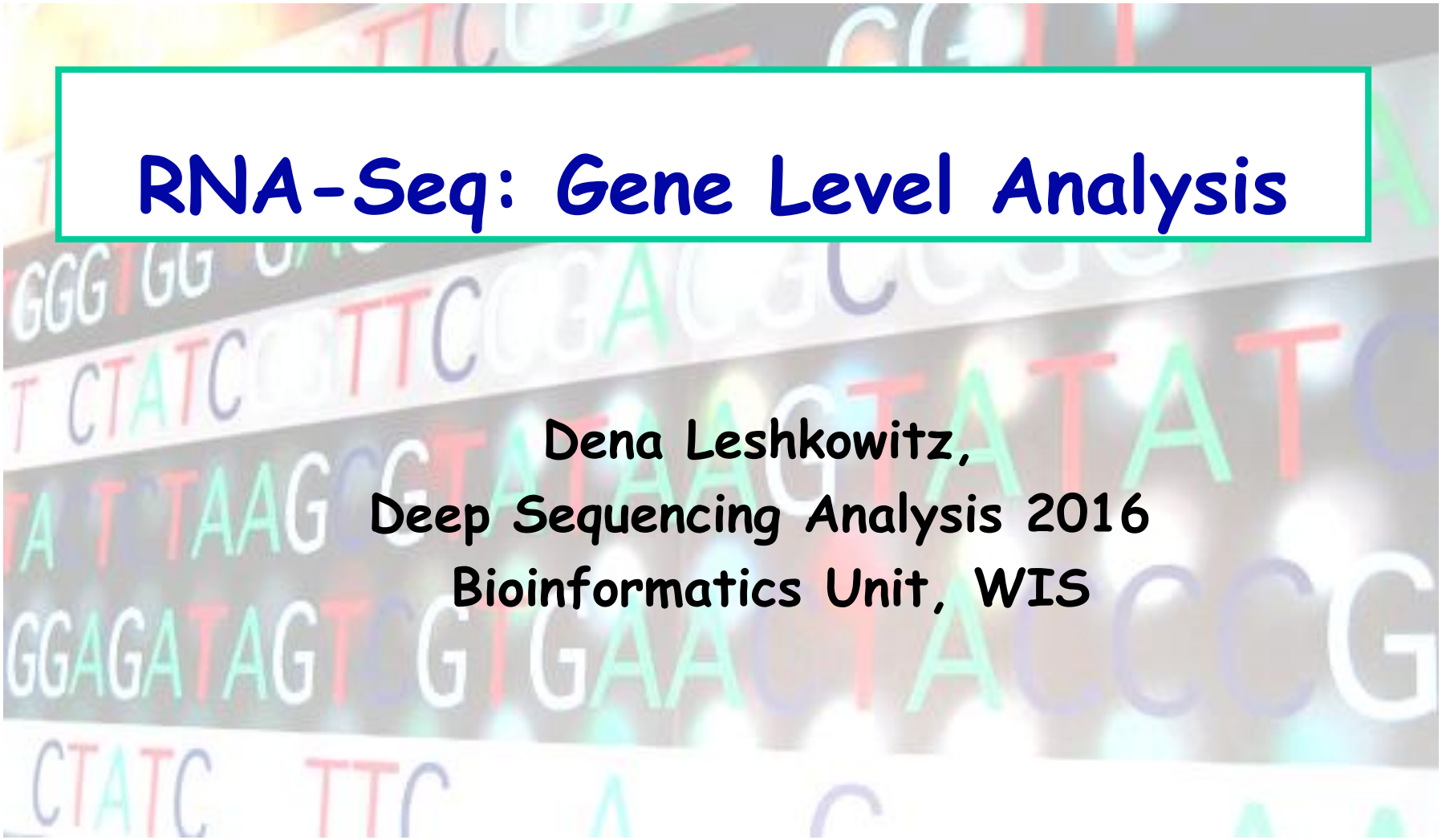
מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE



RNA-Seq: Gene Level Analysis

Dena Leshkowitz,
Deep Sequencing Analysis 2016
Bioinformatics Unit, WIS

RNA-Seq Potential

RNA-Seq: a revolutionary tool for transcriptomics

In theory RNA-Seq can be used to built 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

RNA-Seq: a revolutionary tool for transcriptomics

- Discover novel transcripts
- Determine transcript structure, measure transcripts expression and detect differentially expressed transcripts/isoforms between conditions, treatments…

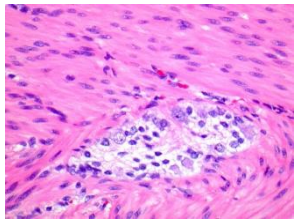
- Measure gene expression and detect differentially expressed genes between conditions, treatments… based on known gene structures (model organisms)

Main Topics

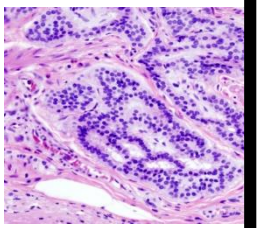
- Introduction
- Experimental design issues
- Analysing RNA-Seq data
  - RNA-Seq pipeline: Tophat- HTSeq-DESeq2
- Chipster

RNA-Seq Workflow

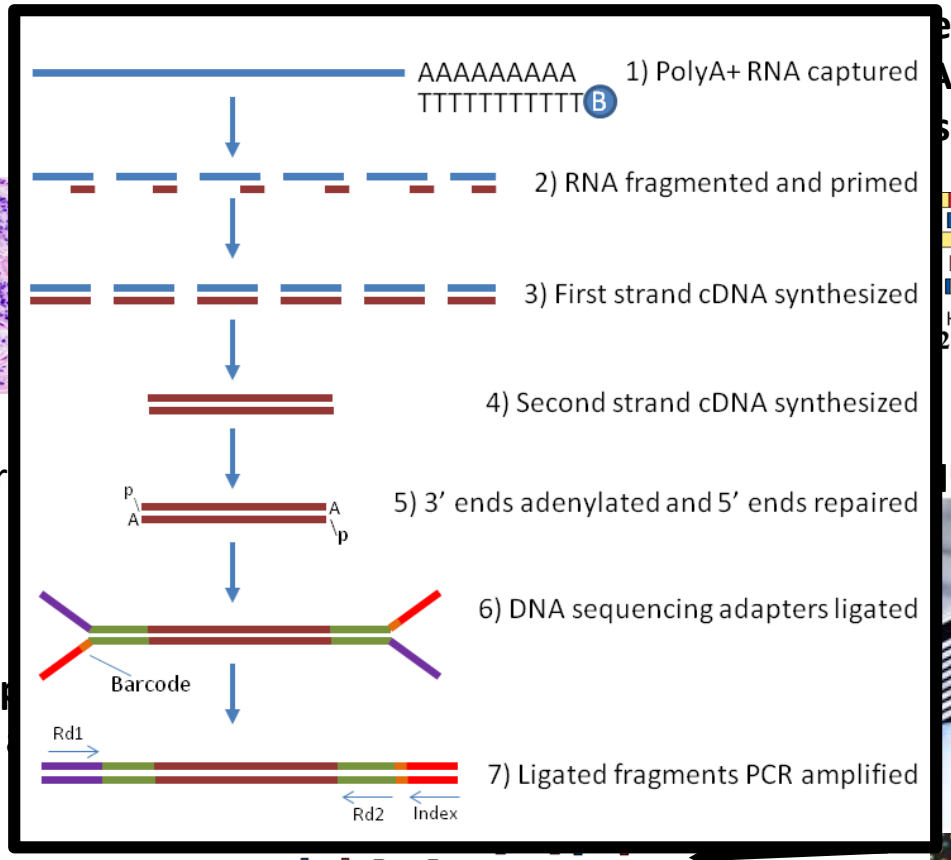
Samples of interest



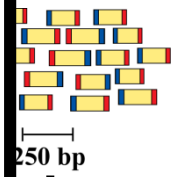
Condition 1
(normal colon)



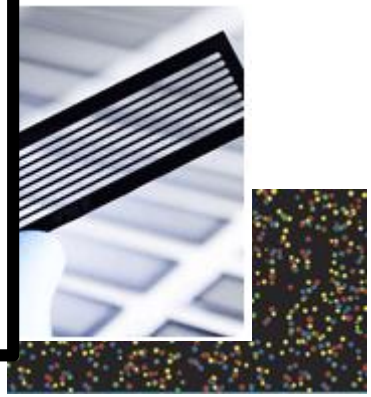
Condition 2
(colon tumor



ed to amplified
A fragments with
s at the ends



quence ends



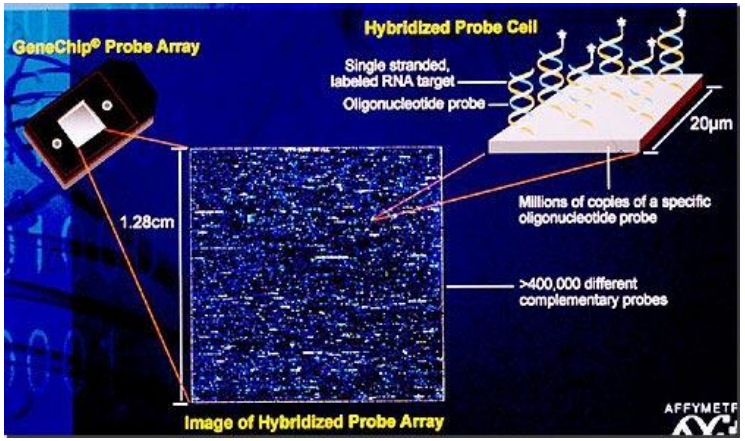
Assemble or map
to genome or to
transcriptome

Quantify and detect differential expression

Million/billions of fragments are sequenced

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

High Throughput Genomics



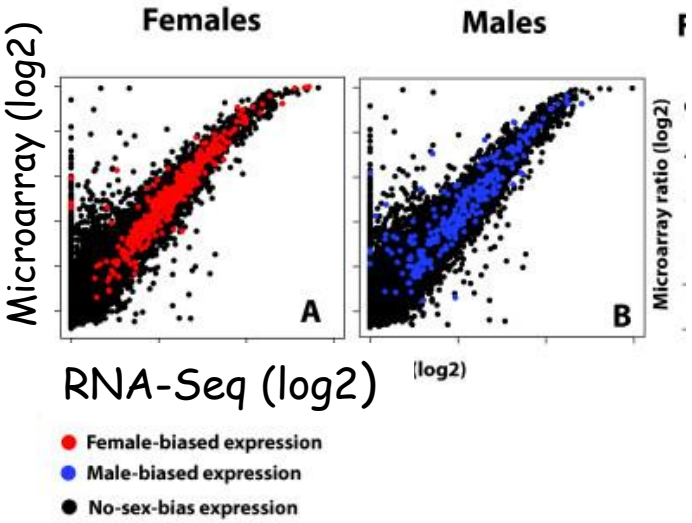
DNA Microarrays

Illumina
HiSeq2500
NextSeq500

Microarrays vs RNA-Seq

Malone et al. BMC Biol. 2011; 9: 34.

