Bioinformatics Workflow

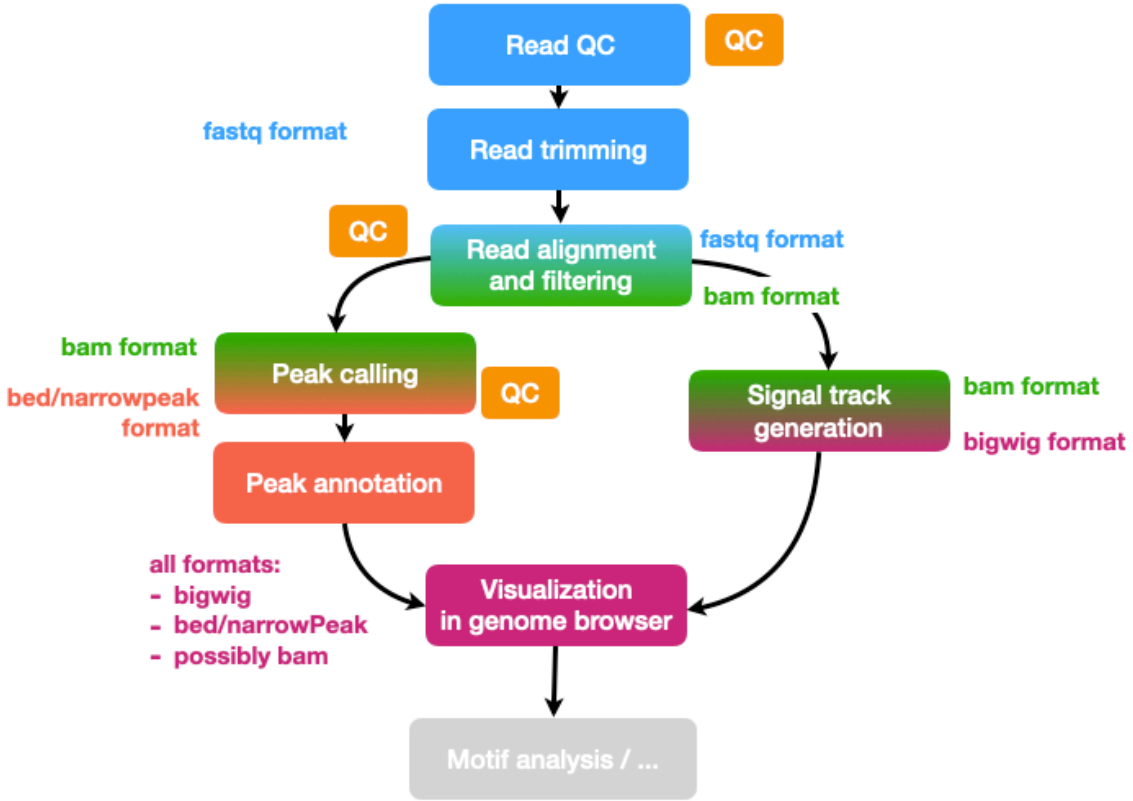
Carl Herrmann

Health Data Science Unit - Medizinische Fakultät Heidelberg

Steps in the ChIP-seq analysis

Medizinische Fakultät Heidelberg

Primary analysis



Secondary analysis

- motif analysis in ChIP-seq / ATAC-seq peaks
- differential analysis between conditions
- integration of various omics: RNA-seq / ChIP-seq / ATAC-seq / DNA-methylation
- definition of chromatin states using multiple histone marks
- ...

Bioinformatics Workflow
- File formats -

Carl Herrmann

Health Data Science Unit - Medizinische Fakultät Heidelberg

File formats?

http://www.genome.ucsc.edu/FAQ/FAQformat.html

A word about file formats

Medizinische Fakultät Heidelberg

- General sequence format: **fasta**

```
>chr1:91424-91556
ACCAGGTGGCAGCAGAGGTCAGCAAGGCAAACCCGAGC
>chr1:181924-182053
CCCGCCTGCTGGCAGCTGGGGACACTGCCGGGCCCTCT
>chr1:267896-268124
AAAGCTTTCCCACATTATACAGCTTCTGAAAGGGTTGC
CATTGTTGTTTAGTTT
>chr1:586064-586228
TTATTCAGCTTCTGAAAGGGTTGCTTGACCCACAGATG
>chr1:778514-778666
TTCAGCCGGCAACACACAGAACCTGGCGGGGAGGTCAC
>chr1:778782-778956
GGAGCGCGCATGAGCGGACGCTGCCTACTGGTGGCCGG
```

- each sequence consists of 2 **lines**
  1. header line (starting with ">") containing some free text (for example identifier of the sequence, or coordinates)
  2. genomic sequence (possibly broken over multiple lines)

A word about file formats

Medizinische Fakultät Heidelberg

- Raw sequencing reads: **fastq** format

single-end

```
@HWI-ST700693_0098:6:1101:1418:2175#ATCACG/1
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATC
+HWI-ST700693_0098:6:1101:1418:2175#ATCACG/1
__aeeeaeggggiiihfgffihihhihibefgghi
@HWI-ST700693_0098:6:1101:1376:2205#ATCACG/1
GCCATCAGAGAGGGCTTCAATCCTCAGGTTACCTGT
+HWI-ST700693_0098:6:1101:1376:2205#ATCACG/1
a_aeeeeeggggiiiiiiihiiiidhighhiiiig
```

