

# Lab 0: HGEN 47900

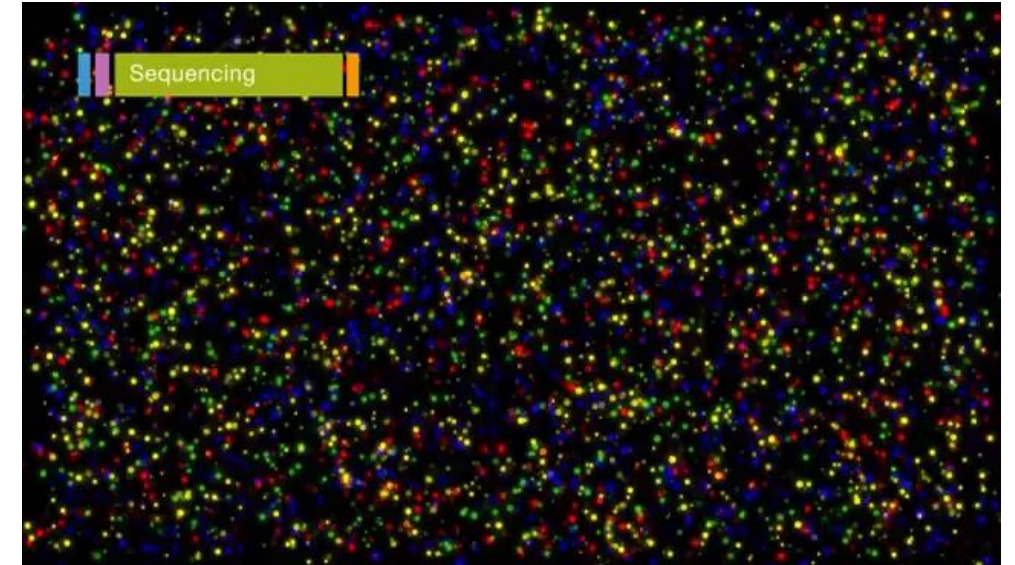
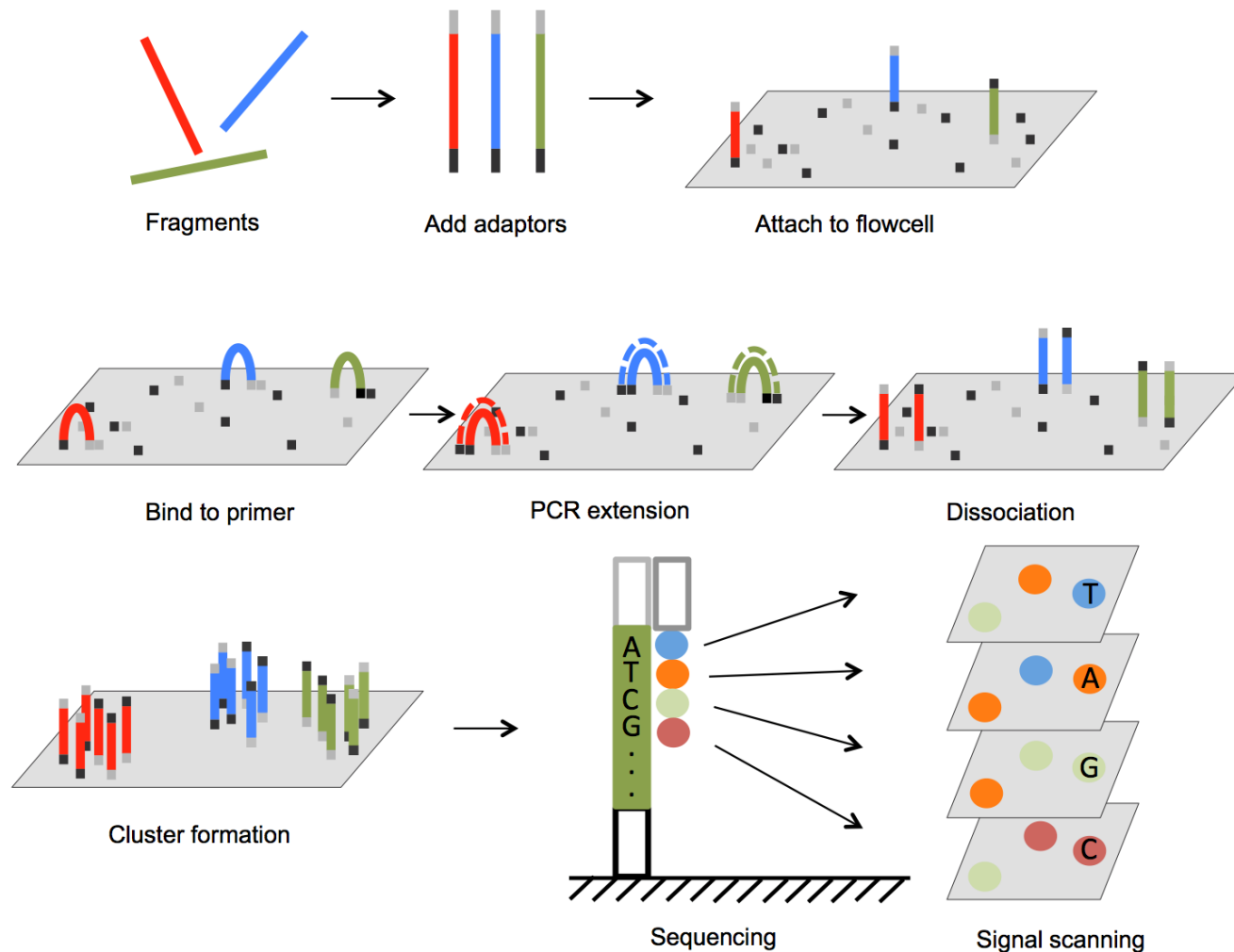
Sequencing data  
processing & R basics

