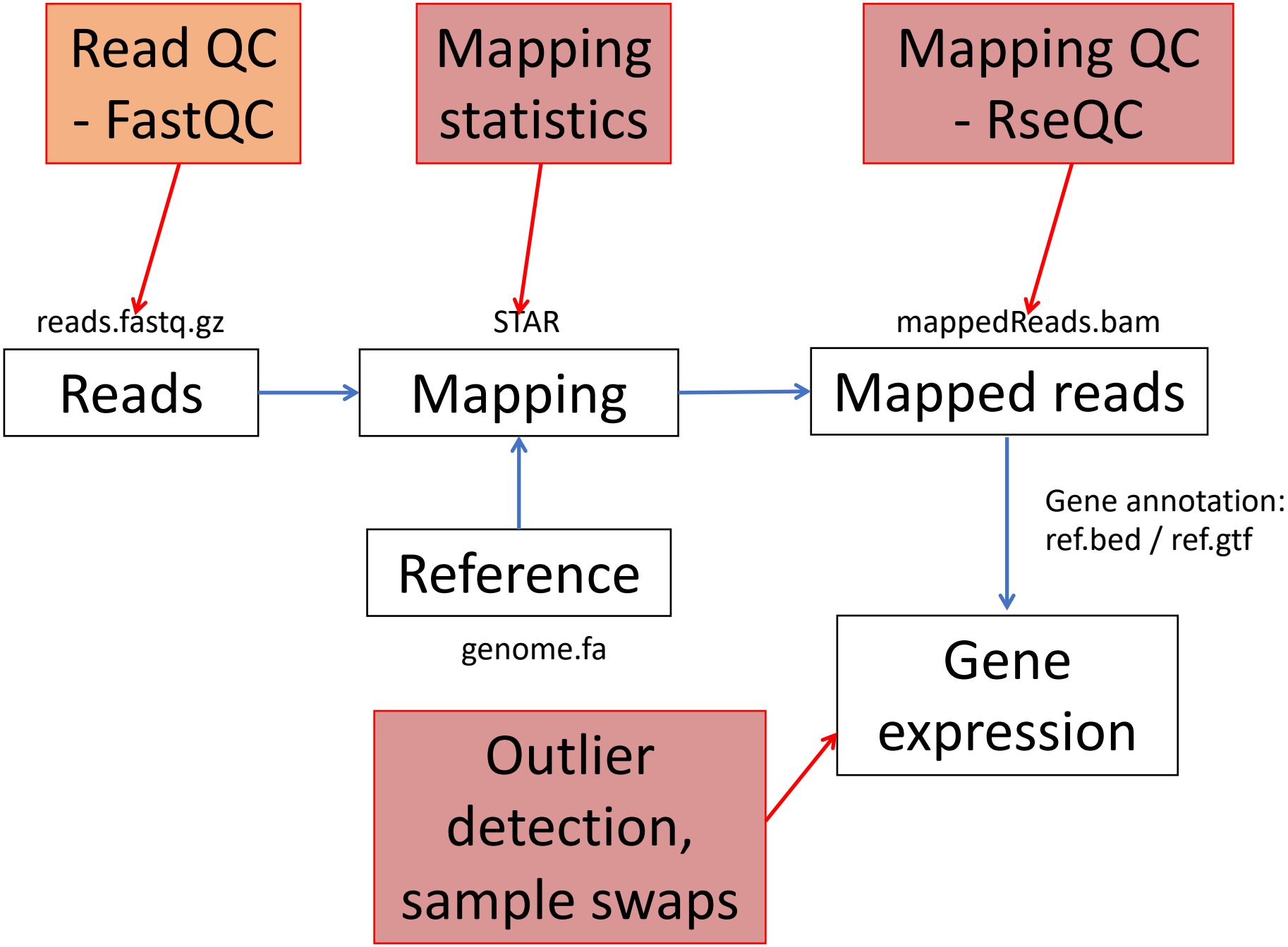
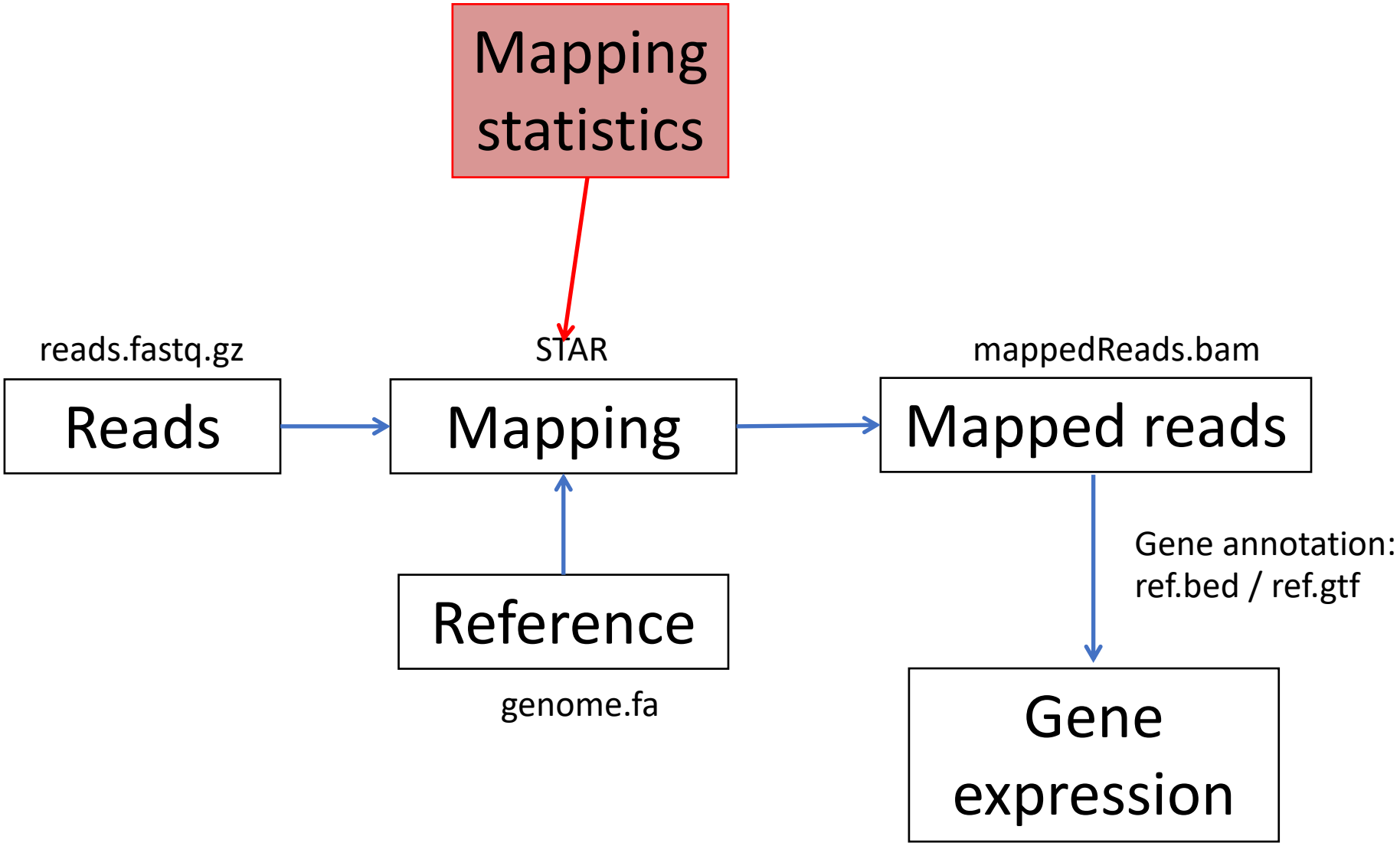
After mapping QC