paired-end

read 1

```
@J00118:569:HGKLCBBXY:5:1101:1489:1261 1:N:0:GTAGAGGA+AGAGTANA
AATCAGCACCCTGTGTCTAGCTCANGGTTTGTAAANATAGCANTCAGCACTCTNTATCTAGCTAAT
+
AA-FFJJJJFJJJJJJFJJJJJJF#FFFJJJ7AJJ#FJJ<<F#-FJJ--FFFJ#JJFJJA<AAFJF
@J00118:569:HGKLCBBXY:5:1101:2422:1261 1:N:0:GTAGAGGA+AGAGTANA
GACCGGAAGGCCCCTTTTCCAGTTNTTCCTAAGATNGCTGCTNCCAGAGGAGTNGAAANGTTNGAT
+
<AAF7AFJFJJJJJJFJJJJJJFJ#FJ-FJJJJ<J#7<F<--#<FJJ<JJ7F<#FFJJ#JFJ#J7J
```

read 2

```
@J00118:569:HGKLCBBXY:5:1101:1489:1261 2:N:0:GTAGAGGA+AGAGTANA
GACCTNGGCGNTGAGTGTTACAGNTCTTAAAGANGCGCTNTGGAGTTGTTCATNCCTCCNCCTGG
+
AAA-F#A-7F#<JF7AJJJFJJJ#J<FFJJJJJ#FJJJFF#JJJF7AF7FAJ<F#F-A-F#-AF--
@J00118:569:HGKLCBBXY:5:1101:2422:1261 2:N:0:GTAGAGGA+AGAGTANA
CTGCCNCCCTNAGGATTCTTCTANGCCCTAGTGNGATGTGNTGCTGAGATCCTTNTTGAANTGATT
+
AAFAF#JFJJ#JFJJJJFFJJJJ#FJ7JJJAJJ#FA<<FJ#AJJJ<JJJ<AJFJ#JJJJJ#FFJF-
```

- each read consists of **4 lines**
  1. read identifier
  2. read sequence
  3. read identifier/empty line
  4. Phred quality scores

Phred scores

Medizinische Fakultät Heidelberg

- Phred scores $Q = -10 \log_{10} P$
- Q = 30 : probability of wrong base calling P = 0.1%
- Q = 10 : probability of wrong base calling P = 10% ...
- Score for each base, encoded using different ASCII encodings

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
.................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                        59   64       73                            104                   126
0........................26...31.......40
                        -5....0........9.............................40
                              0........9.............................40
                                 3.....9..............................41
0.2......................26...31.......41
0..................20........30........40........50..........................................93

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
P - PacBio        Phred+33,  HiFi reads typically (0, 93)
```

Make sure to identify the right encoding

[Wikipedia]

A word about file formats

Medizinische Fakultät Heidelberg

- Aligned reads: **SAM/BAM format**

read ID	flag	chromosome	alignment coordinate	mapping quality	CIGAR string				sequence	Phred scores
SOLEXA-1GA-1_0055_FC629PW:6:76:6410:9673#0/1	16	chr1	17481	35	35M	*	0	0	GCCGAGCCACCCGTCACCCCCTGGCTCCTGGCCTA	<FCC37AGD<DEB@;2=GGGHGH@HHHHGHHGHHH
SOLEXA-1GA-1_0055_FC629PW:6:19:17344:9379#0/1	16	chr1	48159	31	36M	*	0	0	AAACATGTTCACATCGTGTGCGTTCCATTTTCCTAA	BE?>3E3?2BD,DB:8DCBEBG@@G?DBB::@BD::
SOLEXA-1GA-1_0055_FC629PW:6:11:10688:7659#0/1	16	chr1	49246	30	34M	*	0	0	AAGGCAGGAACAGAAATCCAAATACCGCATGTTC	?;>9D,B??DDDB@D=;BB@DDBD@:D=DDBD<B
SOLEXA-1GA-1_0055_FC629PW:6:3:3281:8061#0/1	0	chr1	49262	31	33M	*	0	0	TCCAAATACCGCATGTTCTCACTTATGAGCGTG	GD=GGEEBB=D>G@GGGGBG=GGGGGG,G?ECG

- **Mapping quality**: $MAPQ = -10 \log_{10}$(Probability wrong mapping position)
  how the MAPQ is computed depends on the aligner used!
- **CIGAR**: represents how the read was aligned
  - M = match / I = insertion / S = mismatch / D = deletion

