Whole transcriptome sequencing data analysis workshop

Quality control, trimming, alignment and quantification

Simone Tiberi, University of Zurich

1-8/02/2019

Introduction

- Most of the genome does not appear in RNA-seq data: in humans, genes only constitute approx. 3% of the genome.

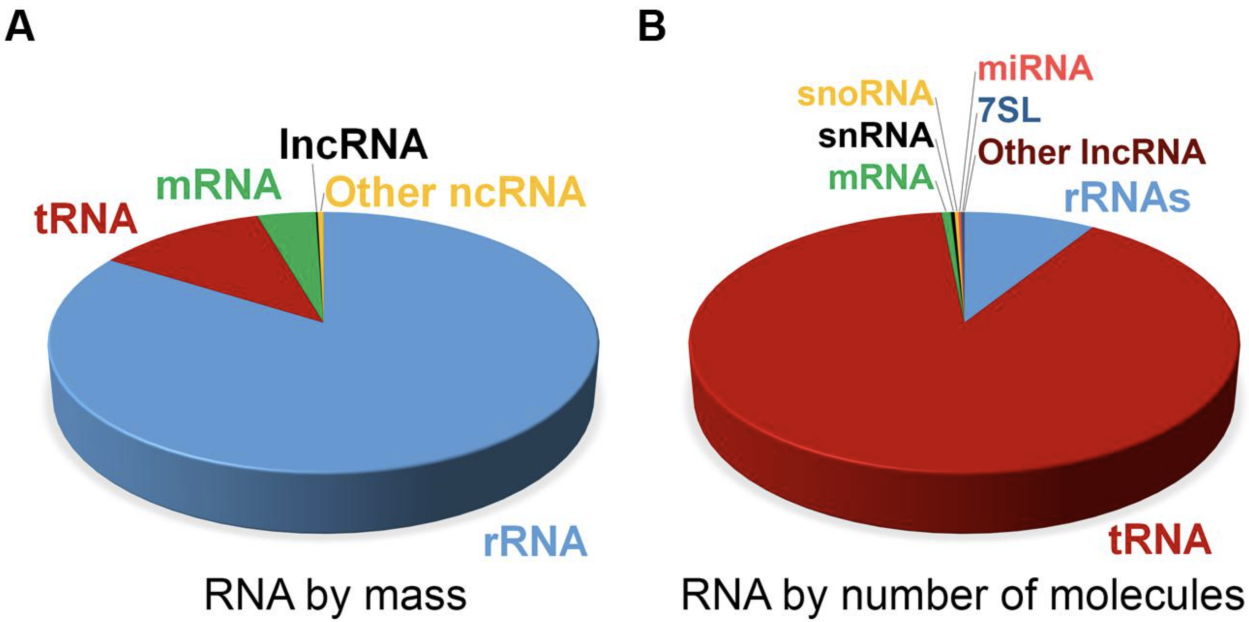


FIGURE 1. Estimate of RNA levels in a typical mammalian cell. Proportion of the various classes of RNA in mammalian somatic cells by total mass (A) and by absolute number of molecules (B). Total number of RNA molecules is estimated at roughly 10^7 per cell. Other ncRNAs in (A) include snRNA, snoRNA, and miRNA. Note that due to their relatively large sizes, rRNA, mRNA, and lncRNAs make up a larger proportion of the mass as compared to the overall number of molecules.

Palazzo & Lee, Front. Genet. 2015

Overview of sequencing technologies

- First generation: microarrays (mostly replaced by RNA-seq).
- **Next generation sequencing or second generation: RNA-seq (mostly Illumina).**
- Third generation: long-reads (mostly PacBio and Oxford nanopore). It is growing in parallel, mostly useful for de-novo assembly, new transcript discoveries and transcript level analyses.
- "Recent" developments of RNA-seq: single-cell RNA-seq (scRNA-seq). Similar to "bulk RNA-seq" (standard RNA-seq) but with a single-cell resolution, i.e. observations are available for single cells instead of aggregating expression over many cells.

Protocol: Single-end vs. paired-end reads

- Single-end:
  - reads are sequenced from one end only;
  - cheaper and faster;
  - reads are approx. 100/150 base pairs (bp) long.
- Paired-end (fragments):
  - reads are sequenced from both ends;
  - more expensive but the most popular protocol nowadays;
  - fragments are approx. 250-800 (bp) long, i.e. approx. 100/150 at each end plus a gap in between (of known length);
  - more accurate as it reduces the number of multi-mapping reads.

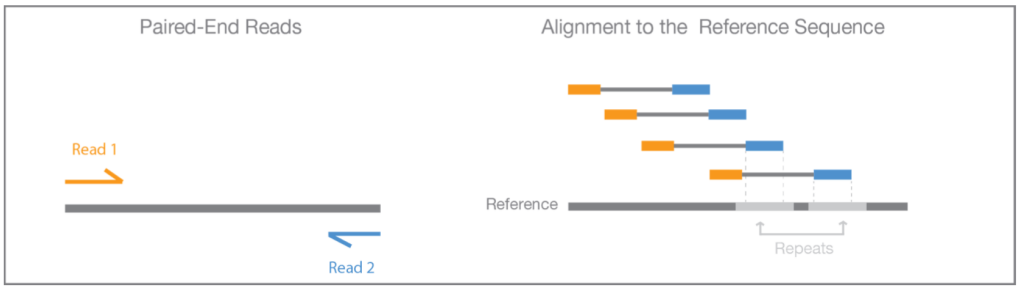


Figure 4: Paired-End Sequencing and Alignment—Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in better alignment of reads, especially across difficult-to-sequence, repetitive regions of the genome.

illumina, An introduction to Next-Generation Sequency Technology

Protocol: Stranded vs. unstranded

- Unstranded:
  - we don't know what strand reads come from.
- Stranded:
  - we know what strand reads come from:
  - it decreases the number of ambiguous reads, when reads fall in regions overlapping between multiple genes.
- Ideal scenario:
  - 150 bps long;
  - paired-end;
  - stranded;
  - "many" samples per condition.

FASTQ file

- The raw reads are stored in a FASTQ file.
- Each read is represented in 4 lines:
  - 1) a header for the read (starting with @);
  - 2) the sequence itself;
  - 3) a header for the quality (starting with +);
  - 4) a quality score as long as the sequence: each base of the sequence is associated to a score.

```
@HWI-ST1034:40:C08PJACXX:2:1101:20681:1994 1:N:0:ATCACG
CTCGNAGACTGGCAACTTGTTCTGGTTTACTGCACCTTCTTTTAAAGGCAGAAAGGC
+
CCCF#2ADHHHGHJJJIJJHIIIJJHIJJJJJJJJJBGIIJJJJJJJJJJJJJJIJ
```