- Both high throughput methods can profile the genes with similar performance
- Microarrays suffer from compression (saturation) at the high end
- Low expression is problematic in both platforms



Microarray & RNA-Seq Pros and Cons

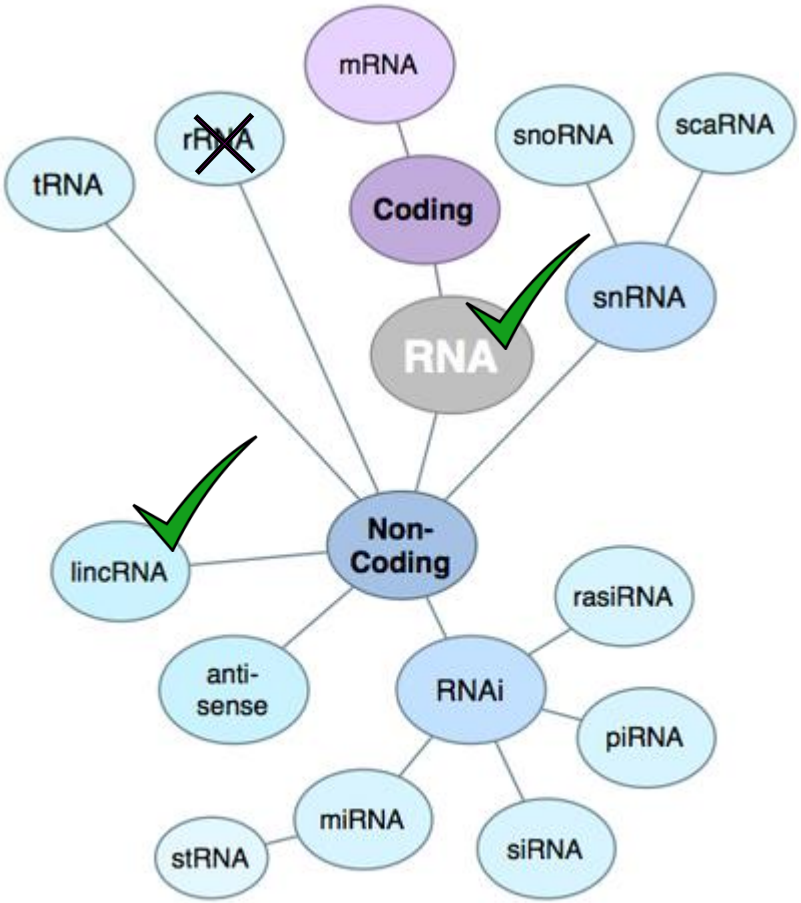
	Microarrays	RNA-Seq
Cost	$$	$$ (gene profiling) or $$$
Biases	Decade of research and solutions	Understanding is evolving
Data sizes	Mb -images	Gb- sequence data
Dynamic range	10^2	10^5
Transcript discovery , isoform identification & Transcript-chimeras	No	Yes
Genome required	Yes	No
Allele specific expression	No	Yes

Main Topics

- Introduction
- **Experimental design issues**
- Analysing RNA-Seq data
  - RNA-Seq pipeline: Tophat-Cufflinks-Cuffdiff
- Challenges

mRNA in the RNA "World"

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



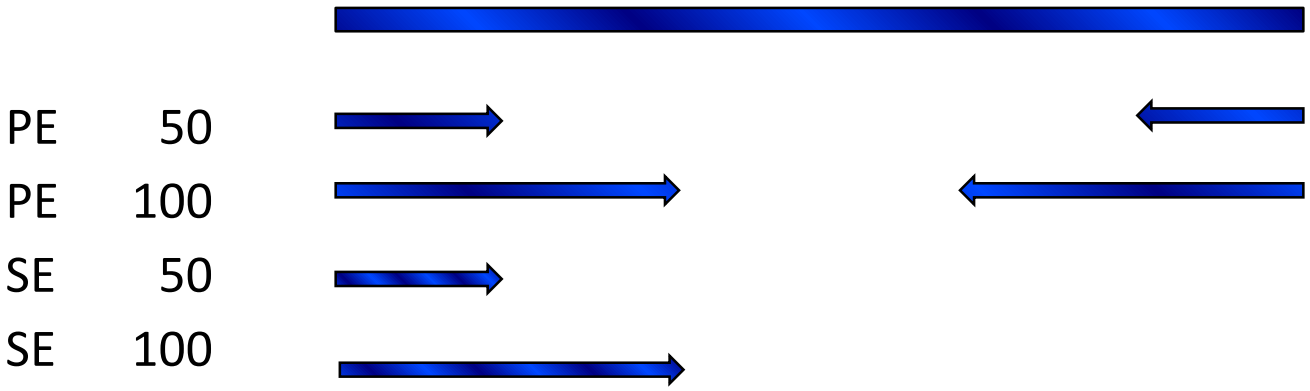
geospiza

Experiment Design
Sequencing Options

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



BIOINFORMATICS DISCOVERY NOTE

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

Gene expression

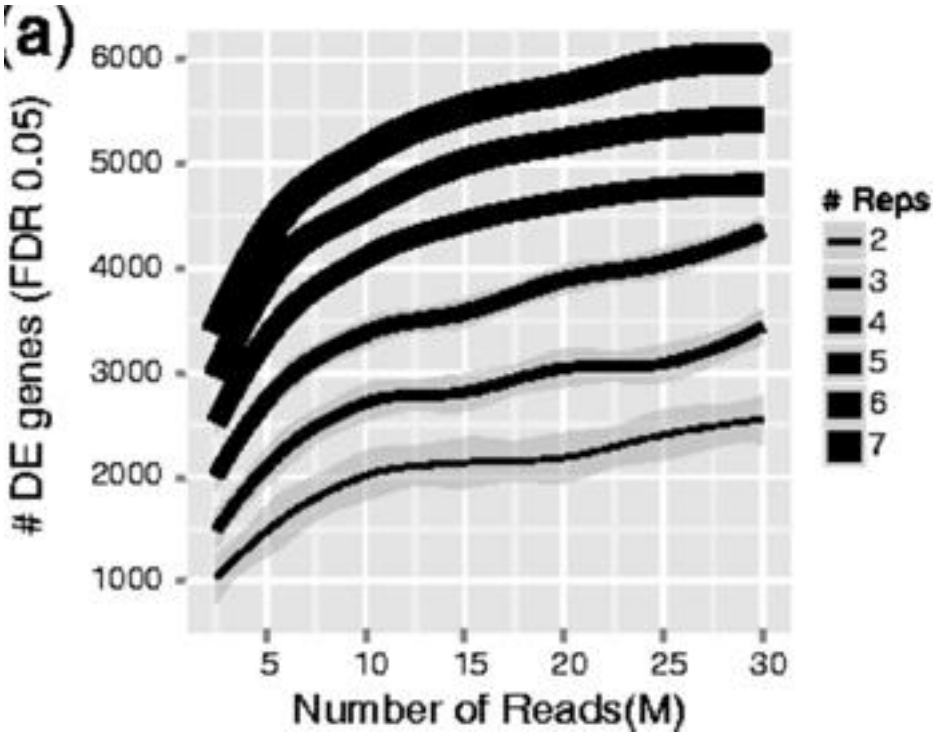
Advance Access publication December 6, 2013

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



ENCODE consortium's *Standards, Guidelines and Best Practices for RNA-Seq*

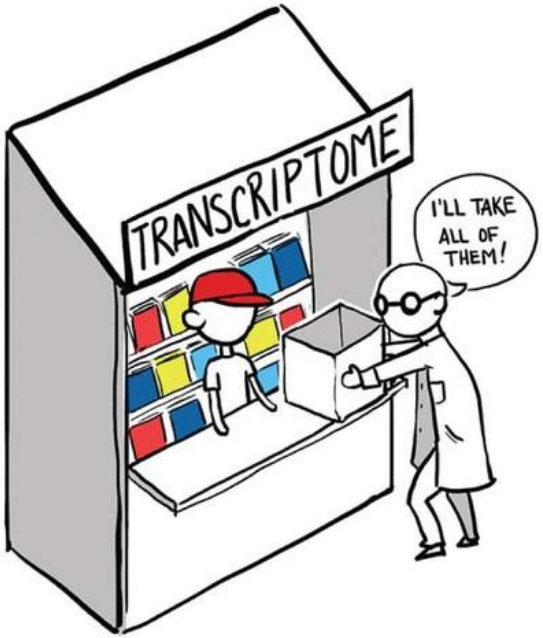
Experimental Design

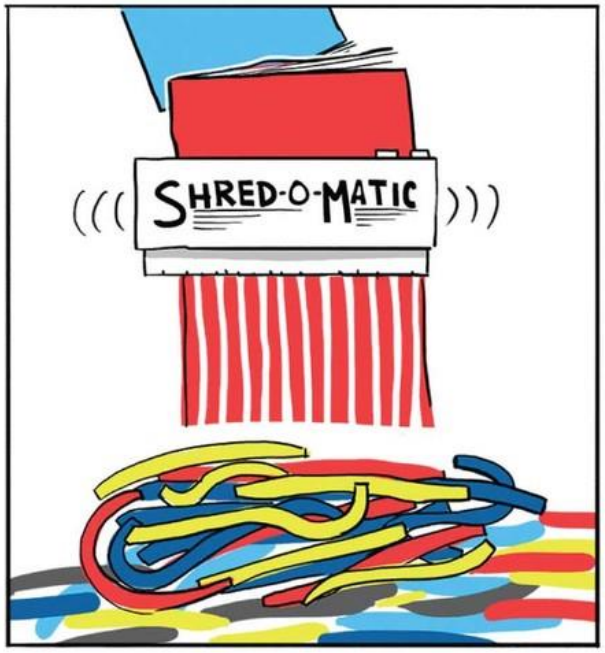
- Library protocol
- Sequencing depth (reads/fragments)
- Sequencing options (length, SE, PE)
- Assessing biological variation requires biological replicates
  - Duplicates (2X2) are a minimum, yet more recommended

    (pooling, avoid batch effect)
- Consult with the person which will analyse the data before performing the experiment – Kick-off meeting

Main Topics

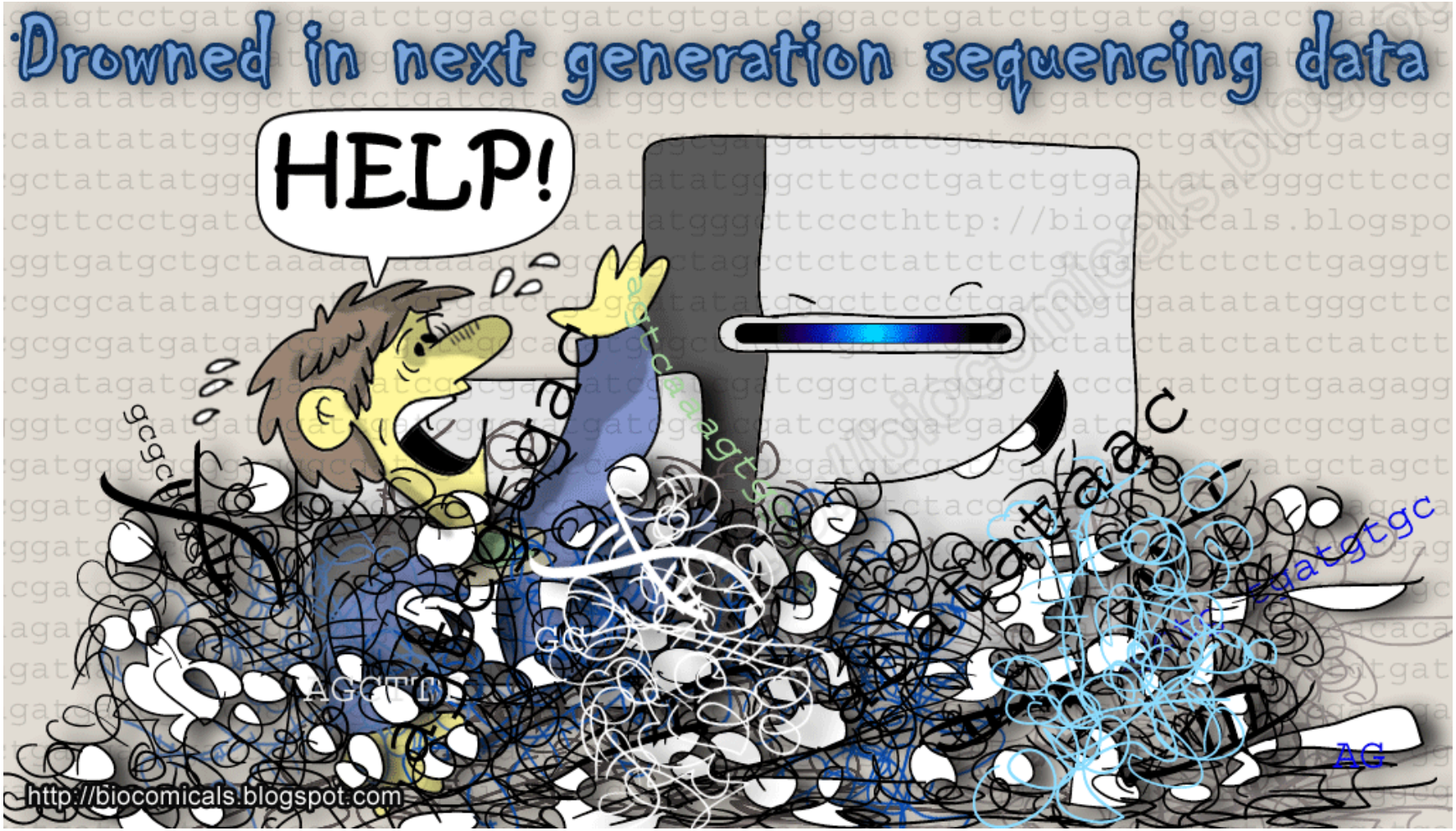
- Introduction
- Experimental design issues
- **Analysing RNA-Seq data**
  - Quality Control
  - **RNA-Seq pipeline:Tophat- HTSeq-DESeq2**
- Chipster





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?