Sébastien Bastide  
March 24<sup>th</sup>, 2023

*HOW TO BECOME A*  
**HACKERMAN**

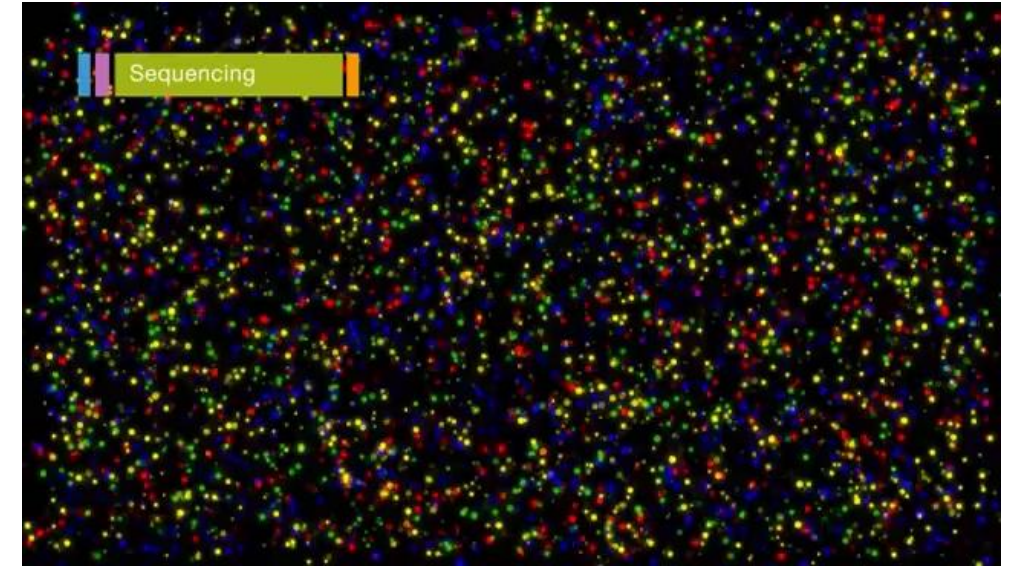
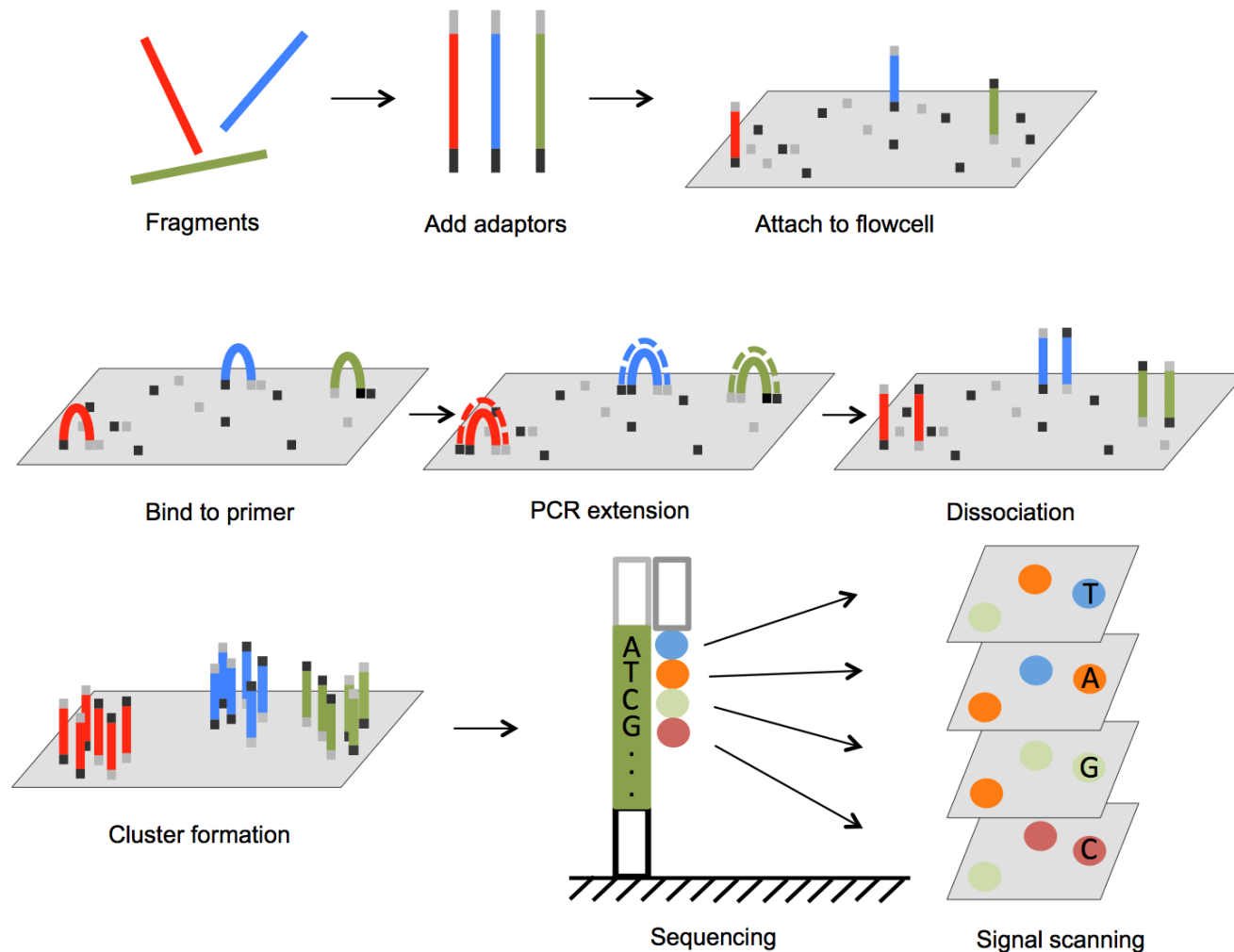


