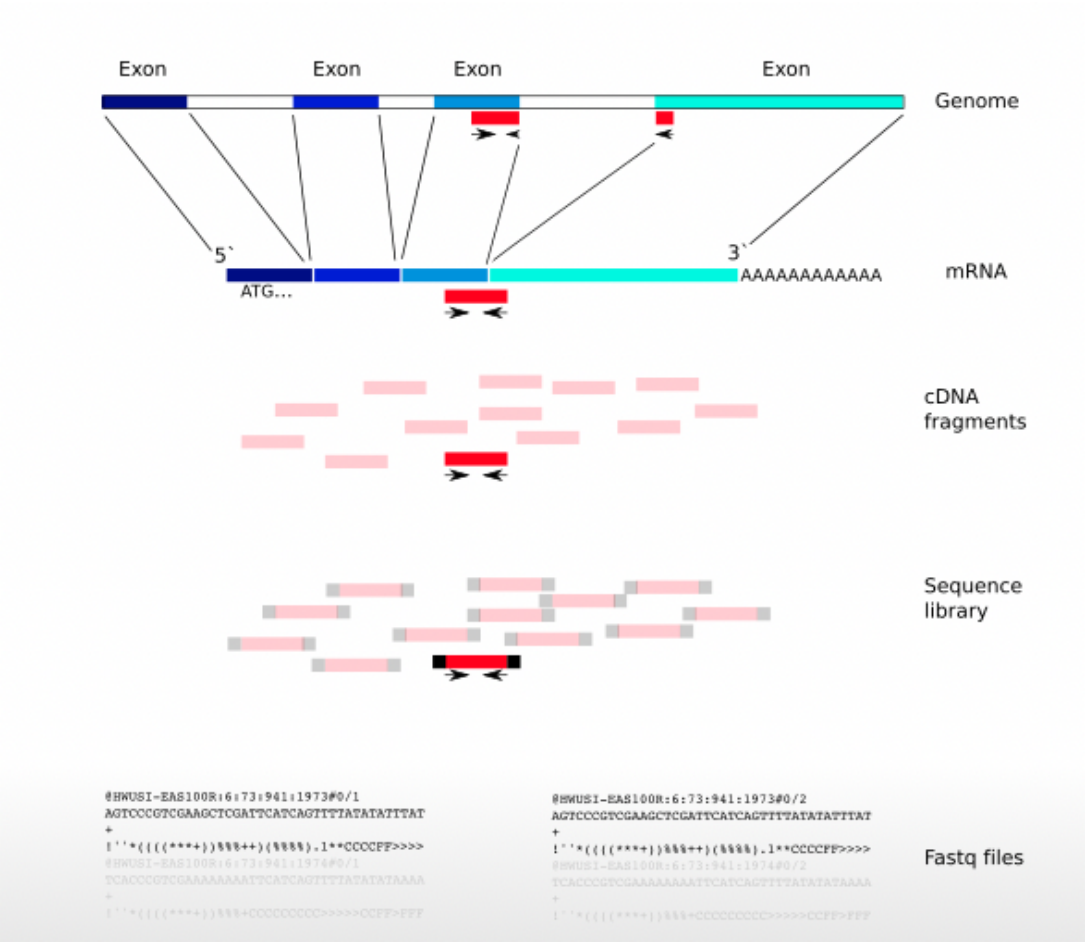


```
#!genome-build GRCm38.p4
#!genome-version GRCm38
#!genome-date 2012-01
#!genome-build-accession NCBI:GCA_000001635.6
#!genebuild-last-updated 2015-07
1       havana  gene    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693"; gene_version "1"; gene_name "493340
1J01Rik"; gene_source "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";
```

