UNIVERSITY OF OXFORD

NUFFIELD DEPARTMENT *of* MEDICINE

Alignment

Irina Pulyakhina

WTCHG, RNA-Seq course, 25-April

What does “alignment” mean

Alignment == mapping.

To align NGS **reads** to a **reference sequence** means to find the locations of the reads on the reference.

Reference sequence

(taking human reference sequence as an example)

- one long string of nucleotides per chromosome;
- non-conventional chromosomes (“GL000192.1”) ;
- data from just **one individual** (important for SNP calling).

Concept

Mapping short (hundreds of nucleotides) reads to long (mln of nucleotides) reference sequences.

Indexing a reference sequence:
 - prior to the alignment, subdivide reference sequence into short seeds and store in a table (has to be done once);
- during the alignment, compare the table of short reads with the table of short reference seeds.

Why do we need alignment at all?

- quality control purposes (technical +biological)

- downstream analysis (quantify gene abundance; identify SNPs; look at allelic imbalance)

How does alignment work

```
...ACTTTTACCGACGCA...


       TTTT
```

```
...ACTTTTACCGACGCA...
     TTTT
       ✓
```

```
...ACTTTTACCGACGCA...
   TTTT
     ✗
```

```
...ACTTTTACCGACGCA...
    TTTT
     ✗
```

How does alignment work

```
...ACTTTTACCGACGCA...


       TTTT
```

```
...ACTTTTACCGACGCA...
     TTTT
        ✓
```

```
...ACTTTTACCGACGCA...
   TTTT
     ✗
```

```
...ACTTTTACCGACGCA...
    TTTT
     ✗
```

Sometimes we will continue looking for matches:

```
...ACTTTTACCGACGCA...
      TTTT ✗
       TTTT ✗
```

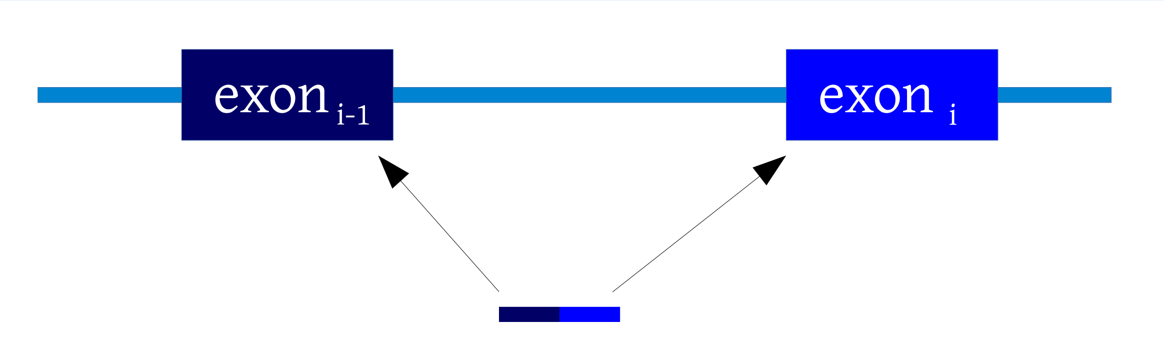
Aligning RNA to the genome

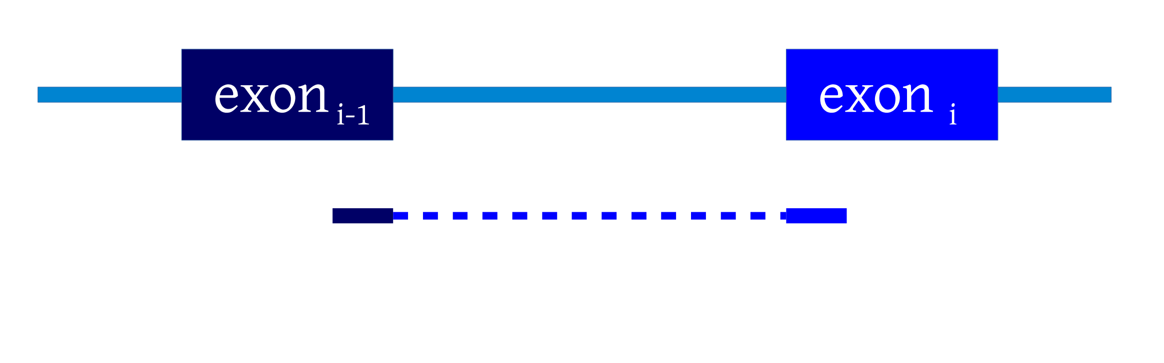
RNA reads are most commonly mapped to the genome, not to the transcriptome.

