Gene Expression Analysis
RNA-Seq & MARS-Seq

Dena Leshkowitz

An introduction to deep-sequencing analysis for biologists 2021

מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

Agenda

- Introduction & Experimental design
- Analysing Gene expression from RNA-Seq data
- Analysing Gene expression from bulk MARS-Seq data

RNA-Seq Potential

RNA-Seq: developed a decade ago has become an indispensable tool for transcriptome analysis

In theory RNA-Seq can be used to build a complete map of the transcriptome across all cell types, perturbations and states (Trapnell C. et al, Nature methods 6 469-477(2011))

RNA-Seq Applications

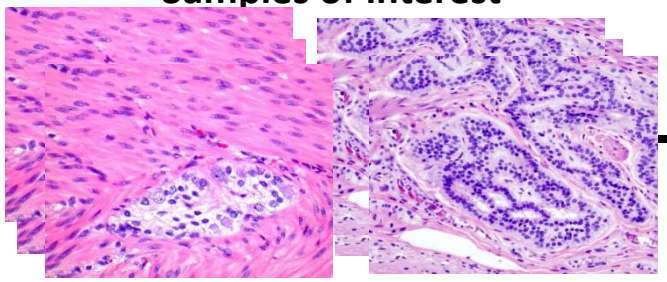
- Transcript level analysis (to be discussed Dec 13th) :
  - Discover novel transcripts
  - Determine transcript structure
  - Measure transcripts expression
  - Detect differentially expressed transcripts/isoforms between conditions, treatments…

- Gene expression analysis for Model Organism:
  - Gene expression quantification
  - Detect differentially expressed genes between conditions, treatments… **based on known gene structures**

RNA-Seq Experiment Workflow

Why do we need to add linkers?

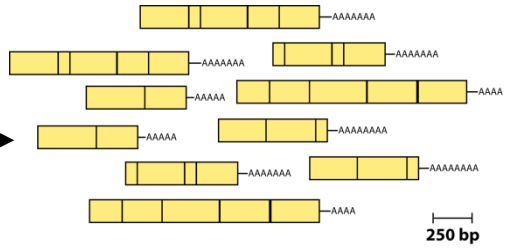
Samples of interest



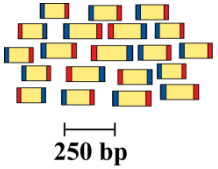
Condition 1 (normal colon)

Condition 2 (colon tumor)

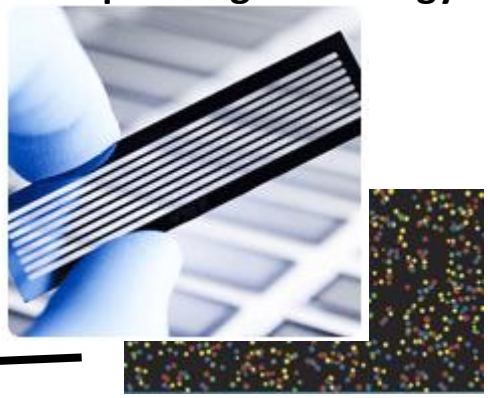
Isolate RNAs



Library : convert mRNA to amplified small cDNA fragments with linkers at the ends

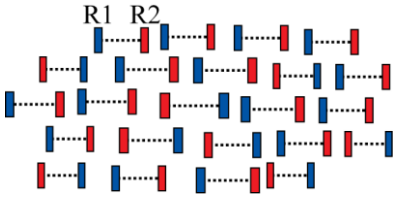


Illumina sequencing technology



Million/billions of fragments are sequenced

Sequencing short reads (R1 and R2)

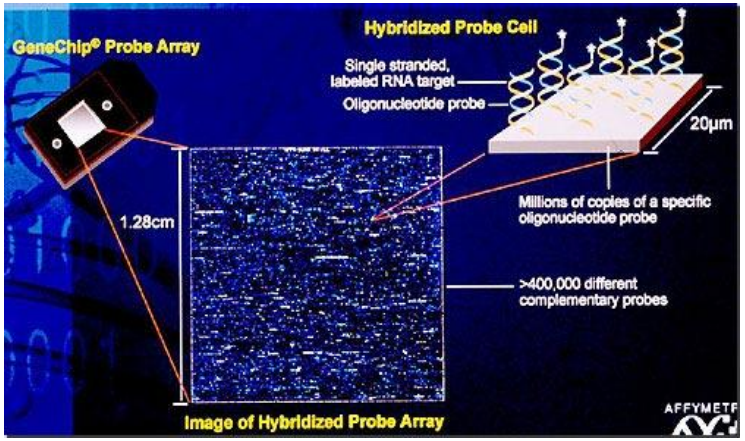


Assemble or map to genome or to a transcriptome

Quantify and detect differentially expressed genes (DEG) and transcripts

Modified - www.med.nyu.edu/rcr/rcr/course/**RNA-seq**.pptx

High Throughput Genomics



DNA Microarrays

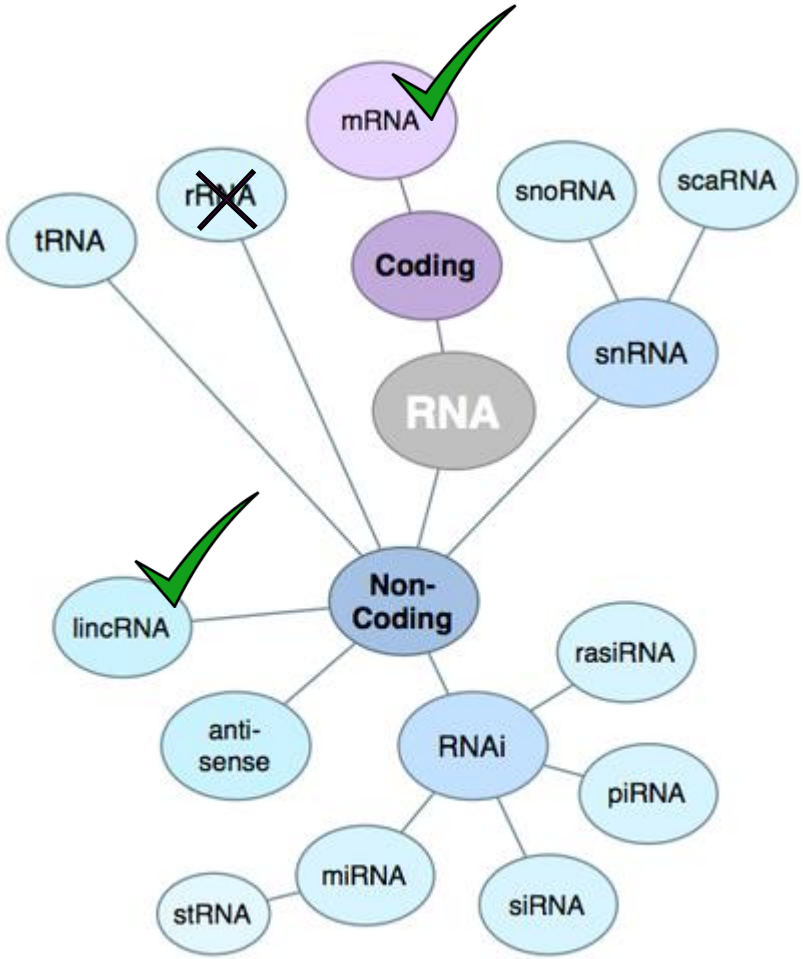
Illumina
NovaSeq
NextSeq500
…

Agenda

- Introduction & Experimental design
- Analysing Gene expression from RNA-Seq data
- Analysing Gene expression from bulk MARS-Seq data

mRNA in the RNA "World"

- Most abundant RNA is rRNA – 98%
- Illumina standard protocol enriches for mRNA by:
  - oligo(dT)-based affinity matrices
  - Sequence: rRNA capture beads (Ribo-Zero)



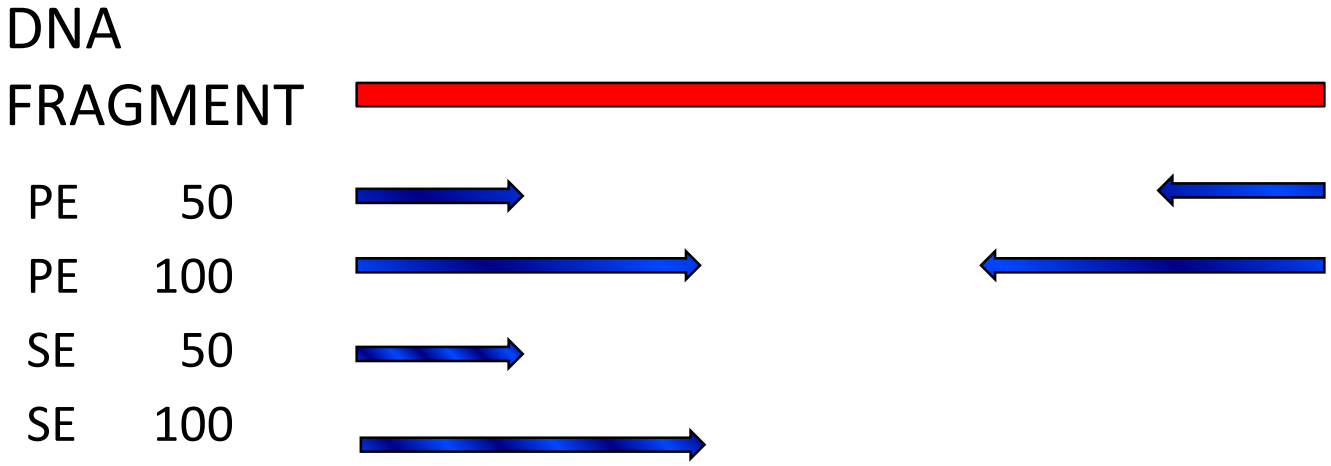
geospiza

Sequencing Options

Illumina NextSeq/NovaSeq Sequencing options:

- Length of sequence (up to 300 bases)
- Paired-end (PE) or single-end (SE)

Both PE and longer length sequencing increase the sensitivity and specificity of the detection of the alternative splicing and novel transcripts



Experimental Design

Mammalian tissue

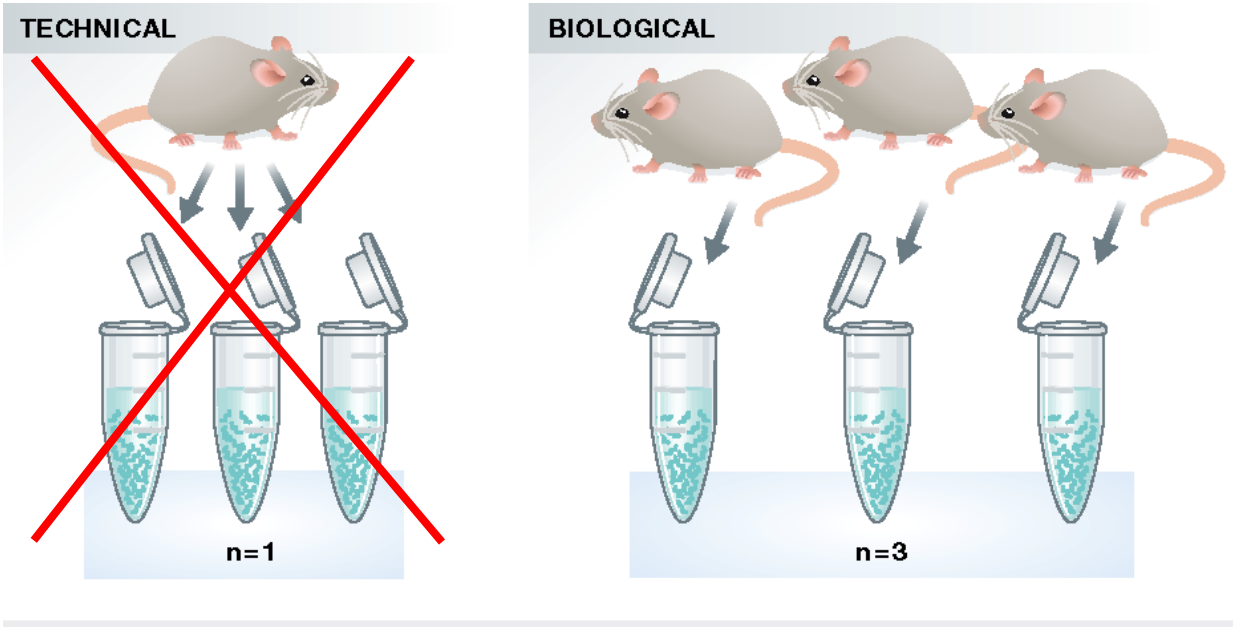
Liu Y. et al., 2014; ENCODE 2011 RNA-Seq

Differential gene expression profiling	**>20M aligned reads**	**>30 base single-end**
Alternative splicing	50-100M	>75 base paired-end
Allele specific expression	50-100M	100 base paired-end
De novo assembly	>100M	100 base paired-end

(5M Bulk MARS-Seq)

Biological Replicates

- Usually our goal in a RNA-Seq experiment is to detect Differentially Expressed Genes (DEGs) between groups.
- For each group we should have several samples, which are also known as replicates
- Assessing biological variation requires biological replicates - three are a minimum, yet more are recommended



BIOINFORMATICS DISCOVERY NOTE

Vol. 30 no. 3 2014, pages 301–304
doi:10.1093/bioinformatics/btt688

Gene expression

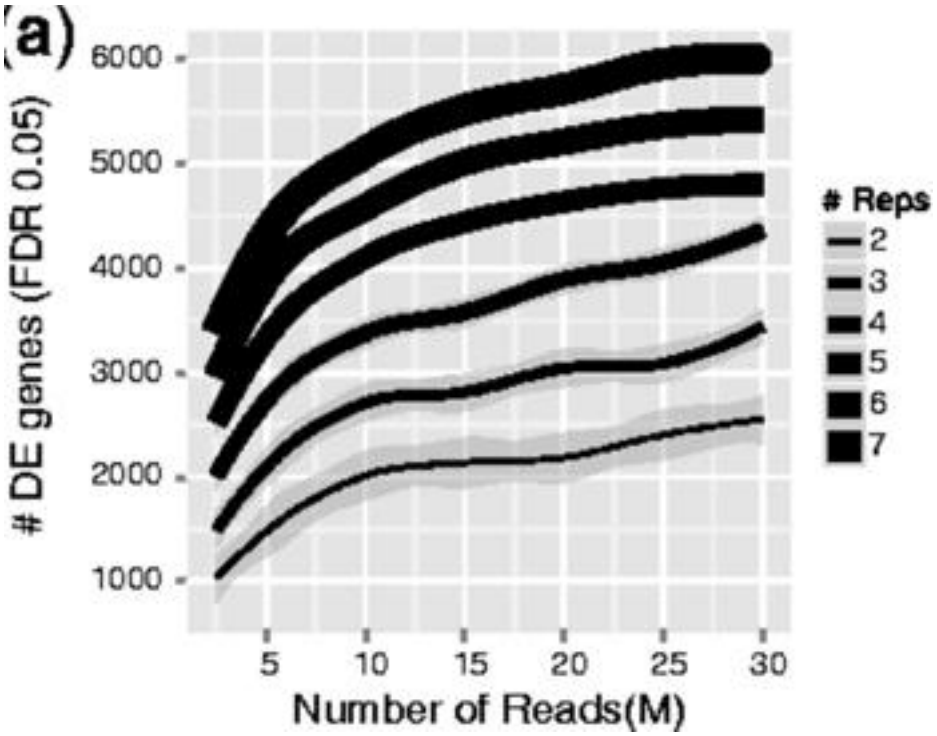
Advance Access publication December 6, 2013

RNA-seq differential expression studies: more sequence or more replication?