# Next Generation Sequencing (Illumina)



**Adaptors** (p5 and p7) are added to the ends of the DNA fragments to be sequenced. Those are used to attach to the flow cell and allow to perform the bridge amplification. This is a simple solid phase PCR where individual fragments are locally amplified and form **clusters**.

# Next Generation Sequencing (Illumina)



Each cycle, all four 3'-modified nucleotides are added to the flow cell and the extension happens for 1 bp. Further extension is blocked by the 3' modification. After washing and imaging, the 3' modification is removed and the next cycle begins.

# Next Generation Sequencing (Illumina): more details

Illumina sequencing kits come in different flavors depending on:

- The machine (MySeq, NextSeq, NovaSeq): this mostly changes the total number of reads,
- The output of the kit (mid vs high): this also changes the total number of reads,
- The number of cycles (up to 2x250 bp): this changes the length of the reads.

## Cluster calling:

The detection of amplicon clusters on the flow cell is called **cluster calling**. On most Illumina machines, it happens only once at the beginning of the sequencing run (usually at cycle 5). This step is dependent on the diversity of nucleotides in the first few cycles of read 1.

In older, 2-channel systems (T is green, C is red, A is green + red and G has no color), having more than 2-3 Gs in the first bp of read 1 could prevent the detection of the cluster. Check what chemistry you use and how diverse you expect your library to be. Similarly, if all your DNA fragments start with the exact same sequence, then the flow cell could be uniformly colored, which makes it difficult to detect clusters.

A potential solution to increase the diversity is to add some Phi X174 phage DNA to your library.

The expected **cluster density** on a NextSeq 500/550 High output kit is ~ 170k – 220k clusters/mm<sup>2</sup>

## Cluster filtering:

After some number of cycles (e.g., 25 on a NextSeq), the clusters are filtered to guarantee the quality of the base call. If clusters overlap too much, they will be discarded. If they are not homogeneous, they will be discarded. The risk of this happening increases with the concentration of DNA in your library.

A good sequencing run has over 90% clusters passing filter.

# Finding data on NCBI (Gene Expression Omnibus)

The screenshot shows the NCBI GEO DataSets search results for the query "Sox9 ChIP-seq nasal". The search bar at the top contains the query and a "Search" button. Below the search bar, there are tabs for "GEO DataSets", "Create alert", and "Advanced". The left sidebar contains various filters such as "Entry type", "Organism", "Study type", "Author", "Attribute name", and "Publication dates". The main content area displays two search results, both titled "Distinct regulatory programs for Sox9 in transcriptional regulation of the developing mammalian chondrocyte". The first result is a ChIP-seq dataset (GSE69109) and the second is an expression array dataset (GSE69108). Both results provide a brief description of the study, the organism (Mus musculus), the platform used, and links to the full text in PMC, similar studies, and download options. The right sidebar contains sections for "Find related data", "Search details", "Important Links", and "Recent activity".