```
24      68M6I24M
24      63M2I36M
```

- The unfiltered BAM file also contains non-aligned reads!

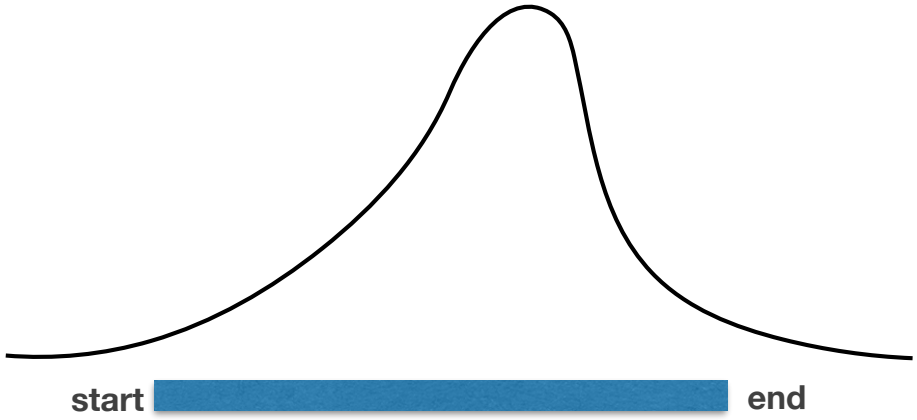
```
J00118:569:HGKLCBBXY:5:1101:1489:1261   77      *       0       0       *       *       0       0
J00118:569:HGKLCBBXY:5:1101:1489:1261   141     *       0       0       *       *       0       0
```

A word about file formats

Medizinische Fakultät Heidelberg

- Genomic regions: **bed** format

chrom	start	end	name	score
chr1	91506	91507	CTCF_peak_1	9.62564
chr1	182036	182037	CTCF_peak_2	8.37175
chr1	268004	268005	CTCF_peak_3	32.81926
chr1	586177	586178	CTCF_peak_4	36.35550
chr1	778611	778612	CTCF_peak_5	5.97809

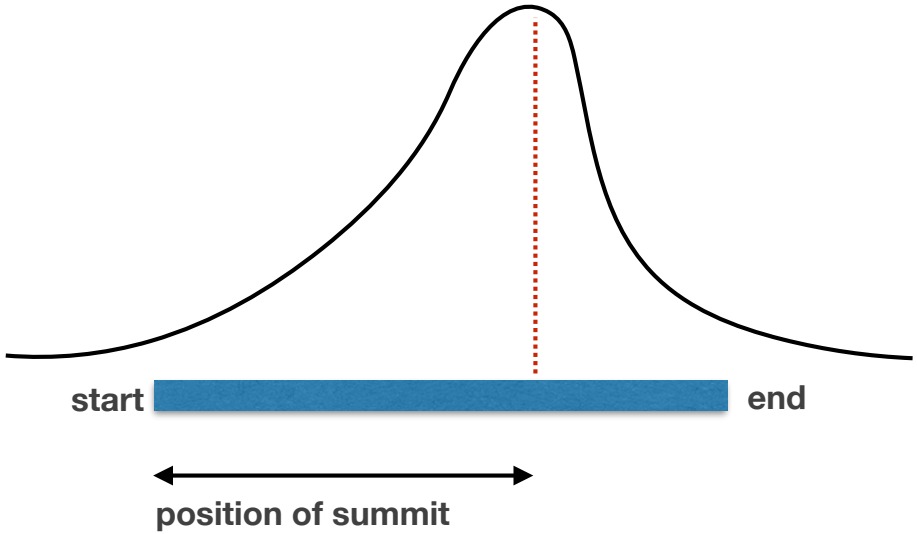


A word about file formats

Medizinische Fakultät Heidelberg

- Genomic regions: **narrow Peak** format

chrom	start	end	name	score	strand	signal	-log10(Pval)	-log10(Qval)	Position of summit
chr1	869712	870081	CTCF_peak_8	1236	.	45.81503	127.31690	123.60920	191
chr1	904629	904937	CTCF_peak_9	1223	.	45.64757	126.01864	122.32471	173
chr1	912894	913115	CTCF_peak_10	177	.	11.26369	20.46802	17.77105	122
chr1	921056	921327	CTCF_peak_11	499	.	23.12020	52.96307	49.93298	153
chr1	938137	938451	CTCF_peak_12	655	.	28.19386	68.75376	65.58743	143
chr1	951461	951678	CTCF_peak_13	360	.	18.20854	38.96448	36.06257	107

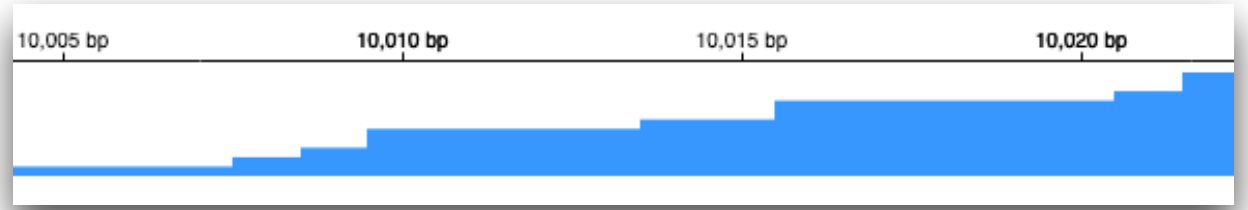


A word about file formats

Medizinische Fakultät Heidelberg

- Continuous signal : **wig/bigwig/bedgraph** format:

```
chr1    10008   10009   1
chr1    10009   10014   4
chr1    10014   10015   5
chr1    10015   10020   8
chr1    10020   10021   10
chr1    10021   10027   13
chr1    10027   10033   17
chr1    10033   10039   21
chr1    10039   10043   22
chr1    10043   10045   23
chr1    10045   10051   26
chr1    10051   10056   29
chr1    10056   10057   30
chr1    10057   10059   33
chr1    10059   10060   32
chr1    10060   10065   29
chr1    10065   10066   28
chr1    10066   10067   25
```



Strength of the signal in bins of variable sizes

File formats - summary

- **fastq, fasta** : raw sequence formats
- **sam, bam**: aligned read format (bam = compressed version of sam)
- **bedGraph, wig, bigwig** : signal tracks
- **bed, narrowPeak, broadPeak** : genomic regions

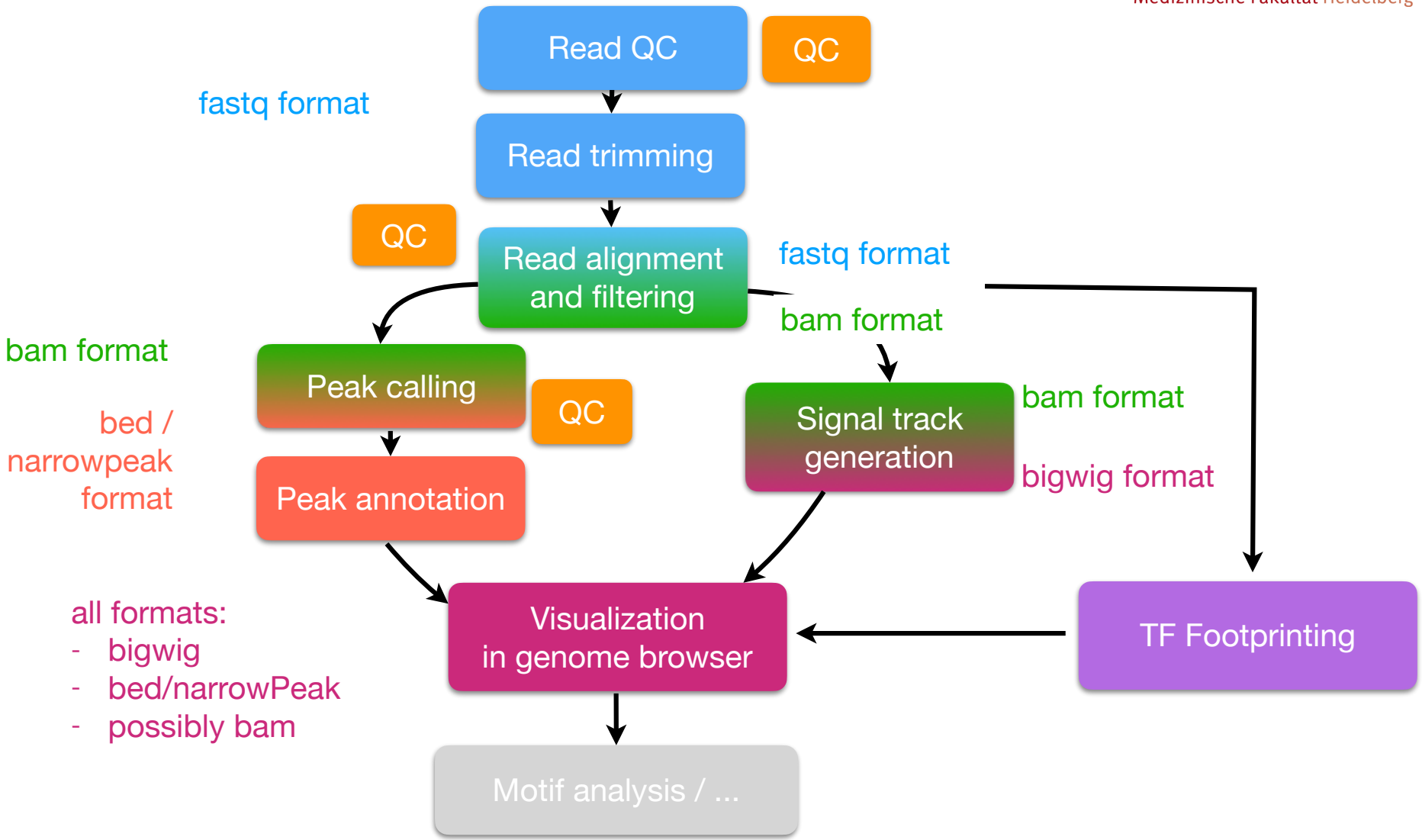
Bioinformatics Workflow
- General Workflow -

Carl Herrmann

Health Data Science Unit - Medizinische Fakultät Heidelberg

General Workflow

Medizinische Fakultät Heidelberg



Bioinformatics Workflow
- Read QC / trimming -

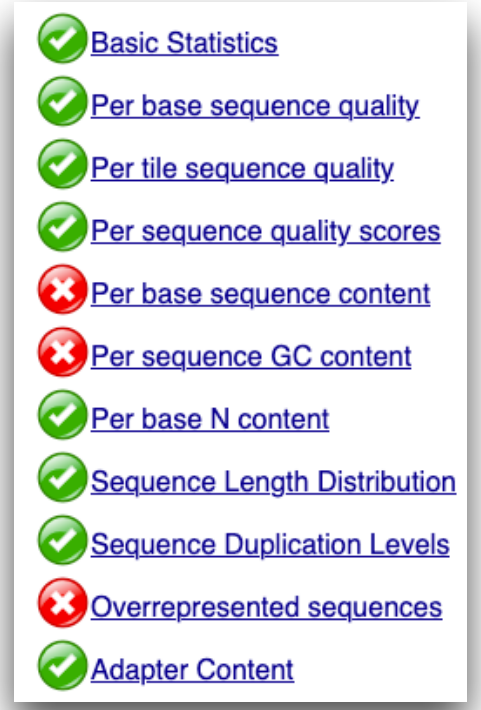
Carl Herrmann

Health Data Science Unit - Medizinische Fakultät Heidelberg

Sequencing QC

Medizinische Fakultät Heidelberg

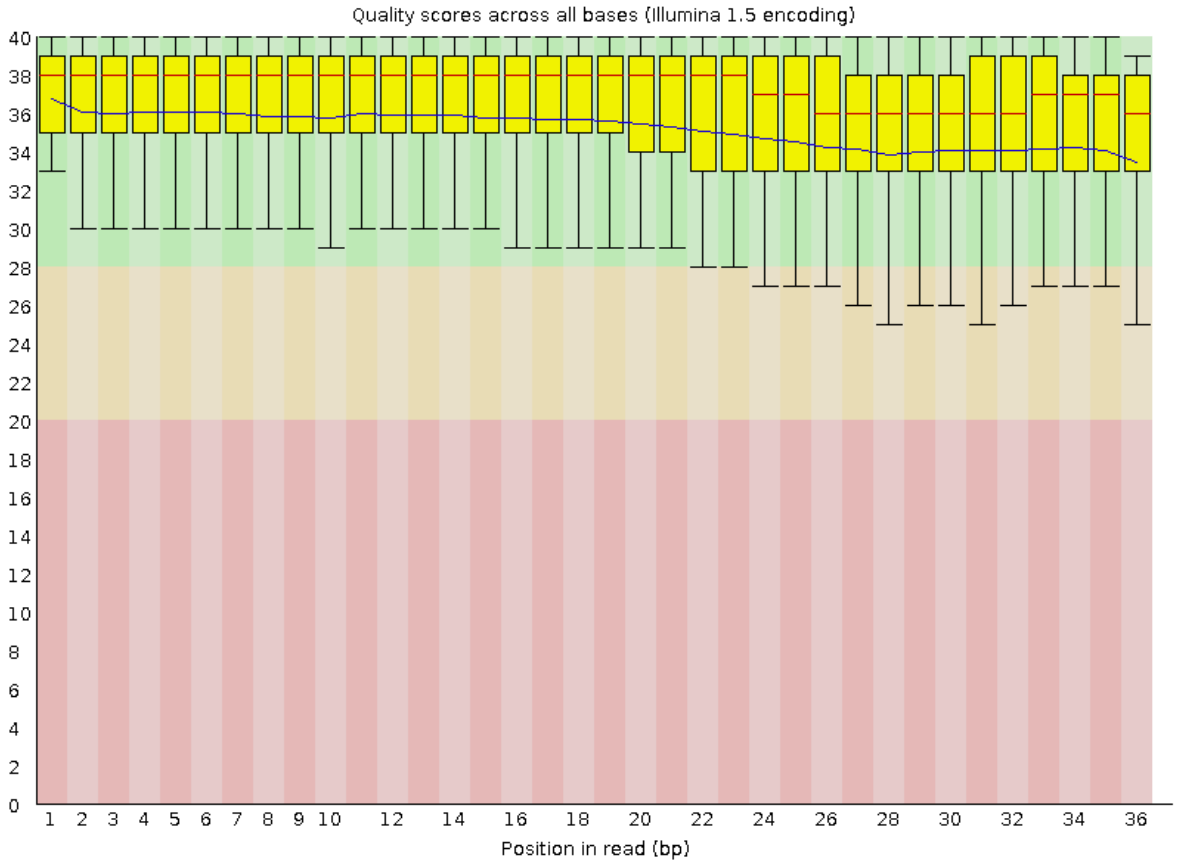
- Reads from high-throughput sequencer are obtained in fastq format
- We first check the **quality of the raw library**
  - sequencing quality?
  - biases in GC content?
  - biases in quality depending on position on flow-cell?