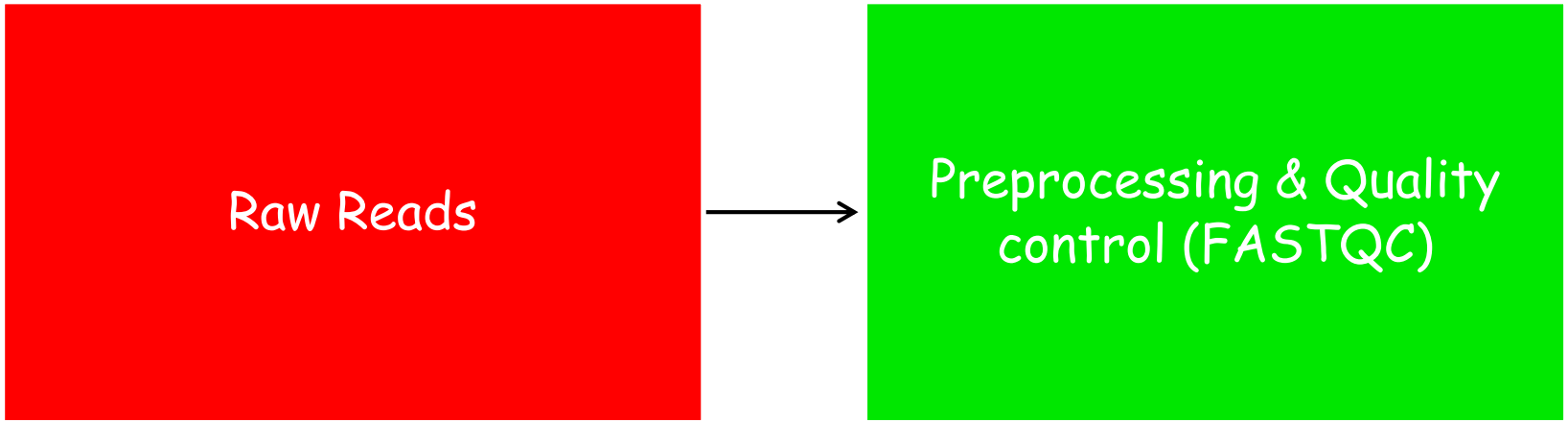
Abigail Yu Nature Methods 10, 1165–1166 (2013)



HELP !!!!

I just got sequence data…

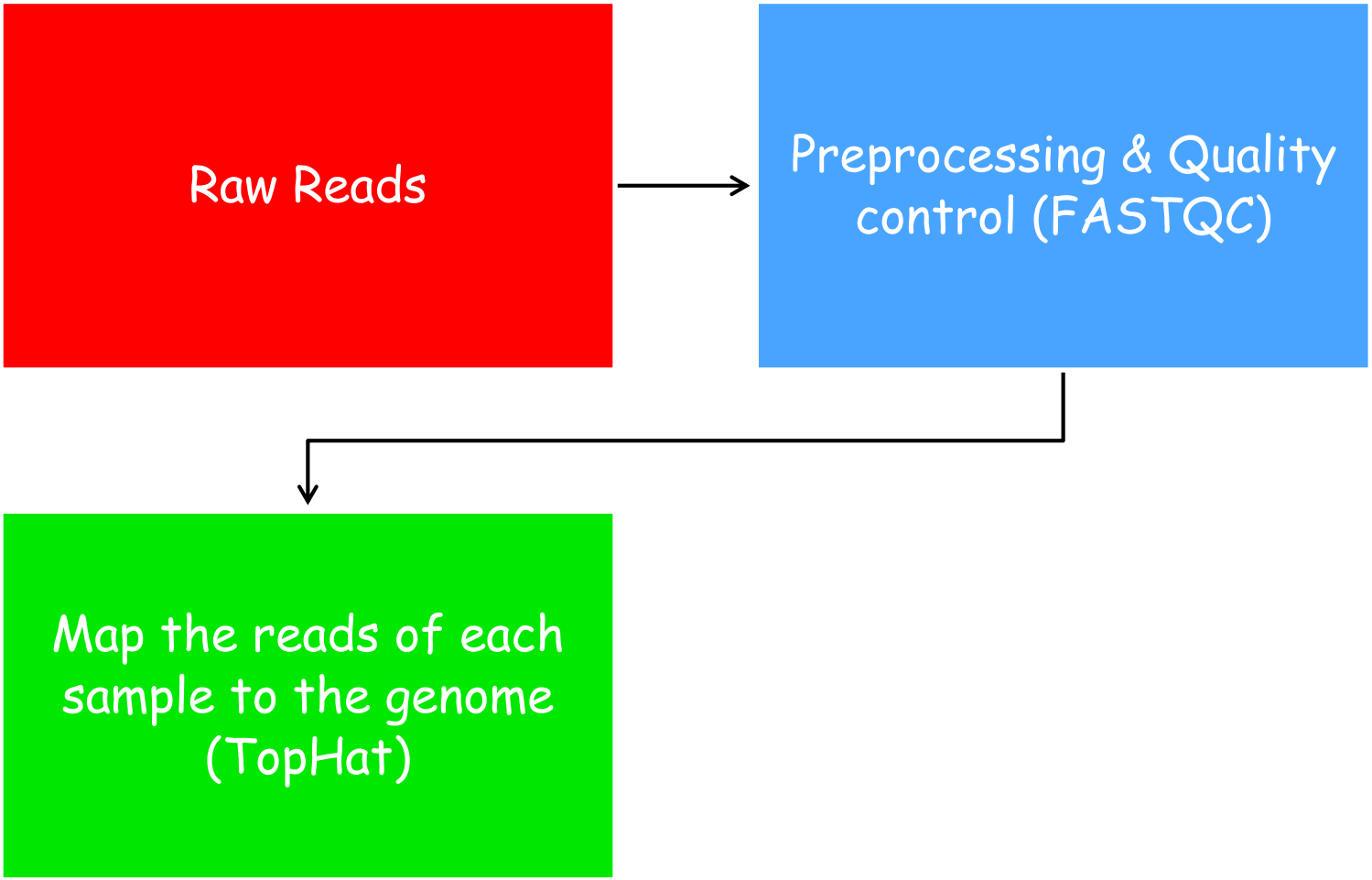
RNA-Seq Workflow



Pre-processing

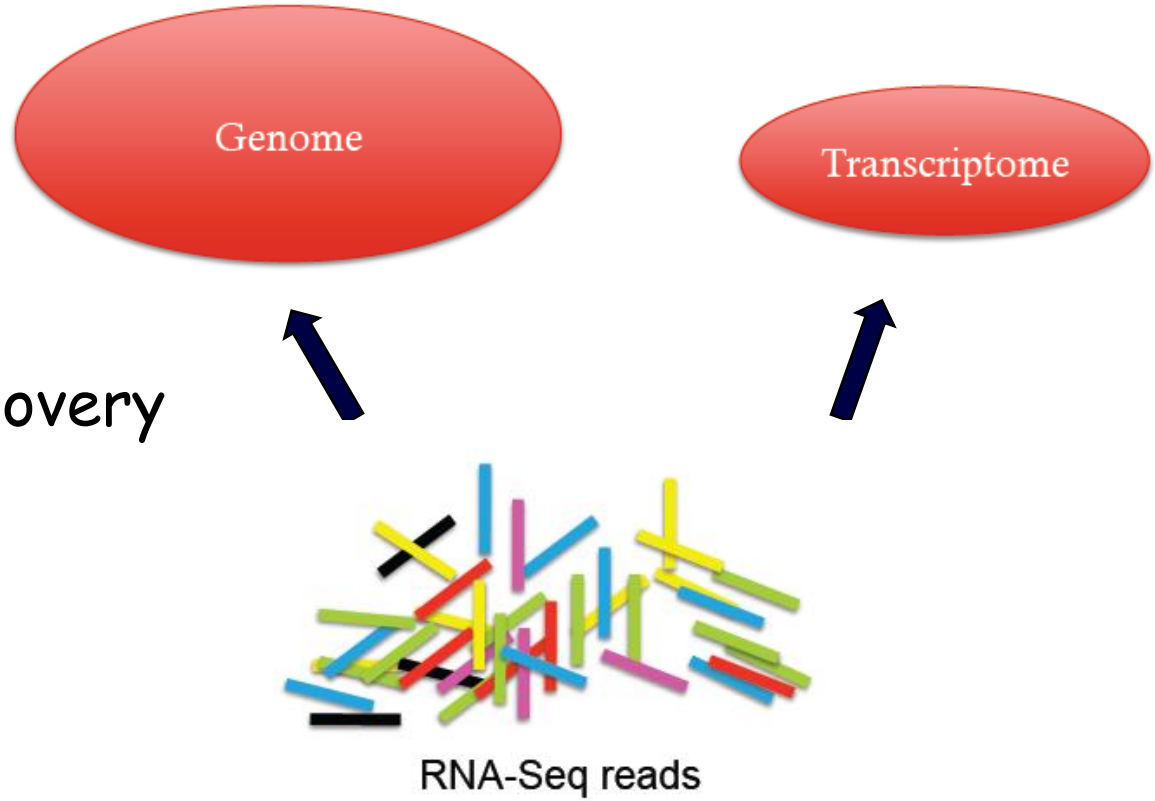
- Need to assess that the sequence data has sufficient quality
- Recommendation is to use the high quality sequence data (more important in de novo assembly):
  - Filter low quality reads
  - Check the amount of read duplication (too much PCR amplification)
  - Trim sequences if the end is of low quality

RNA-Seq Workflow



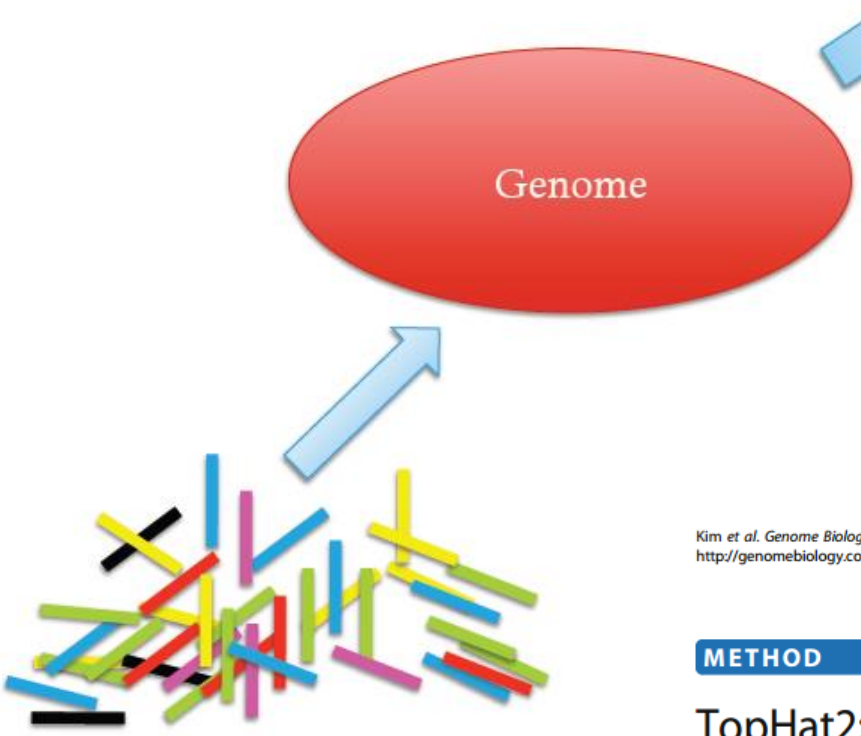
Mapping Short RNA-Seq Reads

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



RNA-Seq mapping with TopHat

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



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]