**GEO DataSets**    [Help](#)

[Create alert](#) [Advanced](#)

**Entry type**  
DataSets (0)  
Series (2)  
Samples (0)  
Platforms (0)

**Organism**  
Customize ...

**Study type**  
Expression profiling by array  
Methylation profiling by array  
Customize ...

**Author**  
Customize ...

**Attribute name**  
tissue (0)  
strain (1)  
Customize ...

**Publication dates**  
30 days  
1 year  
Custom range...

[Clear all](#)  
[Show additional filters](#)

**Summary**

**Filters:** [Manage Filters](#)

**Search results**  
**Items: 2**

☐ [Distinct regulatory programs for Sox9 in transcriptional regulation of the developing mammalian chondrocyte \[ChIP-seq\]](#)  
(Submitter supplied) We compared **Sox9**-association at chondrocyte targets to a broad catalogue of regulatory indicators of chromatin organization and transcriptional activity to determine **Sox9**'s direct regulatory actions in normal developing chondrocytes. **Sox9**-associated regions resolve into two distinct regulatory categories. Class I regions closely associate with transcriptional start sites (TSSs). Their targets reflect general regulators of basal cell activities that **Sox9** engages indirectly though a likely association with the basal transcriptional complex. [more...](#)  
Organism: Mus musculus  
Type: Genome binding/occupancy profiling by high throughput sequencing  
Platforms: GPL9250 GPL13112 13 Samples  
Download data: BED  
Series Accession: GSE69109 ID: 200069109  
[PubMed](#) [Full text in PMC](#) [Similar studies](#) [SRA Run Selector](#)

☐ [Distinct regulatory programs for Sox9 in transcriptional regulation of the developing mammalian chondrocyte \[expression array\]](#)  
(Submitter supplied) We compared **Sox9**-association at chondrocyte targets to a broad catalogue of regulatory indicators of chromatin organization and transcriptional activity to determine **Sox9**'s direct regulatory actions in normal developing chondrocytes. **Sox9**-associated regions resolve into two distinct regulatory categories. Class I regions closely associate with transcriptional start sites (TSSs). Their targets reflect general regulators of basal cell activities that **Sox9** engages indirectly though a likely association with the basal transcriptional complex. [more...](#)  
Organism: Mus musculus  
Type: Expression profiling by array  
Platform: GPL20205 6 Samples  
Download data: CEL  
Series Accession: GSE69108 ID: 200069108  
[PubMed](#) [Full text in PMC](#) [Similar studies](#) [Analyze with GEO2R](#)

**Top Organisms** [\[Tree\]](#)  
Mus musculus (2)

**Find related data**   
Database:

**Search details**  
Sox9[All Fields] AND ChIP-seq[All Fields] AND ("nose"[MeSH Terms] OR nasal[All Fields])  
 [See more...](#)

**Important Links**  
[GEO Home](#)  
[GEO Documentation](#)  
[About GEO DataSets](#)  
[Construct a Query](#)  
[Download Options](#)

**Recent activity**    
Q Sox9 ChIP-seq nasal (2) GEO DataSets

(Has all types of data, not only gene expression)