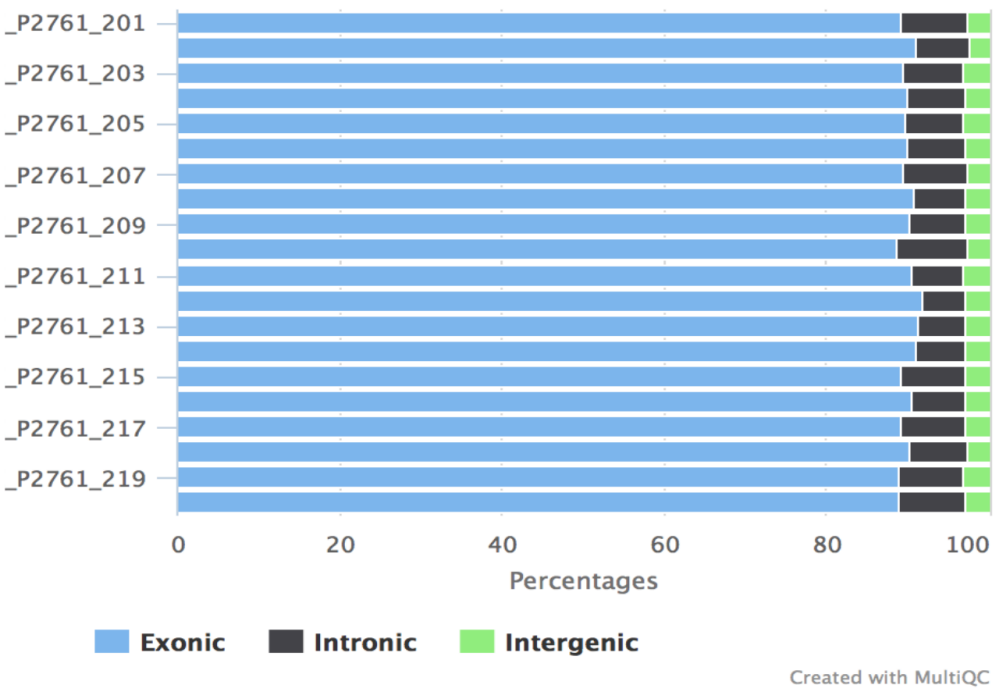
Source

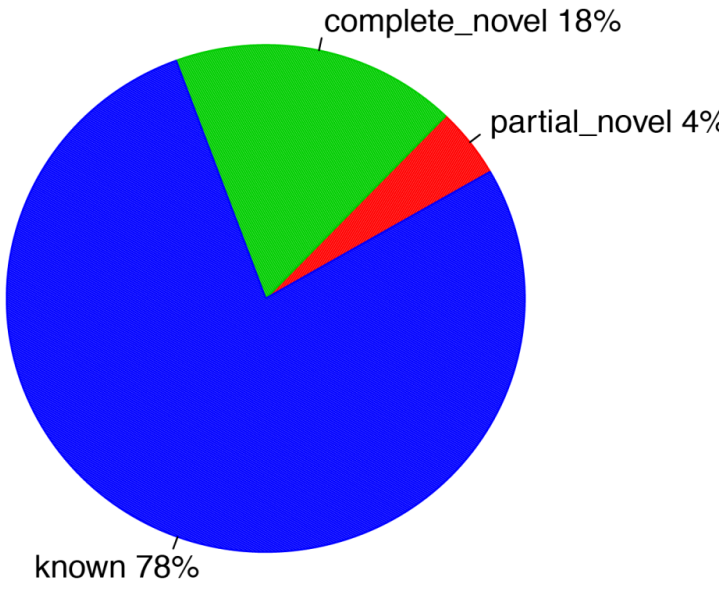
ENSEMBL, NCBI

Aligning reads



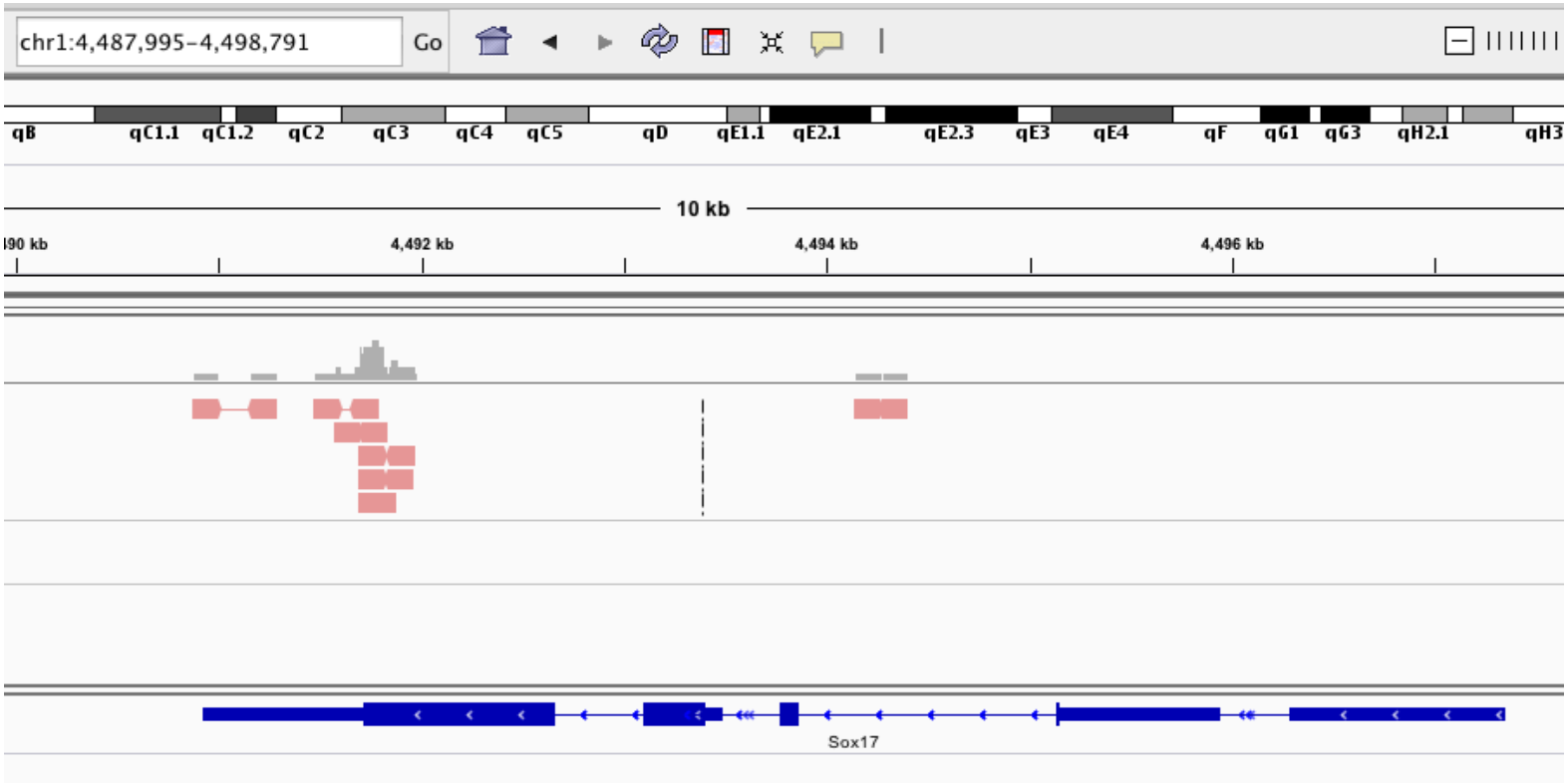
Aligning reads: QC





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

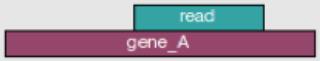
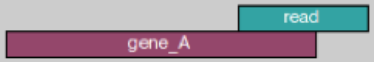
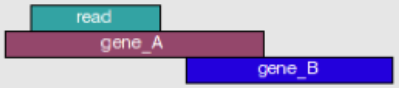
Counting reads