Kim *et al. Genome Biology* 2013, **14**:R36
http://genomebiology.com/2013/14/4/R36

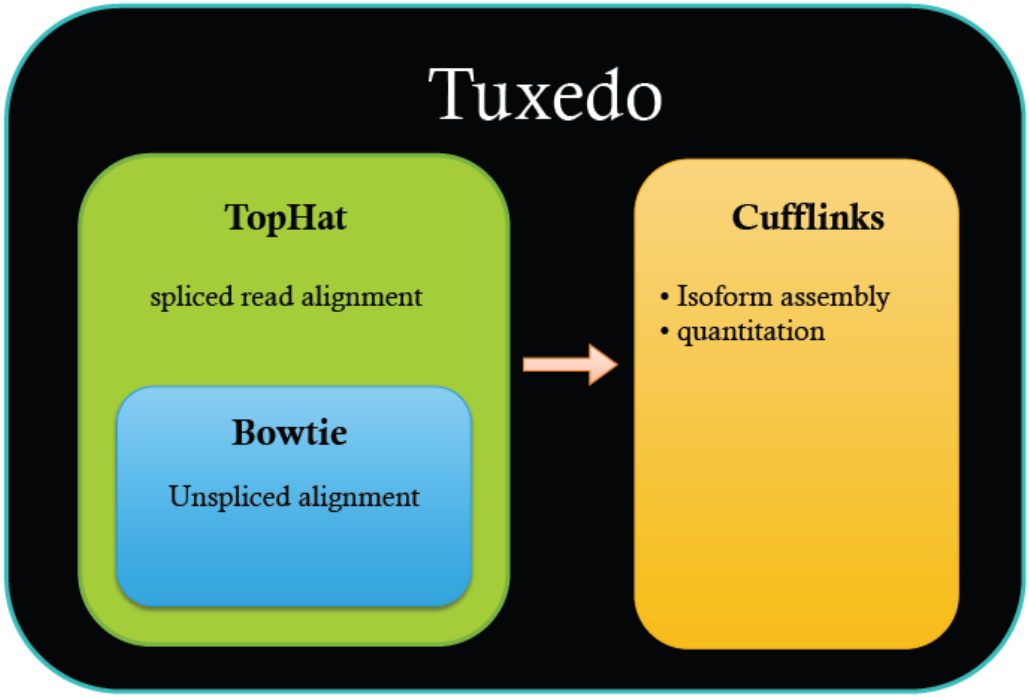
Genome Biology

METHOD Open Access

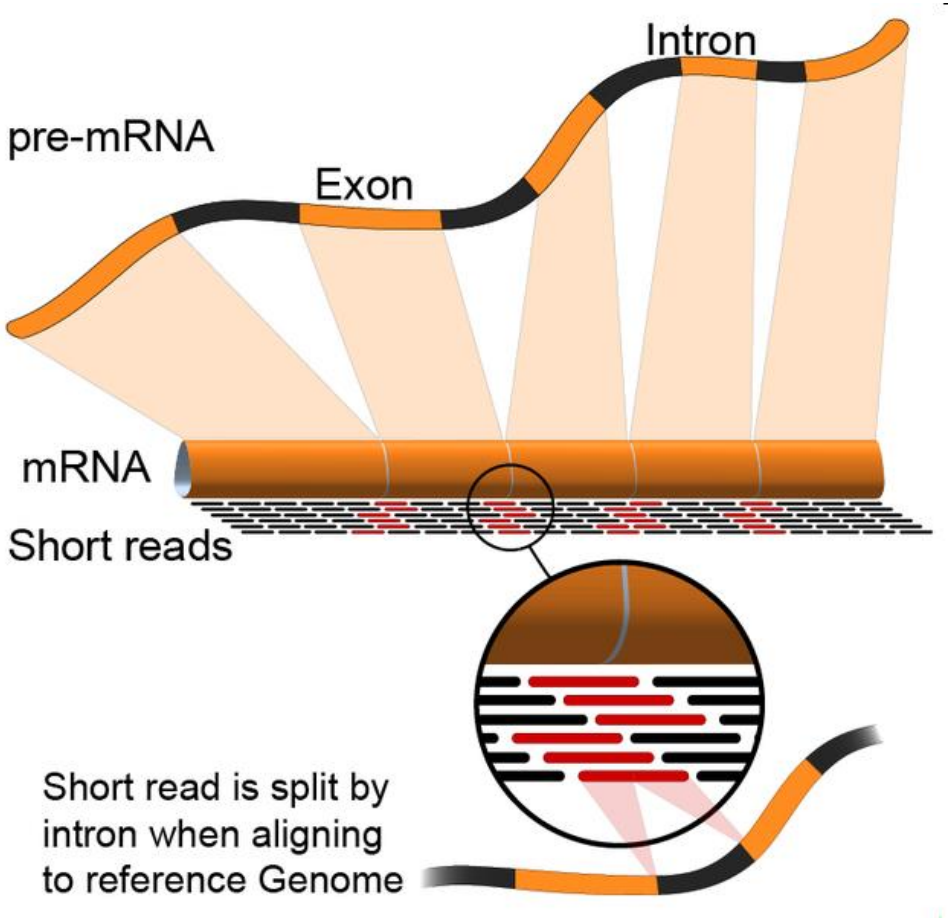
TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions

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

The Tuxedo Tools



Mapping to Genome
How to align reads that span exons?



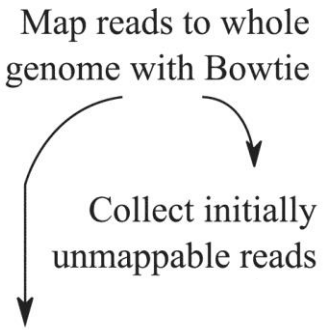
http://en.wikipedia.org/wiki/RNA-Seq

The Challenge –Identifying Novel Junctions

- Reads are short and contain errors
- Rarely transcribed genes have few reads spanning the junctions
- We are interested in discovering novel junctions i.e. we are not relying on annotation of known genes (?)
- Need to perform the task in a timely manner

Tophat: Exon first two step approach

- Mapping to the genome is done with Bowtie
- Extracting unmapped reads (not including low complexity)