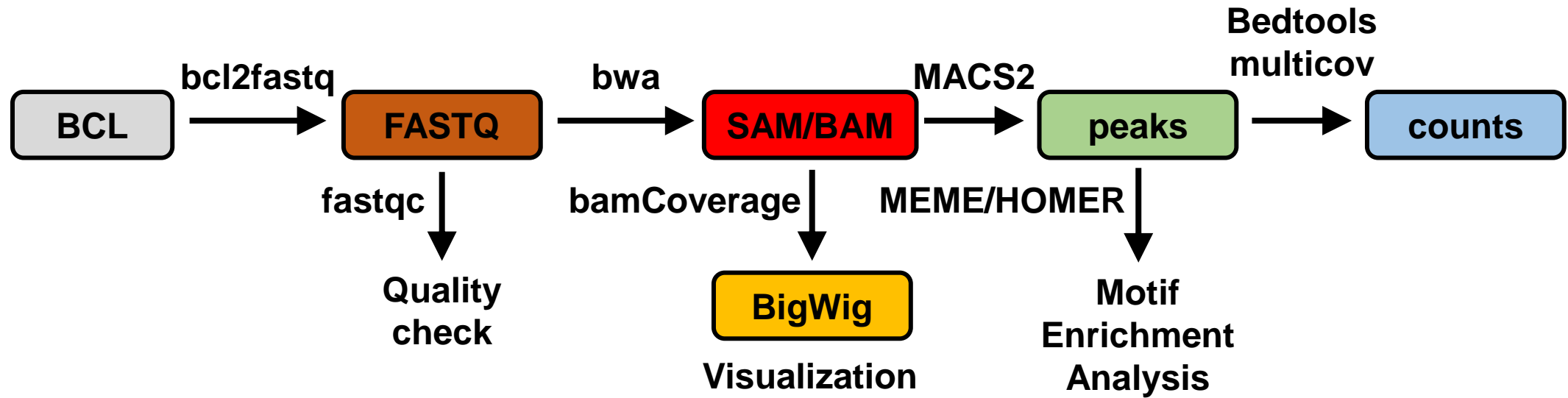
<https://www.ncbi.nlm.nih.gov/geo/>

# Fetching data on Sequence Read Archive (SRA)

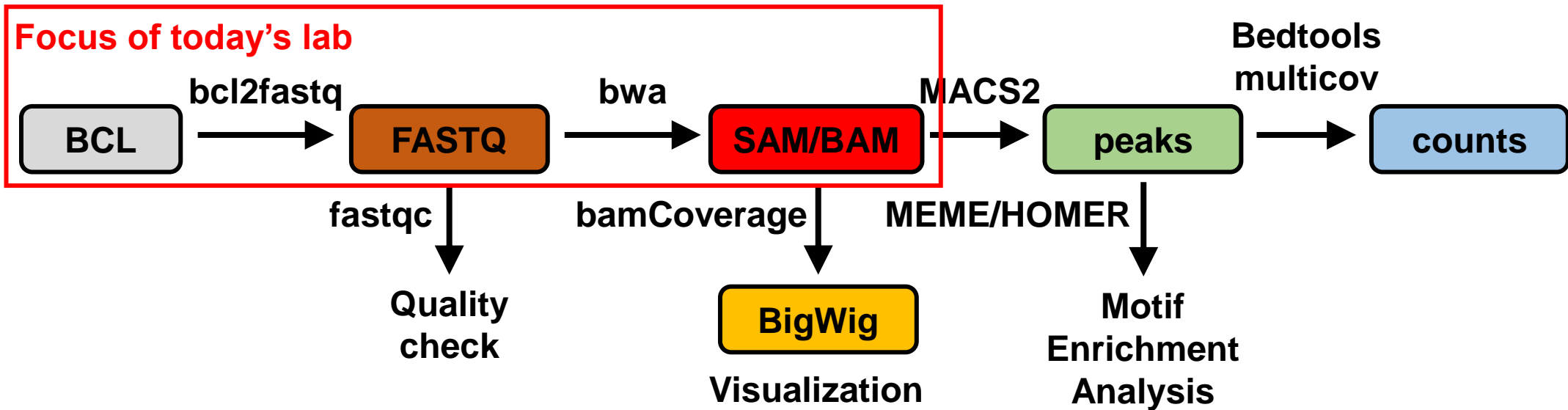
Using **sra-tools** (<https://github.com/ncbi/sra-tools/wiki/>) to fetch the data  
**fastq-dump** -split-files SRRnumber

```
reads read      : 40,527,423
reads written   : 40,527,423
[sbastide@nobel:/data/personal_folders/sbastide/sequencing_data_tutorial$ ls
SRR2034943.fastq
sbastide@nobel:/data/personal_folders/sbastide/sequencing_data_tutorial$
```

# Full pipeline for ATAC-seq/ChIP-seq

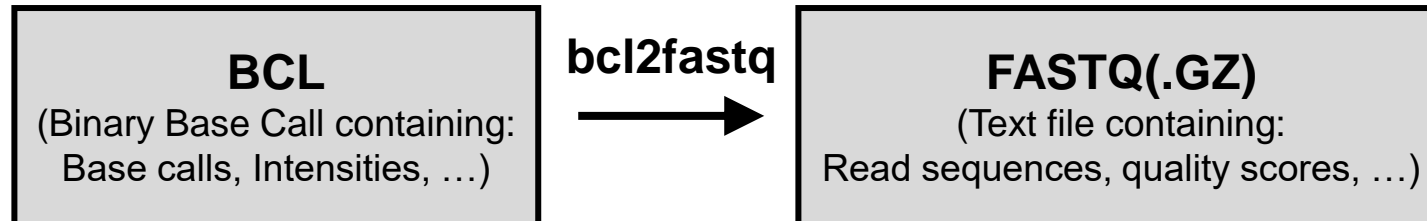


# Full pipeline for ATAC-seq/ChIP-seq





# Fastq files



```
sbastide@nobel:/data/personal_folders/sbastide/sequencing_data_tutorial/Fastqs$ head -n 12 Sample10.fq
@SRR2034943.8532554 ILLUMINA-3AB384:61:FC:8:21:8047:3232 length=42
ACTAGGCAATATTTTACCAATGACAACTTAAGTGTCTTC
+
EHIIIIIIIIIIHIIIIIGIGIIIIIIIIIIIIIIII
@SRR2034943.13764671 ILLUMINA-3AB384:61:FC:8:34:8719:5489 length=42
GACAAGGGCATGCTCTTGTGTGTATGCATGTGGGCCATAGG
+
#####
@SRR2034943.31253095 ILLUMINA-3AB384:61:FC:8:79:13486:4287 length=42
TGCATCTGATCTGTGGCCAGACAGCCTAGACTACTGTGTCCA
+
IIIIIIIIIIIIHIIIIIIIEIHIIIIIIHFFHGIIII
```