RNA-seq data analysis

Johan Reimegård | 13-May-2019


Read QC
- FastQC
Mapping
statistics
Mapping QC
- RseQC
reads.fastq.gz
STAR
mappedReads.bam
Reads
Mapping
Mapped reads
Gene annotation:
ref.bed / ref.gtf
Reference
genome.fa
Gene
expression
Outlier
detection,
sample swaps

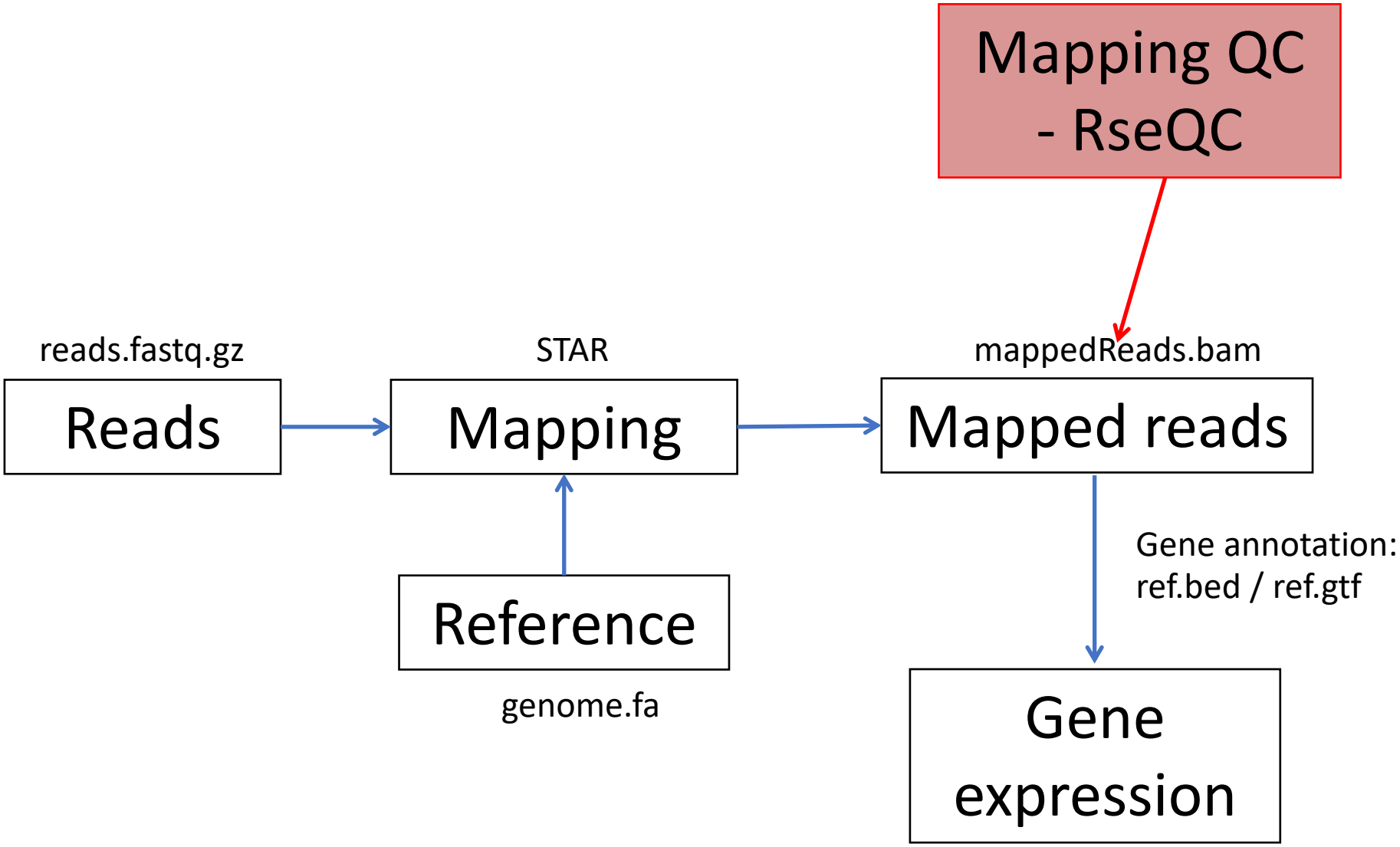

Mapping statistics
reads.fastq.gz
Reads
STAR
Mapping
mappedReads.bam
Mapped reads
Gene annotation:
ref.bed / ref.gtf
Reference
genome.fa
Gene expression

Mapping logs – mapping efficiency

- Program specific how the output will be (STAR, Bowtie, BWA, Tophat…)
- Always gives:
  - % uniquely mapping – ideally around 90% for 100 bp reads
  - % multi-mapping – will depend on read length
  - % unmapped – could indicate contaminations, adaptors
- Also statistics on:
  - Mismatches / indels
  - Splice junctions

Bad mapping – what to do?

- First step – try to figure out why it failed. With the use of FastQC/RseQC/Mapping logs.
  - Perhaps also look for contaminant species
  - Redo library prep controlling for possible errors
- Low mapping, but not completely failed.
  - Figure out why!
  - Is it equal for all samples?
  - Could it introduce any bias in the data?


Mapping QC
- RseQC
reads.fastq.gz
Reads
STAR
Mapping
mappedReads.bam
Mapped reads
Gene annotation:
ref.bed / ref.gtf
Reference
genome.fa
Gene
expression

SAM/BAM file formats

- All mapped reads with location in genome, mapping information etc.
- SAM (Sequence Alignment/Map) format – alignment.sam
- BAM is a compressed sam format – alignment.bam
- A bam-file (always) needs to be indexed and sorted - alignment.bam.bai