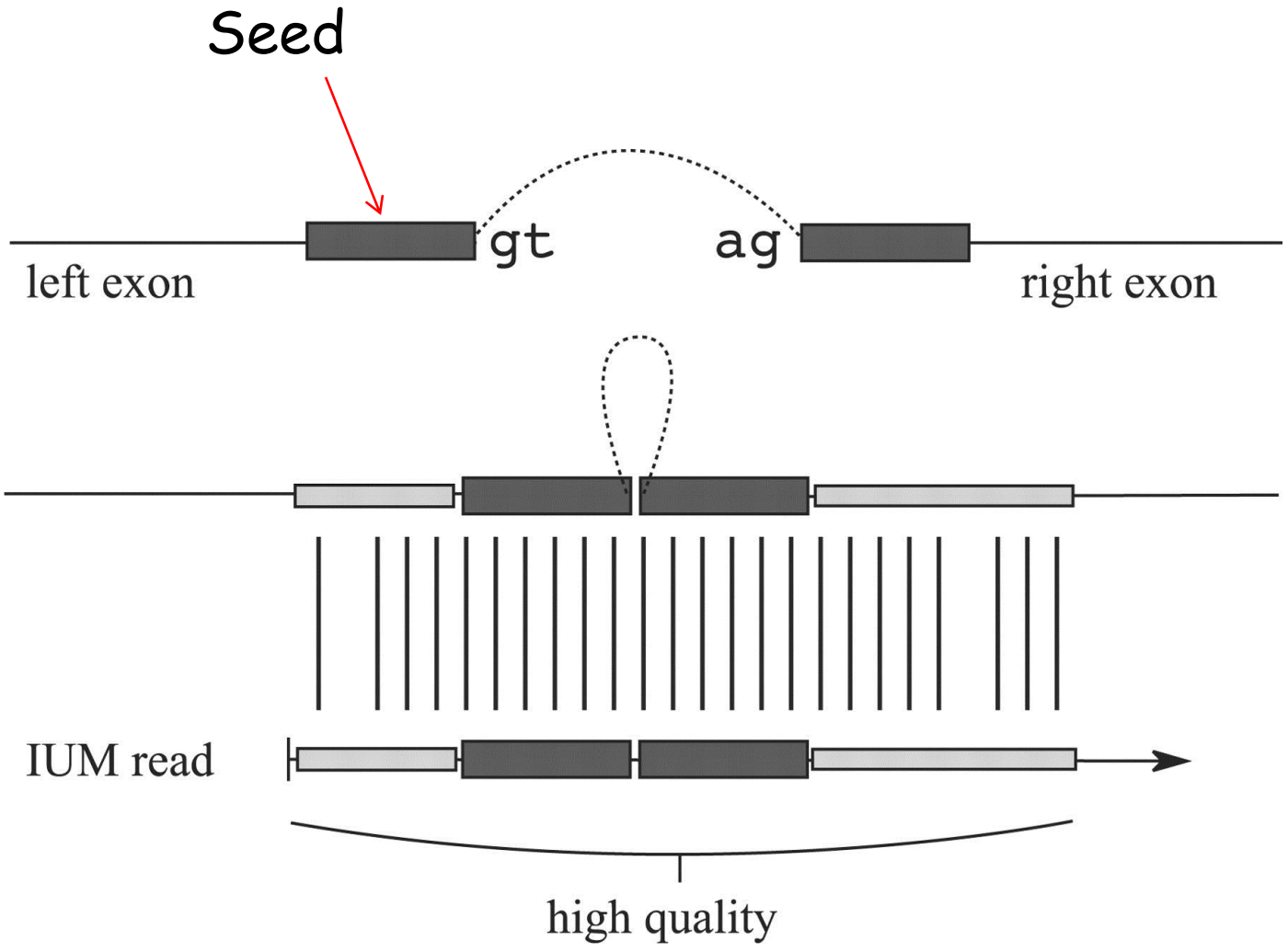


Identifying the transcriptome

identify candidate exons via genomic mapping

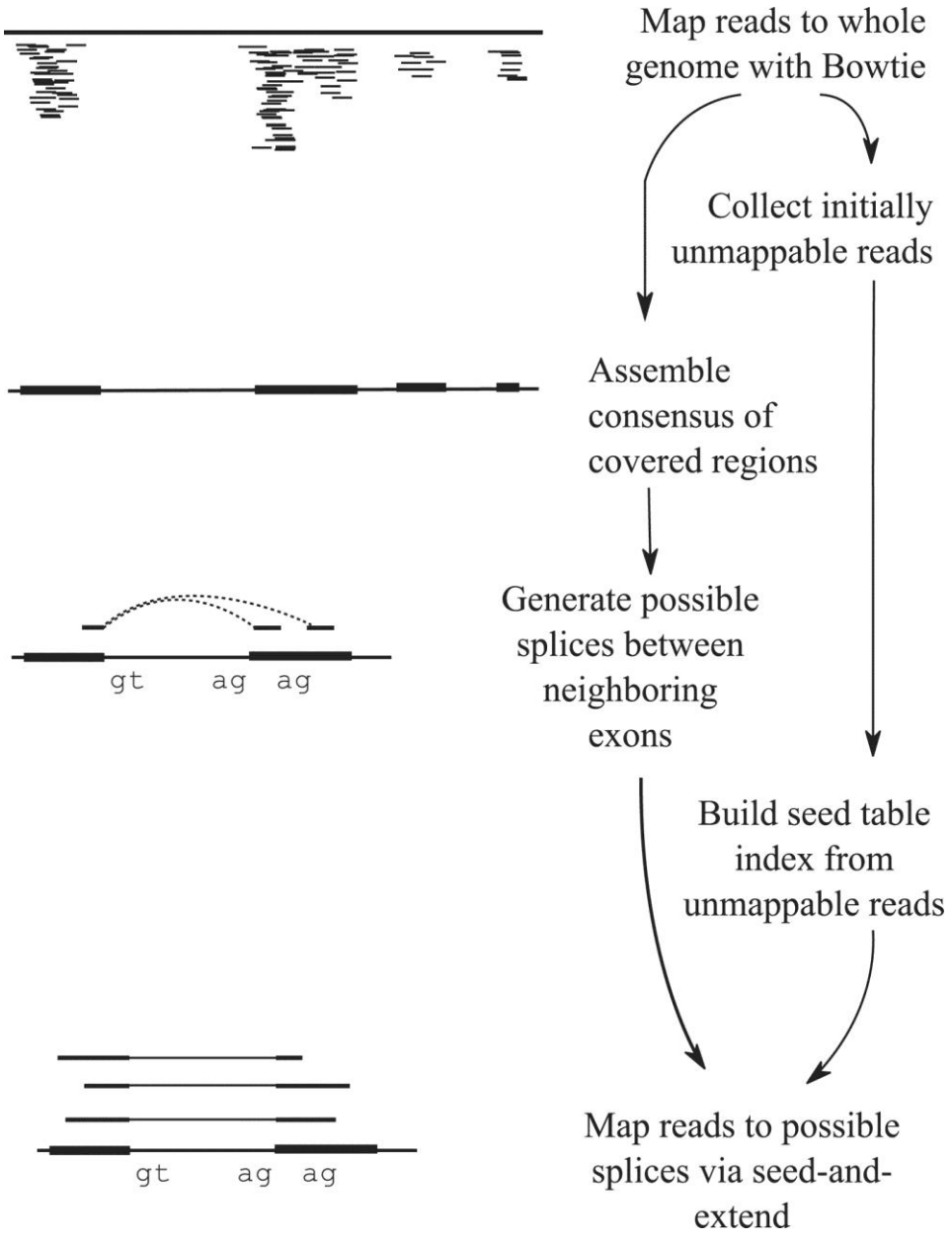
Generate possible pairings of exons

Align “unmappable” reads to possible junctions



Trapnell, C. et al. Bioinformatics 2009 25:1105-1111; doi:10.1093/bioinformatics/btp120

Bioinformatics



Trapnell, C. et al. Bioinformatics 2009 25:1105-1111; doi:10.1093/bioinformatics/btp120

Bioinformatics

Tophat2

TopHat2 combines the ability to identify novel splice sites with direct mapping to known transcripts, producing sensitive and accurate alignments, even for highly repetitive genomes or in the presence of pseudogenes.

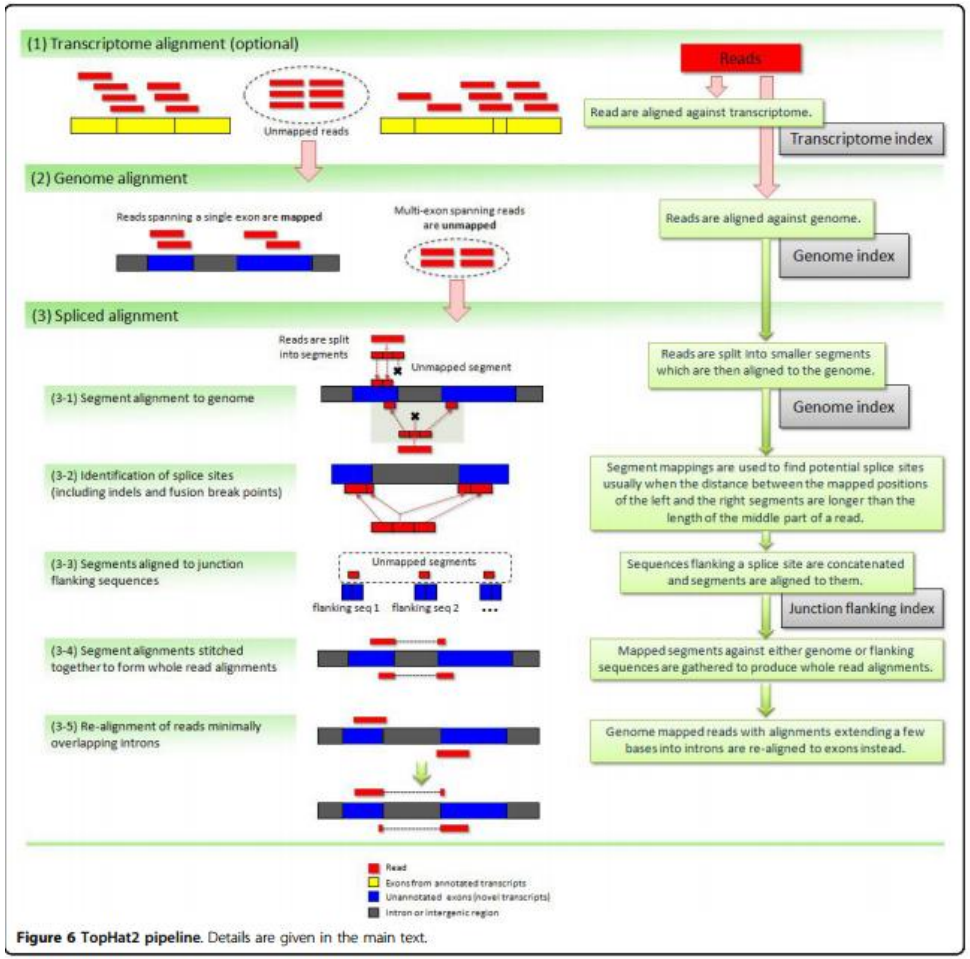


Figure 6 TopHat2 pipeline. Details are given in the main text.

Genome Biology 2013 **14**:R36

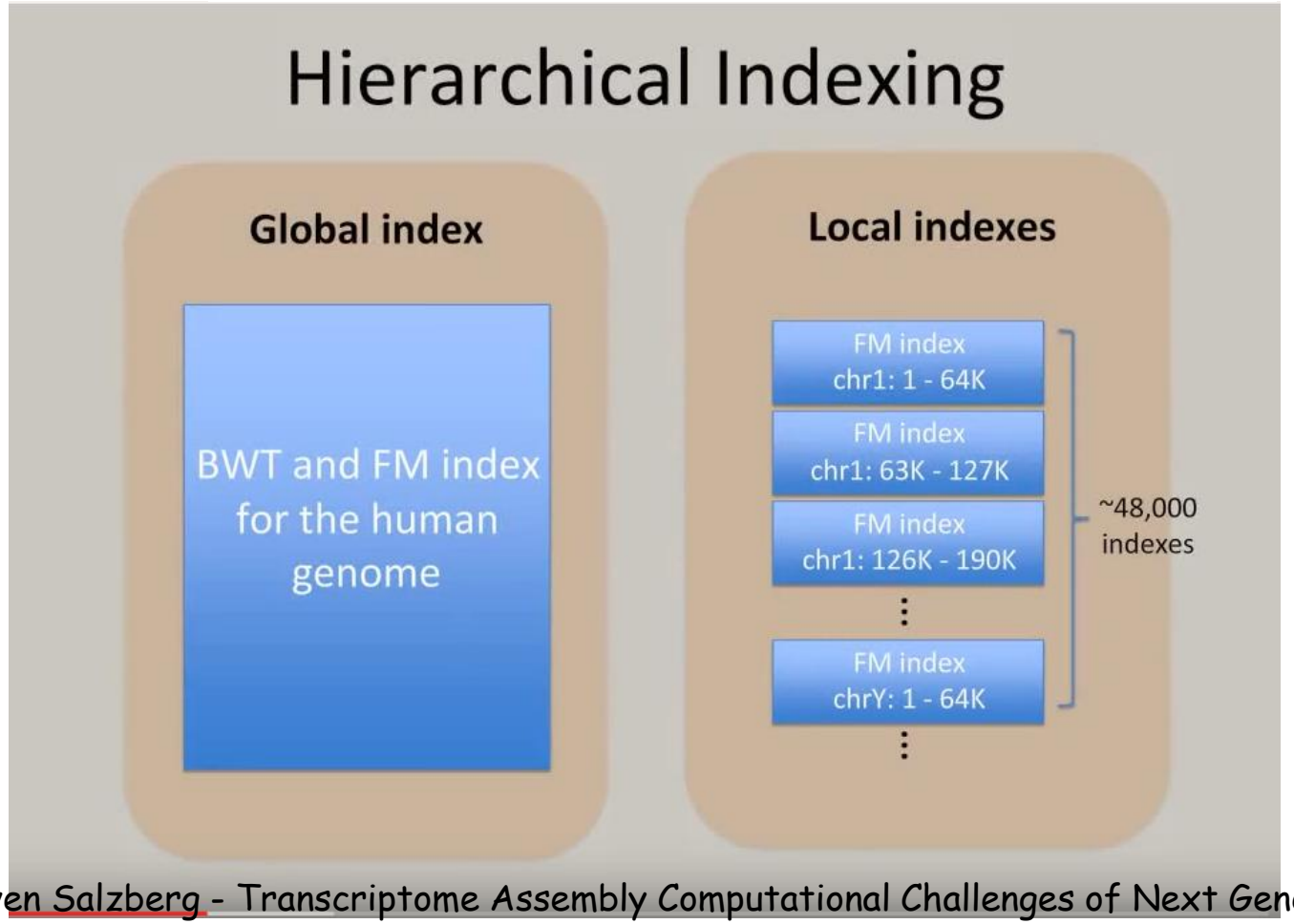
HISAT: a fast spliced aligner with low memory requirements

Daehwan Kim, Ben Langmead & Steven L Salzberg

Affiliations | Contributions | Corresponding authors

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

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



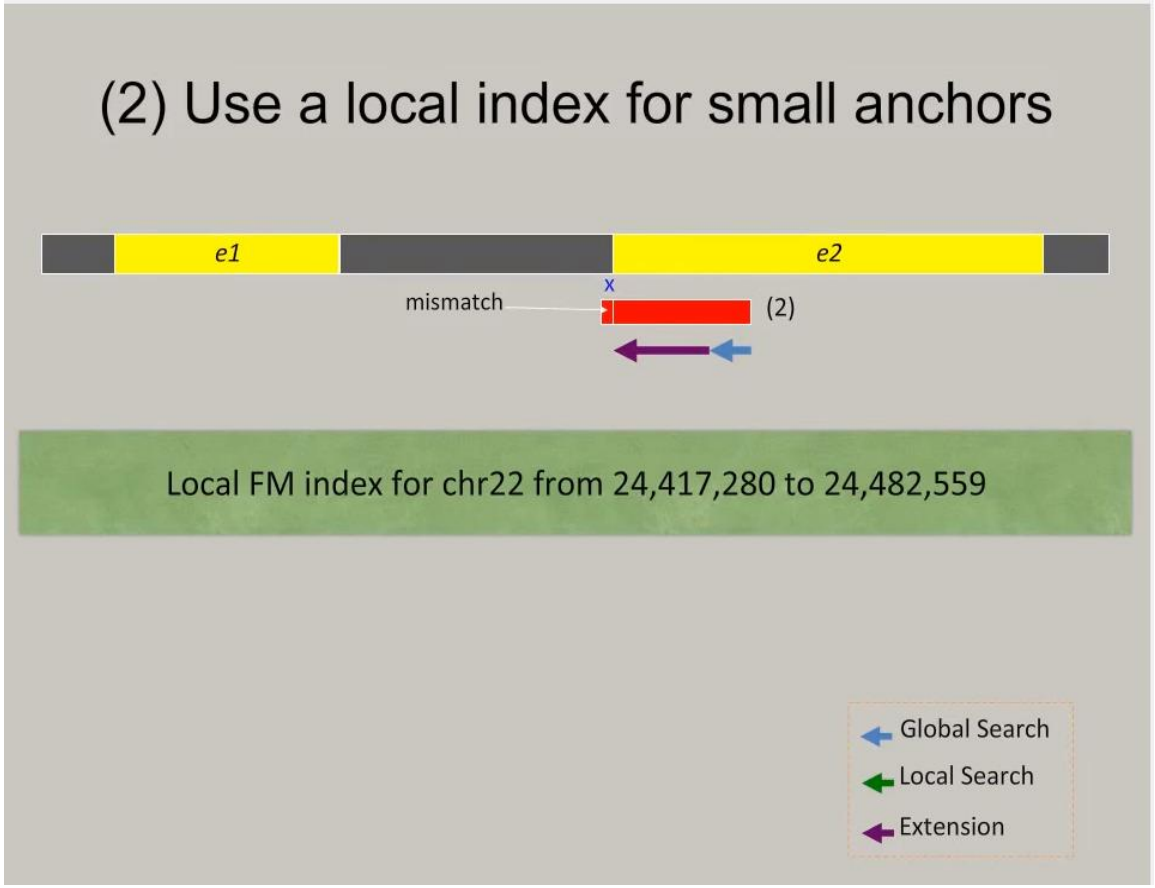
Steven Salzberg - Transcriptome Assembly Computational Challenges of Next Generation Sequence Data
https://www.youtube.com/watch?v=2qGiw4MRK3c

HISAT

To begin processing each read, it first tries to find candidate locations across the target genome

It identifies these locations by first mapping part of each read using the global FM index, which in most cases identifies one or a small number of candidates

HISAT then selects one of the ~48,000 local indexes for each candidate and uses it to align the remainder of the read.