  - presence of repeated sequences?
  - presence of sequencing adapter sequences?
- QC report on fastq files can be obtained using the **FastQC** tool [link here]



Sequencing QC

Medizinische Fakultät Heidelberg

Per base quality

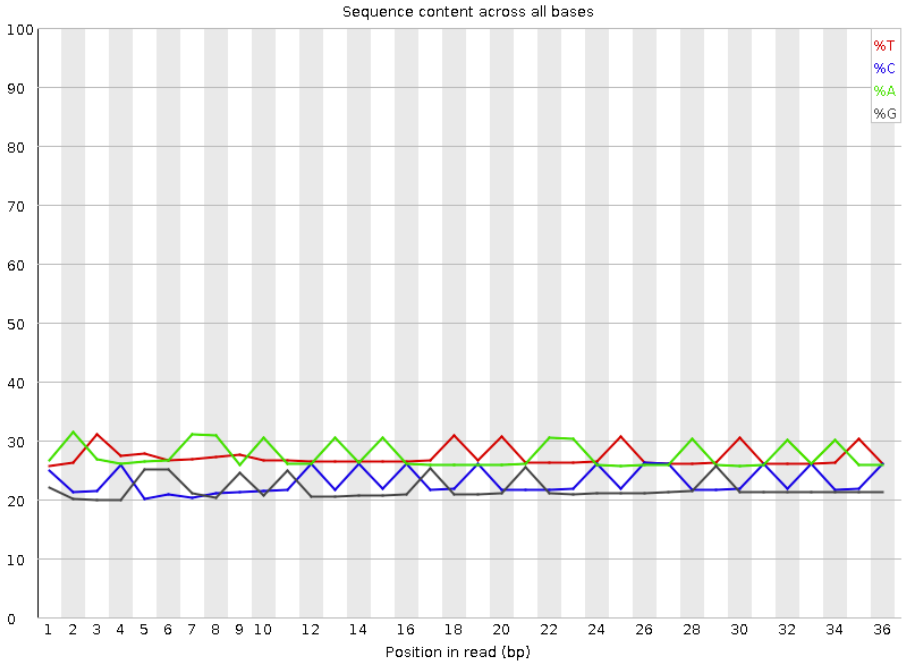


- Displays sequencing quality along the reads
- y-axis displays the Phred score per position

Sequencing QC

Medizinische Fakultät Heidelberg

Checking for adapter contamination



Distribution of bases is
not uniform along the sequences!

Presence of sequencing adapters!

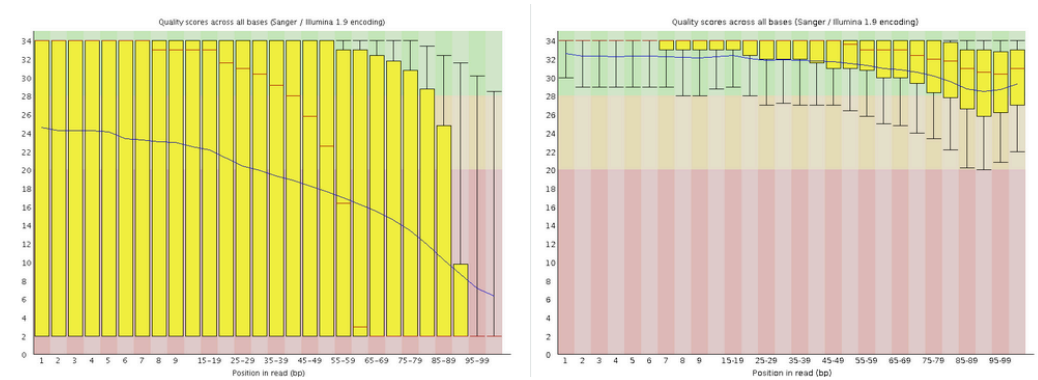
Overrepresented sequences

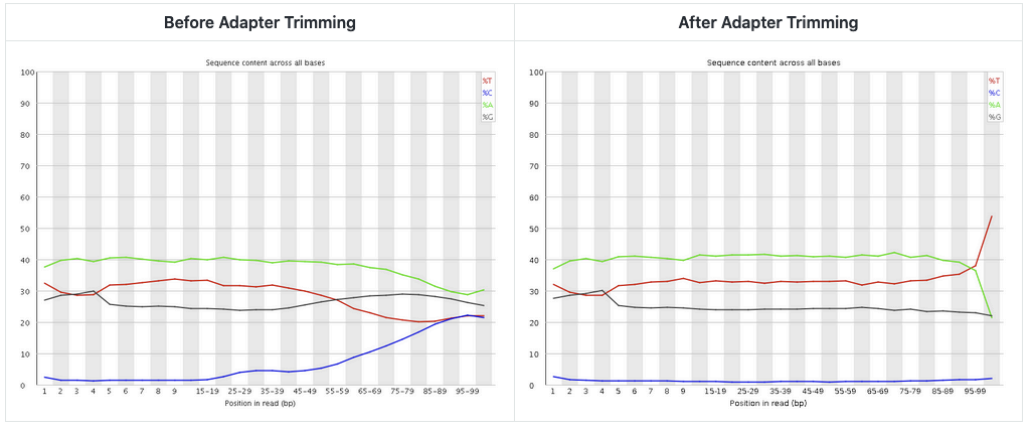
Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATC	810157	4.228608561533904	TruSeq Adapter, Index 1 (100% over 36bp)
ATCGGAAGAGCACACGTCTGAACTCCAGTCACATCA	29842	0.15576010167571813	TruSeq Adapter, Index 1 (100% over 36bp)

Read trimming

Medizinische Fakultät Heidelberg

- Reads can be trimmed at the 5'/3' ends to correct for
  - presence of sequencing adapters
  - poor sequencing quality at the 3' end of the read
- Tool used in this course: **TrimGalore** [link here]





[TrimGalore documentation]

Trimming from 3'end
to remove low quality bases
→ reads which become too short
are removed

Effect of adapter contamination
on base composition
→ trimming improves composition!

Hands-on :
FastQC report and read trimming!

https://hdsu-bioquant.github.io/chipatac2020/02_CHIP_ReadQC.html