- Samtools – a simple program for converting between bam/sam, indexing, sorting, filtering, etc.

Code

```
$ module load bioinfo-tools
$ module load samtools
```

SAM/BAM file format

```
HWI-ST1018:7:1101:1648:2188#0   99      chr1    115275270       255     1S100M  =       115275321       152
NTTCTATATTGGTTGCTCGCTCTAATTTGTCACGTCGGTCTGTTGAAATATTAAACCTAACATGGTCACCTTCCAGCAGGGTCACCTTGGATTTCGTATCT
BS\ccceeeggeggghhhhbghhhhhhhhhhfhhhhhhhfhhhhhhhffhhhffhhhfffgghhgg\Z^ddeeeedbdbdcacbabcbcccbbbcc^abbc]   NH:i:1
HI:i:1  AS:i:194        nM:i:0
HWI-ST1018:7:1101:1648:2188#0   147     chr1    115275321       255     101M    =       115275270       -152
AAACCTAACATGGTCACCTTCCAGCAGGGTCACCTTGGATTTCGTATCTTTGTCTCCAAAGGGAAGTTCTTTAGGGATCACAAAGTCNANTTTGNTNNGTC
BBccbdcccccccccbbccccccddcddeeeeccgggggghihiiiiifhihfgiiihhhhihhihiiiihhiihiiiiihihhihgd]RBRBec]QBQBBbbb   NH:i:1
HI:i:1  AS:i:194        nM:i:0
HWI-ST1018:7:1101:2039:2206#0   99      chr19   14574483        255     1S72M85N28M     =       14574529
232     NCCTTCCGCAACCCTGTCATTGAGAGGATTCCTCGGCTCCGACGGCAGAAGAAAATTTTCTCCAAGCAGCAAGGGAAGGCGTTCCAGCGTGCTAGGCAGAT
BP\ccceefgggghhighiiiiiiiiihhhiiiiihiiiiiiihiggeeeddddbbbcccbbccb^[`aacccccccccccX]accccc^acc]bc^b_a]   NH:i:1
HI:i:1  AS:i:203        nM:i:0  XS:A:+
HWI-ST1018:7:1101:2039:2206#0   147     chr19   14574529        255     26M85N75M       =       14574483
-232    GAAGAAAATTTTCTCCAAGCAGCAAGGGAAGGCGTTCCAGCGTGCTAGGCAGATGAACATCGATGTCGCCACGTGGGTGCGGCTGCTCCGGAGGCTCATCC
]ccdcccbbbacbcaccccbcccccccccccbcccaccccccccdd_dddeeeeeeeggggiiiiihihiiiiiiiihiihiiiiiiigfgggeeeeebbb   NH:i:1
HI:i:1  AS:i:203        nM:i:0  XS:A:+
```