This happens because each organism has one genome.

Trasncriptomes, on the other hand, are rather unstable and depend on cell type/condition/age/...

Splitting reads



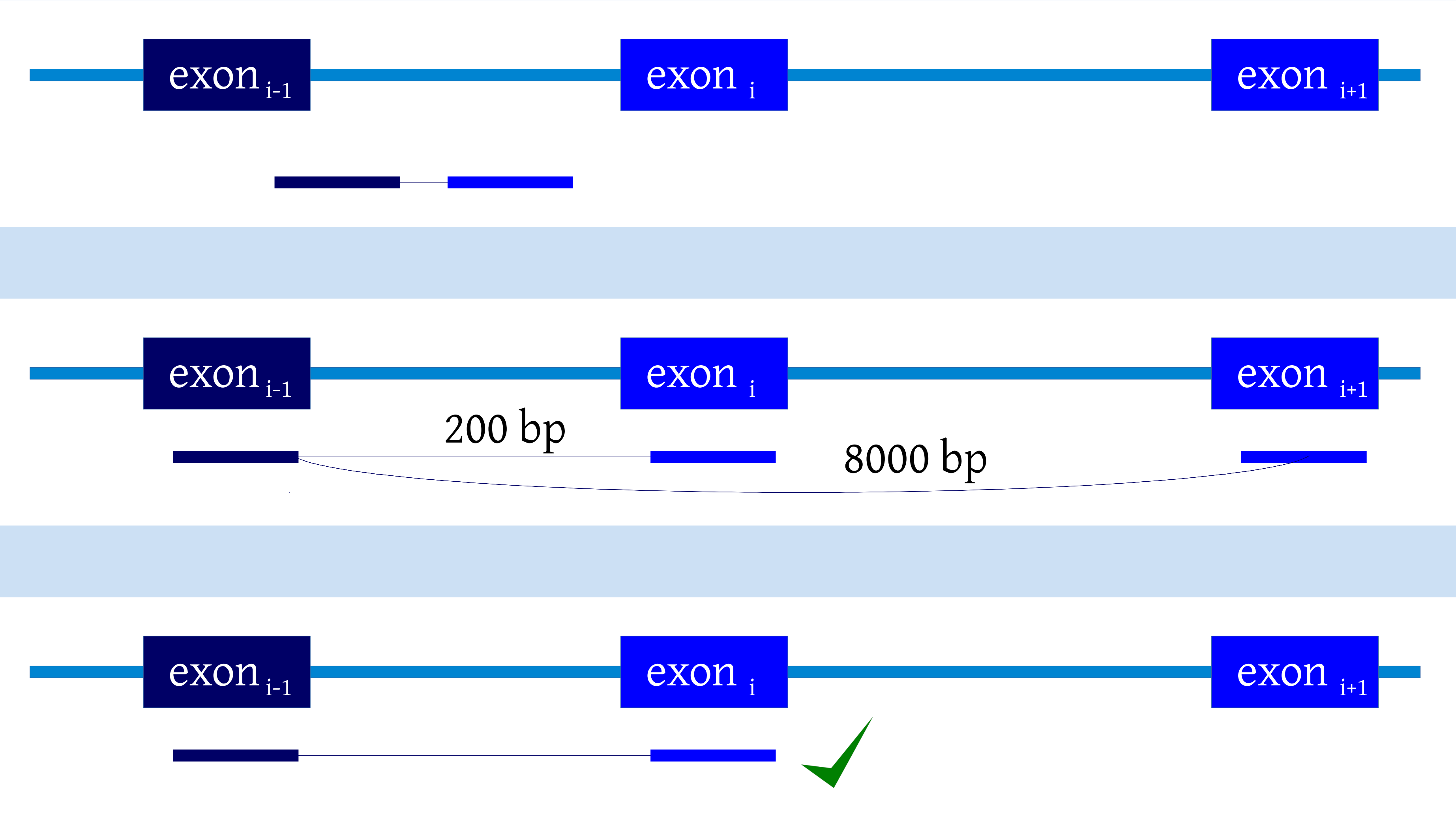


- split reads, but keep the link between the pieces;

- map the two pieces independently;

- split read == exon-exon junction (intron location).

Paired-end reads



Lecture “Alignment”, Irina Pulyakhina, WTCHG, RNA-Seq course, 25-April

Types of alignment

- unique mapping

- multiple hits
  - report all hits
  - report N first hits
  - report N random hits

Types of alignment

- unique mapping

- multiple hits
    - report all hits
    - report N first hits
    - report N random hits

!! “Unique mapping” and “report the first hit” are not the same!! (not always clear in the aligner spec).