Hubert Rehrauer, ETH Zurich

Phred score

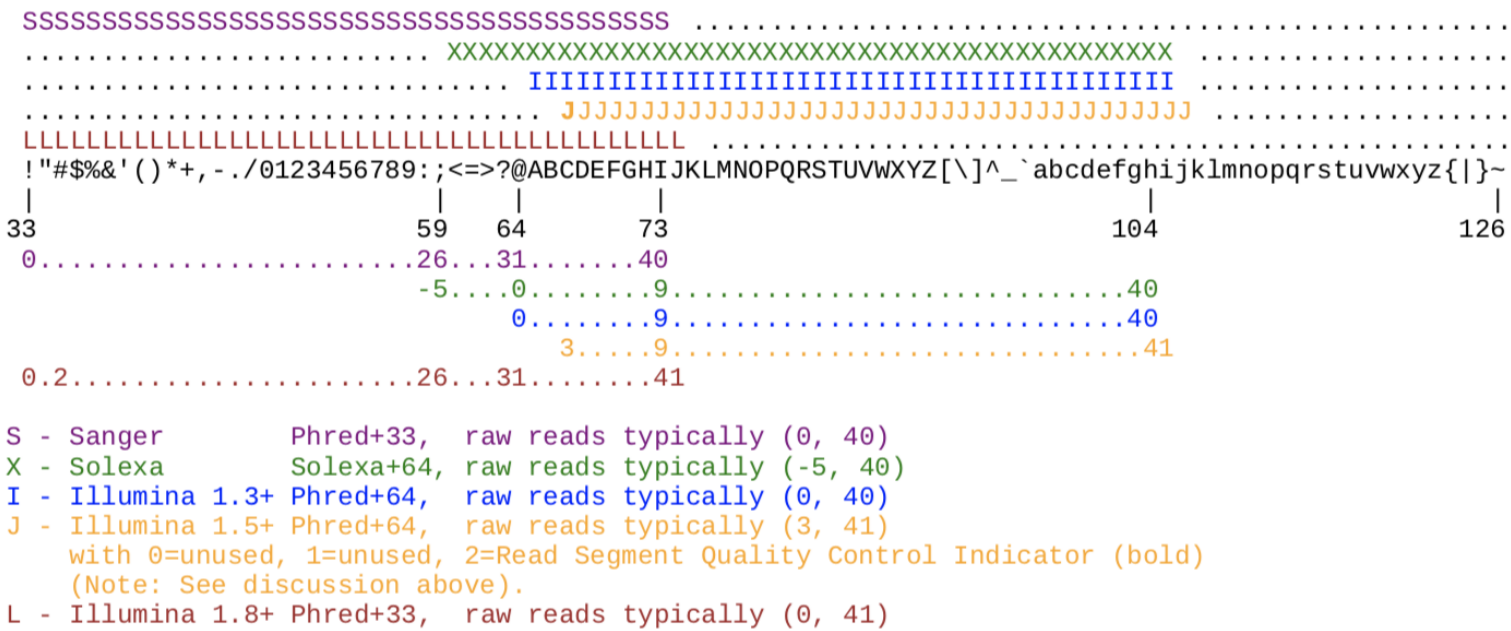
- The Phred score, Q, associates each base with the probability that the base is called incorrectly, P.
- Phred score: $Q = -10log_{10}P$.
- A Phred score of 30 or more is considered to be good enough: the error probability for a single base $< 0.1\%$.

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Wikipedia

Phred score

- The association between the Phred score and the characters of the FASQT file changes with the technology.
- For recent Illumina technologies, capital letters have good quality: error probability for a single base $< 0.1\%$.



Wikipedia

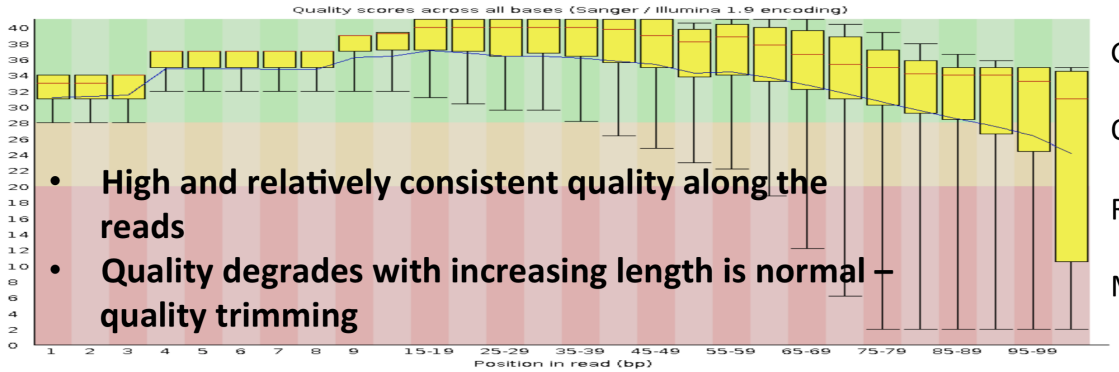
Assessing the quality of reads

- We can assess the quality of the raw reads, before aligning them to the genome/transcriptome, via **FastQC**.
- FastQC outputs one file per sample for single-end reads and two files for paired-end reads.
- To aggregate and jointly visualize the output of FastQC for multiple samples, you can use **MultiQC**.
- In next slides we will inspect the output from FastQC in individual samples.
- Nice interpretation of FastQC output in:
  `https://www.youtube.com/watch?v=GnWSXwQeJ_U`
  How to Check the Quality of Illumina Sequencing Reads with FastQC (Part 2).
- Interpretation of MultiQC output in:
  `https://www.youtube.com/watch?v=qPbIlO_KWNO`.
  Using MultiQC Reports.

FastQC

Per base sequence quality - FastQC

- Range of quality values across all bases at each position



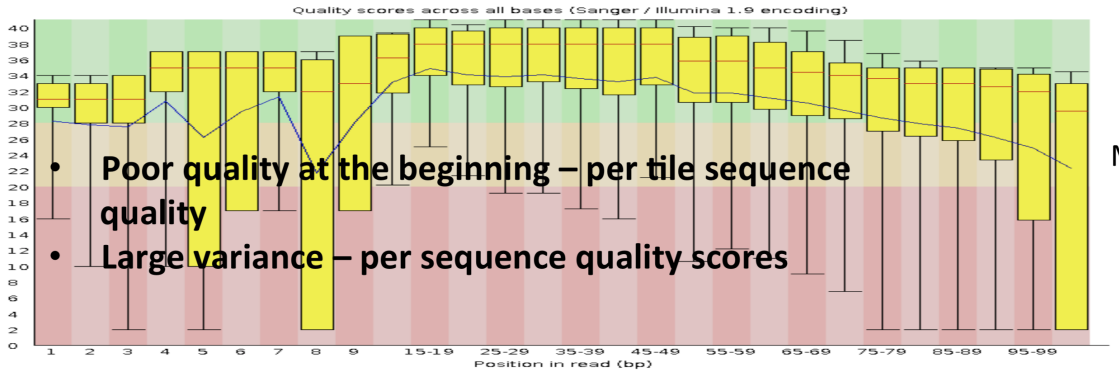
- **High and relatively consistent quality along the reads**
- **Quality degrades with increasing length is normal – quality trimming**

Green: >Q28, good

Orange: >Q20, reasonable

Red:<Q20, poor

Median > Q25



- **Poor quality at the beginning – per tile sequence quality**
- **Large variance – per sequence quality scores**