More details on:
http://samtools.github.io/hts-specs/SAMv1.pdf
http://genome.sph.umich.edu/wiki/SAM

After mapping - RseQC package

- General sequence QC:
  - sequence quality
  - nucleotide composition bias
  - PCR bias and
  - GC bias
- RNA-seq specific QC:
  - evaluate sequencing saturation
  - mapped reads distribution
  - coverage uniformity
  - strand specificity
  - Etc..
- Some tools for file manipulations

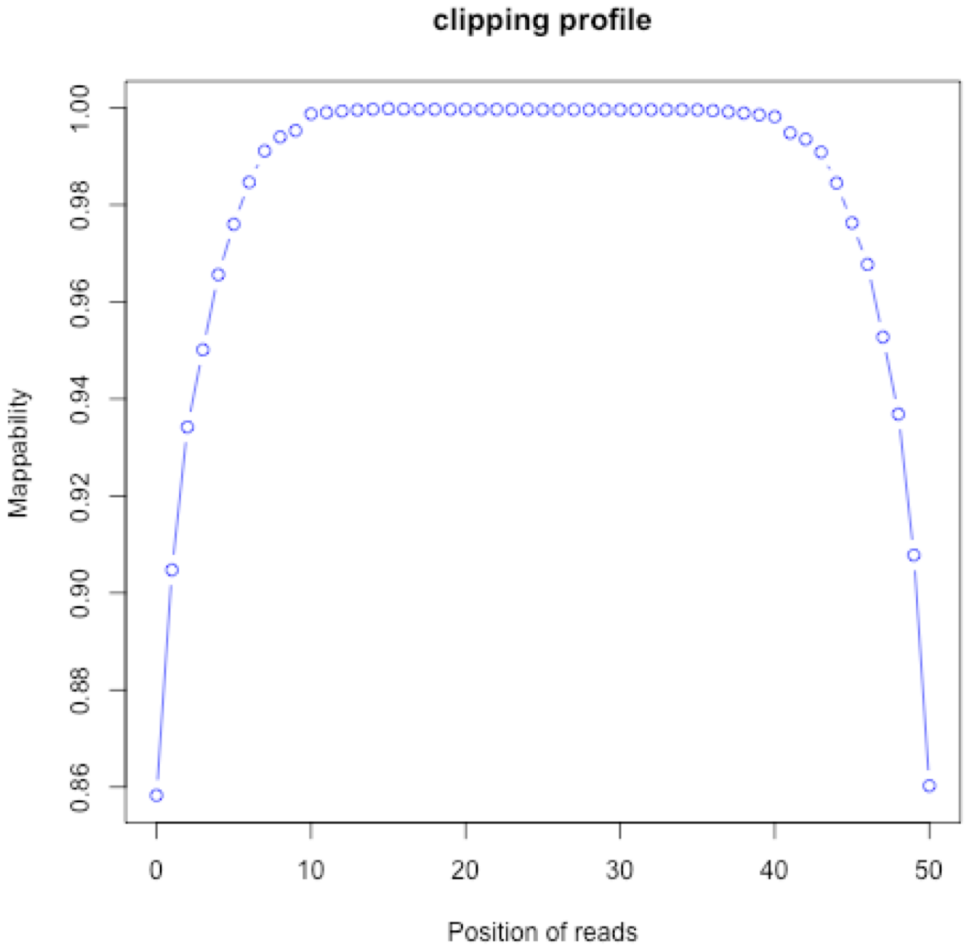
Code

```
$ module load bioinfo-tools
$ module load rseqc/2.4

$ geneBody_coverage.py –r
ref.bed12 –i mappedReads.bam -o
genecoverage
```