Steven Salzberg - Transcriptome Assembly Computational Challenges of Next Generation Sequence Data
https://www.youtube.com/watch?v=2qGiw4MRK3c

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

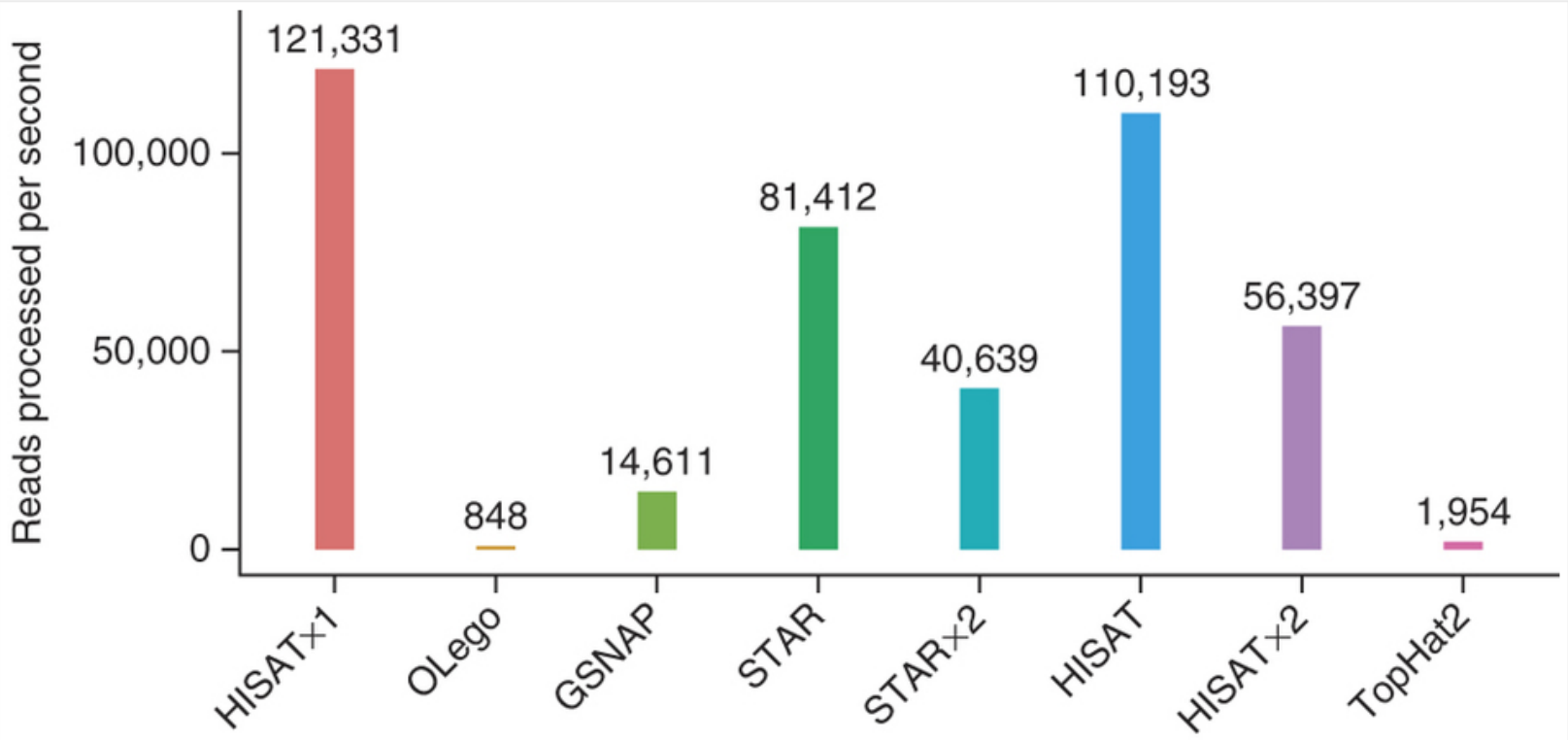
From

HISAT: a fast spliced aligner with low memory requirements

Daehwan Kim, Ben Langmead & Steven L Salzberg

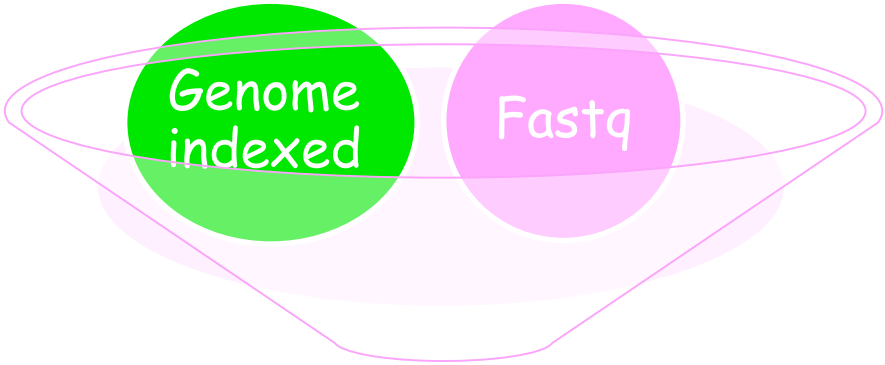
Nature Methods **12**, 357–360 (2015) | doi:10.1038/nmeth.3317

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.

Tophat Outputs



Tophat

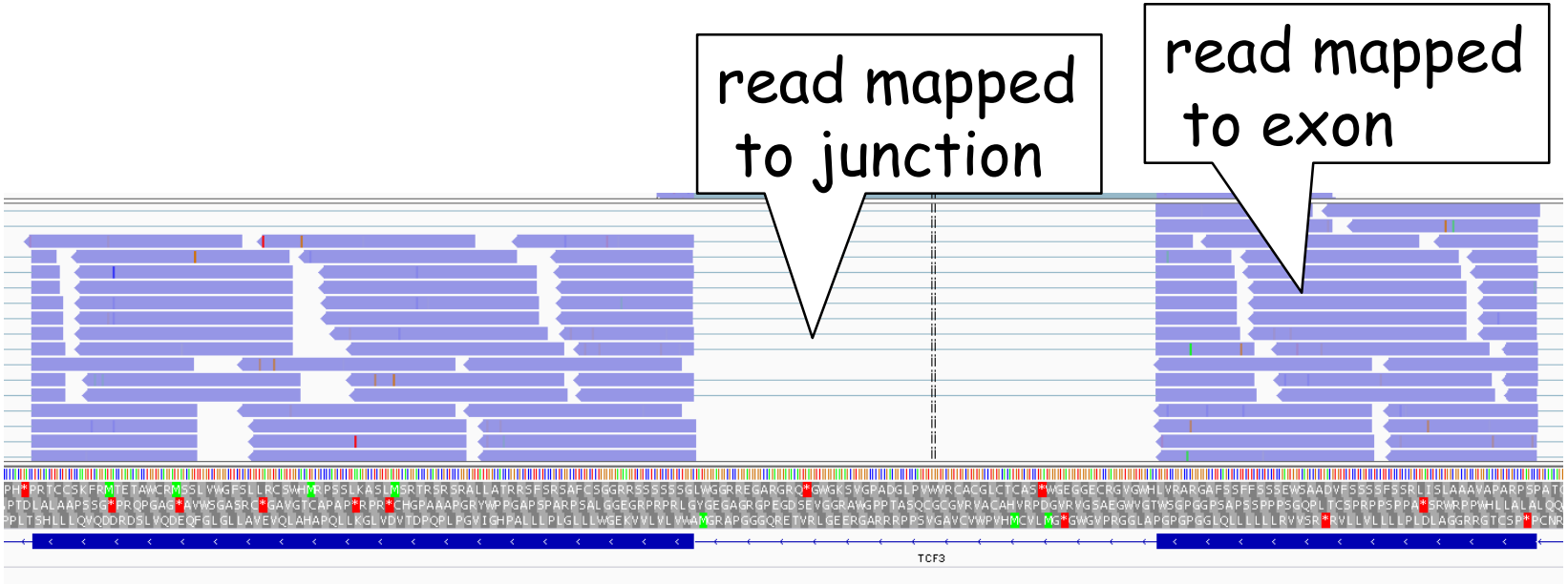
Accepted_hits.bam + Insertion.bed Deletion.bed + Junctions.bed