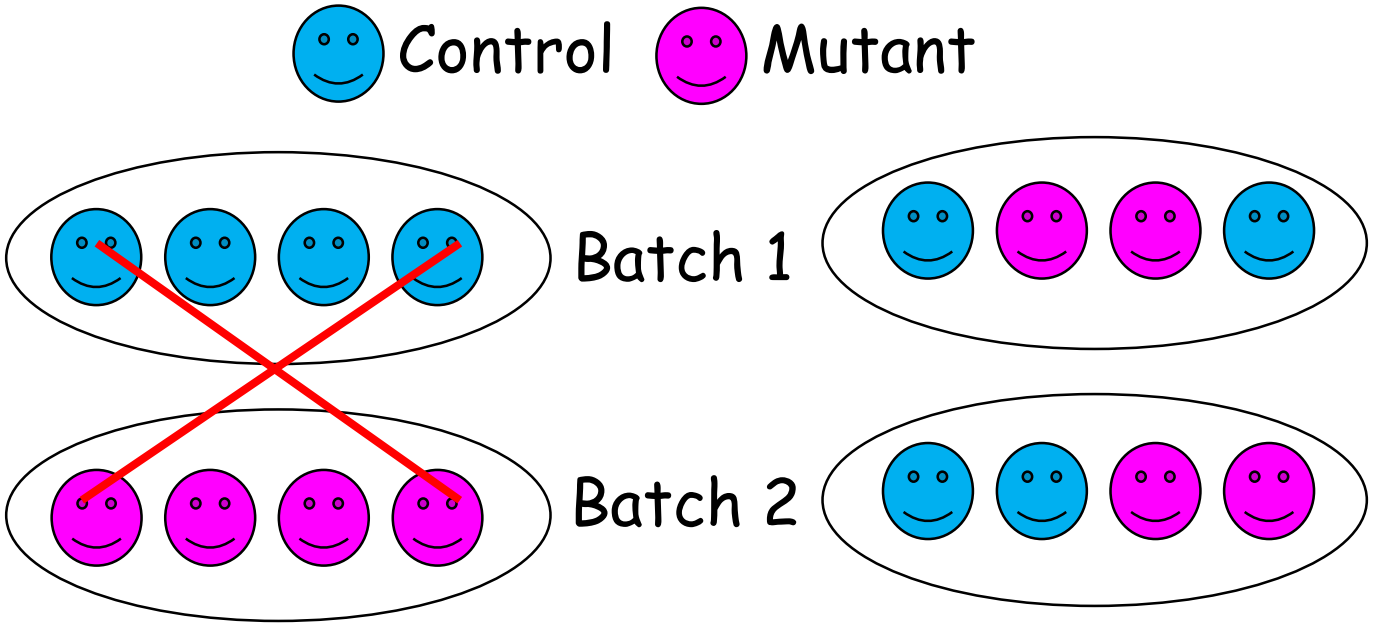
Yuwen Liu[1,2], Jie Zhou[1,3] and Kevin P. White[1,2,3,*]

[1]Institute of Genomics and Systems Biology, [2]Committee on Development, Regeneration, and Stem Cell Biology and [3]Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA

Associate Editor: Janet Kelso



Proper Experimental Design



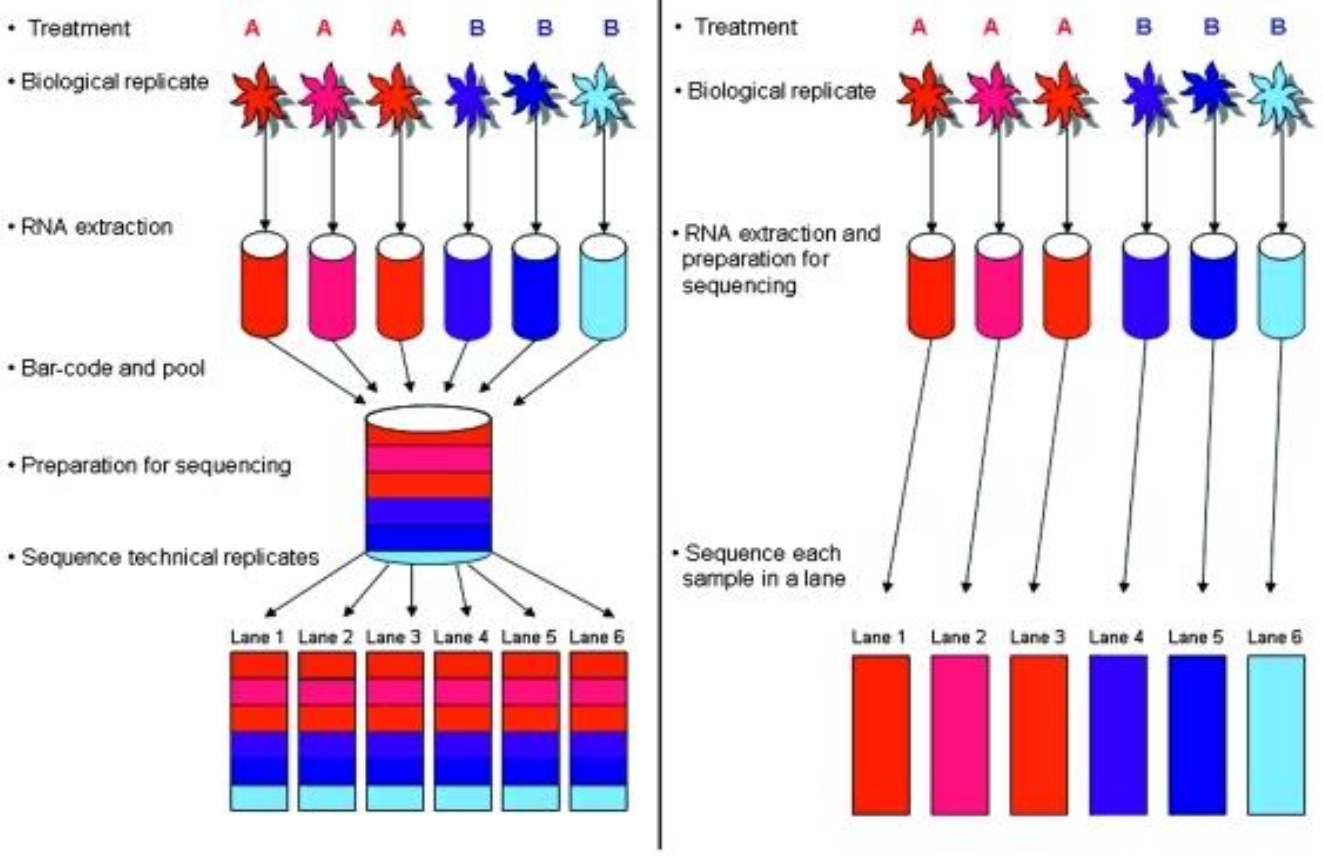
- It is impossible to partition biological variation from technical variation, when these two sources of variation are confounded.
- No amount of statistical sophistication can separate confounded factors after data have been collected.

Batch Effects

Avoid batch effects - Technical sources of variation that can take place during processing the samples:

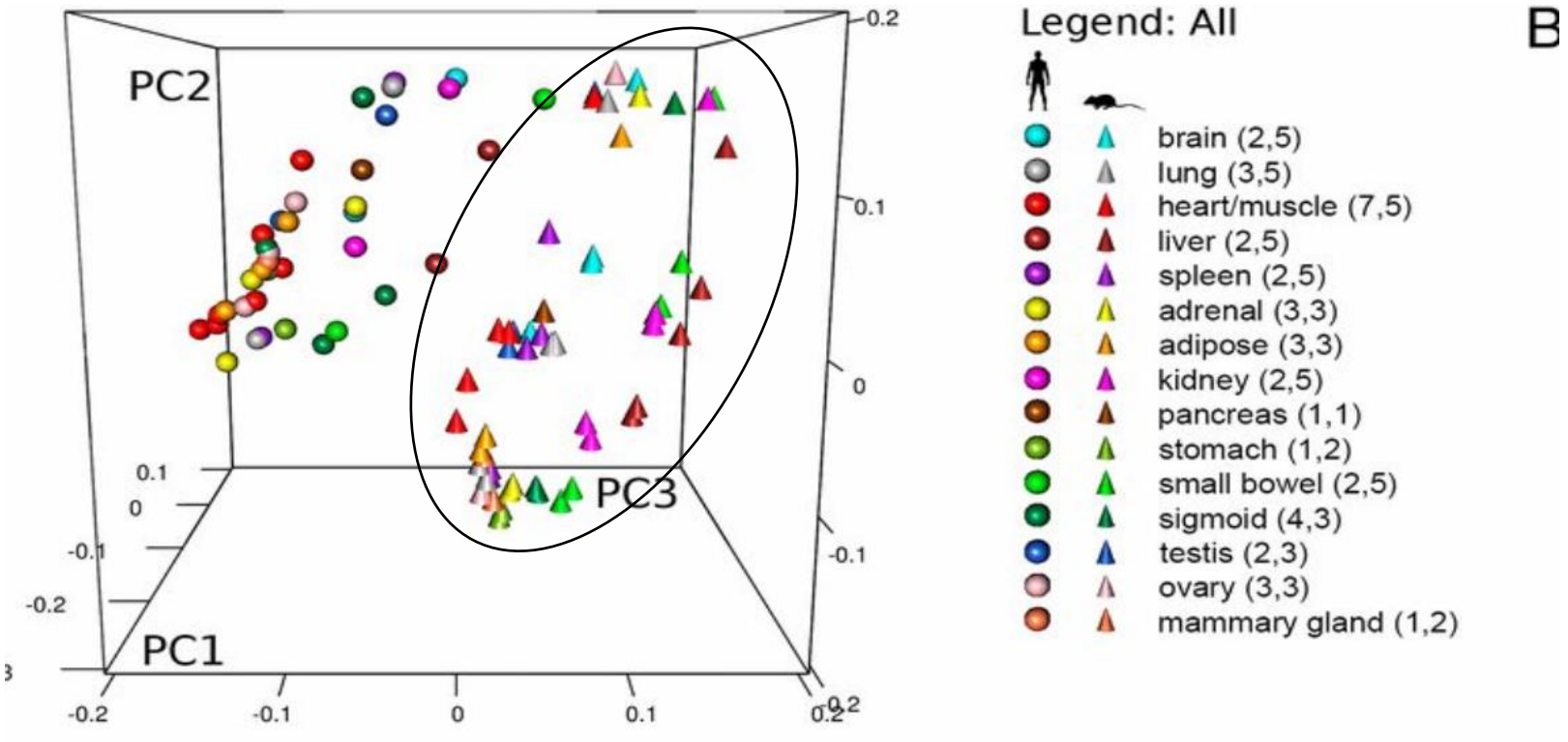
- Extracting RNA with different kits
- Sequencing on different flowcells or lanes

This design avoids the lane and flowcell batch effect



Auer and Doerge GENETICS *June 1, 2010 vol. 185 no. 2*

Encode reported that gene expression was likely to follow a species-specific rather tissue-specific pattern (?)

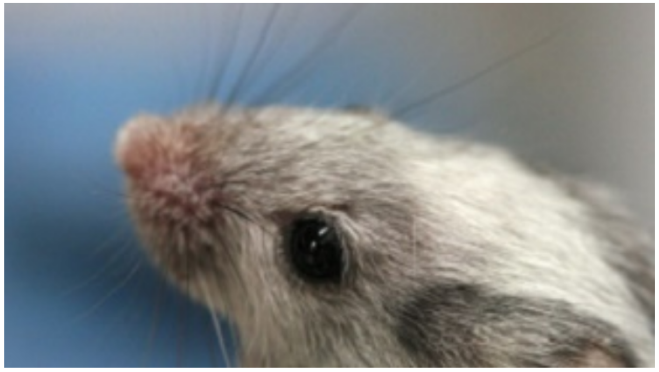


https://doi.org/10.1073/pnas.1413624111

TheScientist
EXPLORING LIFE, INSPIRING INNOVATION

Reanalysis of Mouse ENCODE data suggests mouse and human genes are expressed in tissue-specific, rather than species-specific, patterns.

May 19, 2015
JYOTI MADHUSOODANAN



WIKIMEDIA, RAMA

Late last year, members of the Mouse ENCODE consortium reported in *PNAS* that, across a wide range of tissues, gene expression was more likely to follow a species-specific rather than tissue-specific pattern. For example, genes in the mouse heart were expressed in a pattern more similar to that of other mouse tissues, such as the brain or liver, than the human heart.

But earlier this month, Yoav Gilad of the University of Chicago called these results into question on Twitter. With a dozen or so 140-character dispatches (including three heat maps), Gilad suggested the results published in *PNAS* were an anomaly—a result of how the tissue samples were sequenced in different batches. If this "batch effect" was eliminated, he proposed, mouse and human tissues clustered in a tissue-specific manner, confirming previous results rather than supporting the conclusions reported by the Mouse ENCODE team.

Figure 1. Study design.

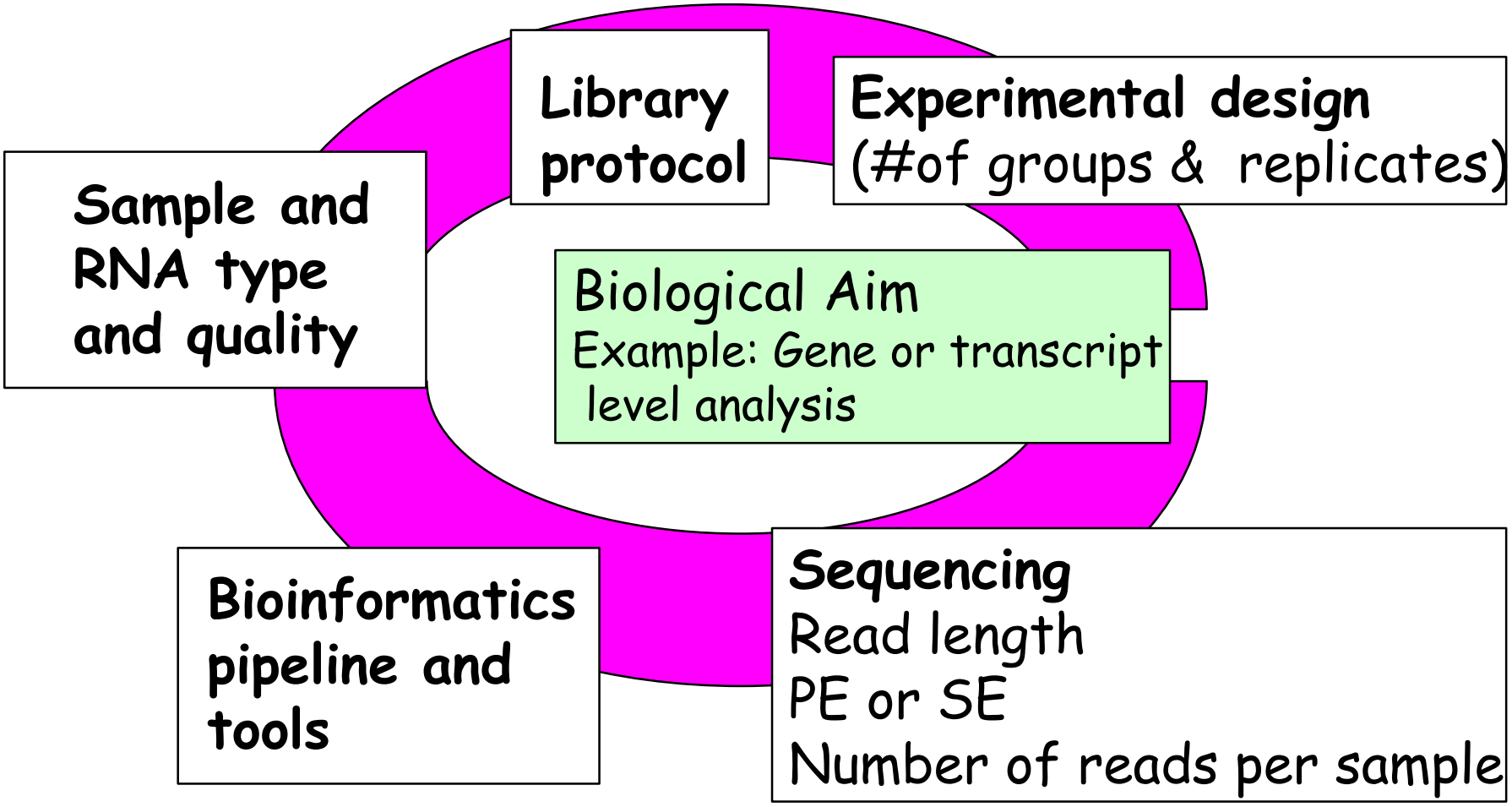
D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX , lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX , lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX , lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	● Human
testis		pancreas		● Mouse

Gilad Y and Mizrahi-Man O. A reanalysis of mouse ENCODE comparative gene expression data [version 1]. F1000Research 2015, 4:121 (doi: 10.12688/f1000research.6536.1)

Analysis done in which the batch effect is accounted for shows clustering is in a tissue specific manner.

F1000Research

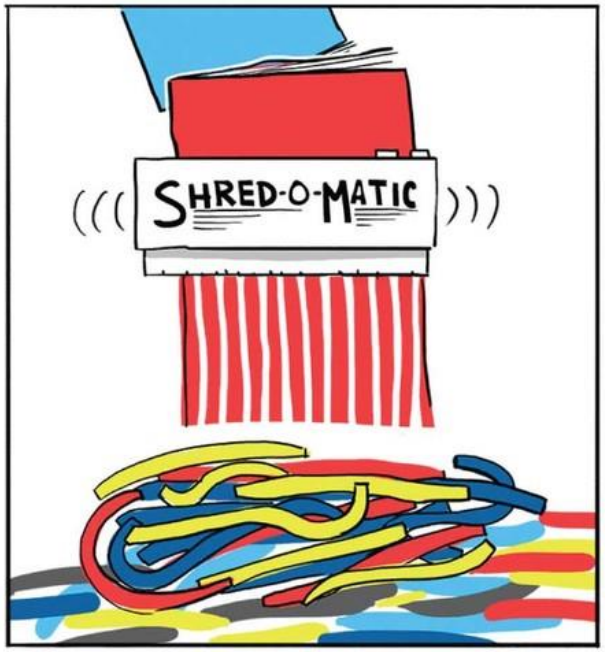
Summary RNA-Seq Experiment Planning



We recommend to come to a kick-off meeting with us, to help **plan** your experiment

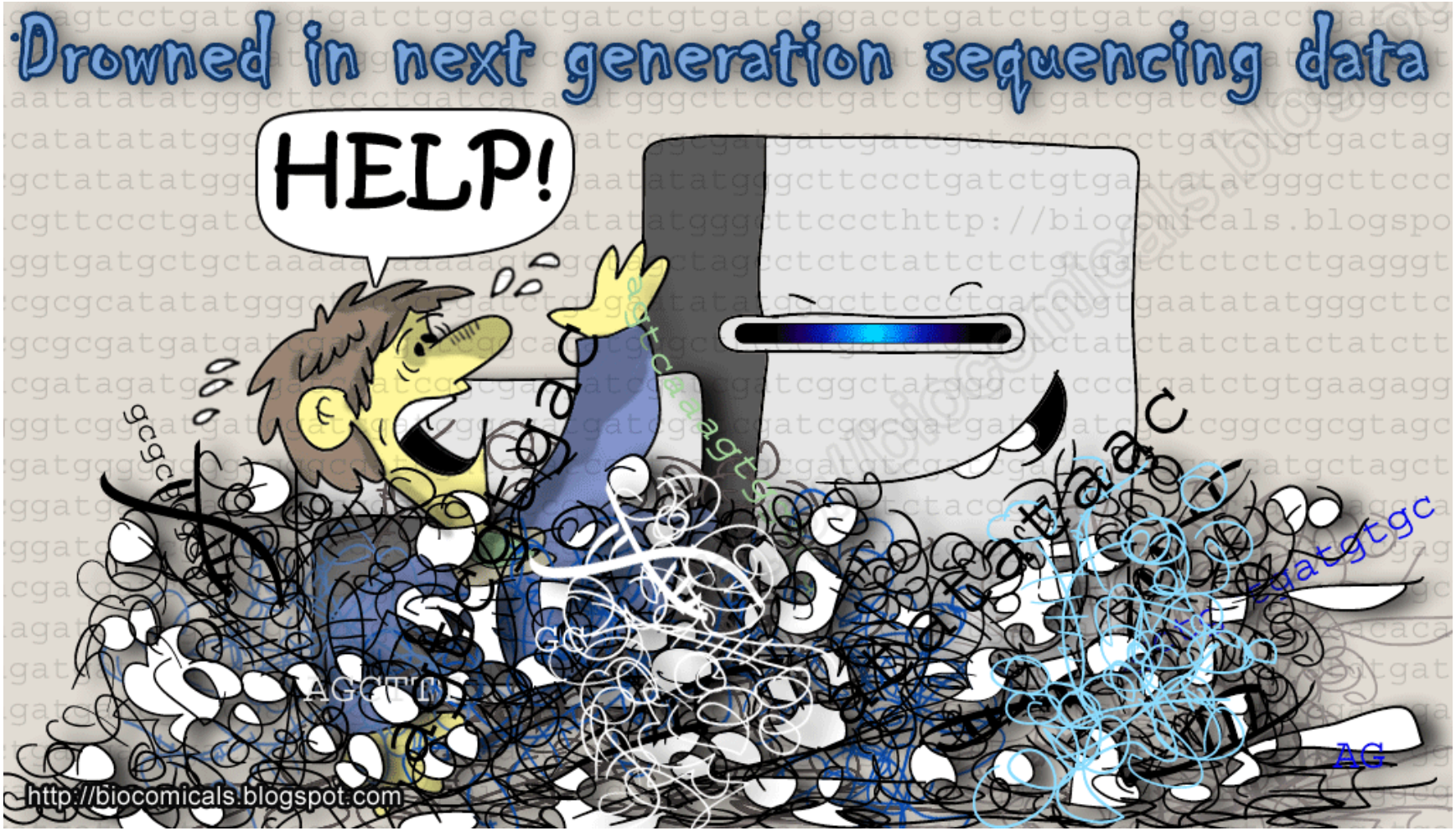
Agenda

- Introduction & Experimental design
- Analysing Gene expression from RNA-Seq data
- Analysing Gene expression from bulk MARS-Seq data



RNA-Seq is a straightforward process: you isolate RNA, sequence it with a high-throughput sequencer, and put it all back together. What is the problem?