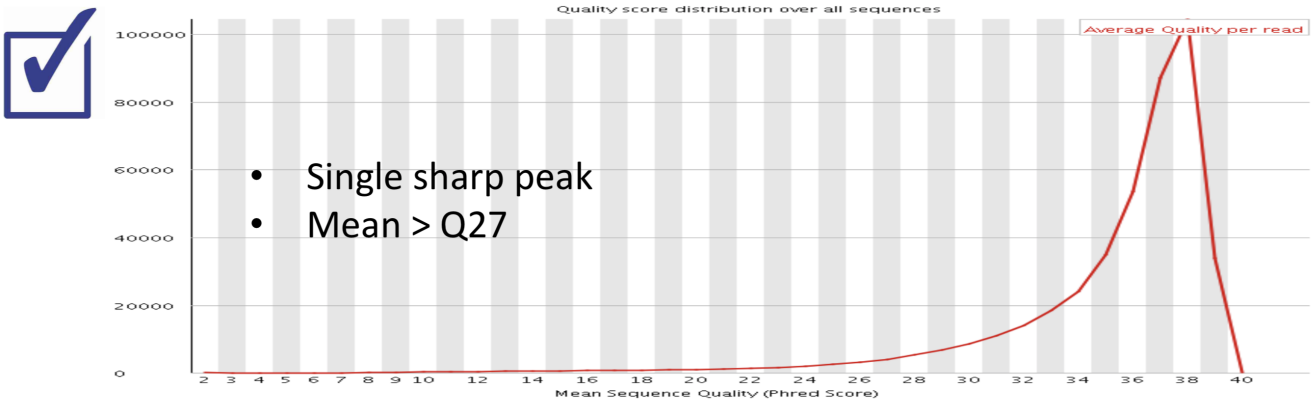
Median < Q20

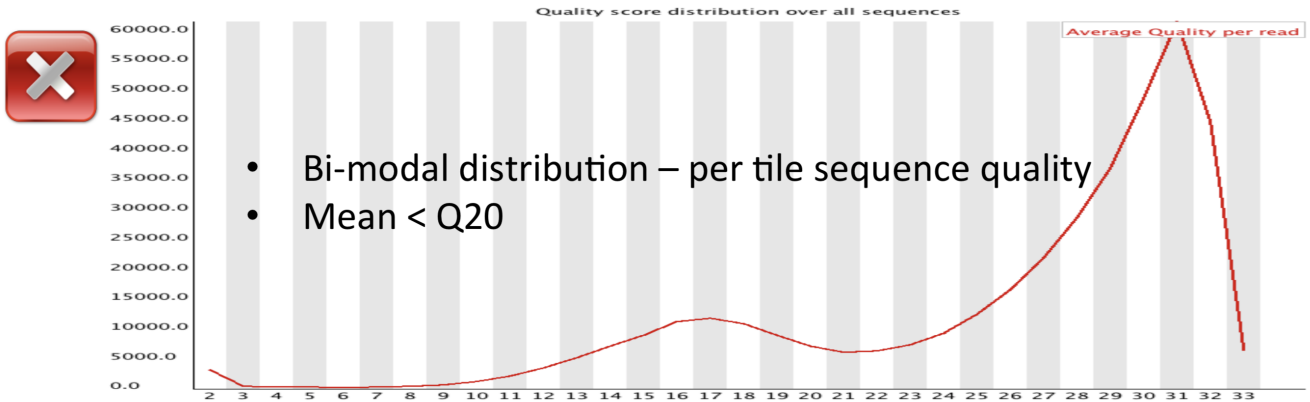
Hubert Rehrauer, ETH Zurich

FastQC

Per sequence quality scores - FastQC

- Subset of sequences with universally low quality values

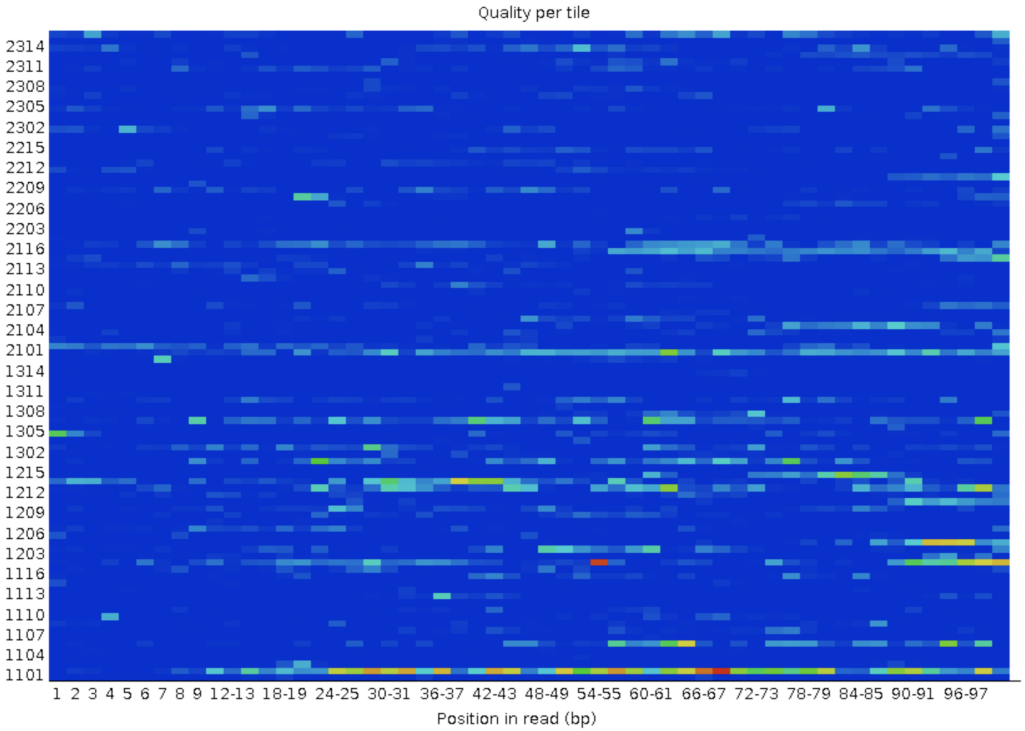




FastQC

Per tile sequence quality - FastQC

- Quality scores from each tile across all bases - loss in quality associated with only one part of the flowcell



Deviation from average quality

Cold colors: ≥ average

Hotter color: worse quality

Good: universal blue

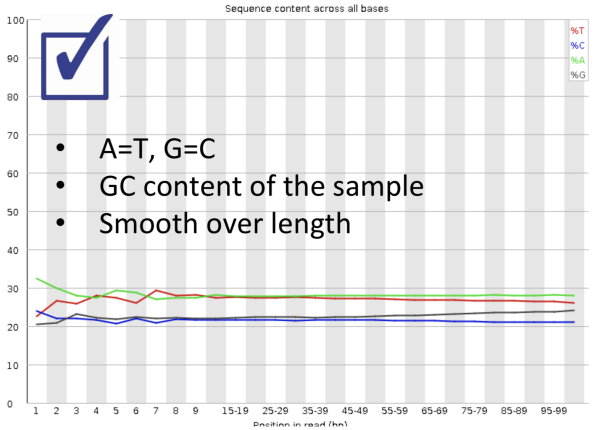
Failure: < average - 5

Hubert Rehrauer, ETH Zurich

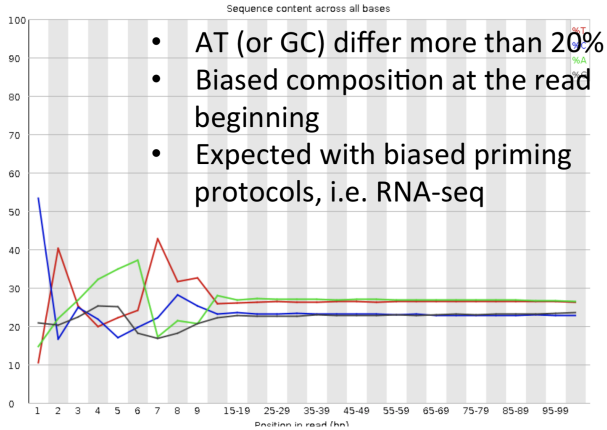
FastQC

Per base sequence content - FastQC

- The portion of A, T, G, and C at each position



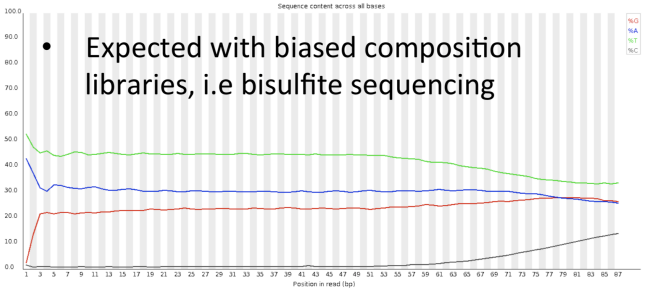
- A=T, G=C
- GC content of the sample
- Smooth over length