When do we use unique/multiple mapping

- unique mapping ← almost always

- multiple hit ← repetitive regions
  regions with high similarity
  gene copies (result of transposition)
  low quality data

Errors/mistakes

```
...ACTTTTACCGACGCA...
     TTAT
```

mismatch

Is this a SNP or a sequencing error? (The answer is in the number of reads containing this mismatch).

```
...ACTT-TTACCGACGCA...
     TTATT
```

insertion/deletion

Is “true” nucleotide in the reference or in the sequencing data? (The answer can be in dbSNP).

```
...ACTTtcagcgacgtgc...ctagctagctataTTACCG...
     TT------------...-------------TT
```

big insertion (intron)

Alignment strategies (1)

First strategy a.k.a. building ***coverage islands****:*

1. trying to map one of the ends without splitting;
2. splitting reads unmapped in p.1;
3. mapping split pieces from p.2 in the vicinity of reads from p.1.

Example: Tophat/Tophat2

Alignment strategies (2)

Second strategy a.k.a. ***seed-based*** *strategy:*

1. splitting both ends of a fragment;
2. mapping the table of split fragments to the table of indexed reference sequence;
3. joining split ends.

Example: GSNAP

Pros and cons of each alignment strategy

- Both produce very similar results (up to 95% overlap).

- Coverage island strategy works better when analyzing pre-mRNA data with high intron content.

Alignment format

SAM format: **S**equence **A**lignment **M**ap

BAM format: **B**inary S**AM** (for the sake of space)

SAMtools, BAMtools

Specification:
https://samtools.github.io/hts-specs/SAMv1.pdf

SAM file (1)

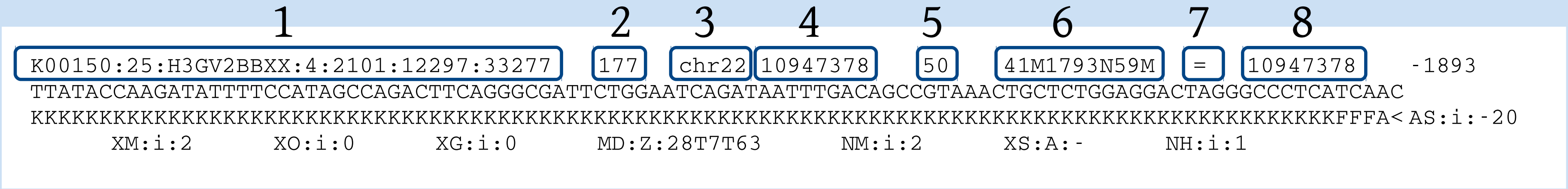
Header – information about the reference.

Contains one line per alignment!

- For one paired-end fragment, there will be at least two lines (one line for each end).

- Multiple alignments for one read == multiple lines in a SAM file.

SAM file (1)



```
K00150:25:H3GV2BBXX:4:2101:12297:33277  177  chr22 10947378   50   41M1793N59M  =  10947378    -1893
TTATACCAAGATATTTTCCATAGCCAGACTTCAGGGCGATTCTGGAATCAGATAATTTGACAGCCGTAAACTGCTCTGGAGGACTAGGGCCCTCATCAAC
KKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKFFFA< AS:i:-20
      XM:i:2      XO:i:0      XG:i:0      MD:Z:28T7T63      NM:i:2      XS:A:-      NH:i:1
```

1 – Read name

2 – Bitwise flag

3 – Reference sequence name (chromosome name)