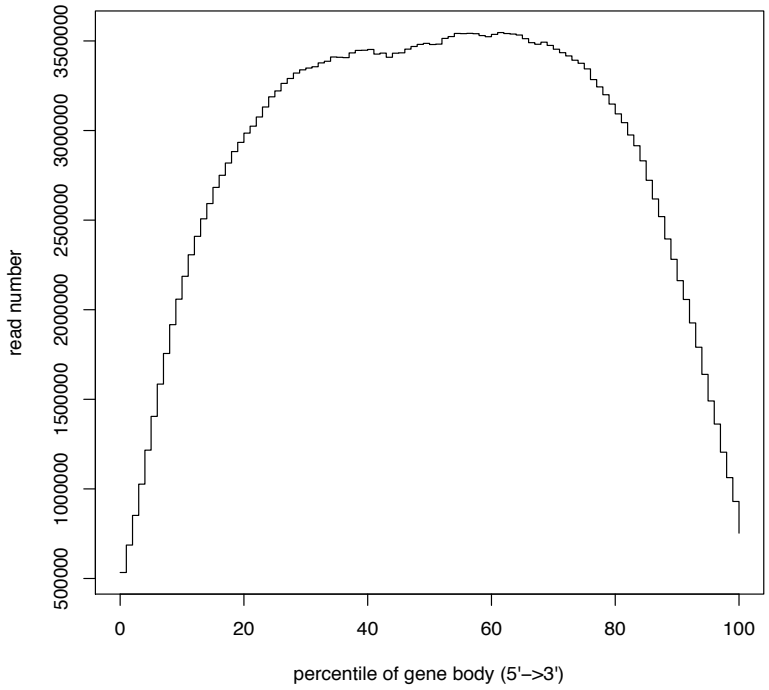
http://rseqc.sourceforge.net/

Soft clipping - clipping_profile.py

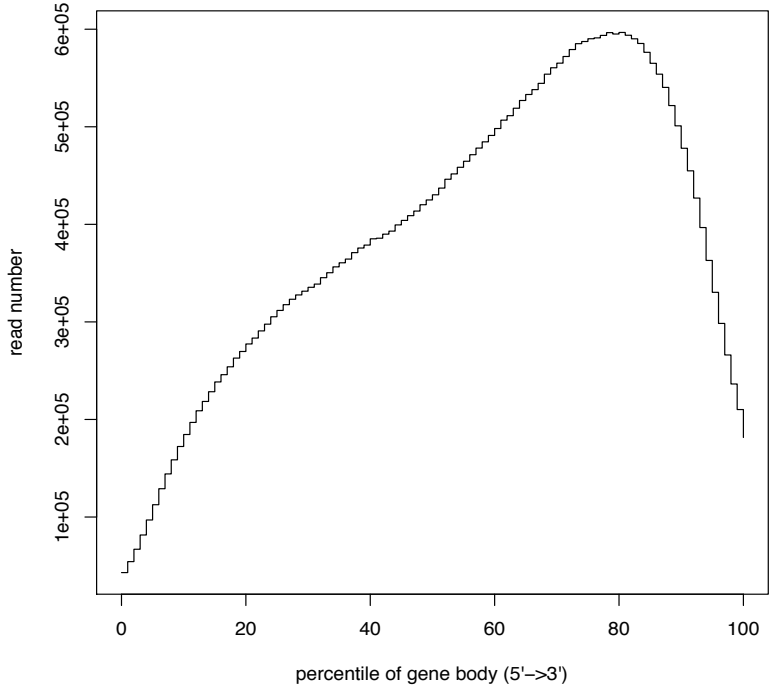


Gene coverage - geneBody_coverage.py

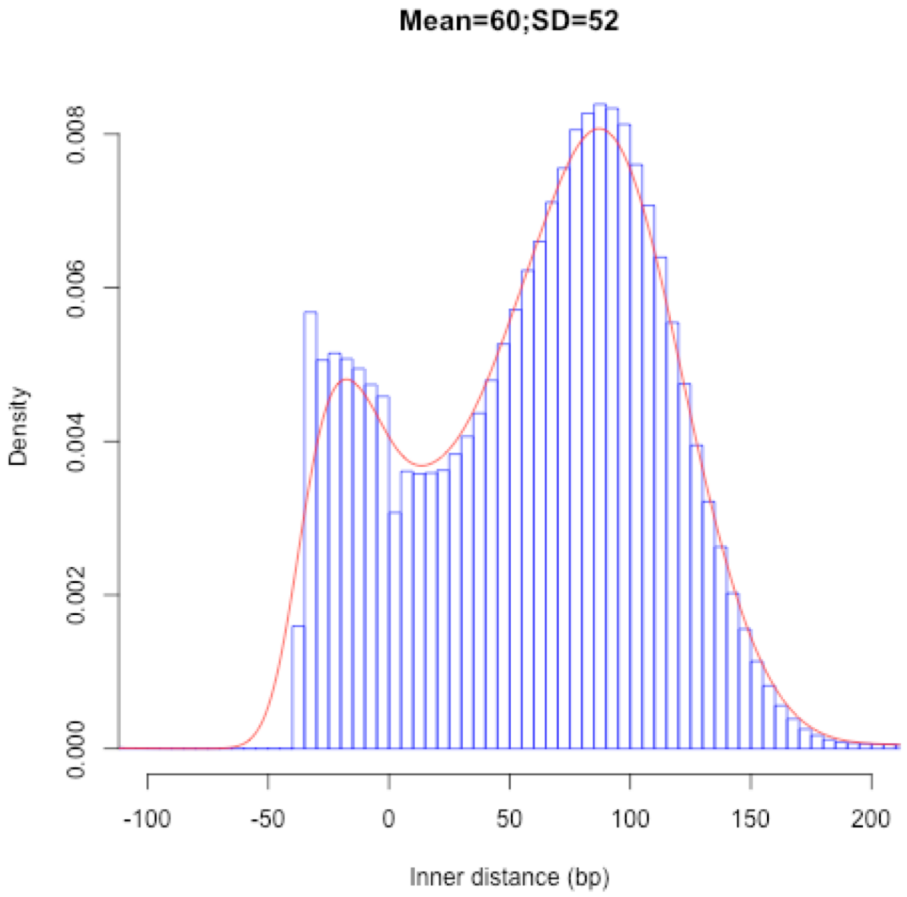