- AT (or GC) differ more than 20%
- Biased composition at the read beginning
- Expected with biased priming protocols, i.e. RNA-seq

Biases in Illumina transcriptome sequencing caused by random hexamer priming

Kasper D. Hansen[1,*], Steven E. Brenner[2] and Sandrine Dudoit[1,3]



- Expected with biased composition libraries, i.e bisulfite sequencing

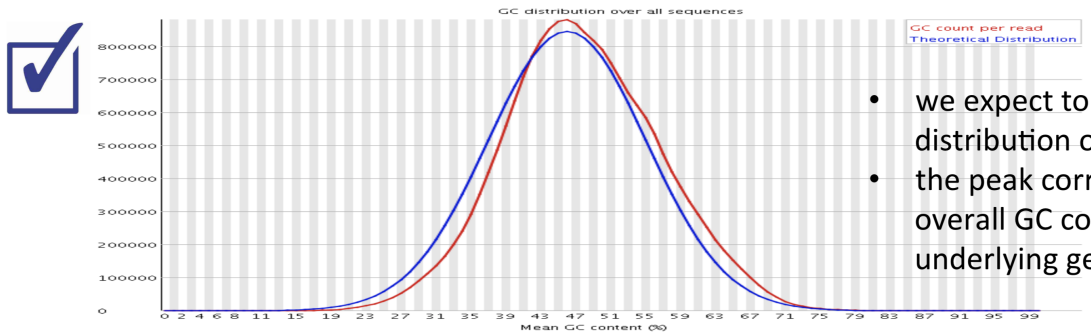
Treatment of DNA with bisulfite converts cytosine to uracil, but leaves methylated cytosine unaffected. Therefore, DNA that has been treated with bisulfite retains only methylated cytosines.

Hubert Rehrauer, ETH Zurich

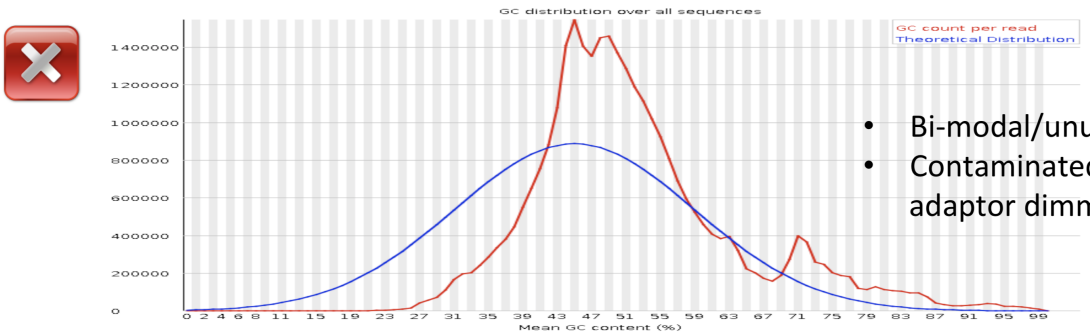
FastQC

Per sequence GC content - FastQC

- Distribution of average GC in all reads



- we expect to see a roughly normal distribution of GC content
- the peak corresponds to the overall GC content of the underlying genome



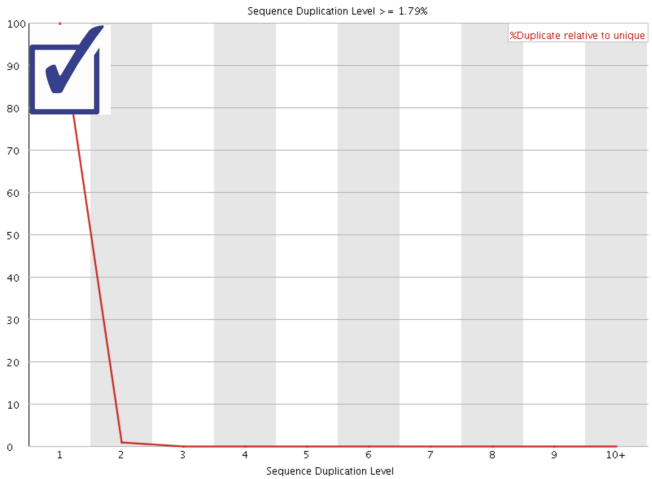
- Bi-modal/unusual distribution
- Contaminated/biased subset, i.e. adaptor dimmers, rRNA etc

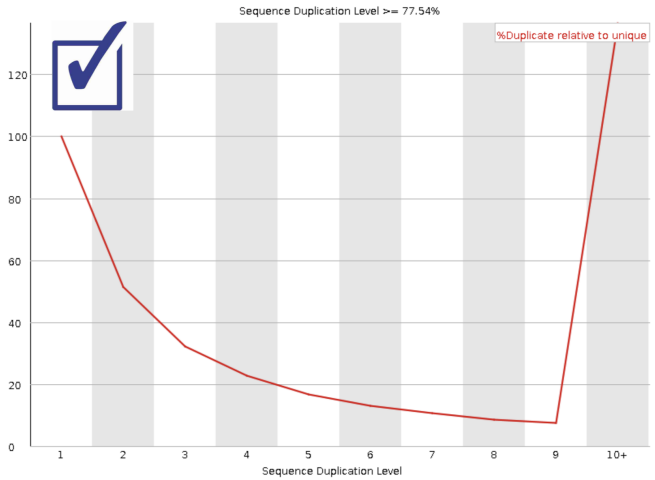
Hubert Rehrauer, ETH Zurich

FastQC

Sequence duplication - FastQC

- Relative number of sequences with different degrees of duplication





- Essentially no duplication

High duplication levels:

- DNA-seq: PCR over amplification, too little input material
- Normal in RNA-seq: high expression

Hubert Rehrauer, ETH Zurich

FastQC

Overrepresented sequences - FastQC

- Sequences make up >0.1 % of the total
- Compare those with a contamination database for finding contamination (i.e. adaptor dimmers)