4 – Leftmost position of the alignment

5 – Mapping quality

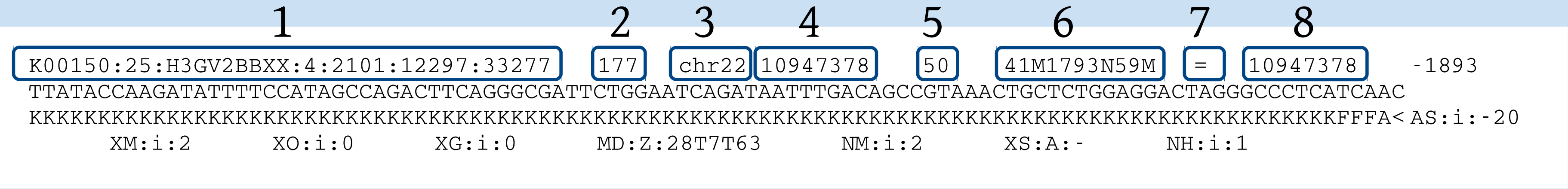
6 – CIGAR string

7 – Ref. name for the alignment of the second end

8 – Leftmost position of the second end

…

SAM file (2)



```
 K00150:25:H3GV2BBXX:4:2101:12297:33277   177   chr22 10947378    50    41M1793N59M   =   10947378    -1893
 TTATACCAAGATATTTTCCATAGCCAGACTTCAGGGCGATTCTGGAATCAGATAATTTGACAGCCGTAAACTGCTCTGGAGGACTAGGGCCCTCATCAAC
 KKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKFFFA< AS:i:-20
         XM:i:2        XO:i:0        XG:i:0        MD:Z:28T7T63        NM:i:2        XS:A:-        NH:i:1
```

6 – CIGAR string

example:

41M1793N59M

M – match
X – mismatch
I – insertion
D – deletion
N – big insertion (defined only for RNA aligners)

Post-alignment quality control (1)

- number of mapped reads (> 80%)

- number of uniquely mapped reads (> 70%)

- number of non-duplicates (> 70%)

- number of properly-paired reads (varies)

- check for contamination:
  number of reads mapped to chrY

Duplicates in RNA-Seq

A **duplicate** is an exact copy of a read.

```
...ACTTTTACCGACGCA...
      TTTT
      TTTT
```

Duplicates are often identified based on the location and regardless of the sequence itself.

Why do we observe duplicates?

In theory, when we are fragmenting the pool of RNA to load it in a sequencer, we can chop two identical transcripts at identical locations. This will produce two identical reads.

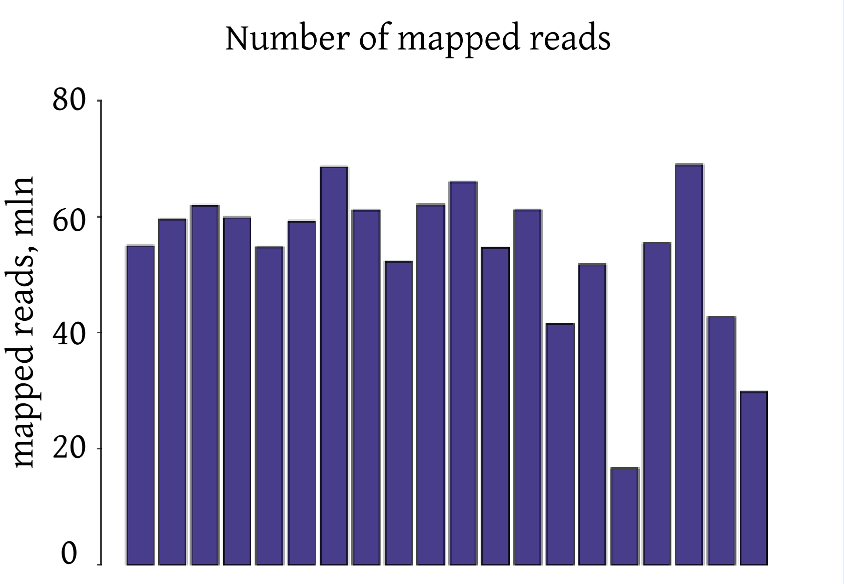
During the amplification, a fragment can be duplicated and produce identical reads.

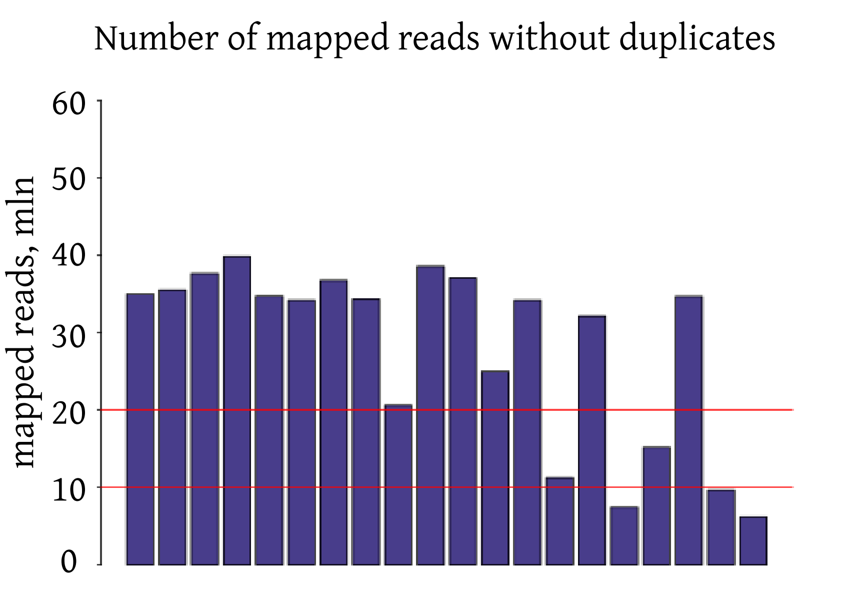
What to do with duplicates

- ChIP-Seq, DNase-Seq, ATAC-Seq, WEX – always remove duplicates.