https://hdsu-bioquant.github.io/chipatac2020/03_CHIP_Trimming.html

Bioinformatics Workflow
- Alignment -

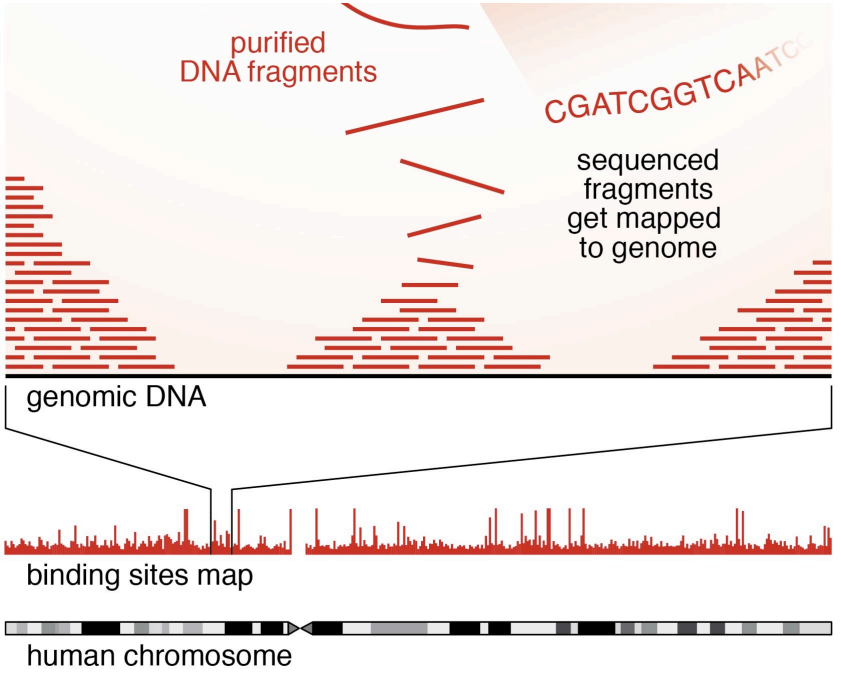
Carl Herrmann

Health Data Science Unit - Medizinische Fakultät Heidelberg

Genome alignement

Medizinische Fakultät Heidelberg

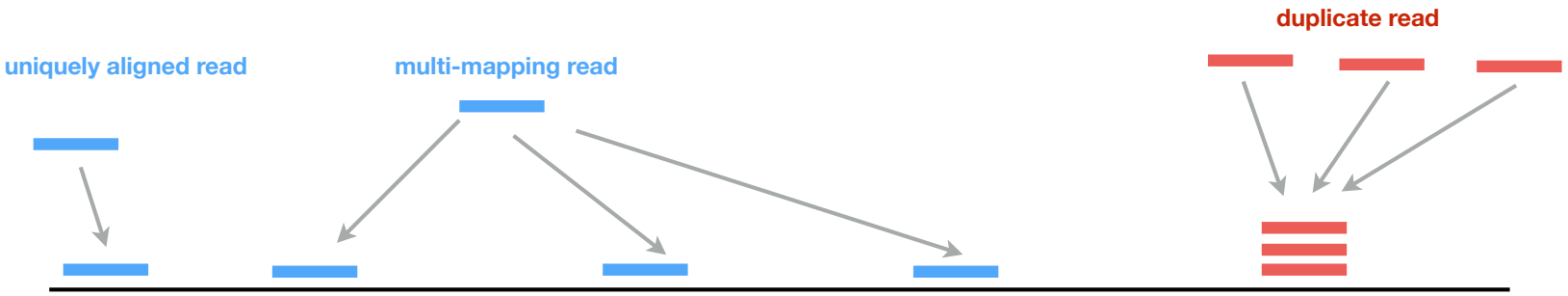
- Raw reads must be aligned to the reference genome
- **fastq → sam/bam** format
- Many tools available which differ in
  - computational efficiency
  - memory requirements
  - handling of split reads,...
- Popular tools
  - STAR
  - BWA
  - Bowtie2
  - ...



Genome alignment

Medizinische Fakultät Heidelberg

- **Challenges**
  - computational efficiency: algorithms use a genome index to identify matching positions
  - multiple matches: short reads / containing repetitive sequences can align multiple times in the genome



  - the mapping quality score (**MAPQ**) combines
    - quality of the aligned bases
    - difference in alignment score of best vs. second-best alignment

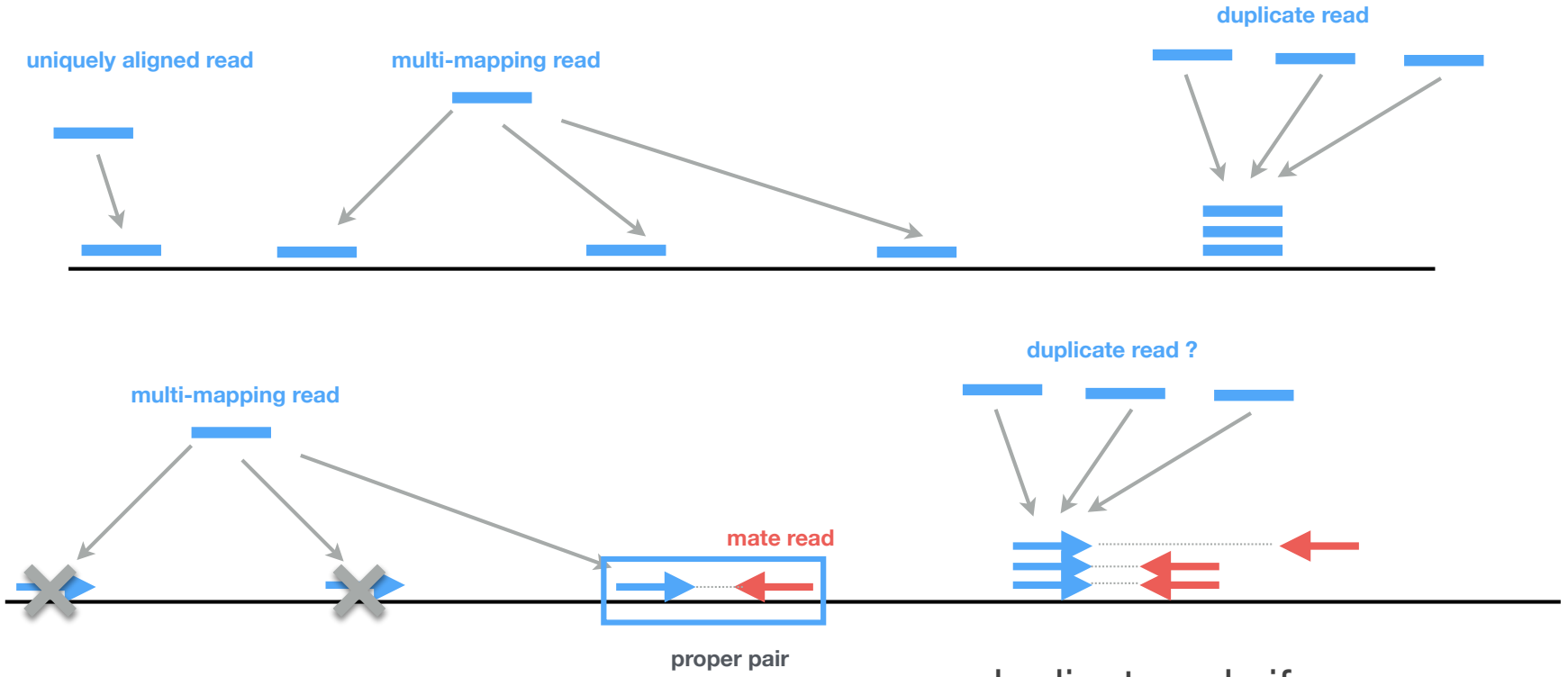
$$MAPQ = -10 \log_{10}(\text{Probability wrong mapping position})$$

  - Different aligners have different definitions of MAPQ!

Genome alignment

Medizinische Fakultät Heidelberg

- **Paired-end** vs **single-end**:
  paired -end sequencing improves the alignment, especially regarding low complexity regions



duplicate only if
both mates are duplicates!

Genome alignment

Medizinische Fakultät Heidelberg

- Typical **Bowtie2** command