Overrepresented sequences

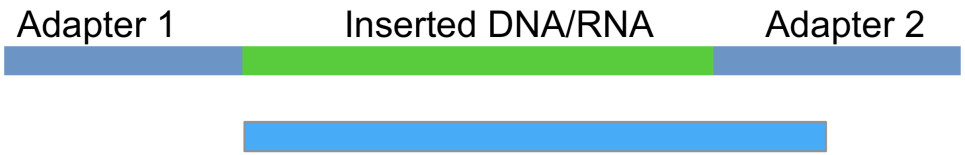
Sequence	Count	Percentage	Possible Source
GGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGCCGTC	75874	1.5613887498682963	TruSeq Adapter, Index 7 (100% over 50bp)
GGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTC	7636	0.15713900010536297	TruSeq Adapter, Index 2 (100% over 50bp)
GGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCCGTC	7539	0.1551428656095248	TruSeq Adapter, Index 5 (100% over 50bp)
GGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGCCGTC	5117	0.10530123933199874	TruSeq Adapter, Index 6 (100% over 50bp)

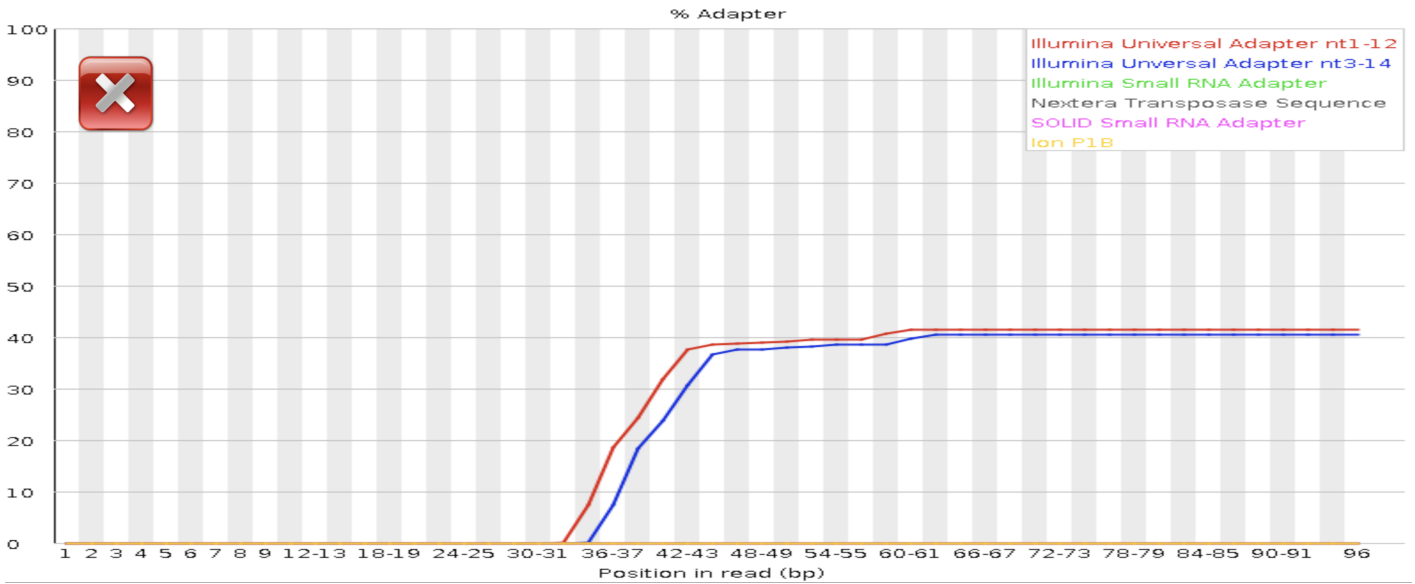
- Can be normal and biologically meaningful
  - highly expressed transcripts
  - high copy number repeats
  - Less diverse library (amplicons)

Hubert Rehrauer, ETH Zurich

FastQC

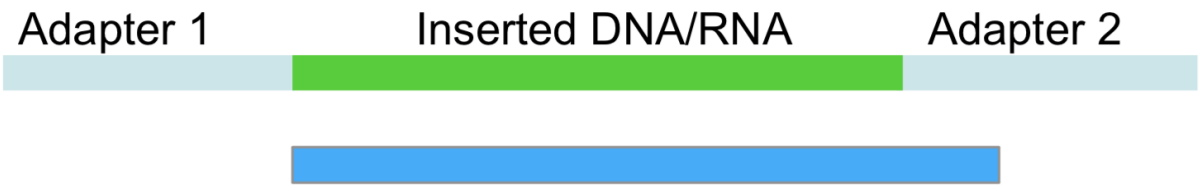
Adapter Content - FastQC





Hubert Rehrauer, ETH Zurich

FastQC



Hubert Rehrauer, ETH Zurich

- If the fragment of DNA/RNA is too short, our read could include part of the adapter.

FastQC

Data preprocessing common tasks

1. Trimming: remove bad bases from (end(s) of) reads
   - Adaptor sequence
   - Low quality bases
2. Filtering: remove bad reads
   - Low quality reads
   - Contaminating sequences
   - Low complexity reads (repeats)
   - Short (<20bp) reads – they slow down mapping software

Hubert Rehrauer, ETH Zurich

Trimming

- Trimming is always recommended.
- Trimming removes "low" quality bases from the end (and eventually the start) of reads and adaptor sequences; it also removes entire reads if, after trimming, they are too short (e.g., < 20 bps).
- Two popular tools for trimming reads are **trimmomatic** and **Trim Galore**.
- Trim Galore user guide: `https://github.com/FelixKrueger/TrimGalore/blob/master/Docs/Trim_Galore_User_Guide.md`
- We can cut the end of the reads when they fall below a minimum quality or we can trim after a fixed number of bases.
- After trimming, we need to check again for the quality of trimmed reads to make sure quality is now adequate.

QC examples

- Visualize FastQC and MultiQC outputs before and after trimming reads.

QC & trimming ex. code

FastQC:

```
fastqc -o out_directory sample_1_R1.fastq.gz
fastqc -o out_directory sample_1_R2.fastq.gz
```

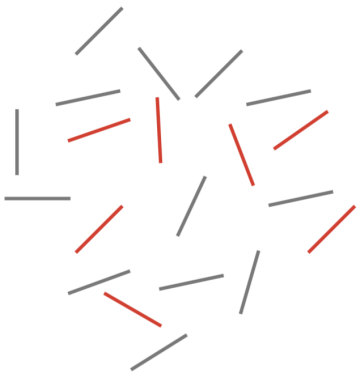
MultiQC:

```
cd out_directory
multiqc . –interactive
% multiqc will merge all fasqQC files present in the folder.
```

Trim Galore:

```
trim_galore -q 20 –phred33 –length 20 –illumina -o out_directory \
–paired sample_1_R1.fastq.gz sample_1_R2.fastq.gz –fastqc –gzip
% -q 20, bps with Phred scores < 20 will be trimmed
% –length 20, remove reads if, after trimming, they become shorted than
20 bps
% –illumina, use illumina standard adapters
% -o, directory where the output will be stored
% –paired, use paired end reads (if one read is removed, also remove the
other one)
% –fastqc, run fastqc on the outputed fastq files
% –gzip, compress the output file
```

Alignment



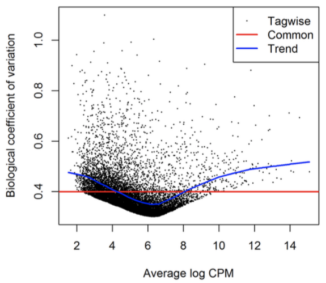
ALIGNMENT

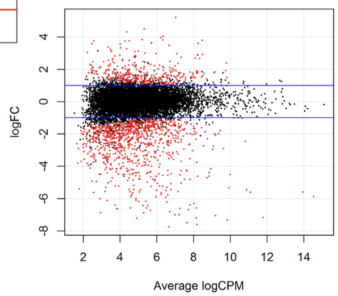
COUNTING

Gene A
Gene B
.
.
.
Gene X

$$\begin{pmatrix} 7 & \cdots \\ 13 & \cdots \\ \cdots & \cdots \\ \cdots & \cdots \\ \cdots & \cdots \end{pmatrix}$$

ANALYSIS





Charlotte Soneson, UZH

Genome alignment

- In most analyses, after checking the quality of our reads, we align them to a reference.
- We can use a full aligner to align reads to a reference genome (**STAR**, **TopHat**, ...).
- Full aligners attempt to align the entire read to a reference genome (splice-aware alignment).
- To align our reads to the genome we need:
  - a reference genome (DNA), usually in a fasta format;
  - a gene transfer format (GTF) file which contains the location of genes on the reference genome.
- Both can be downloaded from `http://www.ensembl.org/info/data/ftp/index.html` under the DNA and Gene sets columns.