Abigail Yu Nature Methods 10, 1165–1166 (2013)

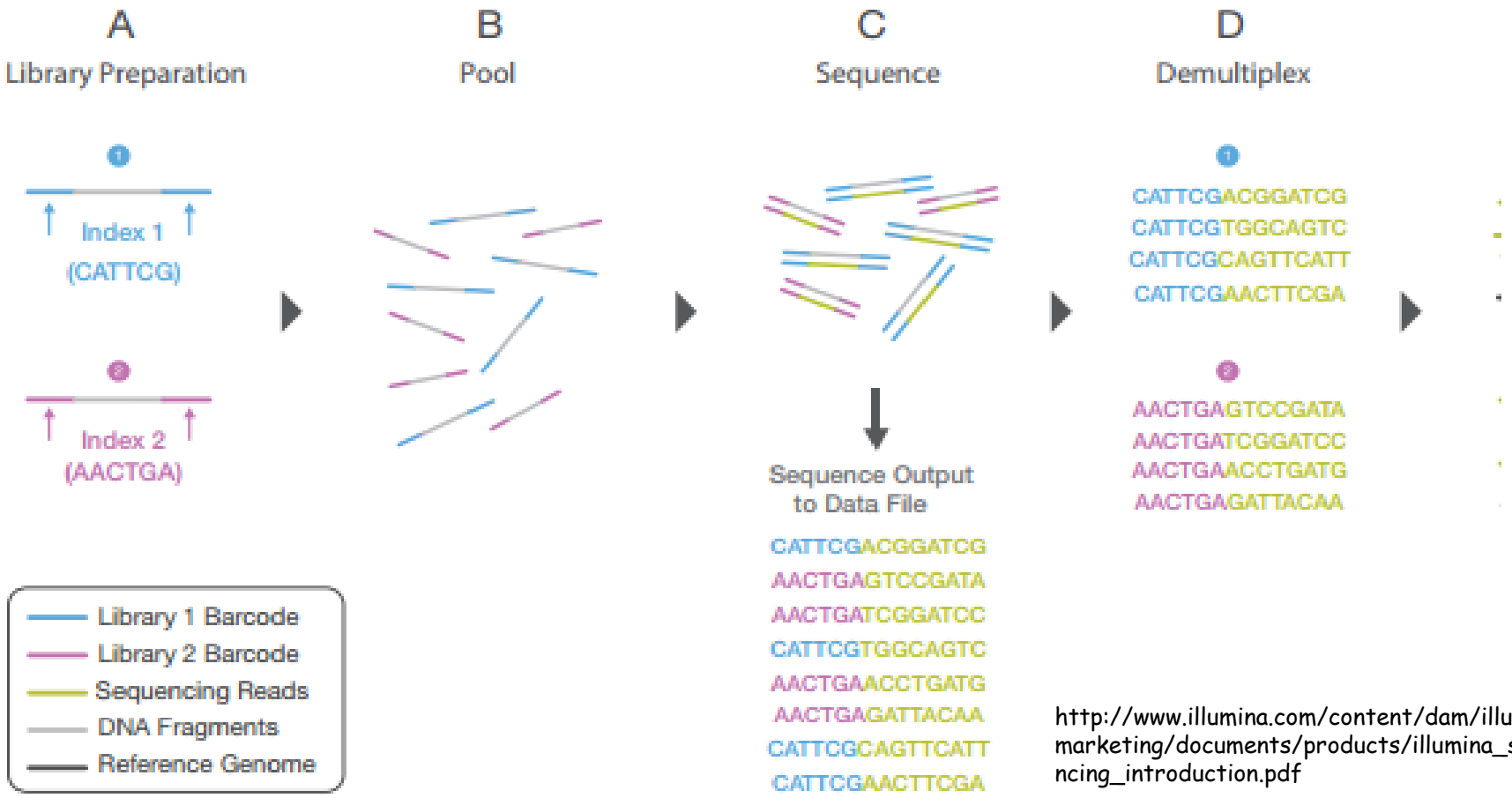


HELP !!!!

I just got sequence data…

Library Barcode

- Multiplexing: the process of pooling samples together and sequencing them simultaneously
- Demultiplexing: separating reads using the library barcode to identify the origin sample



http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Sample Index (barcodes)

Setting up the NGS workflow includes loading the information that matches between the index and the sample name

Sequence Output Format

- FASTQ

  Line 1: Unique ID for a sequencing read

  Line 2: Sequences

  Line 3:+

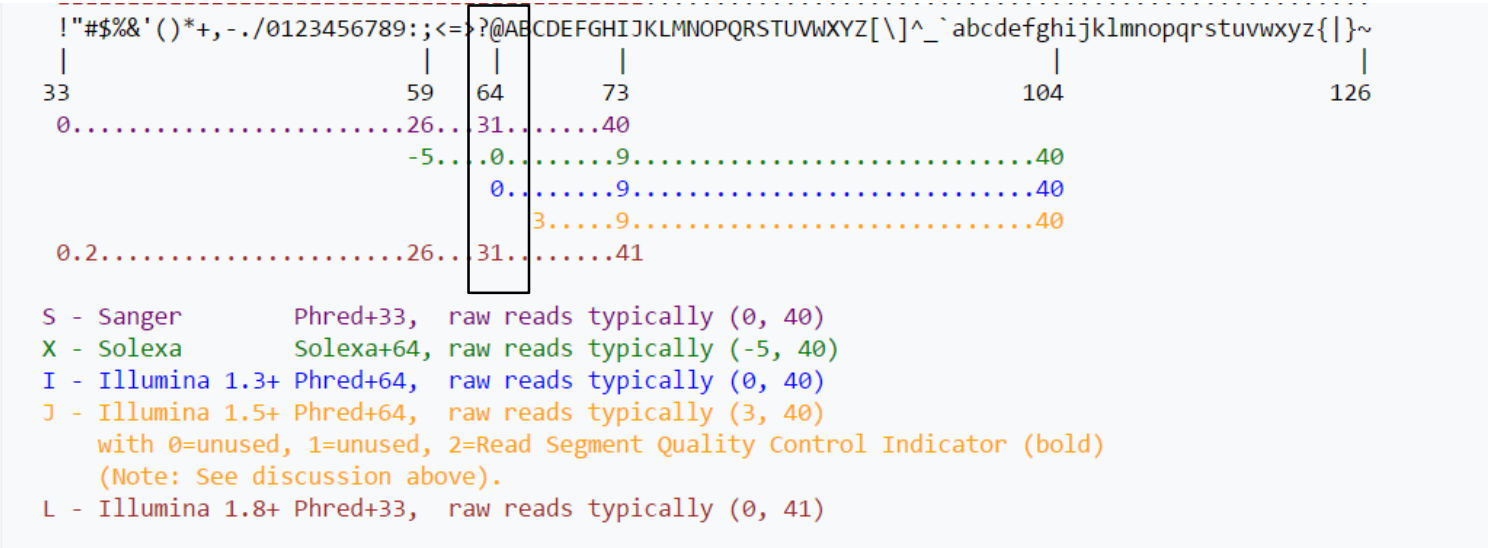
  Line 4: Base calling quality score (Analogous to Phred scores but in ASCII value)

  Example:

```
@HISEQ:126:H14YJADXX:1:1101:1118:2101 1:N:0:ATCACG
CTCCATAGTCAGAAACTTCAGCATGACAGTACCTCATGCTGCATCAGGTGATCATGAAAAGATTACAGGCTTTCTAAAATTATCAGCAAGATATGG
+
@@?ADDDD?ADHDIIIIIIIEIIIGEFHC<?FH4C9E9BGAFIGH<DG9BD?@DGGEGHHG<DCBBCC8C>FHCGEHIGEEE>EEHEEEEC>A>;;
```

Quality Score Representation

- Quality scores are represented as ASCII characters in order to save space, so that there is one ASCII character per base.
- Converting between the ASCII value and the phred score



https://en.wikipedia.org/wiki/FASTQ_format

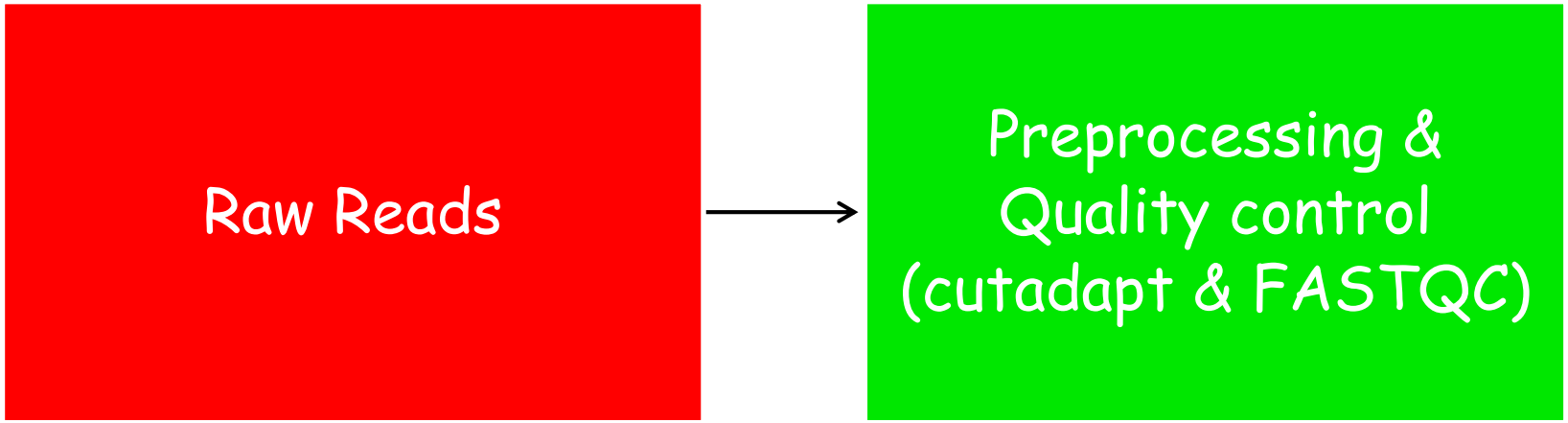
Quality Scores

- Quality scores are used to measure base accuracy
- Illumina uses Phred scores

ASCII Quality Score	Probability of Incorrect Based Call	Base Call Accuracy	Q-score
+	1 in 10	90%	Q10
5	1 in 100	99%	Q20
?	1 in 1000	99.9%	Q30
I	1 in 10000	99.99%	Q40

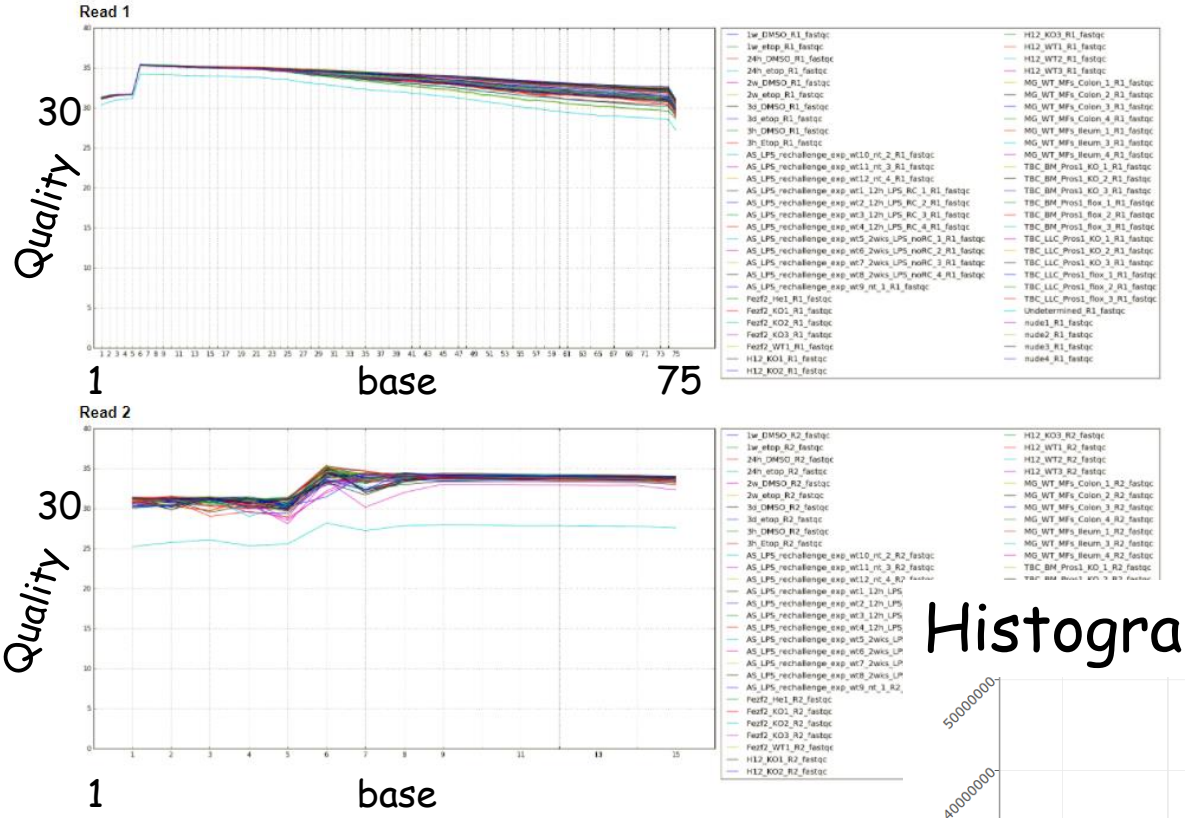
$$\text{Q score} = -10 \log_{10} p$$

RNA-Seq Workflow

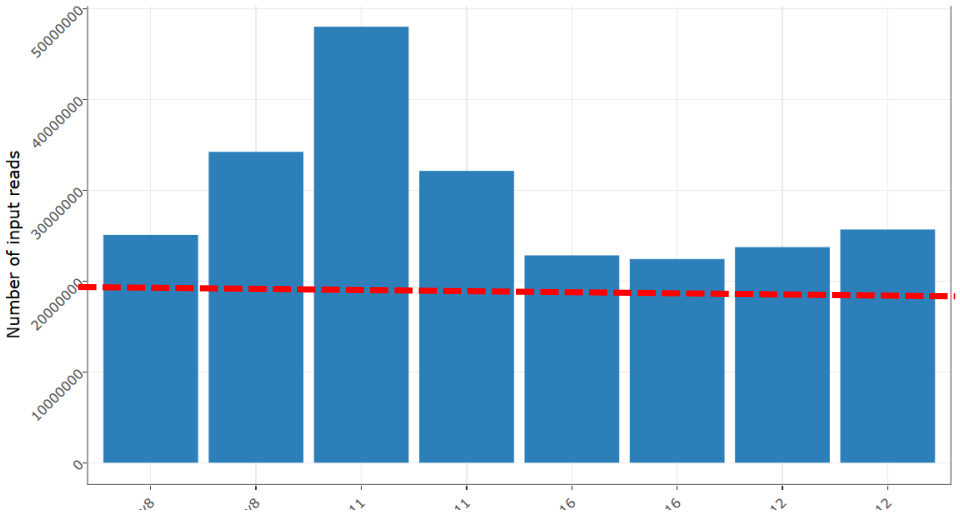


Example of QC report

Mean Per Base Sequence Quality



Histogram: # Reads per Sample



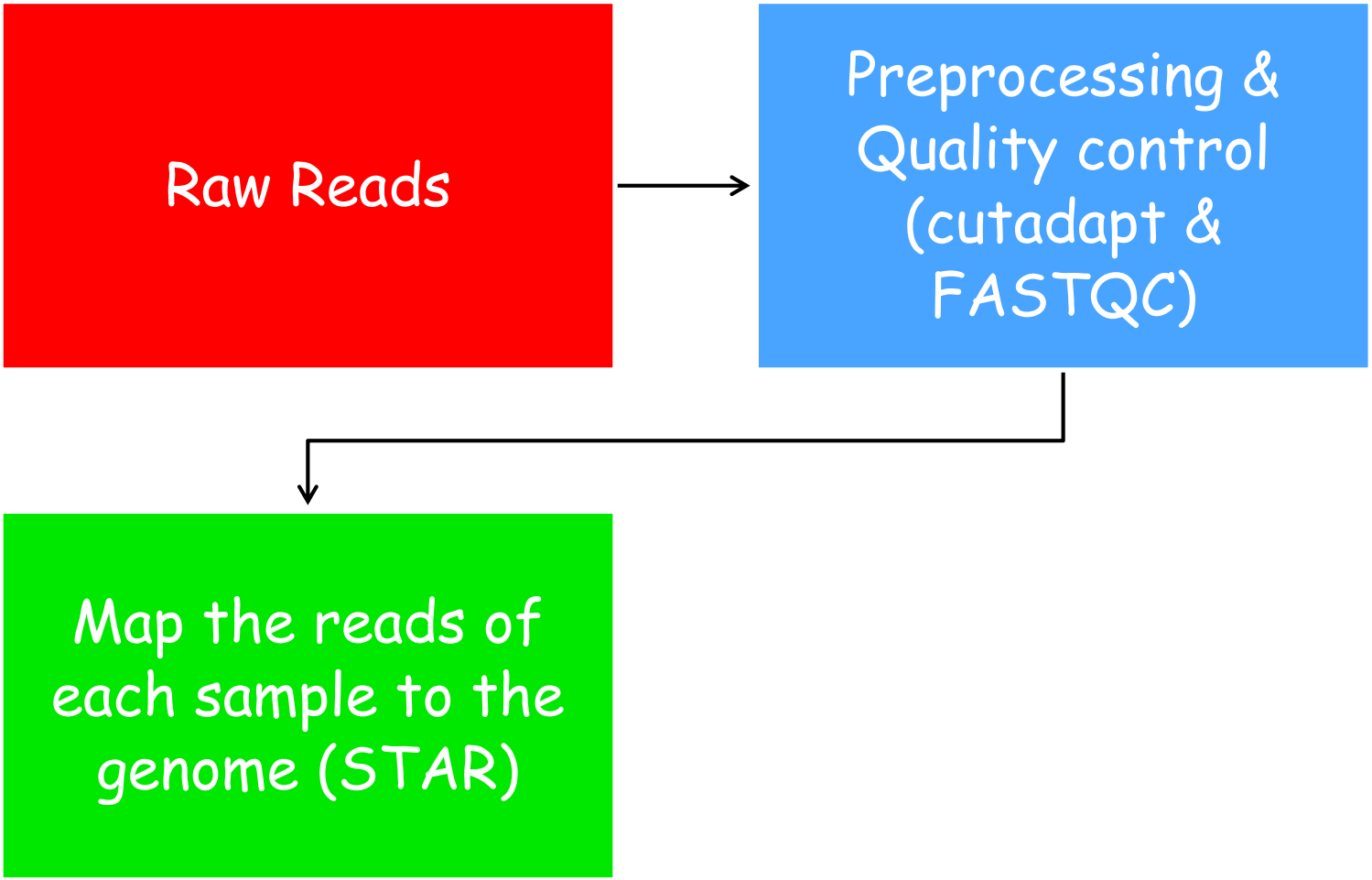
Pre-processing

- Recommendation is to use the high quality sequence data (is critical for de novo assembly), pre-processing includes:

cutadapt

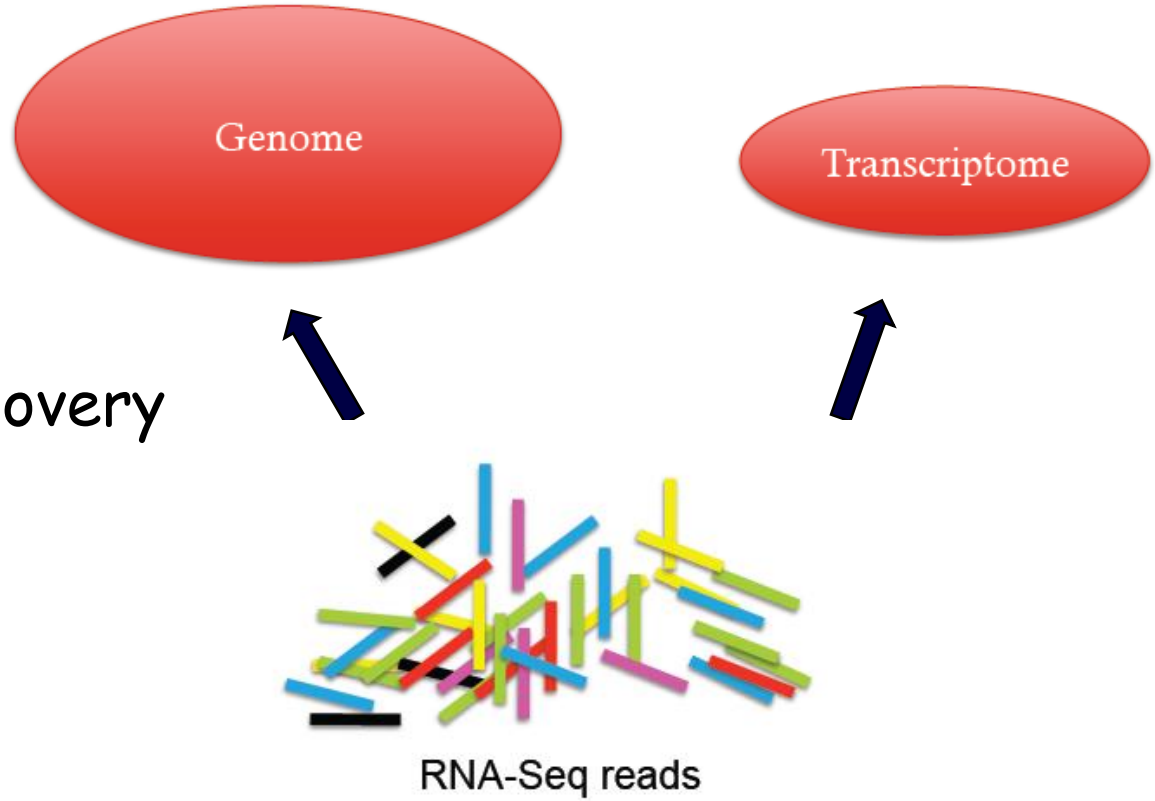
  - Trim sequences if: the end is of low quality, contain adapter or polyA or polyT
  - Filter low quality reads
  - Avoid using samples with insufficient number of reads

RNA-Seq Workflow



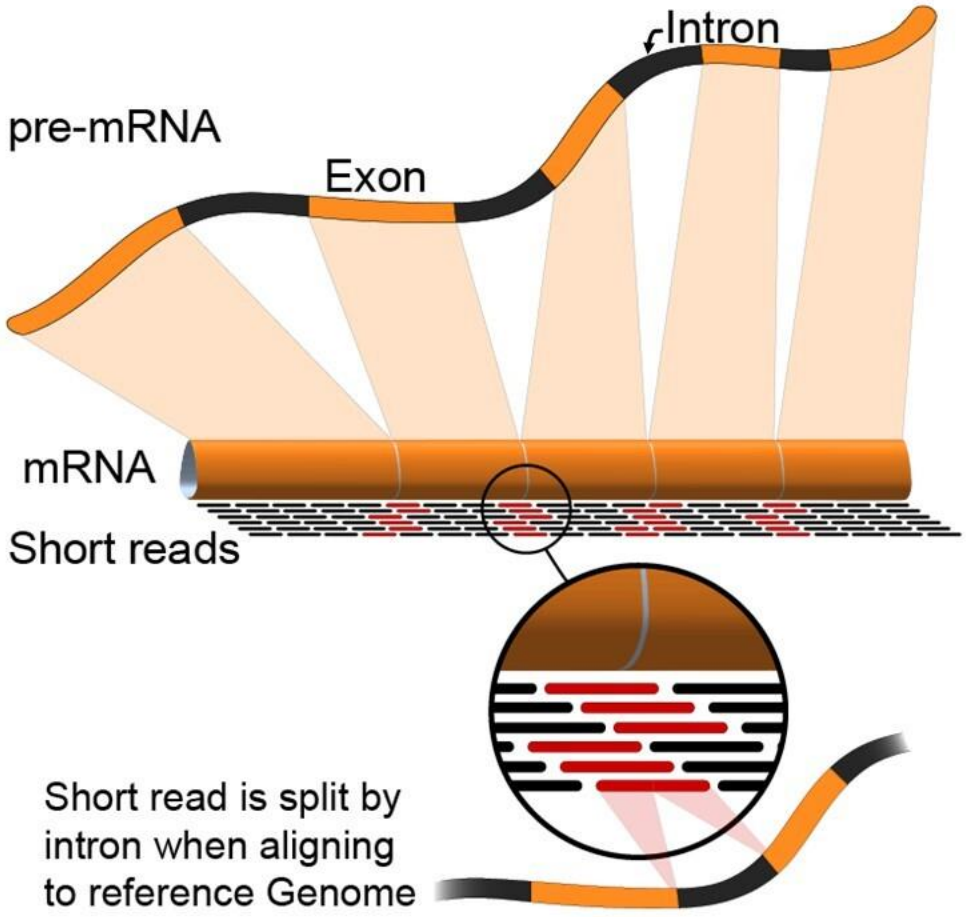
Mapping Short RNA-Seq Reads

Do I align the reads to the genome or to the transcriptome?



Mapping to Genome

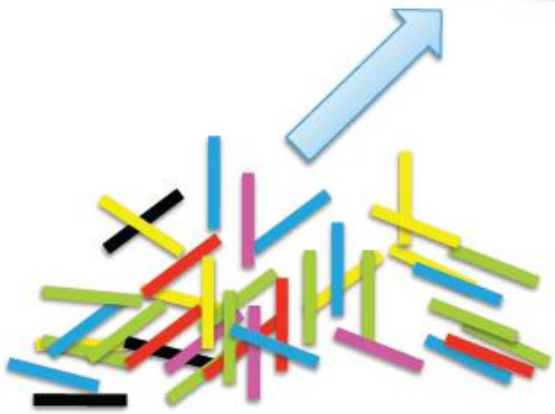
What is the challenge when mapping to the genome?



http://en.wikipedia.org/wiki/RNA-Seq

RNA-Seq mapping with TopHat

Goal: **identify** all transcripts and estimate relative amounts from RNA-Seq data



RNA-Seq reads

Bioinformatics. 2009 May 1; 25(9): 1105–1111. PMCID: PMC2672628
Published online 2009 March 16. doi: 10.1093/bioinformatics/btp120

TopHat: discovering splice junctions with RNA-Seq

Cole Trapnell,[1*] Lior Pachter,[2] and Steven L. Salzberg[1]

[1]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742 and [2]Department of Mathematics, University of California, Berkeley, CA 94720, USA
*To whom correspondence should be addressed.
Associate Editor: Ivo Hofacker

Genome Biology

METHOD Open Access

TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions

Daehwan Kim[1,2,3*], Geo Pertea[3], Cole Trapnell[5,6], Harold Pimentel[7], Ryan Kelley[8] and Steven L Salzberg[3,4]

Newer Aligners – Improving speed

Figure 2: Alignment speed of spliced alignment software for 20 million simulated 100-bp reads.

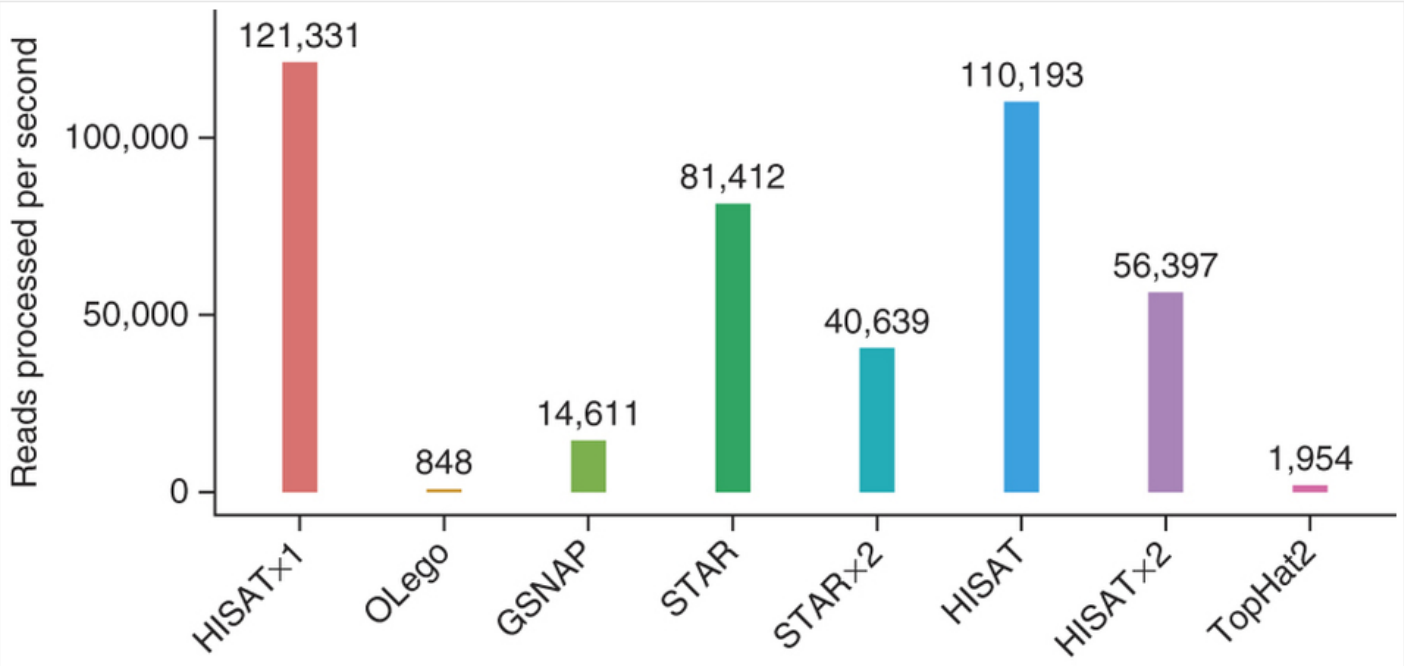
From

HISAT: a fast spliced aligner with low memory requirements

Daehwan Kim, Ben Langmead & Steven L Salzberg

Nature Methods **12**, 357–360 (2015) | doi:10.1038/nmeth.3317

Received 07 August 2014 | Accepted 16 January 2015 | Published online 09 March 2015



Alignment speed for all read types (defined in Fig. 1) combined, measured as the number of reads processed per second by the indicated tools. Supplementary Figure 2 provides the alignment speed for each type of read separately.

STAR-Spliced Transcripts Alignment to a Reference

- In the first step, the algorithm finds the Maximal Exact Match (*MMP*) starting from the first base of the read
- Next, the *MMP* search is repeated for the unmapped portion of the read
- Speed of search is achieved since the genome suffix array is not compressed and therefore requires **increased memory usage (30-60Gb)**
- Last stage is Clustering, stitching and scoring

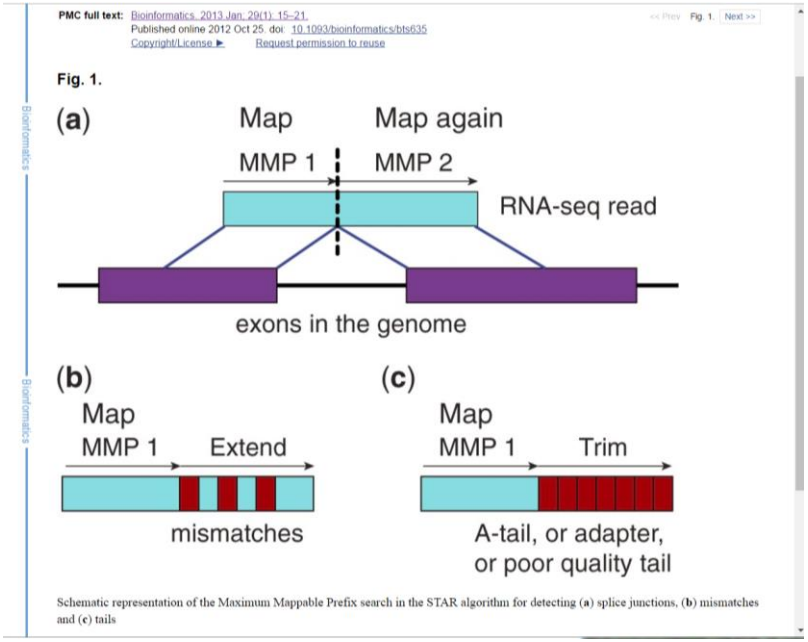
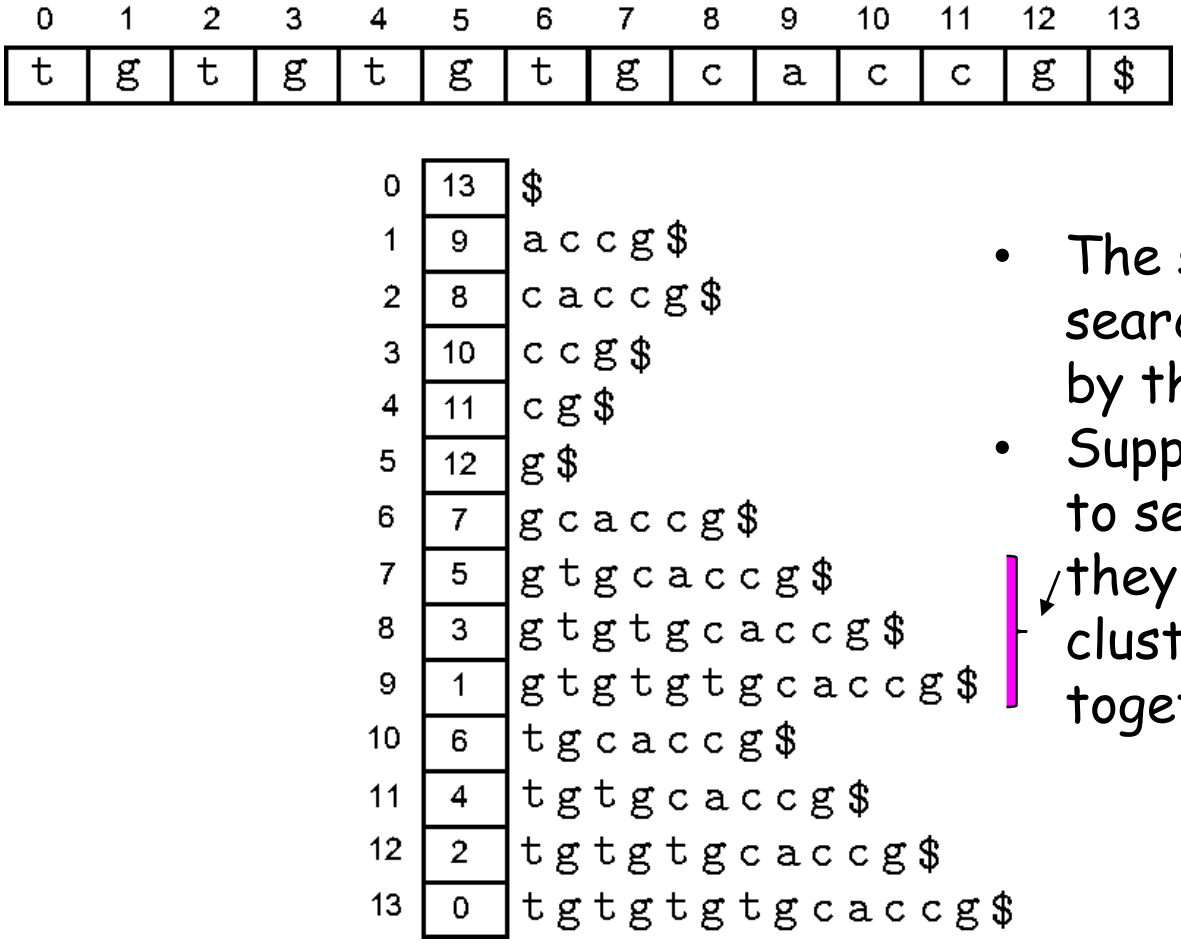


Figure 1: A string (above) and its suffix array (shown vertically) along with the position index on the left and the ...