Fastq files contain records for all individual reads (i.e. sequences of nucleotides derived from a single cluster)  
Each record has 4 lines

- Line 1 (**header**): starts with a @ followed by the sequence label
- Line 2: DNA sequence (may contain Ns when the base calling is bad)
- Line 3: Line break (contains a + by convention)
- Line 4: Quality (Phred) scores in ASCII

**fastqc** SRR2034943.fastq

Creates a .html file that contains a bunch of information about the sequencing run

# Optional: read trimming using Cutadapt (for ATAC-seq)

<https://cutadapt.readthedocs.io/en/stable/guide.html>

**Not necessary** if using modern aligners (STAR, BWA-MEM, ...)

Be careful, by default, Bowtie2 uses end-to-end alignment which does not allow soft-clipping of the beginning or end of the read sequences

Other tools: **Trimmomatic**, **Trim Galore**

(Tn5 from the Nextera kit introduces CTGTCTCTTATACACATCT)

```
cutadapt -a AACCGGTT -o output.fastq input.fastq
```

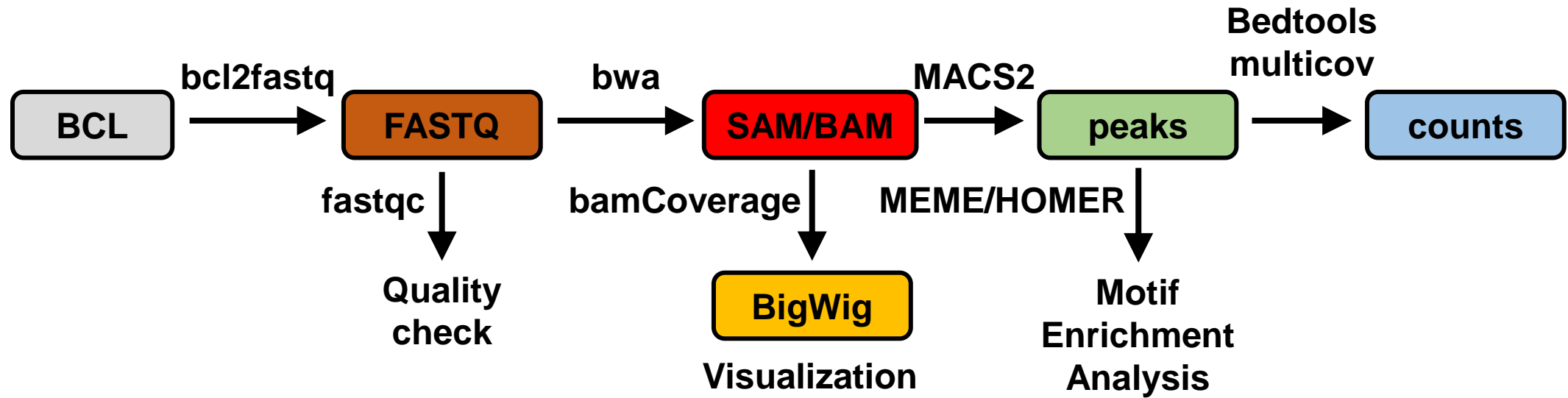
```
AACCGGTTACTGCATCGTAGCTGCA  
AACCGGTTATCGCTGCGATCGCATG  
AACCGGTTCCGCGGACTCGCGATAC
```

**cutadapt**



```
ACTGCATCGTAGCTGCA  
ATCGCTGCGATCGCATG  
CCGCGGACTCGCGATAC
```

# Full pipeline for ATAC-seq/ChIP-seq



# Mapping reads to a Genome using BWA-MEM

```
sbastide@nobel:/data/personal_folders/sbastide/sequencing_data_tutorial/FastQC$ bwa mem
Usage: bwa mem [options] <idxbase> <in1.fq> [in2.fq]
```

<idxbase> corresponds to the BW transformed index (reference genome)

**Burrows-Wheeler Transform (BWT)** is useful for compression:

- Completely reversible without loss of information
- Only the output is required to uncompress
- Same characters tend to cluster together (e.g.  $BWT(ACAACG\$) = GC\$AAAC = GC\$3AC$ )

Transformation				
1. Input	2. All rotations	3. Sort into lexical order	4. Take the last column	5. Output
<div>^BANANA  </div>	<div>^BANANA     ^BANANA A   ^BANAN NA   ^BANA ANA   ^BAN NANA   ^BA ANANA   ^B BANANA   ^</div>	<div>ANANA   ^B ANA   ^BAN A   ^BANAN BANANA   ^ NANA   ^BA NA   ^BANA ^BANANA     ^BANANA</div>	<div>ANANA   ^B ANA   ^BAN A   ^BANAN BANANA   ^ NANA   ^BA NA   ^BANA ^BANANA     ^BANANA</div>	<div>BNN^AA   A</div>