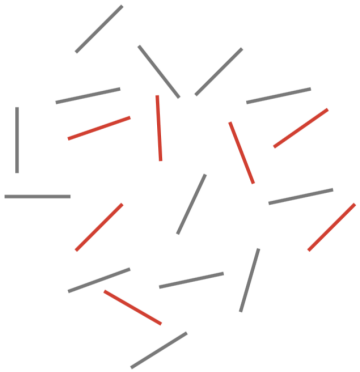
Transcriptome alignment

- Alternatively, we can align our reads directly to a reference transcriptome.
- Most transcript aligners are pseudo/quasi aligners (**Salmon**, **kallisto**, **RSEM**, ...).
- Pseudo/quasi aligners don't align the full reads, instead they use a low cost pseudo alignment:
  - from each read they create many k-mers (substrings of the read of k base pairs);
  - they map the k-mers to the reference transcriptome and check their compatibility with the transcripts.
- To align our reads to the transcriptome we need a reference transcriptome (cDNA) alone, which can be downloaded from `http://www.ensembl.org/info/data/ftp/index.html` under the cDNA column.

Considerations on the alignment

- Aligners take into account mismatches with respect to the reference, due to sequencing error (approx. 1/500) or mutations (approx. 1/10,000 in humans).
- A read could align in multiple positions of the reference genome/transcriptome, with the same or similar alignment scores. The output is represented by both unique aligning reads and multi mapping reads. Some reads (a minority) remain unmapped because they don't align well enough to any location of the genome.
- Alignment is an optimization problem: for every read, it looks for the alignment with the highest score.
- Aligners do not run a complete search of all possible alignments of all read, the optimization is heuristic, not optimal (yet very sophisticated).

Counting



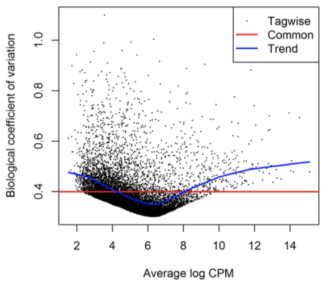
ALIGNMENT

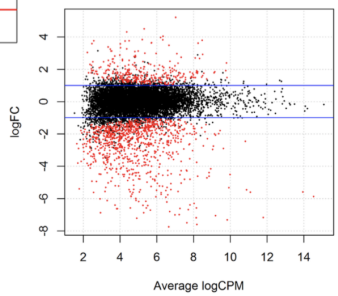
COUNTING

Gene A
Gene B
.
.
.
Gene X

$$\begin{pmatrix} 7 & \cdots \\ 13 & \cdots \\ \cdots & \cdots \\ \cdots & \cdots \\ \cdots & \cdots \end{pmatrix}$$