Briefings in Bioinformatics, Volume 15, Issue 2, March 2014, Pages 138–154, https://doi.org/10.1093/bib/bbt081

Examples of Input and Output

Sequences – fastq

```
@HISEQ:226:C95PJANXX:2:1106:7378:57379 1:N:0:CGCTATGT
CCCGTCTCAAAAGGGAAACAGGTTGATATTCCTGTGCCAATAGTATTATGAGTTTCTTAGA
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGEECGGGGDDGGGGGGGGGGGGG
@HISEQ:226:C95PJANXX:2:1106:7467:57387 1:N:0:CGCTATGT
CTCAGGTGCGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAA
+
CCCCCGGGGGGGGBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@HISEQ:226:C95PJANXX:2:1106:7460:57432 1:N:0:CGCTATGT
CTTGCATTGTTATTTCTTGTCACTACCTCTCTTCTTTAGAATTGGGTAATTTACGCGCCTG
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@HISEQ:226:C95PJANXX:2:1106:7269:57438 1:N:0:CGCTATGT
GTGGGGAGTTTGACTGGGGCGGTACATCTGTTAAATATTAACGCAGATGTCCAAAGACAAG
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
```

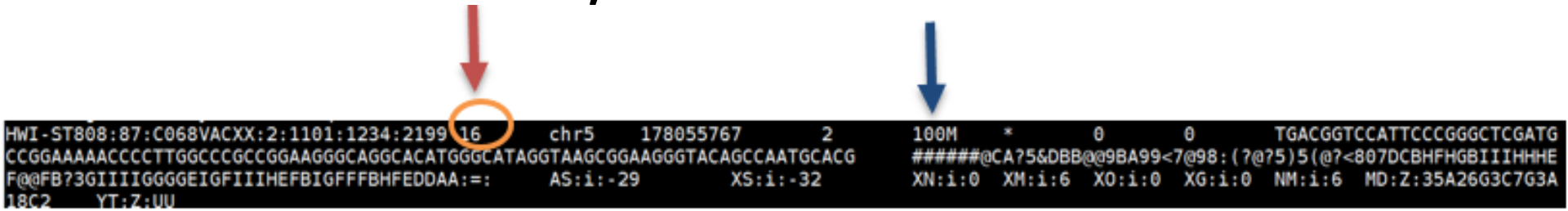
Mapping to genome

```
HISEQ:153:H8PNPADXX:1:1106:16824:79365  83      chr1    4839877 255     101M    =       4839864 -114
GGTAACTTTCTGTAGTTGGTAAGCCTACCAAGAATTCATGGTTACATCAGGCATCTTATTTTTAAGATATTTCTTTGAC DDEEDDDEDDDDBFFFFHHHHH
IGEJJJJJJJJJJJJJJIIGJIJJJIJJJJJJJJJJJJJJHHHHHFFFFFCCC NH:i:1  HI:i:1  AS:i:200        nM:i:0  MD:Z:1
```

Mapped Reads – bam format

Mapping Output - Alignment file

- SAM or binary BAM file



1.4 The alignment section: mandat

In the SAM format, each alignment line typicall
has 11 mandatory fields. These fields always appe
can be '0' or '*' (depending on the field) if the co
gives an overview of the mandatory fields in the

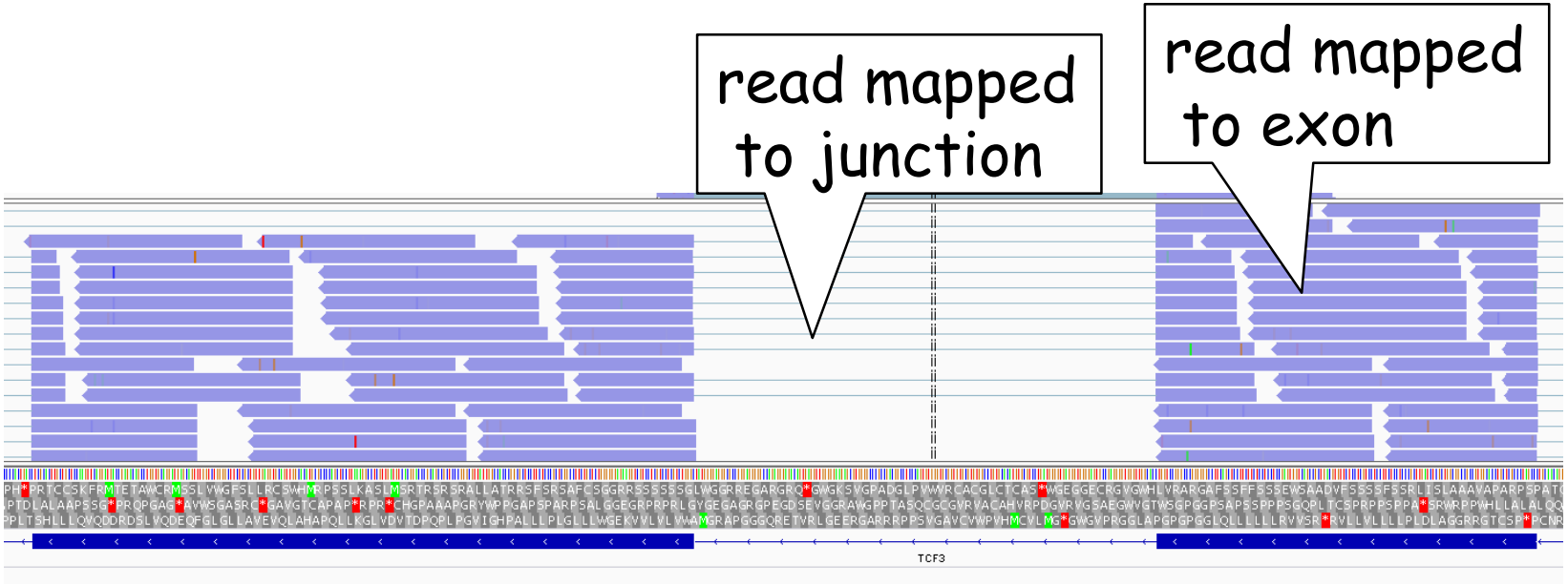
Col	Field	Type	Regexp/Range
1	QNAME	String	[!-?A-~]{1,254}
2	FLAG	Int	[0,2^{16}-1]
3	RNAME	String	*|[!-()+-<>-~]
4	POS	Int	[0,2^{31}-1]
5	MAPQ	Int	[0,2^{8}-1]
6	CIGAR	String	*|([0-9]+[MIDN
7	RNEXT	String	*|=|[!-()+-<>-
8	PNEXT	Int	[0,2^{31}-1]
9	TLEN	Int	[-2^{31}+1,2^{31}-1]
10	SEQ	String	*|[A-Za-z=.]+
11	QUAL	String	[!-~]+

understand the FLAG code of SAM format

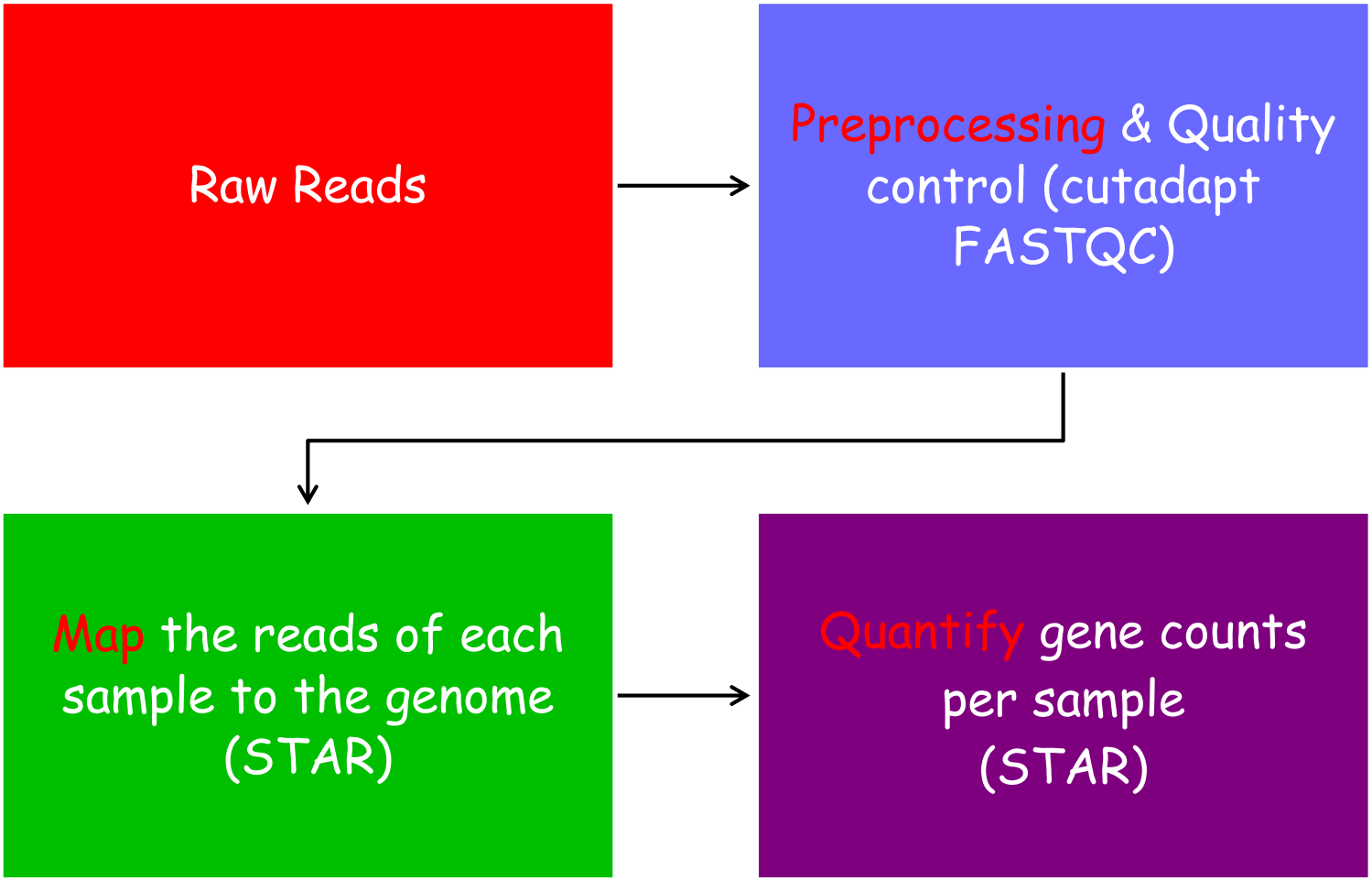
Here is what SAM specification stated for the FLAG column:

```
FLAG: bitwise FLAG. Each bit is explained in the following table:
Bit Description
0x1 template having multiple segments in sequencing (1)
0x2 each segment properly aligned according to the aligner (2)
0x4 segment unmapped (4)
0x8 next segment in the template unmapped (8)
0x10 SEQ being reverse complemented (16)
0x20 SEQ of the next segment in the template being reversed (32)
0x40 the rst segment in the template (64)
0x80 the last segment in the template (128)
0x100 secondary alignment (256)
0x200 not passing quality controls (512)
0x400 PCR or optical duplicate (1024)
```

Visualization of Bam outputs in a Genome Browser (IGV)



RNA-Seq Workflow



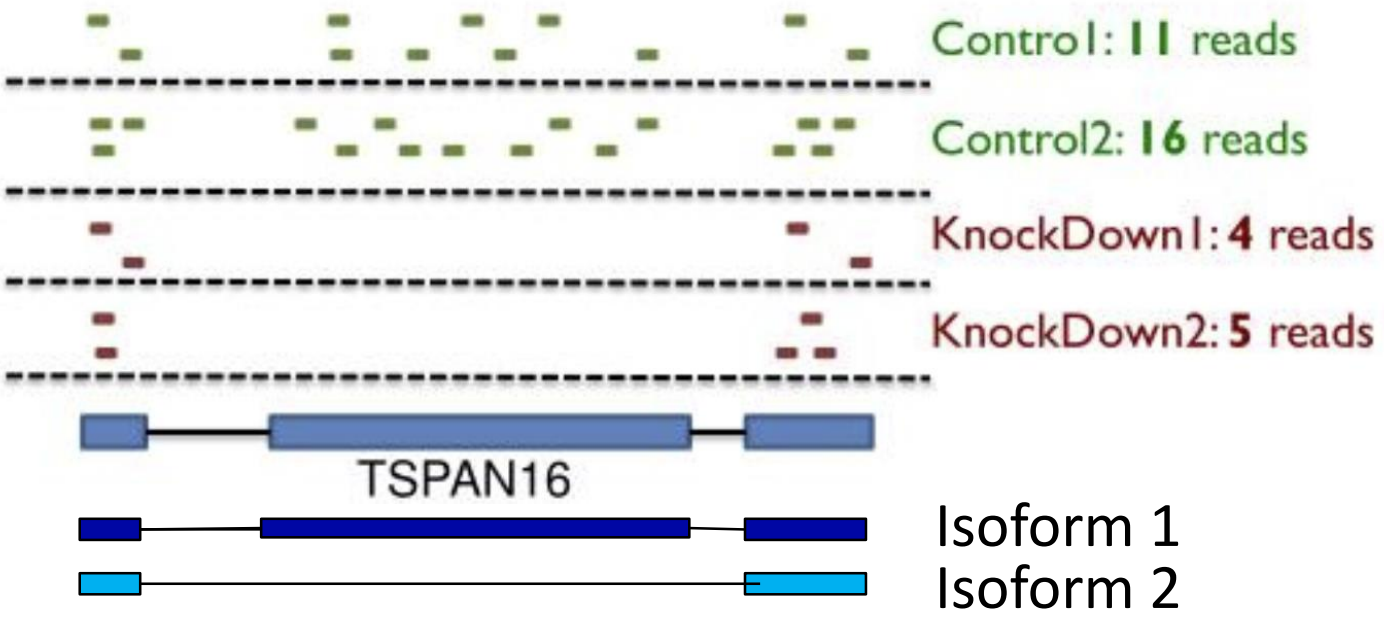
Gene Annotation File
Gene Transfer Format (GTF)

- GTF file is used to capture gene structure information (tab-delimited text file)
- Example:

```
chr1    unknown exon        3214482 3216968 .       -       .       gene_id "Xkr4"; gene_name "Xkr
chr1    unknown stop_codon  3216022 3216024 .       -       .       gene_id "Xkr4"; gene_n
chr1    unknown CDS         3216025 3216968 .       -       2       gene_id "Xkr4"; gene_name "Xkr
chr1    unknown CDS         3421702 3421901 .       -       1       gene_id "Xkr4"; gene_name "Xkr
chr1    unknown exon        3421702 3421901 .       -       .       gene_id "Xkr4"; gene_name "Xkr
chr1    unknown CDS         3670552 3671348 .       -       0       gene_id "Xkr4"; gene_name "Xkr
chr1    unknown exon        3670552 3671498 .       -       .       gene_id "Xkr4"; gene_name "Xkr
chr1    unknown start_codon 3671346 3671348 .       -       .       gene_id "Xkr4"; gene_n
chr1    unknown exon        4290846 4293012 .       -       .       gene_id "Rp1"; gene_name "Rp1"
chr1    unknown stop_codon  4292981 4292983 .       -       .       gene_id "Rp1"; gene_na
```

Gene Quantification

- A gene is quantified by counting the number of fragments/reads which align uniquely to all its exons.
- When performing gene level analysis we do not need to determine from which transcript the read was derived