Transcript discovery/Transcript or Gene Quantification/View in a Genome Browser

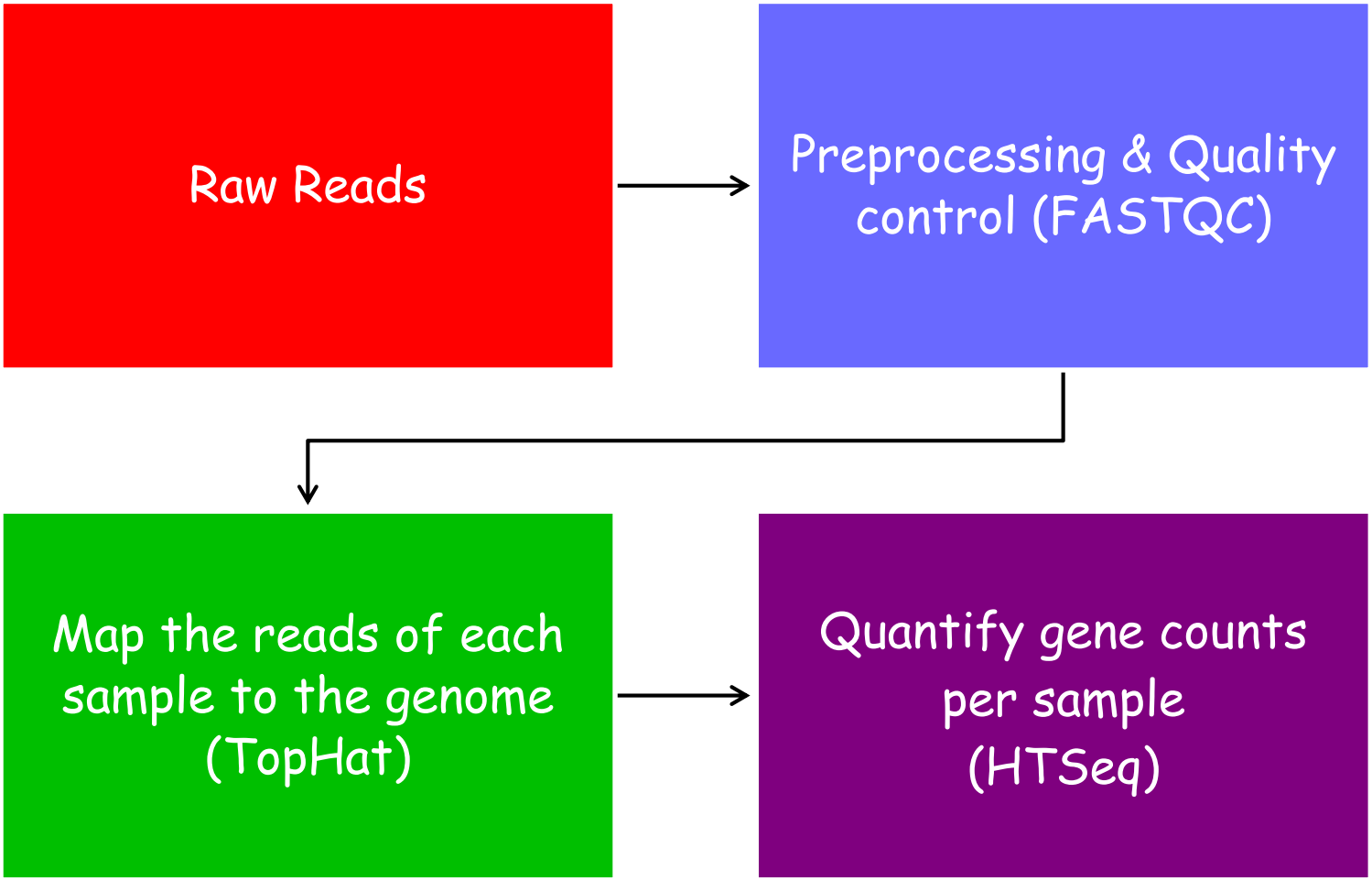
Junction.bed

5	chr1	94935734	94936354	JUNC00022394	457	-	11,33	0,587
6	chr1	94962730	94963662	JUNC00022414	456	+	32,34	0,898
7	chr10	28642081	28642574	JUNC00005781	38	-	18,35	0,458
8	chr10	76067101	76067623	JUNC00006282	74	+	30,34	0,488
9	chr10	118650871	118652271	JUNC00007170	20	-	16,22	0,1378
10	chr10	126684651	126685114	JUNC00007305	87	-	26,35	0,428
11	chr10	127986002	127986178	JUNC00007509	96	-	25,35	0,141
12	chr10	128006919	128007139	JUNC00007511	63	+	27,34	0,186
13	chr11	23326211	23326669	JUNC00002682	31	+	34,35	0,423
14	chr11	59637921	59639825	JUNC00003176	20	-	35,28	0,1876
15	chr11	69161988	69162244	JUNC00003397	49	-	35,33	0,223
16	chr11	69652041	69652451	JUNC00003476	24	+	34,30	0,380
17	chr11	97048737	97048877	JUNC00004353	22	+	18,24	0,116
18	chr12	113031876	113032578	JUNC00009692	45	+	34,31	0,671
19	chr14	53158445	53159099	JUNC00012157	20	-	31,29	0,625
20	chr14	53160268	53160487	JUNC00012158	30	-	32,32	0,187
21	chr15	76454881	76455044	JUNC00010523	21	-	30,27	0,136
22	chr15	89289222	89289365	JUNC00011107	27	-	16,29	0,114

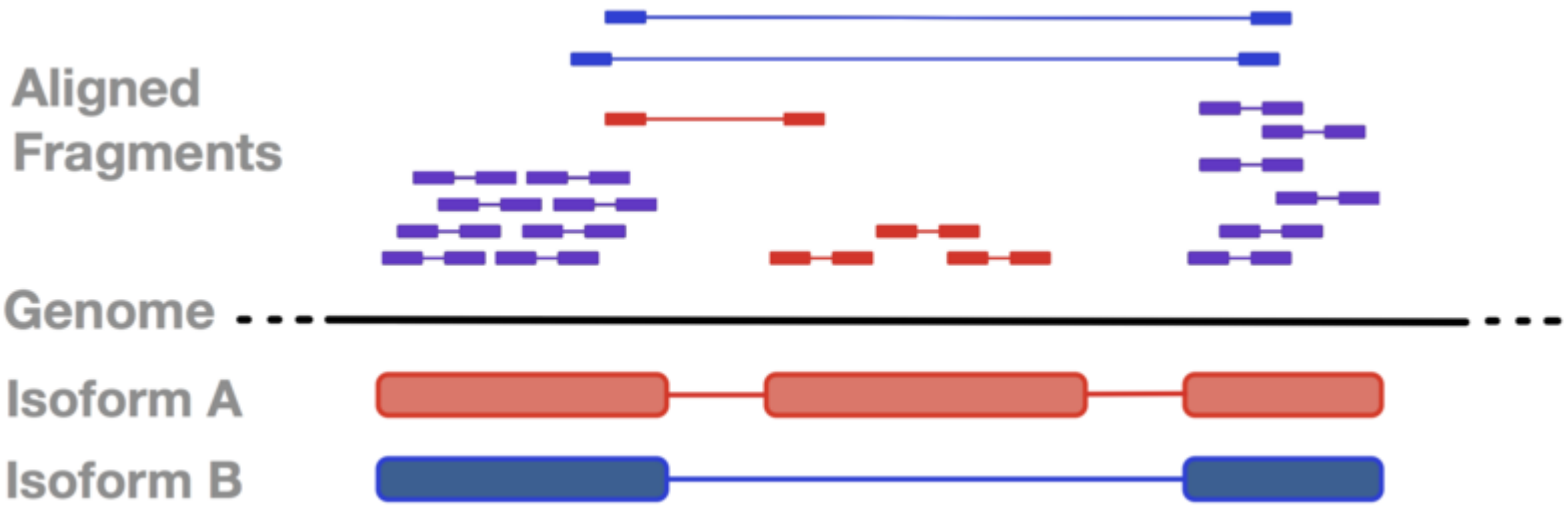
Visualization of Tophat outputs in a Genome Browser (IGV)



RNA-Seq Workflow



Basic Quantification Step



Count the number of reads aligned to each gene (do not need to determine to from which transcript the read was derived)

Trapnell et al. Nat Biotechnol. 2010 May;28(5):511-5.

HTSeq

A gene is quantified by counting the number of fragments/reads which align to all its exons

Discard a read if it is non-uniquely mapped to a gene

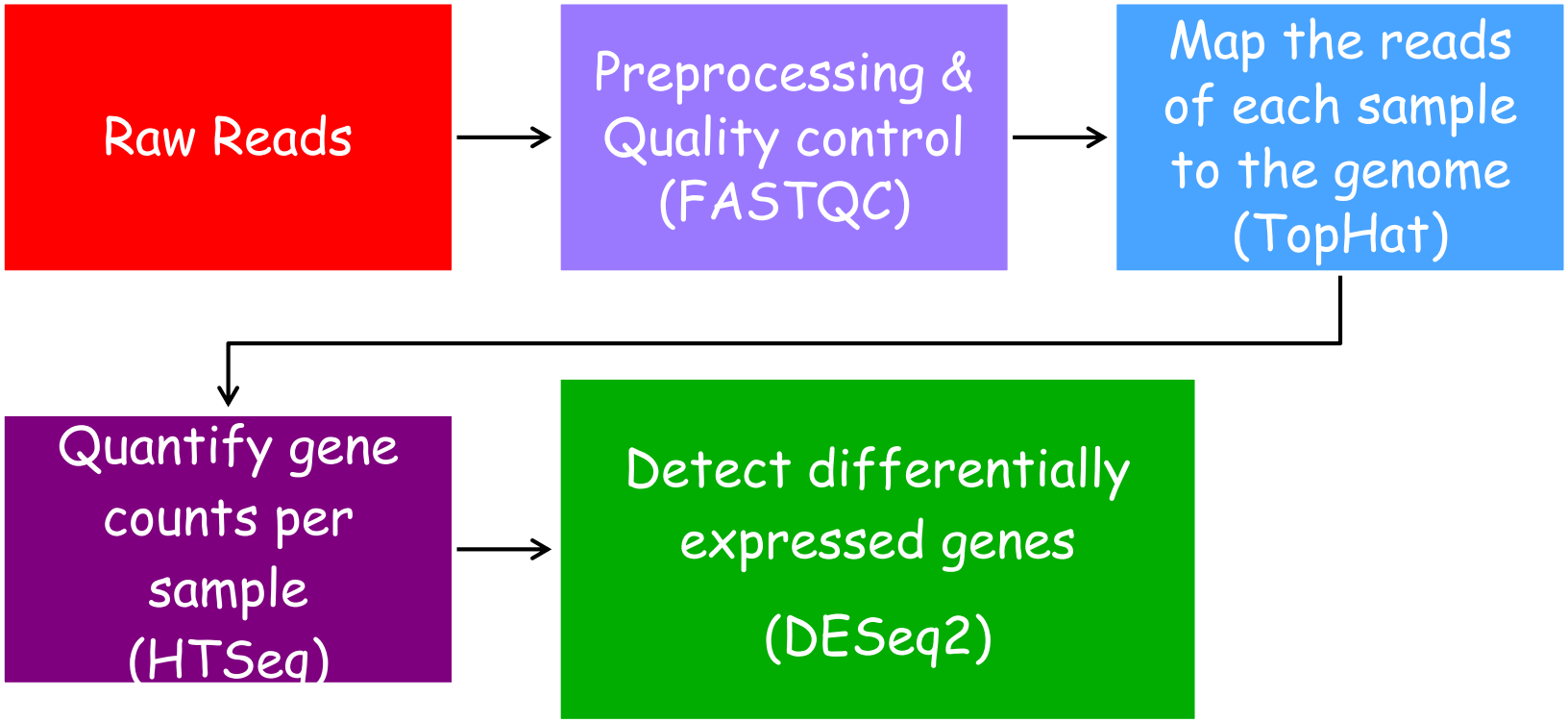
http://www-huber.embl.de/users/anders/HTSeq/doc/count.html

HTSeq Result
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	5897	7537	10029
SUM	10M	50M	30M	20M

Need to account for the differences in sequence amount between the samples

RNA-Seq Workflow



DESeq2 Normalization

Need to normalize between 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.

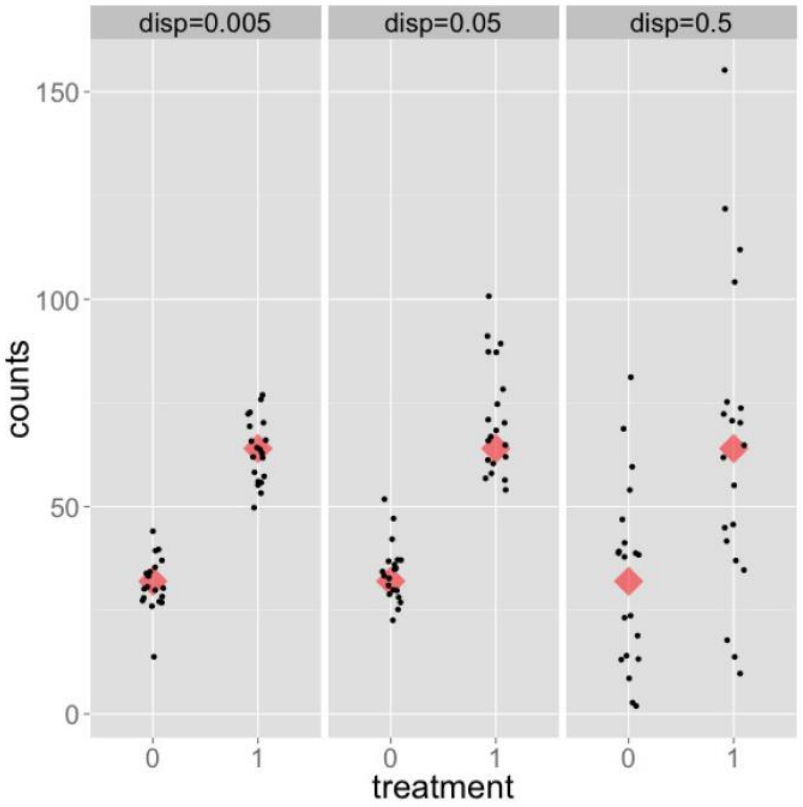
This procedure corrects for library size and RNA composition bias, which can arise for example when only a small number of genes are very highly expressed in one experiment condition but not in the other.

Example-DESeq Normalization

	sample_1	sample_2	sample_3	sample_4	geometric mean		ratio sample_1
gene_1	15	9	11	18	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

Determining Differentially Expressed Genes

Discover genes showing different average expression levels across two groups



The advantage of having many replicates allows to learn about the biological variation within the conditions tested.
Aim: finding genes which have a difference in expression which is larger than the "noise"

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

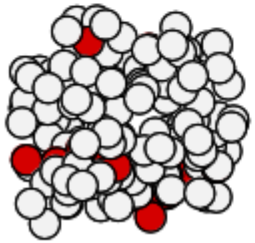
RNA-Seq Noise

Suppose we sequence the same library twice to the same depth. Will we get the same gene counts?