Suffix Array

Read: GCA

Read: GCA

Read: GCA

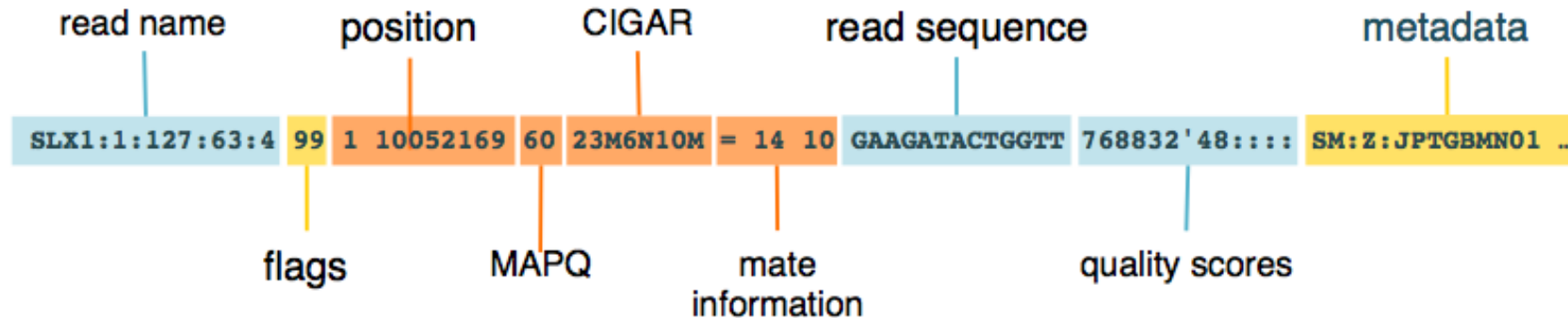
12.	0.	\$	.	.	.	T	0.	\$	.	.	.	T	0.	\$	.	.	.	T
6.	1.	A	.	.	.	C	1.	A	.	.	.	C	1.	A	.	.	.	C
2.	2.	A	.	.	.	G	2.	A	.	.	.	G	2.	A	.	.	.	G
5.	3.	C	.	.	.	G	3.	CA	.	.	.	G	3.	CA	.	.	.	G
7.	4.	C	.	.	.	A	4.	C	.	.	.	A	4.	C	.	.	.	A
0.	5.	C	.	.	.	\$	5.	C	.	.	.	\$	5.	C	.	.	.	\$
8.	6.	C	.	.	.	C	6.	C	.	.	.	C	6.	C	.	.	.	C
1.	7.	G	.	.	.	C	7.	G	.	.	.	C	7.	G	.	.	.	C
4.	8.	G	.	.	.	T	8.	G	.	.	.	T	8.	GCA	.	.	.	T
9.	9.	G	.	.	.	C	9.	G	.	.	.	C	9.	G	.	.	.	C
10.	10.	G	.	.	.	G	10.	G	.	.	.	G	10.	G	.	.	.	G
11.	11.	T	.	.	.	G	11.	T	.	.	.	G	11.	T	.	.	.	G
3.	12.	T	.	.	.	A	12.	T	.	.	.	A	12.	T	.	.	.	A



# SAM file format

SAM files contain a lot of information about reads and their mapping stats  
When reads map to multiple locations, all of those are indicated in separate lines

Thus, SAM files are big files (1.6 Mb for 10,000 reads).  
They can be compressed to a binary format called BAM using samtools (475 Kb for 10,000 reads).



```
bwa mem /data/genomes/mm10.fa input.fastq | samtools sort -O bam -T output.tmp -o output.bam
```

`samtools sort` will sort the alignments: first by chromosome, then by position  
| is called a “pipe” and allows to use the output of a first command as the input of a second command