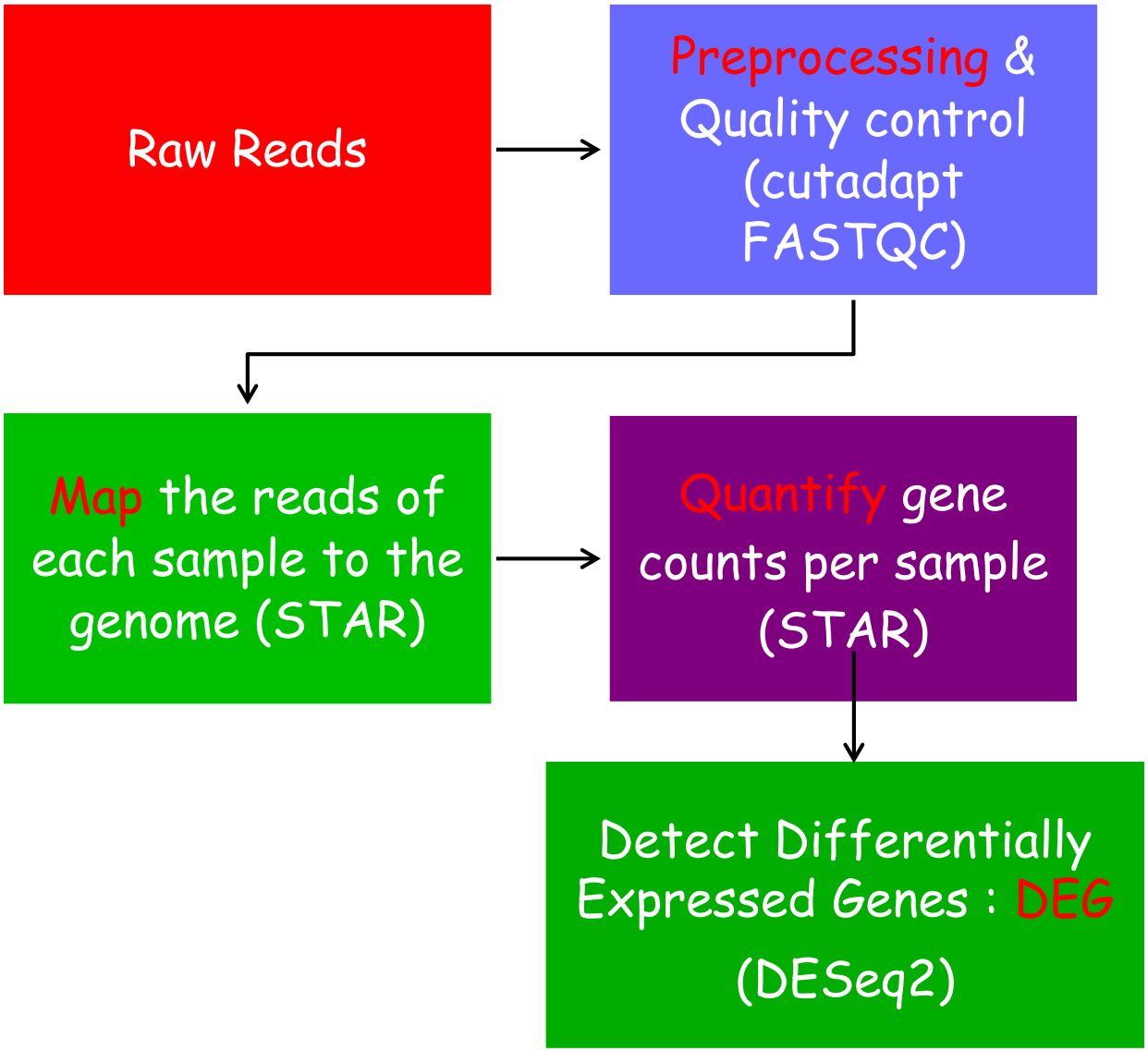


STAR/HTSeq Results- a count matrix

	sample_1	sample_2	sample_3	sample_4
gene_1	15	9	11	18
gene_2	19	21	21	40
gene_3	106	114	153	207
gene_4	569	565	756	992
gene_5	1029	1260	1559	1968
gene_6	5049	10029	7537	200
SUM	**10 M**	**30 M**	**20 M**	**10 M**

- Can we compare gene level expression using this matrix,for instance can we conclude the expression of gene_4 is similar in sample_1 and sample_2?

RNA-Seq Workflow



DESeq2 Normalization
median-of-ratios method

- Create a "virtual reference sample" by taking, for each gene, the geometric mean of counts over all samples
- Normalize each sample to this reference, to get one scaling factor ("size factor") per sample

Example-DESeq Normalization

	sample_1	sample_2	sample_3	sample_4	geometric mean		ratio sample_1
gene_1	15	9	11	18	12.79		15/12.79 = 1.17
gene_2	19	21	21	40	24.06		0.79
gene_3	106	114	153	207	139.87		0.76
gene_4	569	565	756	992	700.73		0.81
gene_5	1029	1260	1559	1968	1412.26		0.73
gene_6	5049	5897	7537	10029	6887.68		0.73
						Median	0.77

The scaling factor for sample 1 is 0.77

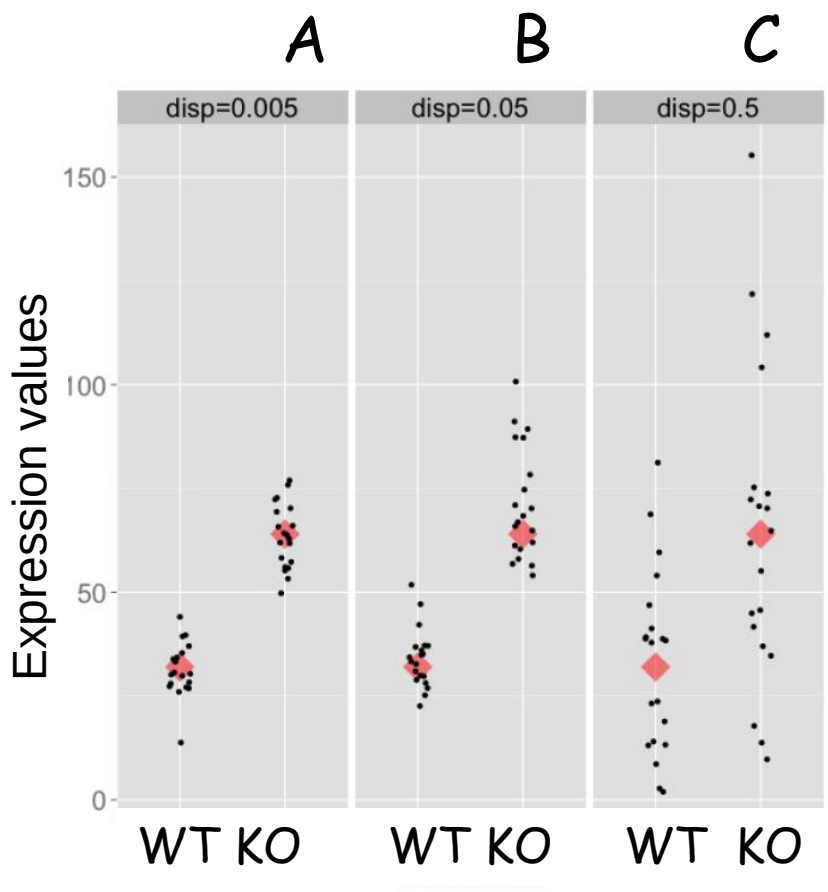
DESeq2 Normalization

Need to normalize the amount of sequence data between the samples

1. Geometric mean is calculated for each gene across all samples.
2. The counts for a gene in each sample is then divided by this mean.
3. The median of these ratios in a sample is the size factor for that sample.
4. The counts are divided by the sample-specific size factors

This procedure corrects for library size and RNA composition bias, which can arise for example when only a small number of genes are very highly expressed in one experiment condition but not in the other.

Determining Differentially Expressed Genes



http://www.molgen.mpg.de/1242892/rnaseq.pdf

- Aim: finding genes which have a significant difference between the groups which is larger than the "noise" – variation within the groups
- The advantage of having many replicates - allows us to learn about the biological variation within the groups tested
- In which of the groups is the gene expression difference between WT and KO more reliable?

Determining Differentially Expressed Genes (DEG)

- Our input are genes counts, i.e. discrete values
- In order to determine the DEG genes we need to module the data i.e. make assumptions on the statistical properties
- Incorrect assumptions can lead to poor false discovery rate (FDR) control and inaccurate true positive identification in the DEG calls.

RNA-Seq Noise

Suppose we sequence the same library twice to the same depth. For instance sequence it on two different lanes?

Which kind of replicates are these ?

Will we get the same gene counts?

RNA-Seq a Sampling Experiment

- A typical RNA library is estimated to have 2.408×10^{12} different molecules.
- If we sequence 30 million reads –this means 30M molecules are sampled.
- Our sample represents approximately 0.0013% of the total number of available molecules.
- It is therefore clear that when we sample twice we will observe a variance in the gene counts

RNA-Seq Noise

- In the case of sequencing the same library twice, since we have a large total number of reads and only a small fraction of reads mapping to each gene, the observed read counts for an individual gene can be well approximated by a Poisson distribution.
- Poisson distribution is sometimes called the law of small numbers because it is the probability distribution of the number of occurrences of an event that happens rarely but has very many opportunities to happen.

Poisson Distribution

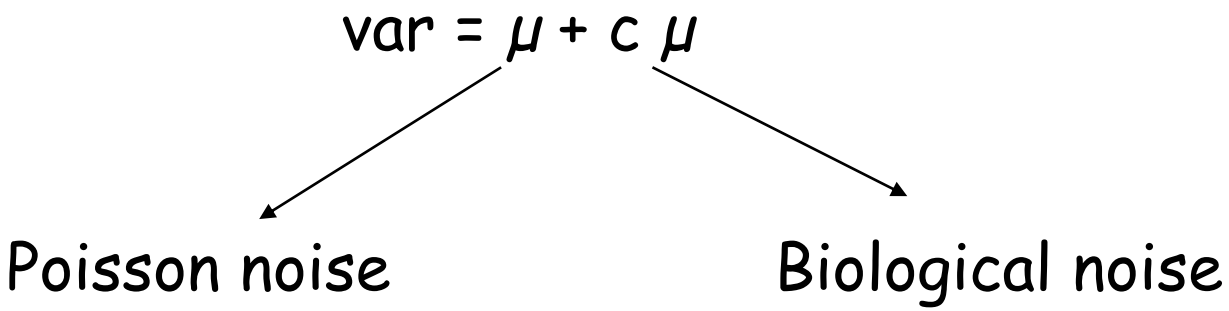
- Assuming the gene counts in a RNA-Seq experiment follow a Poisson distribution we would expect that the average gene count and the variance of the counts are equal.

  var = μ

- https://youtu.be/fxtB8c3u6l8
- https://youtu.be/HK7WKsL3c2w

Biological Variation

- When we sequence biological replicate samples the concentration of a given gene will vary around a mean value with a certain standard deviation
- This standard deviation needs to be to be estimated from the data, in the case of RNA-Seq we need to estimate it with a limited number of replicates



Negative Binomial

- In RNA-Seq analysis the negative binomial distribution is used as an alternative to the Poisson since it takes into account variance that exceeds the gene mean
- The count data is used to estimate the variance, using all genes with a similar expression value
- Orange line: the fitted observed curve for the variance

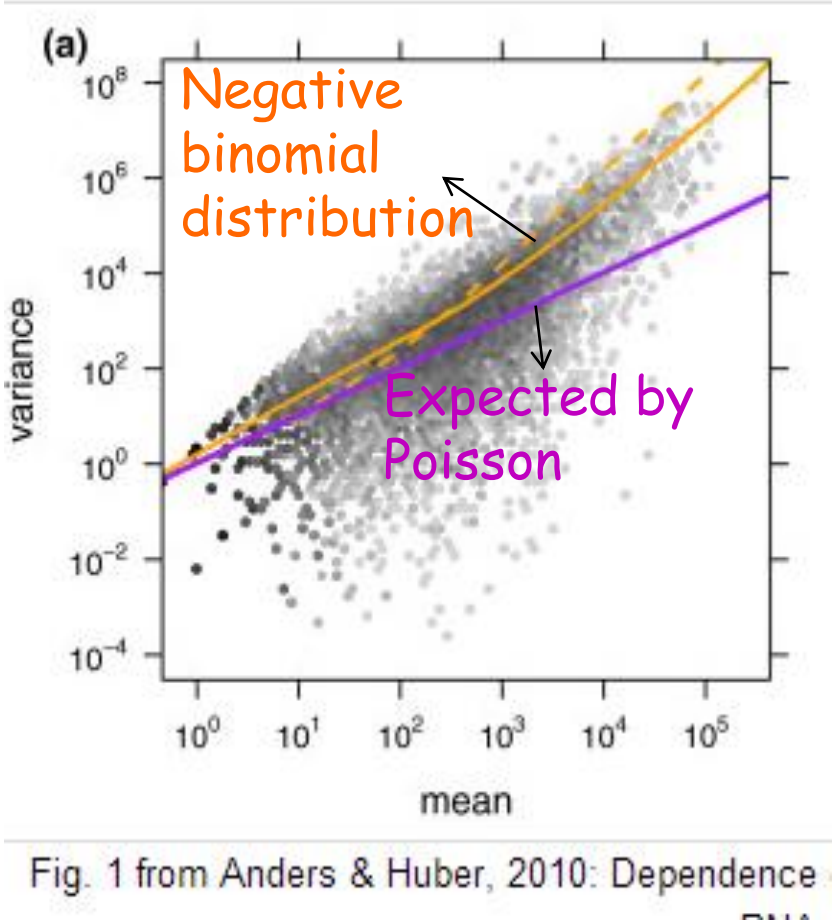
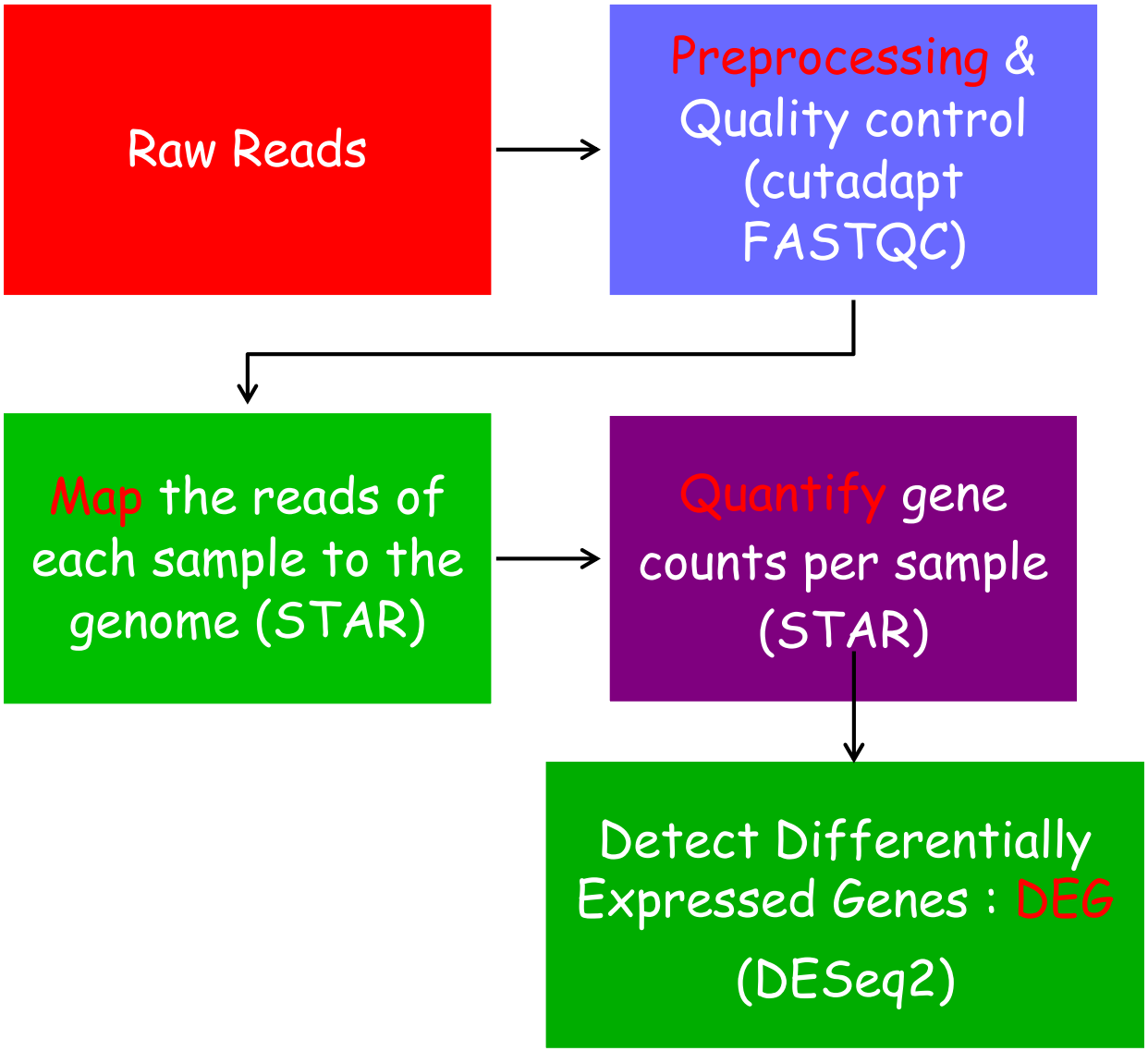


Fig. 1 from Anders & Huber, 2010: Dependence

Detecting Differentially Expressed Genes

- DESeq2 tests for differential expression by the use of negative binomial generalized linear models
- The output consists of:
  - Log fold change (treatment/control)
  - p-value - indicates the probability that the observed difference between treatment and control will be observed even though the there is no true treatment effect
  - Adjusted p value – multiple test correction
    - In the RNA-Seq study we simultaneously tested all genes

RNA-Seq Workflow



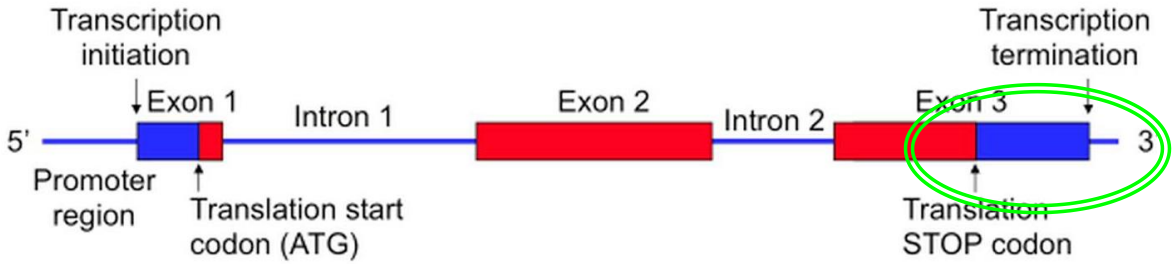
Agenda

- Introduction & Experimental design
- Analysing Gene expression from RNA-Seq data
- Analysing Gene expression from bulk MARS-Seq data