ANALYSIS

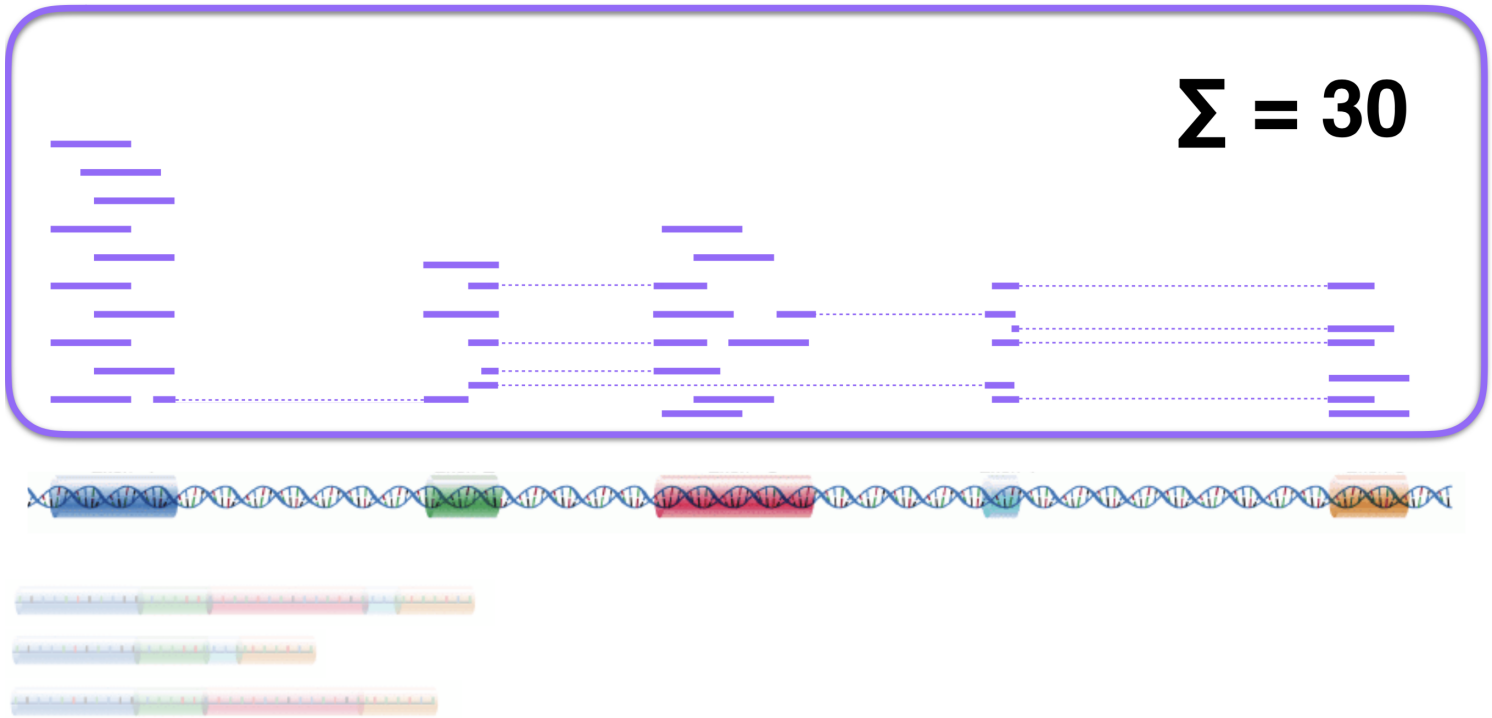




Charlotte Soneson, UZH

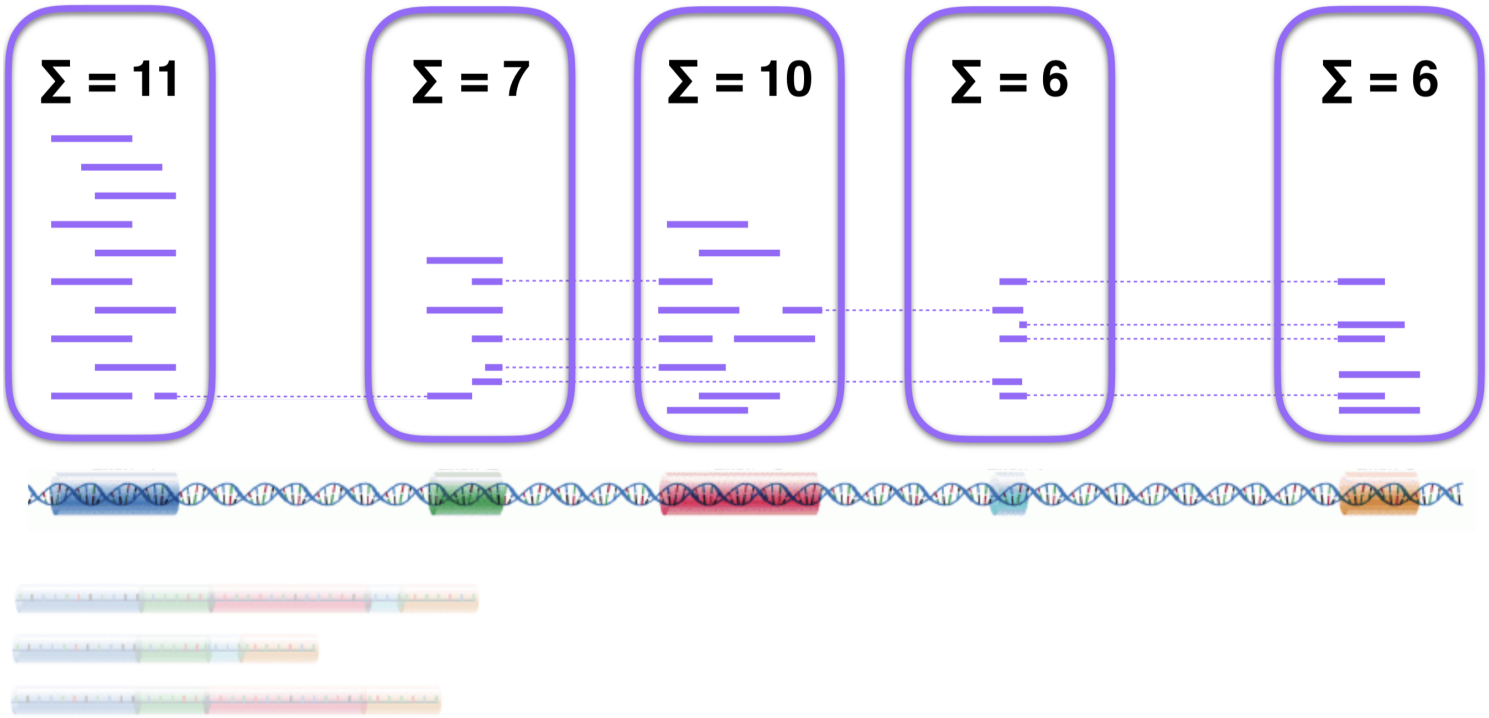
Quantification with genome aligners: gene counts

- Once the reads are aligned to the genome, we need to sort them (for STAR use the *SortedByCoordinate* option).
- We can then count how many reads overlap each gene.
- **STAR** and **TopHat** are the most popular genome aligners (we'll see STAR in the tutorial).
- Remember that RNA-seq reads only map to exons!



Quantification with genome aligners: exon counts

- The counts of each individual transcript cannot be obtained due to the overlapping regions between transcripts.
- To study the transcript, the attention sometimes shifts towards the exons, for which we can observe the counts.



Quantification with genome aligners: junction counts

- An alternative is to only consider junction counts: counts of reads that span over two exons (the ones with the dotted lines in the previous image).
- Pros:
  - we know what exons reads connect;
  - useful for discovering new transcripts with non annotated exon junctions (STAR also outputs non-annotated junction reads).
- Cons:
  - we still miss what transcript each read maps to (there could be > 1 transcripts associated to the two exons the reads spans over);
  - we only use a sub-set of the data.
- STAR outputs junction counts in the SJ.out.tab file.

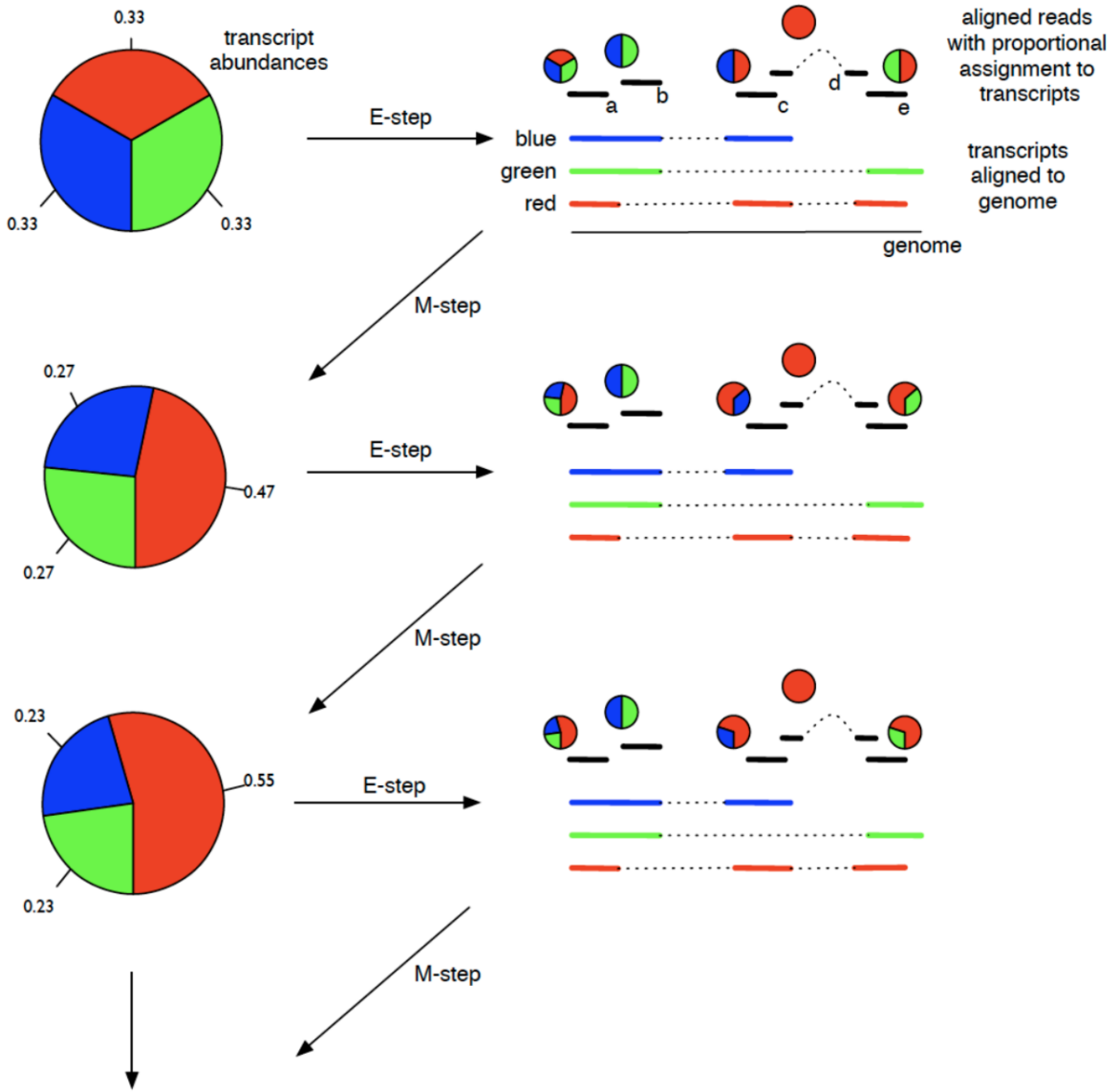
Quantification with transcriptome aligners