Not degraded



Degraded



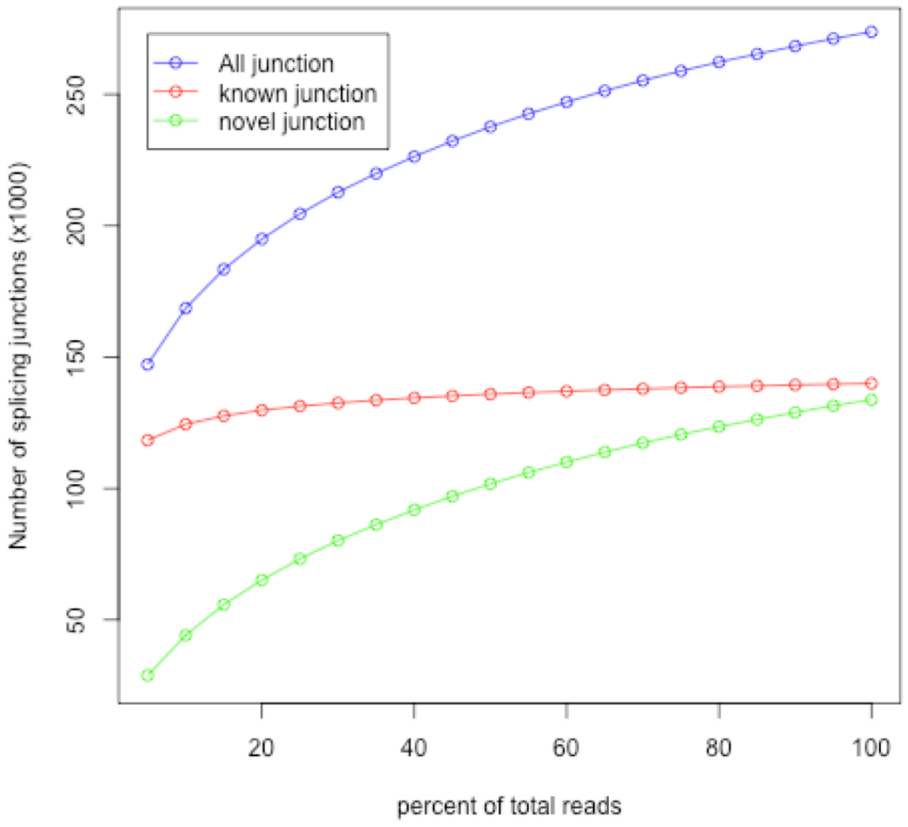
Distance between PE-reads - inner_distance.py



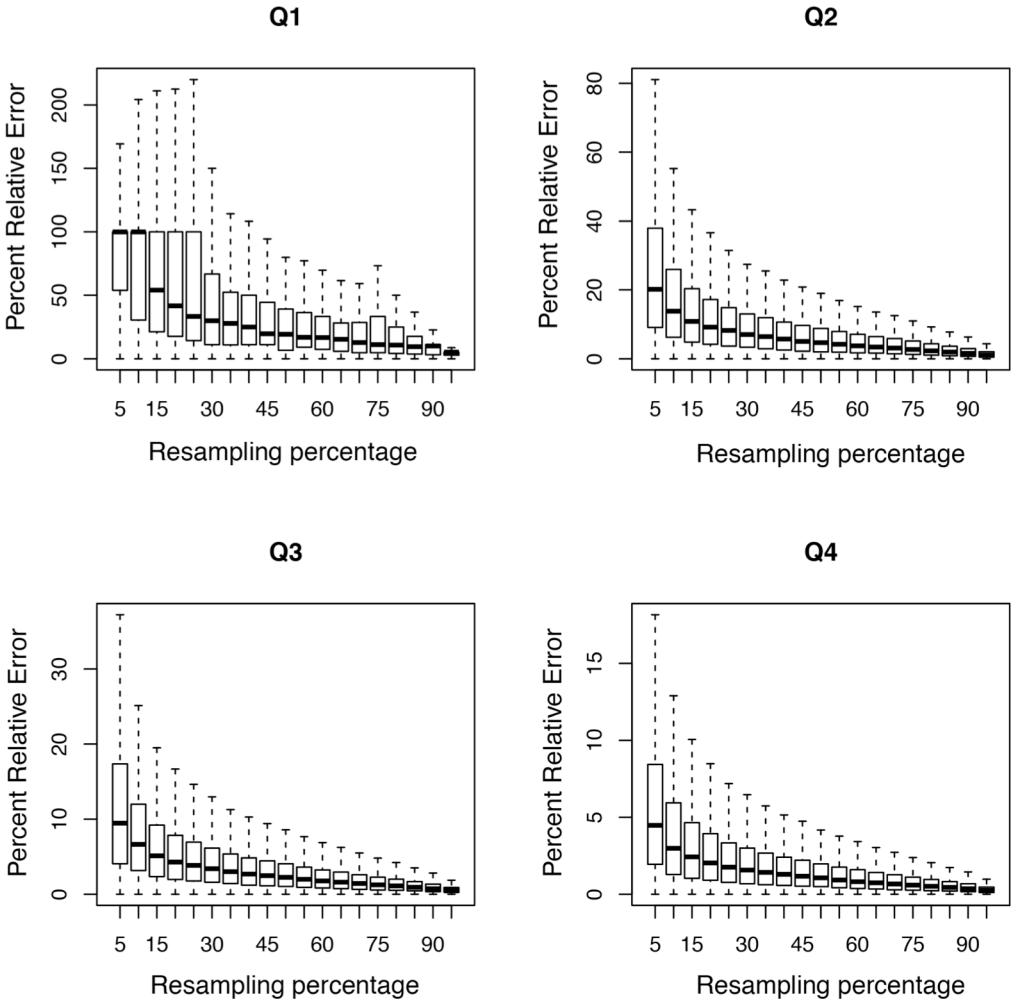
Where in the genome do your reads map? - read_distribution.py