MARS-Seq

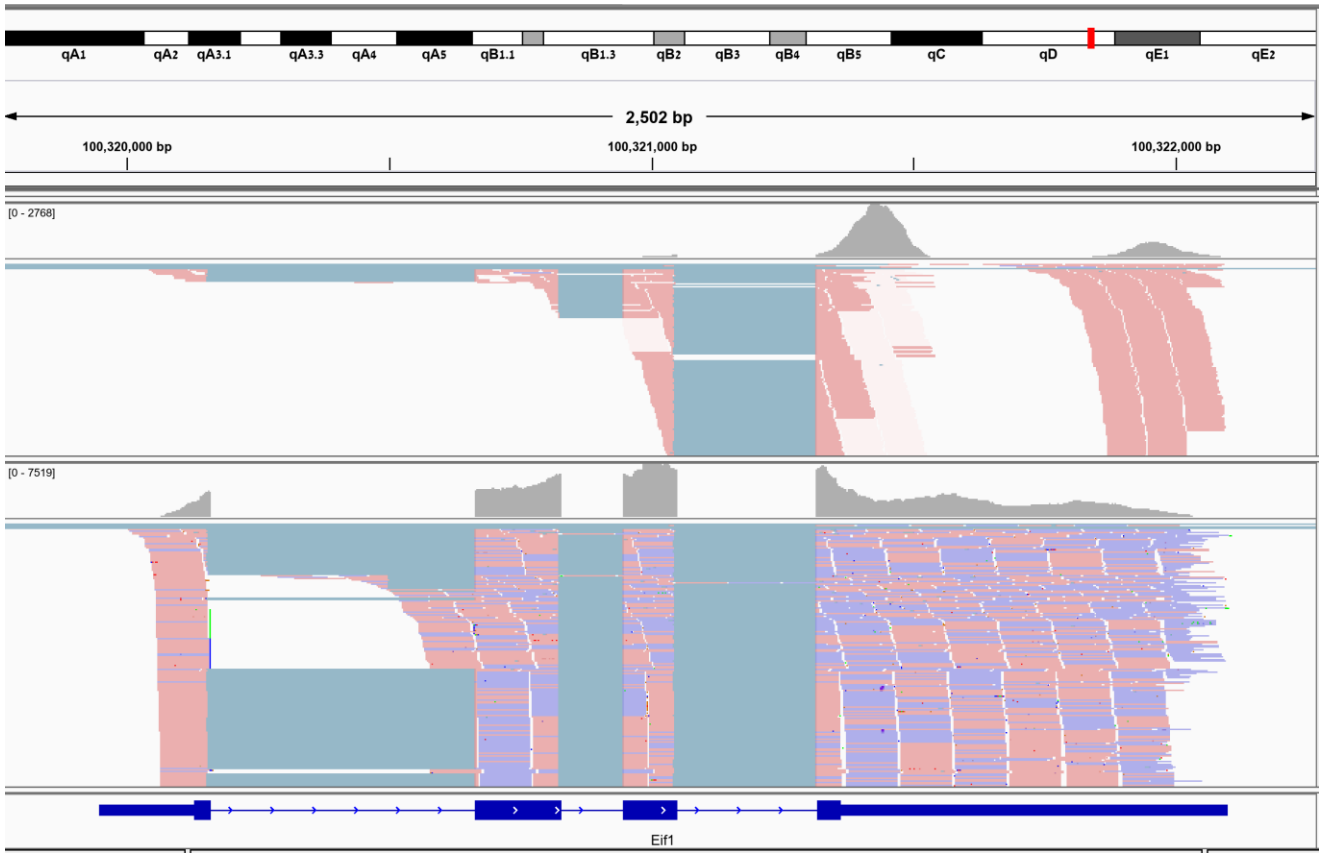
Differences between RNA-Seq and bulk MARS-Seq:

- Library generated contains only 3' end of the transcripts



- Low input material - UMI (Unique Molecular Identifier)

Genome Browser View

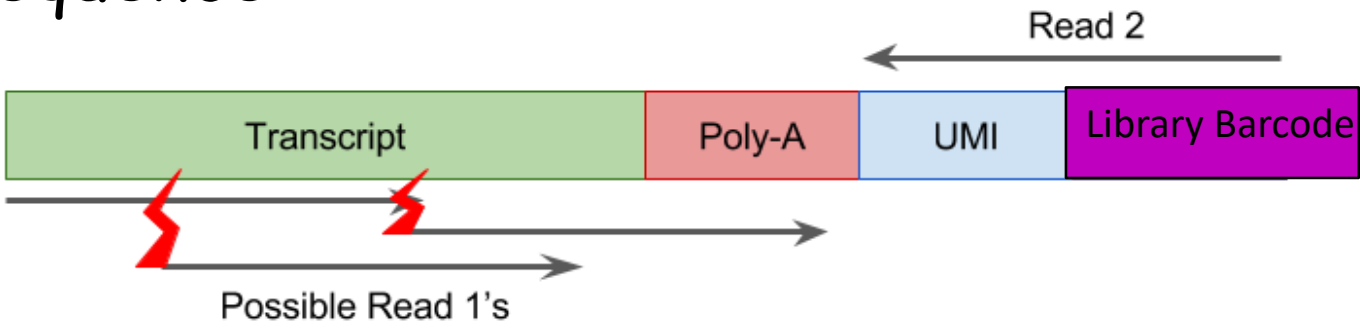


MARS-Seq

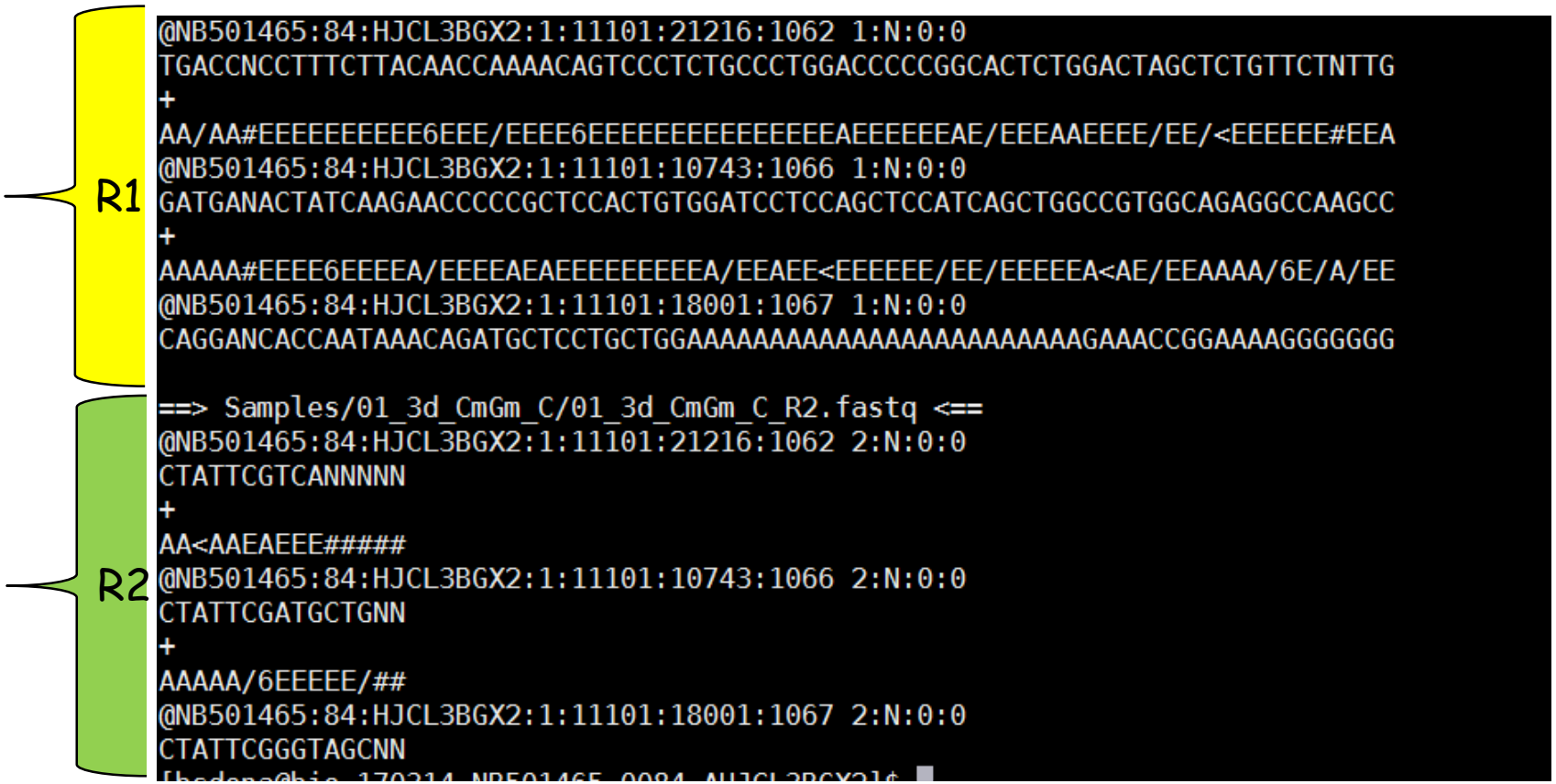
RNA-Seq

MARS-Seq – Paired end sequencing

- Read 1 (R1) contains 3′ cDNA insert sequence
- Read2 (R2) contains the library barcode and the UMI (Unique Molecular Identifier) sequence

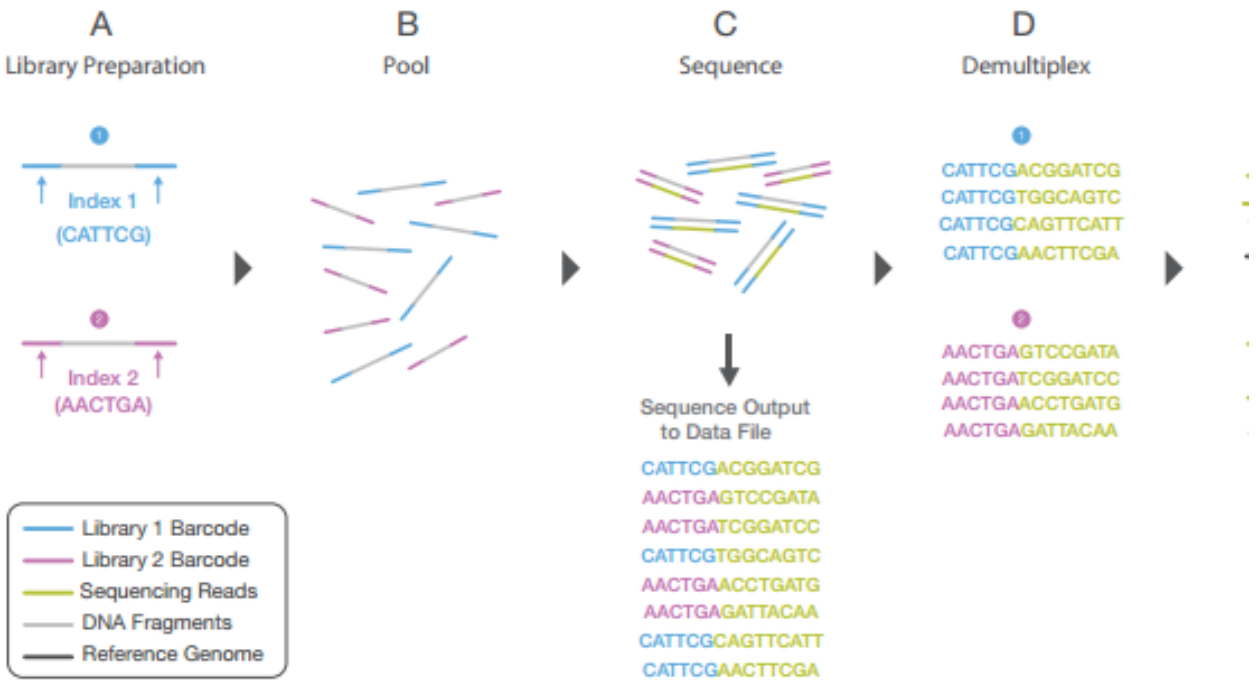


MARS-Seq FASTQ Outputs



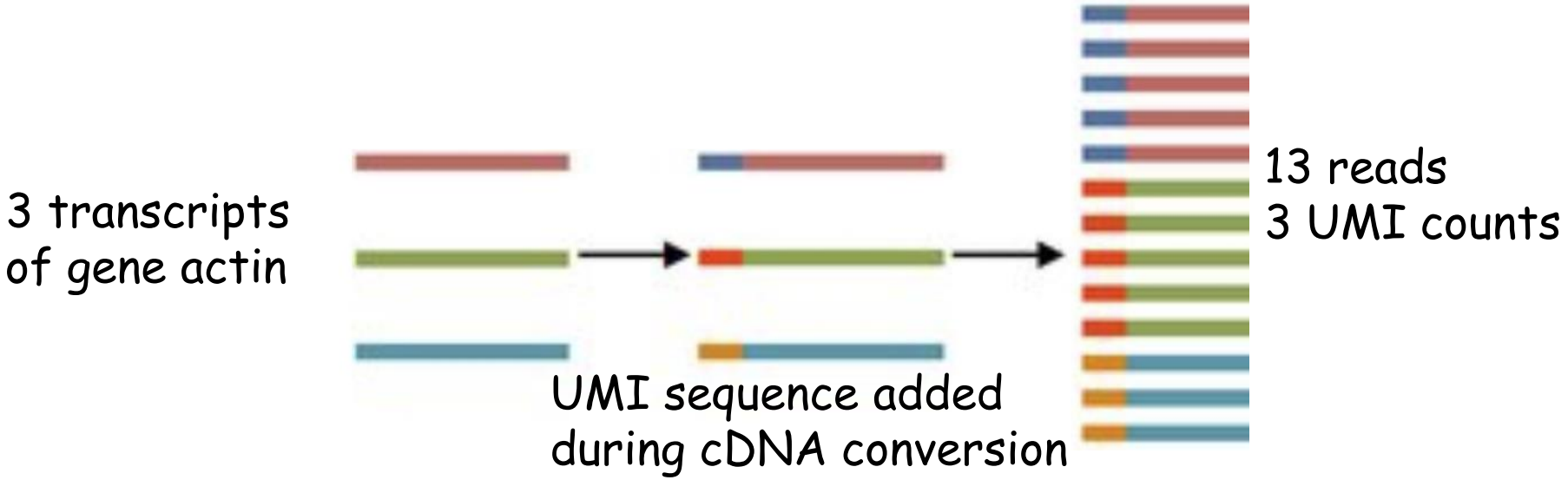
Library Barcode

- Multiplexing: the process of pooling samples together and sequencing them simultaneously
- Demultiplexing: separating reads using the library barcode to identify the origin sample

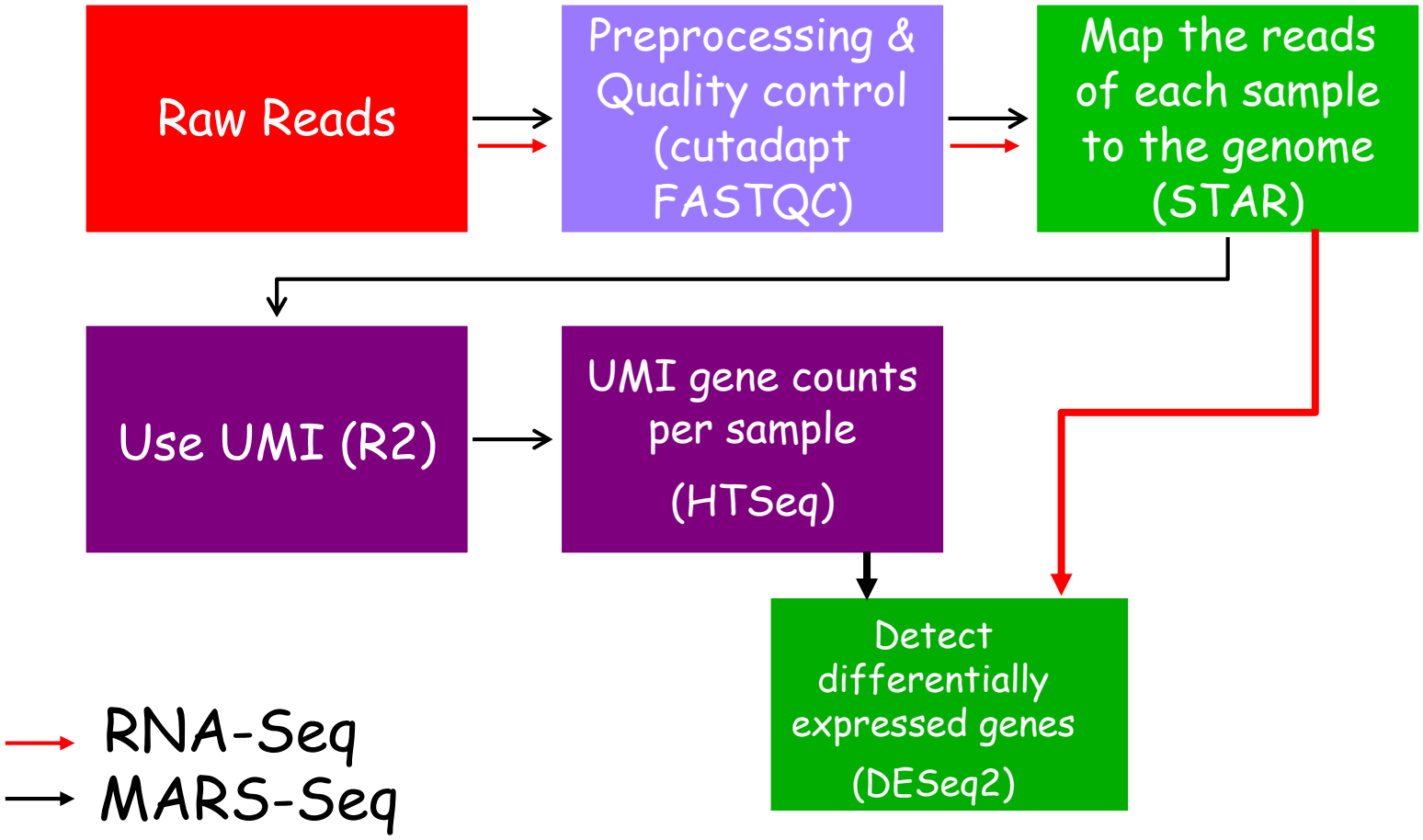


UMI – Unique Molecular Identifier

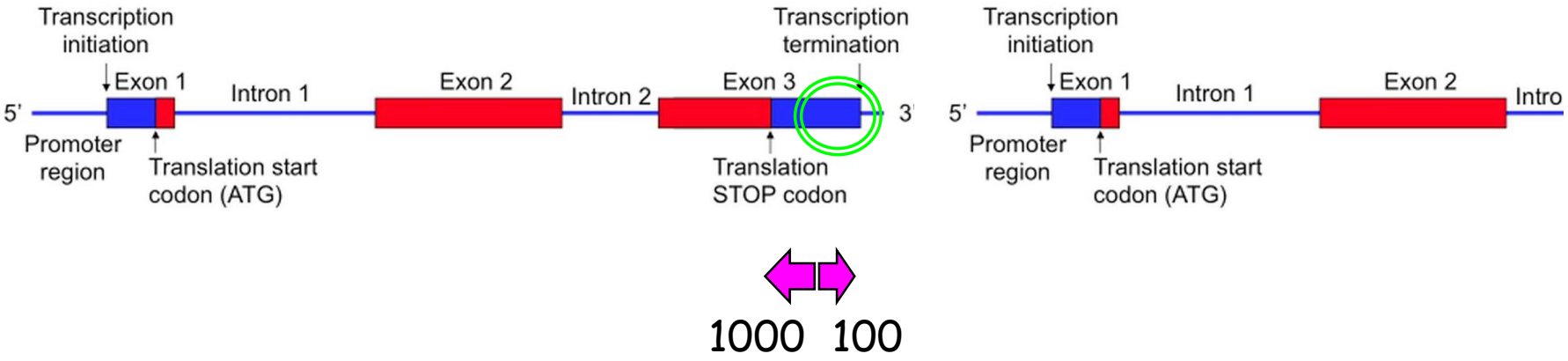
- The 8 base UMI is used as an identifier of a specific transcript molecule
- 4^8 = 65536 theoretical possibilities
- Reads are considered PCR duplicated, if they map to the same gene and have the same UMI
- Instead of counting reads we will count number of unique UMIs per gene