```
bowtie2
--phred33
--maxins 2000
--very-sensitive
--threads 10
-x hg38.idx
-1 my_data_R1.fq.gz
-2 my_data_R2.fq.gz
| samtools view -h -b - >
my_data.bam
```

which Phred encoding?

maximal insert size (paired-end)

alignment option

number of computer-cores to use

index file for genome version hg38 (needs to be provided)

input file (read 1) in compressed fastq format

input file (read 2) in compressed fastq format

converts bowtie2 output (sam) into bam format

output file

Remember: the resulting bam file contains both' aligned and non-aligned reads! → needs to be filtered!

Filtering bam files

Medizinische Fakultät Heidelberg

- Filter out non-aligned reads and poorly mapped reads

single-end

```
samtools view -h -b \
-F 4 \
-q 30
-@ 10 \
-o my_data.filtered.bam \
my_data.bam
```

include bam header in output (-h); output bam format (-b)

filter OUT (-F) unmapped reads (4, for single-end)

filter OUT reads with a mapping quality < 30

use 10 cores

name of the output file

name of the input file

- Mark or remove duplicates

single-end

```
samtools sort -O BAM -@ 10 \
my_data.bam \
| samtools markdup -s -@ 10 -\
my_data.mkdup.bam
```

sort reads by coordinates; use 10 cores

initial bam file

mark duplicate reads and report stats (-s)

output file with marked duplicates

Genome alignment

Medizinische Fakultät Heidelberg

- Single-end bam (BAM)