Group	Total_bases	Tag_count	Tags/Kb
CDS_Exons	33302033	20002271	600.63
5'UTR_Exons	21717577	4408991	203.01
3'UTR_Exons	15347845	3643326	237.38
Introns	1132597354	6325392	5.58
TSS_up_1kb	17957047	215331	11.99
TSS_up_5kb	81621382	392296	4.81
TSS_up_10kb	149730983	769231	5.14
TES_down_1kb	18298543	266161	14.55
TES_down_5kb	78900674	729997	9.25
TES_down_10kb	140361190	896882	6.39

Known and novel splice junctions – junction_saturation.py or junction_annotation.py



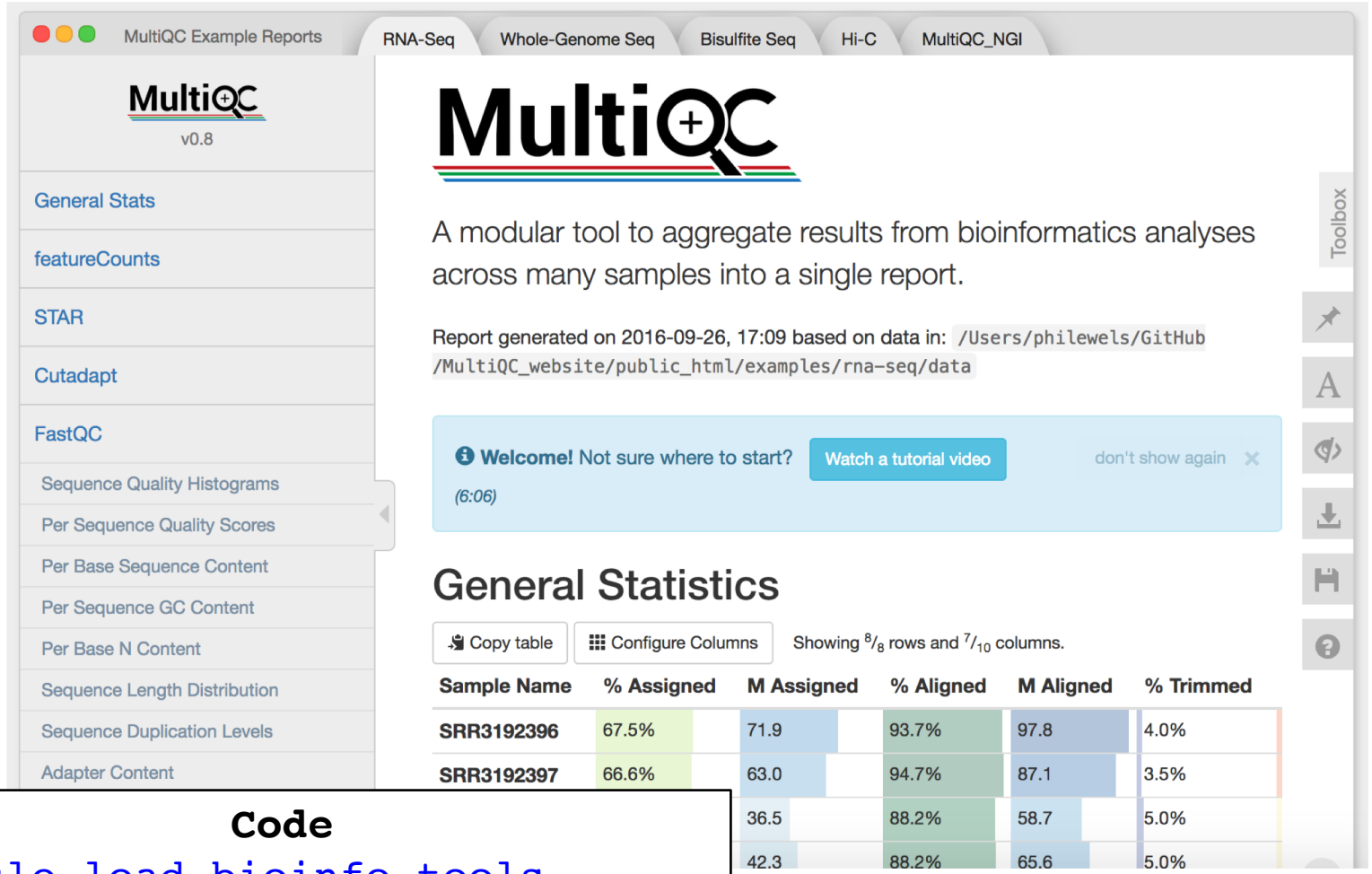
Gene detection subsampling - RPKM_saturation.py
How deep do you need to sequence?