Bioinformatics Workflow



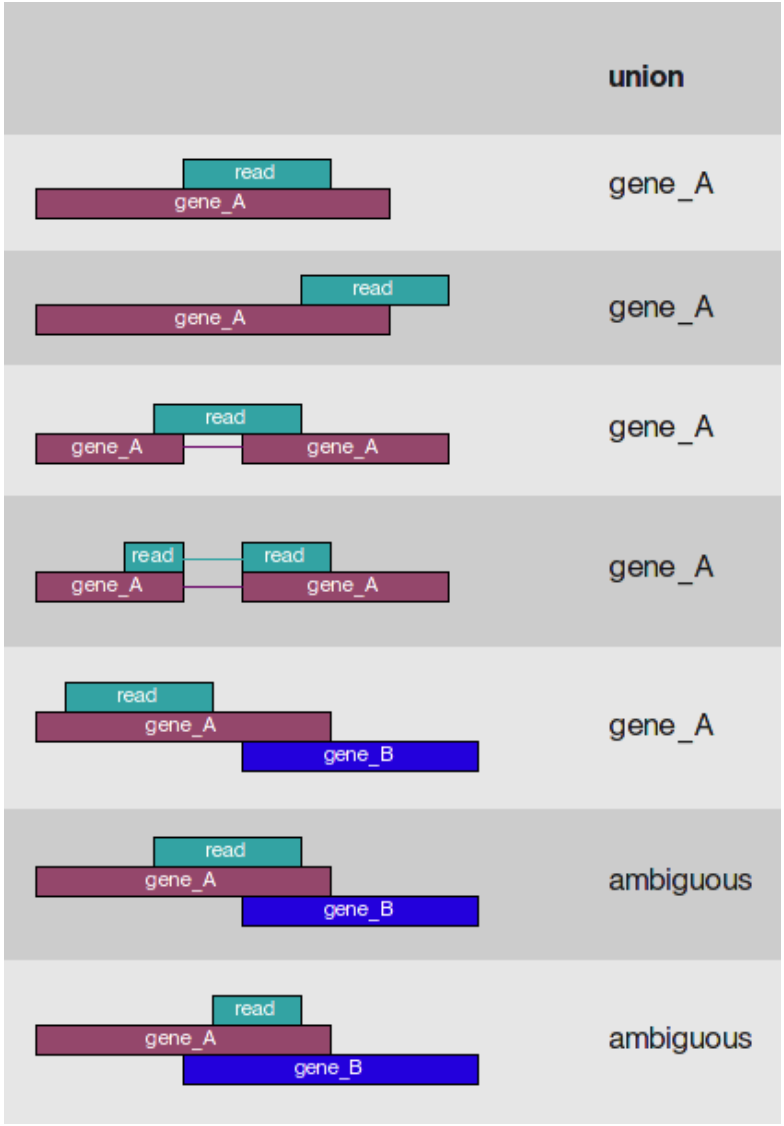
MARS-Seq: Sequencing the transcripts 3'



- Count reads overlapping exons at a limited window in the 3' end of the transcript, using annotation of Genecode or RefSeq.
- Can be used for organisms with a well annotated geneome (human, mouse & Arabidopsis)

MARS-Seq Gene Quantification

- HTSeq criteria to count reads: uniquely mapped to the 3' of gene (using a modified annotation file - gtf)



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

In BAM File Add Feature Tag and Mark Duplicates

UMI barcode

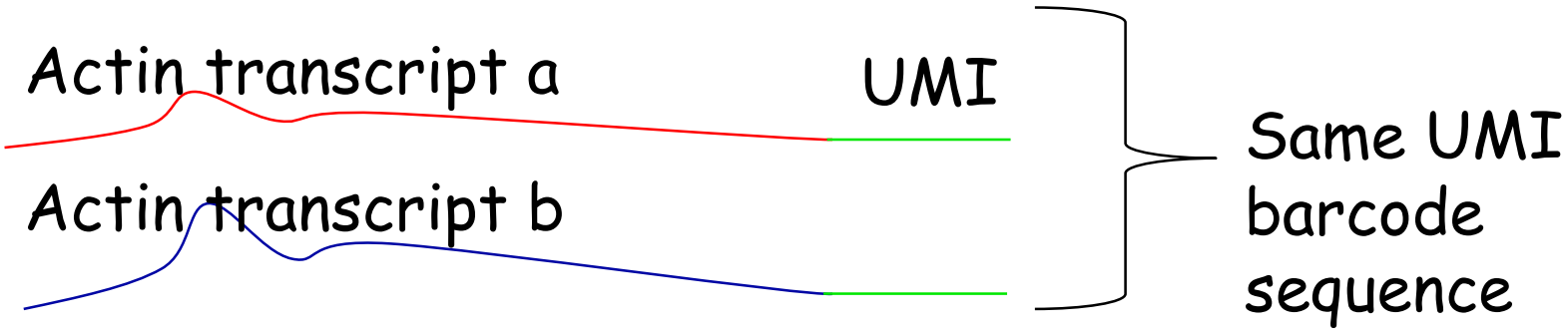
Mark OK

```
K00279:102:HJK2GBBXX:1:1104:3995:7222  0     chr1    164437130       255     101M    *       0       0
GTCCGTTGTGTTAACCTGACTGACACAAGATCAACAAGAGCACTGTACTCCTGGCAATTATTACATATGCTAGAACATGGATTTGCACTGTAGACAACACT
-AFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFJJFJJFJJFFJ   NH:i:1  HI:i:1  AS:i:99 nM:i:0  RX:Z:ATTTTTTT
QX:Z:JJJJJJJJ   XF:Z:Nme7       rp:Z:K00279:102:HJK2GBBXX:1:1104:3995:7222      nd:i:8
```

Mark AS DUPLICATE

```
K00279:102:HJK2GBBXX:1:1102:23399:31734 1024    chr1    164437180       255     100M    *       0       0
CTGGCAATTATTACATATGCTAGAACATGGATTTGCACTGTAGACAACACTTAACACCAGTCTATGGGGTACTGCATTGCTTTTTATAAAGTTCAAAATA
<A-FFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ   NH:i:1  HI:i:1  AS:i:98 nM:i:0  RX:Z:ATTTTTTT
QX:Z:JJJJJJJJ   XF:Z:Nme7       rp:Z:K00279:102:HJK2GBBXX:1:1104:3995:7222      nd:i:8
K00279:102:HJK2GBBXX:1:1102:30919:42724 1024    chr1    164437198       255     101M    *       0       0
GCTAGAACATGGATTTGCACTGTAGACAACACTTAACACCAGTCTATGGGGTACTGCATTGCTTTTTATAAAGTTCAAAATAAAGATTTTTATTTTCAAAC
AAAFFJJJJJJJJJJJJJJJJJJJJFJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJFJJAJJJFJFJJJJJF<JFJJJJFJJJJJJJJ   NH:i:1  HI:i:1  AS:i:99 nM:i:0  RX:Z:ATTTTTTT
QX:Z:JJJJJJJJ   XF:Z:Nme7       rp:Z:K00279:102:HJK2GBBXX:1:1104:3995:7222      nd:i:8
K00279:102:HJK2GBBXX:1:1103:31994:17438 1024    chr1    164437199       255     100M    *       0       0
CTAGAACATGGATTTGCACTGTAGACAACACTTAACACCAGTCTATGGGGTACTGCATTGCTTTTTATAAAGTTCAAAATAAAGATTTTTATTTTCAAAC
AAFAFJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJFFJJJJJJJJJJJFJJJF<JJFJJJAF7<-AFFF77A<FJJF-7FJJJJJJJJJ<FFJF   NH:i:1  HI:i:1  AS:i:98 nM:i:0  RX:Z:ATTTTTTT
QX:Z:JJJJJJJJ   XF:Z:Nme7       rp:Z:K00279:102:HJK2GBBXX:1:1104:3995:7222      nd:i:8
K00279:102:HJK2GBBXX:1:1105:22343:7468  1024    chr1    164437210       255     98M     *       0       0
ATTTGCACTGTAGACAACACTTAACACCAGTCTATGGGGTACTGCATTGCTTTTTATAAAGTTCAAAATACAGATTTTTATTTTCAAACAAGTGACTT
AA<A-<A7A-FJ<FFJA-FJFAFAJJJ-FA-7---F--A--77F-77<FFFJ<7--JF7FF<-<FJA7<--<A<-<7AF-FA-7FJJFAAF7-7F7FA   NH:i:1  HI:i:1  AS:i:94 nM:i:1  RX:Z:ATTTTTTT
QX:Z:JJFJFJJJ   XF:Z:Nme7       rp:Z:K00279:102:HJK2GBBXX:1:1104:3995:7222      nd:i:8
K00279:102:HJK2GBBXX:1:1102:4564:36200  1024    chr1    164437211       255     101M    *       0       0
TTTGCACTGTAGACAACACTTAACACCAGTCTATGGGGTACTGCATTGCTTTTTATAAAGTTCAAAATAAAGATTTTTATTTTCAAACAAGTGACTTTGGT
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ   NH:i:1  HI:i:1  AS:i:99 nM:i:0  RX:Z:ATTTTTTT
QX:Z:JJJJJJJJ   XF:Z:Nme7       rp:Z:K00279:102:HJK2GBBXX:1:1104:3995:7222      nd:i:8
K00279:102:HJK2GBBXX:1:1103:13869:49089 1024    chr1    164437214       255     101M    *       0       0
GCACTGTAGACAACACTTAACACCAGTCTATGGGGTACTGCATTGCTTTTTATAAAGTTCAAAATAAAGATTTTTATTTTCAAACAAGTGACTTTGGTTTT
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJFJFJFF   NH:i:1  HI:i:1  AS:i:99 nM:i:0  RX:Z:ATTTTTTT
QX:Z:JJJJJJJJ   XF:Z:Nme7       rp:Z:K00279:102:HJK2GBBXX:1:1104:3995:7222      nd:i:8
K00279:102:HJK2GBBXX:1:1104:13362:2563  1024    chr1    164437214       255     101M    *       0       0
GCACTGTAGACAACACTTAACACCAGTCTATGGGGTACTGCATTGCTTTTTATAAAGTTCAAAATAAAGATTTTTATTTTCAAACAAGTGACTTTGGTTTT
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJAJJJJJFJJJJJJJJJJJFFAJFAFFJFAFF   NH:i:1  HI:i:1  AS:i:99 nM:i:0  RX:Z:ATTTTTTT
QX:Z:JJJJJJJJ   XF:Z:Nme7       rp:Z:K00279:102:HJK2GBBXX:1:1104:3995:7222      nd:i:8
```

UMI Count Correction



- UMI barcodes are connected to cDNA randomly, so it might happen that two independent transcripts derived from the same gene get assigned the same UMI barcode -> clash
- Genes that are highly gene expressed, have a higher chance of clashing
- Correction is applied to UMI counts taking into account the chance of clashing

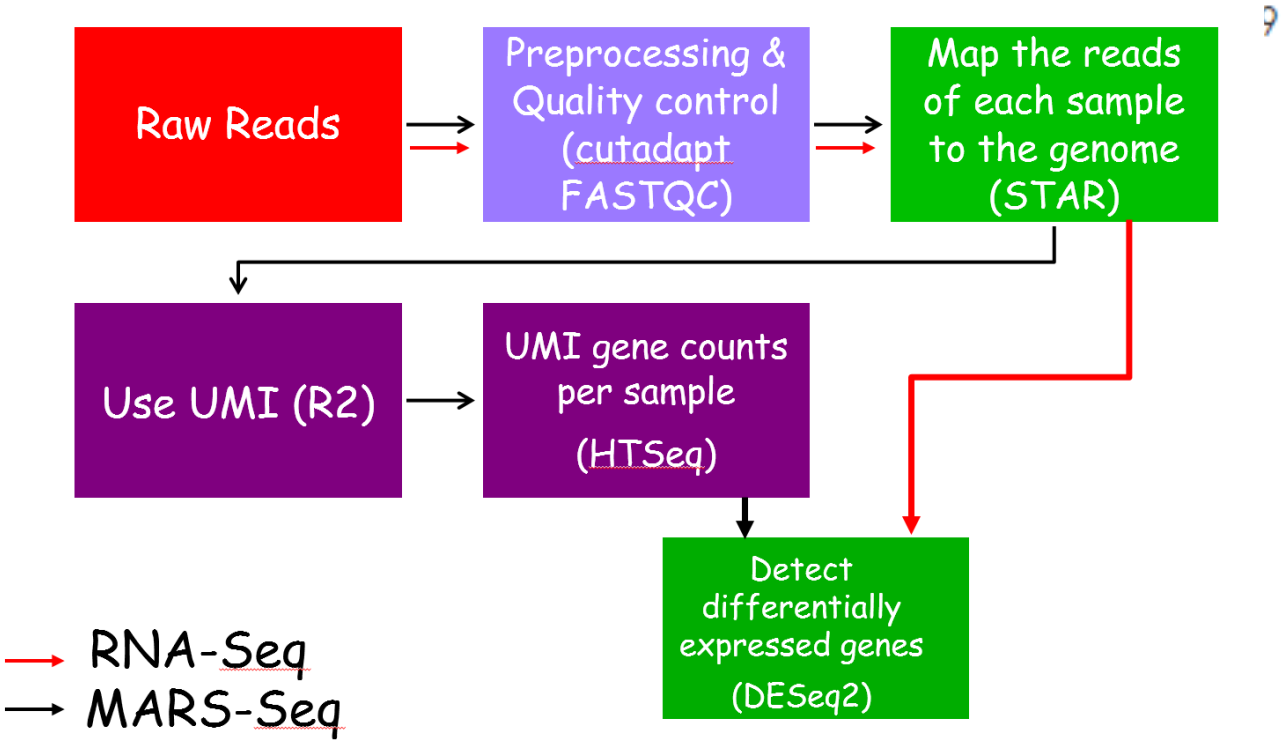
Differences Between Bulk MARS-Seq & RNA-Seq

	MARS-Seq	RNA-Seq
Gene coverage	3′	The whole transcript
Annotation file (GTF) RefSeq/Genecode	Modified GTF that contains the 3′ of the transcripts	Full length transcripts
#READS per sample	5M	20M
Sequencing protocol	PE	SE or PE
Location of library index	R2	Illumina index (i5 & i7)
Location of UMI barcode	R2	No UMI

UTAP: User-friendly Transcriptome Analysis Pipeline

Refael Kohen, Jonathan Barlev, Gil Hornung, Gil Stelzer, Ester Feldmesser, Kiril Kogan, Marilyn Safran and Dena Leshkowitz

BMC Bioinformatics 2019 **20**:154
https://doi.org/10.1186/s12859-019-2728-2



Tools References

1. Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc. 2013;8(9):1765-86. doi: 10.1038/nprot.2013.099. PubMed PMID: 23975260. (DESeq2)
2. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105-11. doi: 10.1093/bioinformatics/btp120. PubMed PMID: 19289445; PubMed Central PMCID: PMCPMC2672628.
3. Anders S, Pyl PT, Huber W. HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166-9. doi: 10.1093/bioinformatics/btu638. PubMed PMID: 25260700.
4. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635.
5. MARS-Seq: Jaitin D. A., Kenigsberg E., Keren-Shaul H., et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science. 2014;343:776-779

מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

LIFE SCIENCE
CORE FACILITIES

Come and consult us **before** performing your experiment

Bioinformatics unit
NGS data analysis

Noa Wigoda

Ester Feldmesser

Dena Leshkowitz
Unit head

Bareket Dassa

Gil Stelzer

Contact us:
Dena.Leshkowitz@weizmann.ac.il

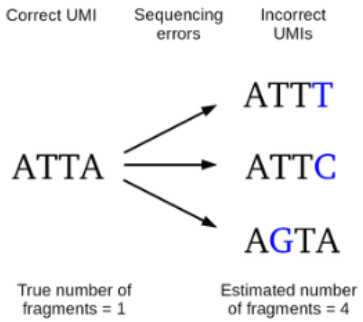
https://utap.wexac.weizmann.ac.il
e-learning tool

The END

THANKS FOR LISTENING

QUESTIONS?

Correction For Barcode Error Currently Not Implemented in UTAP



Sequencing errors inflate the apparent numbers of unique fragments sequenced

https://cgatoxford.wordpress.com/2015/08/14/unique-molecular-identifiers-the-problem-the-solution-and-the-proof/