```
SOLEXA-1GA-1_0055_FC629PW:6:76:6410:9673#0/1    16      chr1    17481   35      35M     *       0       0       GCCGAGCCACCCGTCACCCCCTGGCTCCTGGCCTA      <FCC37AGD<DEB@;2=GGGHGH@HHHHGHHGHHH
SOLEXA-1GA-1_0055_FC629PW:6:19:17344:9379#0/1   16      chr1    48159   31      36M     *       0       0       AAACATGTTCACATCGTGTGCGTTCCATTTTCCTAA     BE?>3E3?2BD,DB:8DCBEBG@@G?DBB::@BD::
SOLEXA-1GA-1_0055_FC629PW:6:11:10688:7659#0/1   16      chr1    49246   30      34M     *       0       0       AAGGCAGGAACAGAAATCCAAATACCGCATGTTC       ?;>9D,B??DDDB@D=;BB@DDBD@:D=DDBD<B
SOLEXA-1GA-1_0055_FC629PW:6:3:3281:8061#0/1     0       chr1    49262   31      33M     *       0       0       TCCAAATACCGCATGTTCTCACTTATGAGCGTG        GD=GGEEBB=D>G@GGGGBG=GGGGGG,G?ECG
```

- Paired-end bam (BAMPE)

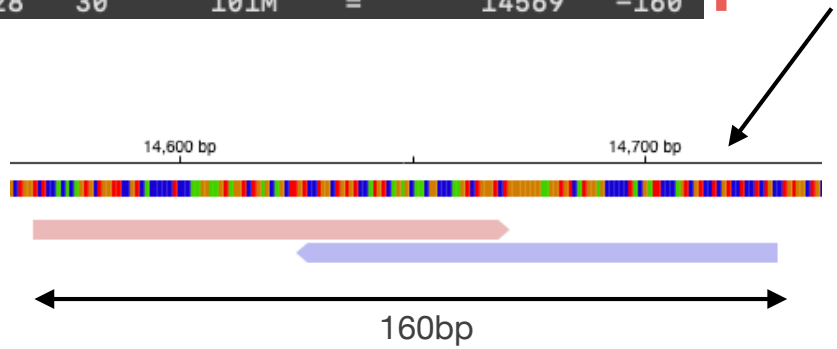
fragment size

```
J00118:569:HGKLCBBXY:5:2101:26565:4690  99      chr1    10509   30      84M     =       10509   84
J00118:569:HGKLCBBXY:5:2101:26565:4690  147     chr1    10509   30      84M     =       10509   -84
J00118:569:HGKLCBBXY:5:1115:14509:20621 99      chr1    10562   30      55M     =       10562   55
J00118:569:HGKLCBBXY:5:1115:14509:20621 147     chr1    10562   30      55M     =       10562   -55
J00118:569:HGKLCBBXY:5:1215:8907:44464  99      chr1    14569   30      100M    =       14628   160
J00118:569:HGKLCBBXY:5:1215:8907:44464  147     chr1    14628   30      101M    =       14569   -160
```

fragment 1

fragment 2

fragment 3



Genome alignment

Medizinische Fakultät Heidelberg

- Aligned reads are stored in **BAM** file
- Statistics can be obtained using the **samtools flagstat** command

total number of reads (R1 + R2)

duplicates, marked by samtools markdup

alignment rate

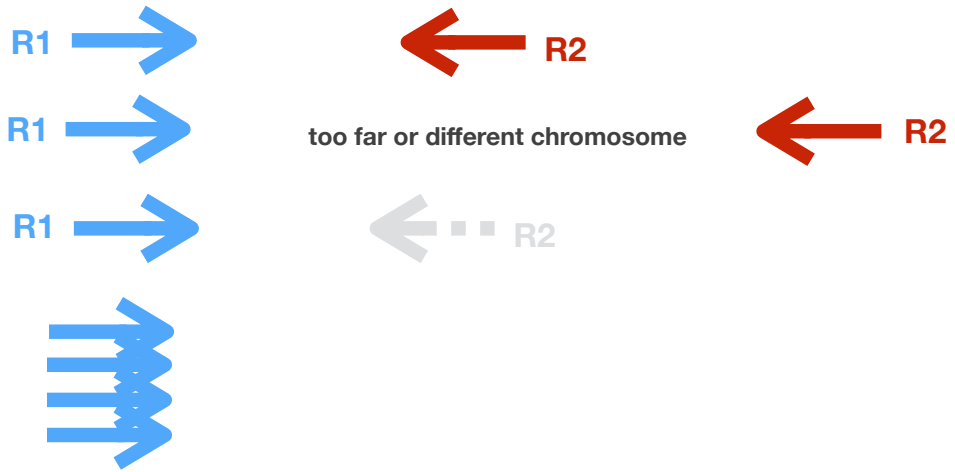
Total number of paired reads

Total number of reads R1

Total number of reads R2

```
25928860 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
0 + 0 supplementary
5726500 + 0 duplicates
15644630 + 0 mapped (60.34% : N/A)
25928860 + 0 paired in sequencing
12964430 + 0 read1
12964430 + 0 read2
15483910 + 0 properly paired (59.72% : N/A)
15513664 + 0 with itself and mate mapped
130966 + 0 singletons (0.51% : N/A)
15622 + 0 with mate mapped to a different chr
6793 + 0 with mate mapped to a different chr (mapQ>=5)
```

- Properly paired (= 1 fragment)
- Both aligned, not properly paired
- Singletons
- Duplicates



Hands-on :
Alignment results and flagstat!

https://hdsu-bioquant.github.io/chipatac2020/04_CHIP_Alignment.html