Bad RseQC output – what to do?

- Try to figure out what went wrong.
  - Redo library prep controlling for possible errors
  - Is it equal for all samples?
  - Could it introduce any bias in the data?
- RNA-degradation in some samples
  - Possible to use a region at 3’ end for expression estimates.

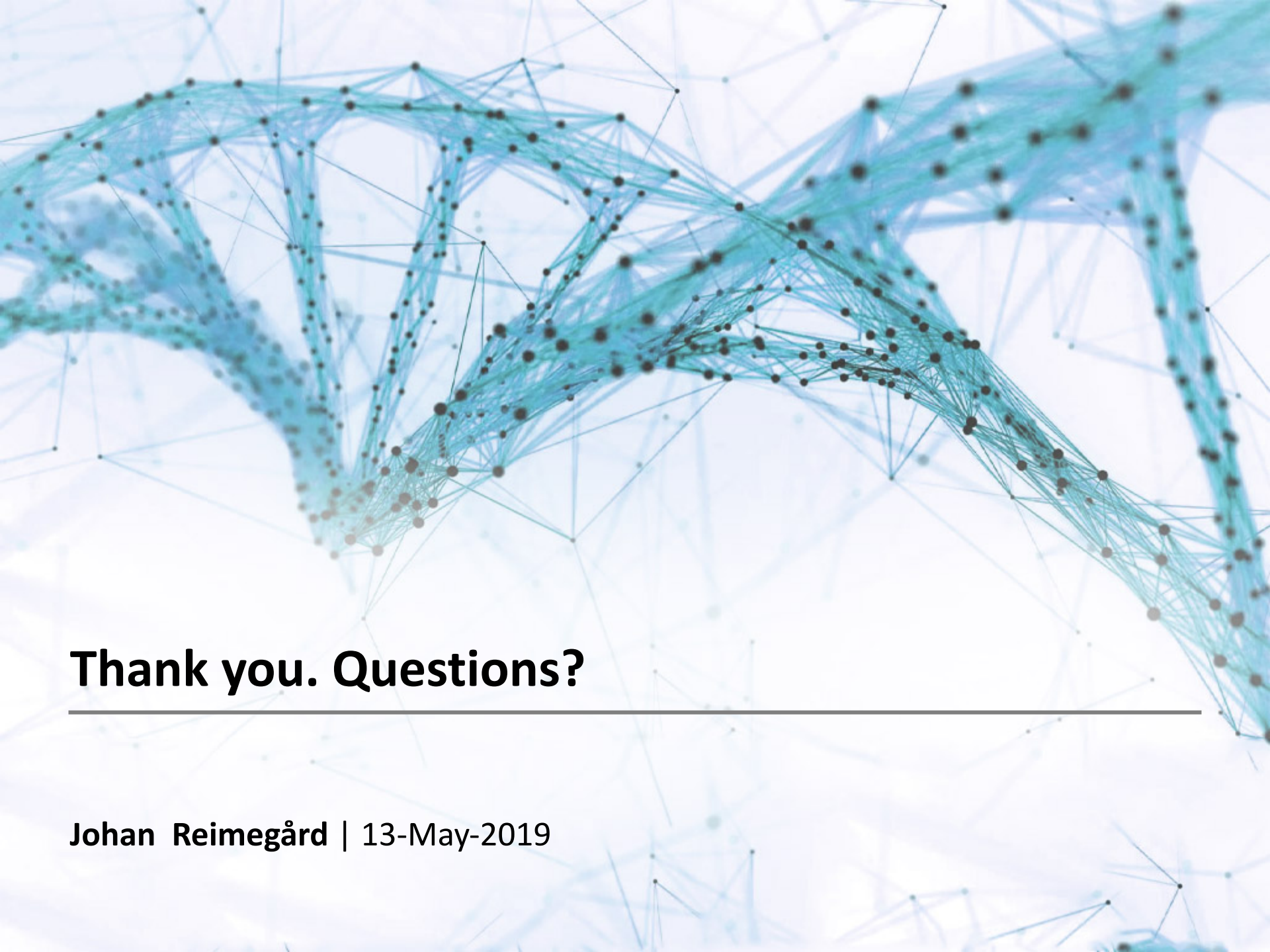
MultiQC – summary of QC stats



Code

```
$ module load bioinfo-tools
$ module load MultiQC
$ multiqc .
```

(http://multiqc.info/)



Thank you. Questions?

Johan Reimegård | 13-May-2019