- Reads can be mapped directly to the transcriptome, skipping genome mapping.
- A well annotated transcriptome is required: if a transcript is missing in the reference, the reads coming from that transcript will either be mapped to other transcripts (if compatible), or remain unmapped.
- Multi-mapping reads are a lot more frequent when mapping to the transcriptome: for every exon, there can be multiple transcripts that contain it.
- Pseudo aligners don't actually map the full reads, but k-mers built from the reads.
- **Salmon** and **kallisto** are the most popular transcript aligners.

Quantification with transcriptome aligners

- After mapping reads to transcripts, an expectation-maximization (EM) algorithm is used to estimate the expected number of reads mapping each transcript.
- The algorithm outputs estimated transcript level counts, from which we can easily obtain gene level counts.
- Note that these numbers are estimates, not real counts.
- Salmon and kallisto provide bootstrap replicates to measure the uncertainty in the estimated counts.

Quantification with transcriptome aligners



Reference-free assembly

- An alternative to genome and transcript mapping is to use a reference-free approach and assemble the reads into transcripts, build a reference transcriptome and then map them back to the reference.
- Very challenging to reconstruct full transcripts from Illumina short reads, even if paired-end reads.
- **Trinity** is the most popular tool for reference-free assembly: Grabherr et al. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data, Nature Biotechnology.
- In this context it is highly suggested to have:
  - paired-ended,
  - stranded,
  - "long" RNA-seq reads (e.g., 150 bps).
- Trinity will use all reads from all samples (across all experimental conditions) when building the transcriptome.
- Rarely used in well studied organisms.

Trinity wiki

- Trinity wiki:
  `https://github.com/trinityrnaseq/trinityrnaseq/wiki`
  `https://github.com/trinityrnaseq/trinityrnaseq/wiki/Running-Trinity`
- Trinity tutorial: `https://biohpc.cornell.edu/lab/doc/trinity_workshop_part1.pdf`.

Importing the counts

- For each sample, the main output of the quantification will be a matrix of genes/transcripts and respective counts, i.e. the number of reads mapping to a gene/transcript.
- Remember that counts are discrete, not continuous.
- The estimated counts from Salmon or kallisto are continuous though because they are expected counts.
- We can import the counts from STAR/TopHat output (stored in a .bam or .sam file) in R via featureCounts (**Rsubread**), see tutorial.
- We can import the counts from Salmon/kallisto output (stored in a .bam or .sam file) in R via tximport (**tximport**), see tutorial.

Genome alignment ex. code I

% STAR requires a genome annotation (in a fasta format) and the location of the annotated transcripts on the genome (in a gtf file).
STAR:
% Generate genome index:
star –runMode genomeGenerate –runThreadN 20 –genomeDir genome_dir \
–genomeFastaFiles ensembl_Homo_sapiens.GRCh37.71.fa \
–sjdbGTFfile Homo_sapiens.GRCh37.71.gtf –sjdbOverhang 100
% –runMode genomeGenerate, to generate the genome index.
% –runThreadN 20, runs on 20 cores.
% –genomeDir genome_dir, specifies the output directory for the genome index.
% –sjdbGTFfile specifies the path to the file with annotated transcripts in the standard GTF format: STAR will extract splice junctions from this file and use them to greatly improve accuracy of the mapping. % While this is optional, and STAR can be run without annotations, using annotations is highly recommended whenever they are available. % sjdbOverhang should be equal to the (average) read length - 1 (i.e., 100 for 101 bp long reads).

Genome alignment ex. code II

STAR:
% Align (paired-end) reds; for each sample:
star –runMode alignReads –runThreadN 20 –genomeDir genome_dir \
–readFilesIn sample_1_R1_trim.fastq.gz sample_1_R2_trim.fastq.gz \
–outFileNamePrefix sample1 –outSAMtype BAM SortedByCoordinate
% STAR download: `https://github.com/alexdobin/STAR`
% STAR manual: `https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf`

Transcriptome alignment ex. code

% salmon documentation:
`https://salmon.readthedocs.io/en/latest/`
% salmon release:
`https://github.com/COMBINE-lab/salmon/releases`
% salmon only requires a transcript annotation (transcriptome.fasta).
Salmon:
% Build the index:
salmon index -i index_dir -t transcriptome.fasta -p 30 –type quasi -k 31
% For every sample:
salmon quant -i index_dir -l A \
-1 sample_1_R1_trim.fastq.gz -2 sample_1_R2_trim.fastq.gz \
-p 20 -o sample1_out_dir –dumpEq –numBootstraps 100 \
–seqBias –gcBias
% -p 20 specifies that salmon will use 20 cores
% –numBootstraps 100, computes 100 bootstrap replicates (used by some differential tools)
% -l A specifies the library type (A = authomatic detection, if unknown).
% –seqBias and –gcBias are options to correct for sequence-specific and GC content biases.

De novo assembly ex. code

% Trinity relies on salmon and bowtie2: both need to be installed.
Trinity:
Trinity –seqType fq –max_memory 400G \
–left sample_1_R1_trim.fastq.gz,sample_2_R1_trim.fastq.gz \
–right sample_1_R2_trim.fastq.gz,sample_2_R2_trim.fastq.gz \
–CPU 40 –output out_dir –SS_lib_type RF
% By default, reads are treated as not strand-specific.
% –SS_lib_type, specifies the strand orientation (RF = reverse forward).
% –genome_guided_bam: genome guided mode, in case a (reliable) genome is available.

References I

- **RNA-seq overview**: Koen Van den Berge et al. (2018). RNA sequencing data: hitchhiker's guide to expression analysis, PeerJ Preprints.
- **RNA-seq overview**: Conesa at el. (2016). A survey of best practices for RNA-seq data analysis, Genome Biology.
- **STAR**: Dobin et al. (2013). STAR: ultrafast universal RNA-seq aligner, Bioinformatics.
- **TopHat**: Trapnell et al. (2009). TopHat: discovering splice junctions with RNA-Seq, Bioinformatics.
- **Salmon**: Patro et al. (20017). Salmon provides fast and bias-aware quantification of transcript expression, Nature methods.
- **kallisto**: Bray et al. (2016). Near-optimal probabilistic RNA-seq quantification, Nature biotechnology.
- **RSEM**: Li and Dewey (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC Bioinformatics.

References II

- **Trinity**: Grabherr et al. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data, Nature Biotechnology.
- **tximport**: Soneson et al. (2016). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences, F1000Research.

Questions?