- RNA-Seq – opinions vary.

Number of mapped reads





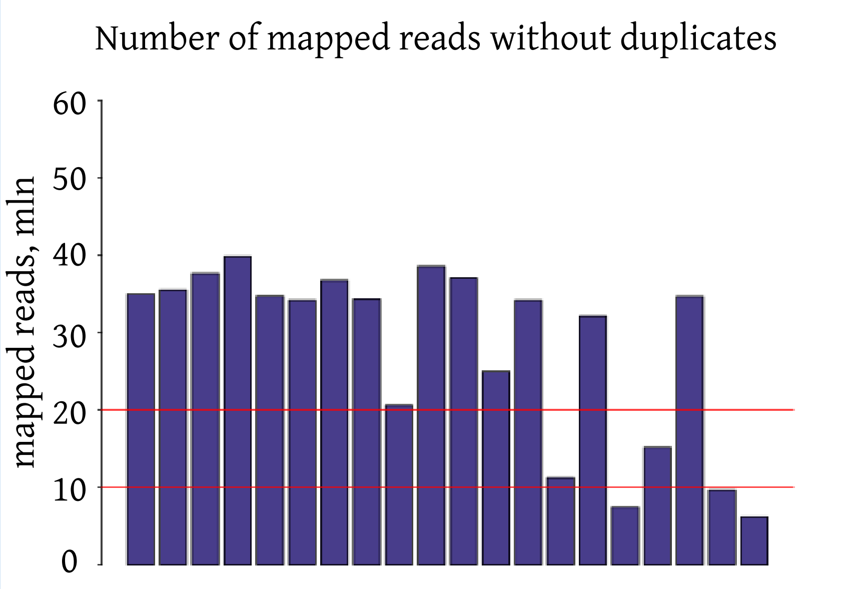
Number of reads remaining after all filtering steps should be > 20 mln.

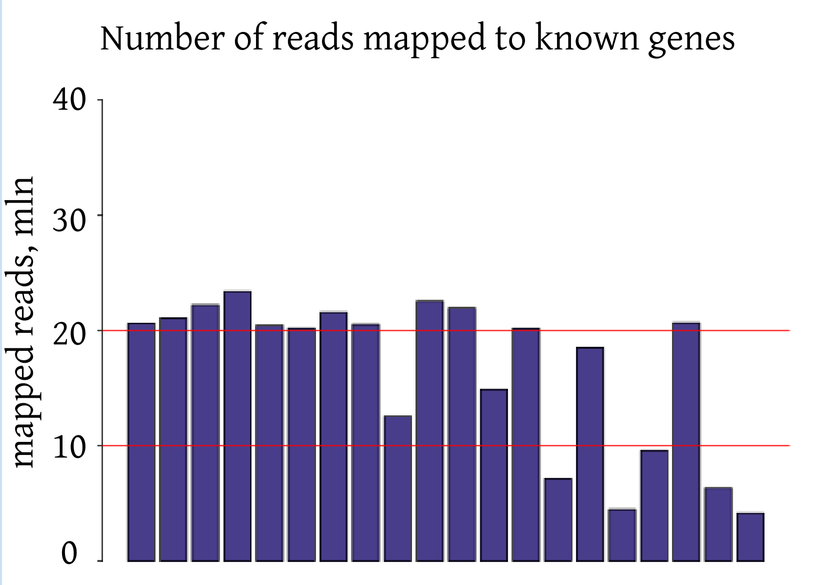
Post-alignment quality control (2)

- number of reads mapped to known genes:
  estimate the contaminations with rRNA
  (poly-A capture did not work well?)

- number of reads mapped to known exons
  estimate the contamination with intronic RNA
  (DNA? Pre-mRNA? DNase treatment did not
  work well?)

Number of reads mapped to known genes





Potential rRNA contamination?
Potential DNA contamination?

Biases in alignments

- mappability bias

- sequencability bias

- 5'-3' bias

- reference bias