- No

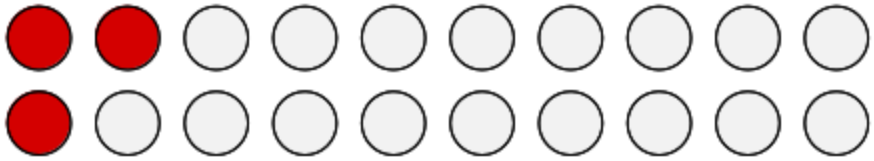
- Poisson noise – the variance in counts that persists even if everything is exactly the same

The Poisson distribution

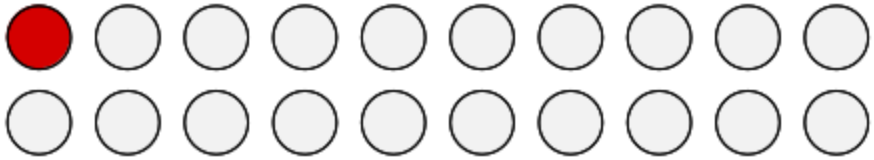


- This bag contains very many small balls, 10% of which are red.
- Several experimenters are tasked with determining the percentage of red balls.
- Each of them is permitted to draw 20 balls out of the bag, without looking.

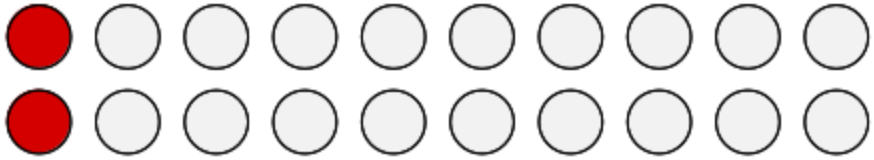
http://www.bioconductor.org/help/course-materials/2013/CSAMA2013/tuesday/morning/Anders__DESeq_DEXSeq.pdf



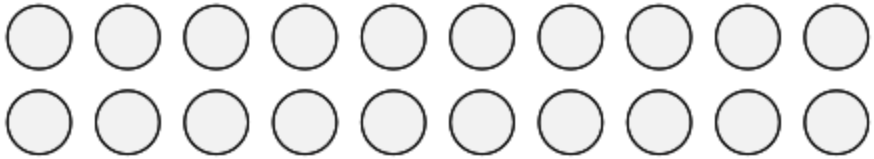
3 / 20 = 15%



1 / 20 = 5%



2 / 20 = 10%



0 / 20 = 0%

$7 / 100 = 7\%$

$10 / 100 = 10\%$

$8 / 100 = 8\%$

$11 / 100 = 11\%$

Poisson Distribution

- The actual number k of red balls follows a Poisson distribution, and hence k varies around its expectation value μ with standard deviation $\sqrt{\mu}$ (variance μ)
- Assuming the gene counts in a RNA-Seq experiment follow a Poisson distribution we would expect the standard deviation to be equal to the square root of the average gene count

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 cannot be calculated, it has to be estimated from the data

$$var = \mu + c\,\mu$$

Poisson noise

Biological noise

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 sample mean
- The count data is used to estimate the variance

  Orange line: the fitted observed curve for the variance

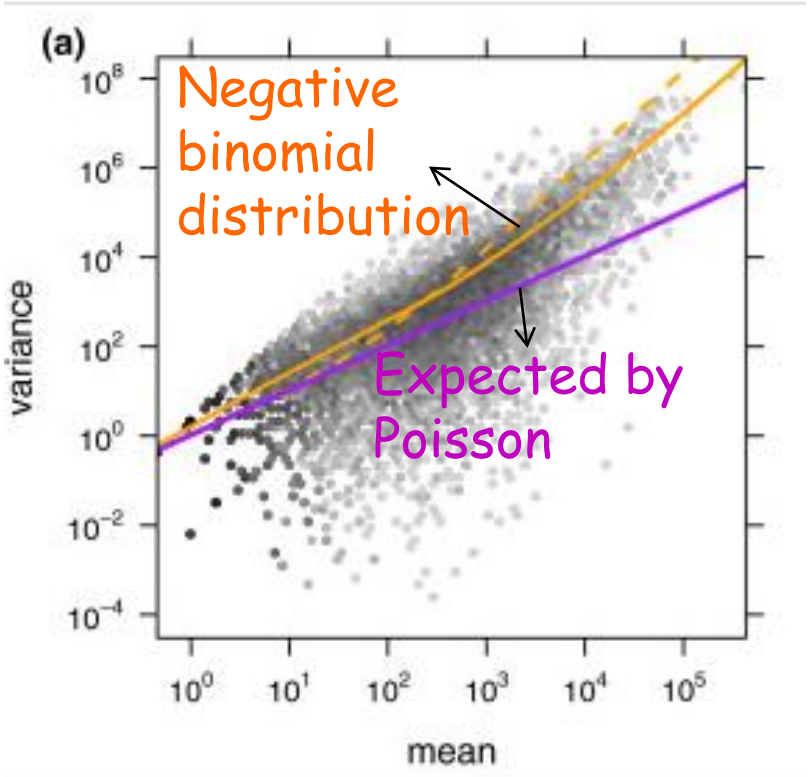
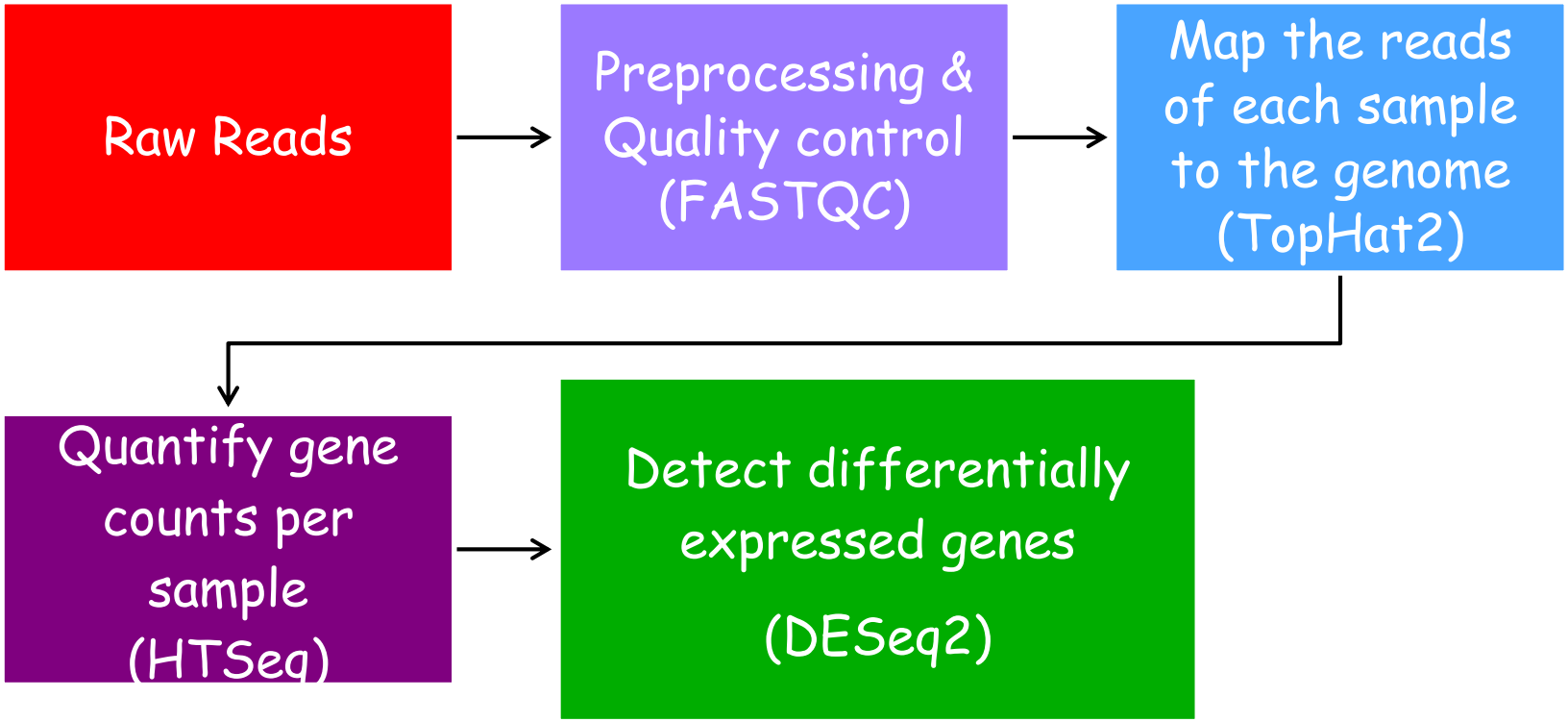


Fig. 1 from Anders & Huber, 2010: Dependence

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



Main Topics

- Introduction
- Experimental design issues
- Analysing RNA-Seq data
  - RNA-Seq pipeline: Tophat- HTSeq-DESeq2
- In the exercise we will use Chipster to run the pipeline

THE END

Thanks

Questions??

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)

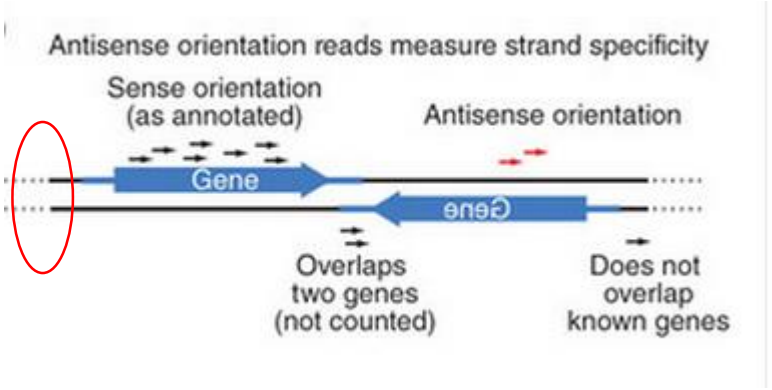
2. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105-11. doi: 10.1093/bioinformatics/btp120. PubMed PMID: 19289445; PubMed Central PMCID: PMCPMC2672628.

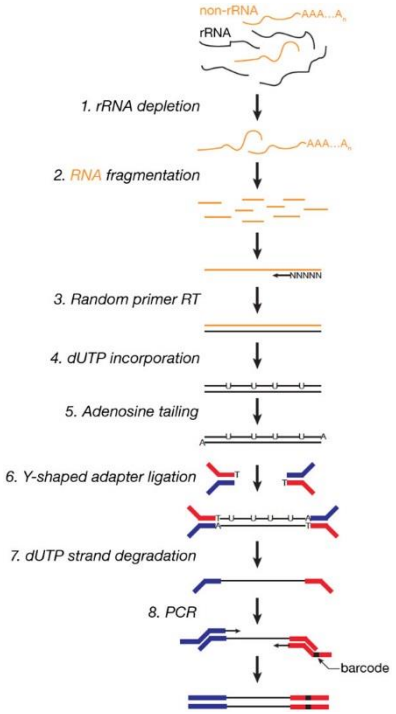
3. Anders S, Pyl PT, Huber W. HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166-9. doi: 10.1093/bioinformatics/btu638. PubMed PMID: 25260700.

4. Kallio MA, Tuimala JT, Hupponen T, Klemelä P, Gentile M, Scheinin I, et al. Chipster: user-friendly analysis software for microarray and other high-throughput data. BMC Genomics. 2011;12:507. doi: 10.1186/1471-2164-12-507. PubMed PMID: 21999641; PubMed Central PMCID: PMCPMC3215701.

Experiment Design
Strand specific protocol

- Why is strand information important?
- How is the stranded library made?





Nature Methods 7, 709–715 (2010)

Expression Values

Fragments (Reads) Per Kilobase of exon per Million mapped fragments

Nat Methods. 2008, Mapping and quantifying mammalian transcriptomes by RNA-Seq. Mortazavi A et al.

$$FPKMi = 10^6 \times 10^3 \times \frac{Ci}{NLi}$$

C= the number of fragments mapped onto the gene's exons

N= total number of (mapped) fragments in the experiment

L= the length of the transcript (sum of exons)