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

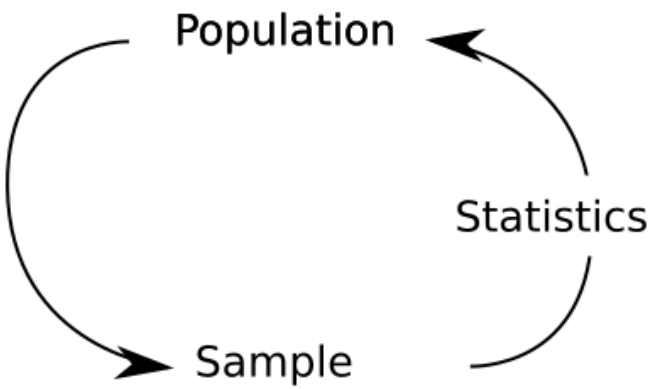
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Transcript	P1822_1	P1822_2	P1822_3	P1822_4	P1822_5	P1822_6	P1822_7	P1822_8	P1822_9	P1822_10	P1822_11	P1822_12	P1822_13	P1822_14	P1822_15	P1822_16
2	ENSMUSG00000102693	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	ENSMUSG00000088000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	ENSMUSG00000103265	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

Differential gene expression

Differential expression analysis

- means taking the normalized read count data &
- performing 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.
- or simply: checking for differences in distributions

Differential gene expression

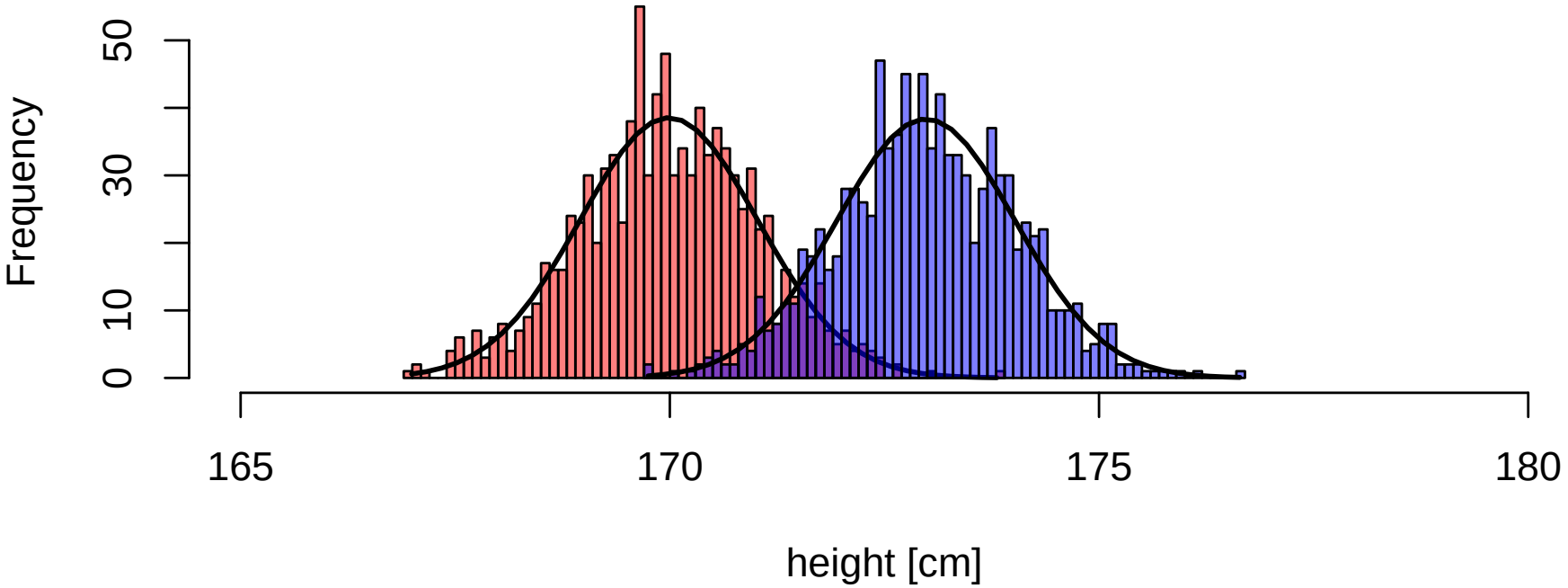


$$Outcome_i = (Model_i) + error_i$$

- 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

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



Differential gene expression

Simple recipe

- model e.g. gene expression with random error
- fit model to the data and/or data to the model, estimate model parameters
- use model for prediction and/or inference

Implications

- the better model fits to the data the better statistics