# Alignment statistics and indexing

```
samtools stats -in ./Bams/SRR2034943.bam
```

```
sbastide@nobel:/data/personal_folders/sbastide/sequencing_data_tutorial$ bamtools stats -in ./Bams/SRR2034943.bam
*****
Stats for BAM file(s):
*****

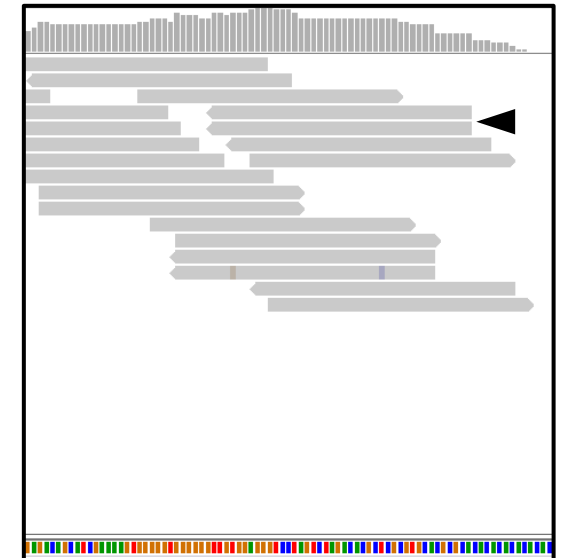
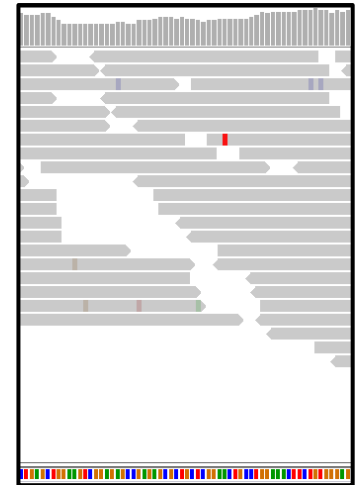
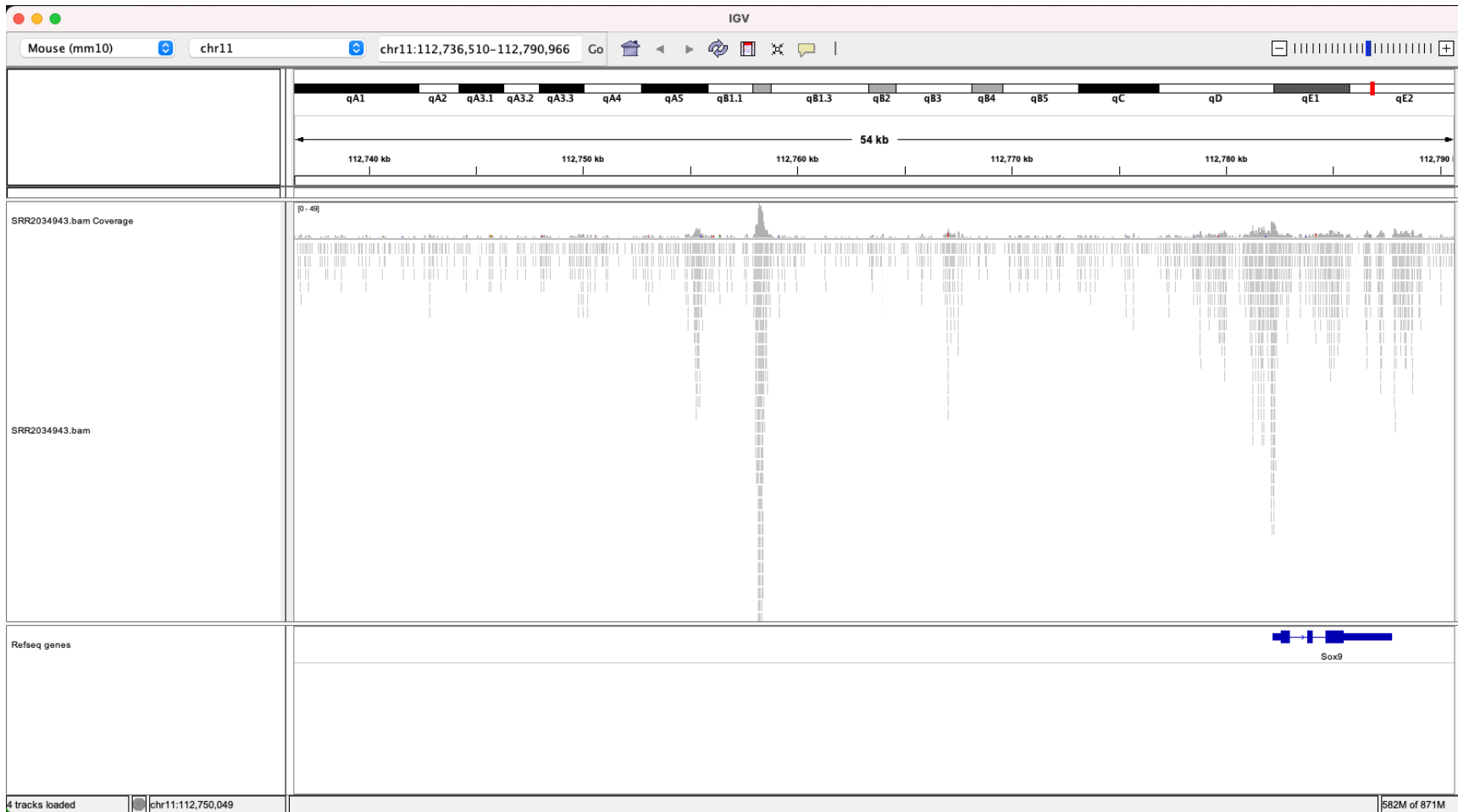
Total reads:      40527423
Mapped reads:    33973845      (83.8293%)
Forward strand:  23575568      (58.1719%)
Reverse strand:  16951855      (41.8281%)
Failed QC:       0      (0%)
Duplicates:      0      (0%)
Paired-end reads: 0      (0%)
```

```
samtools index ./Bams/SRR2034943.bam
```

Creates a .bam.bai file that allows to quickly and efficiently search the BAM file without having to read through every line from the top

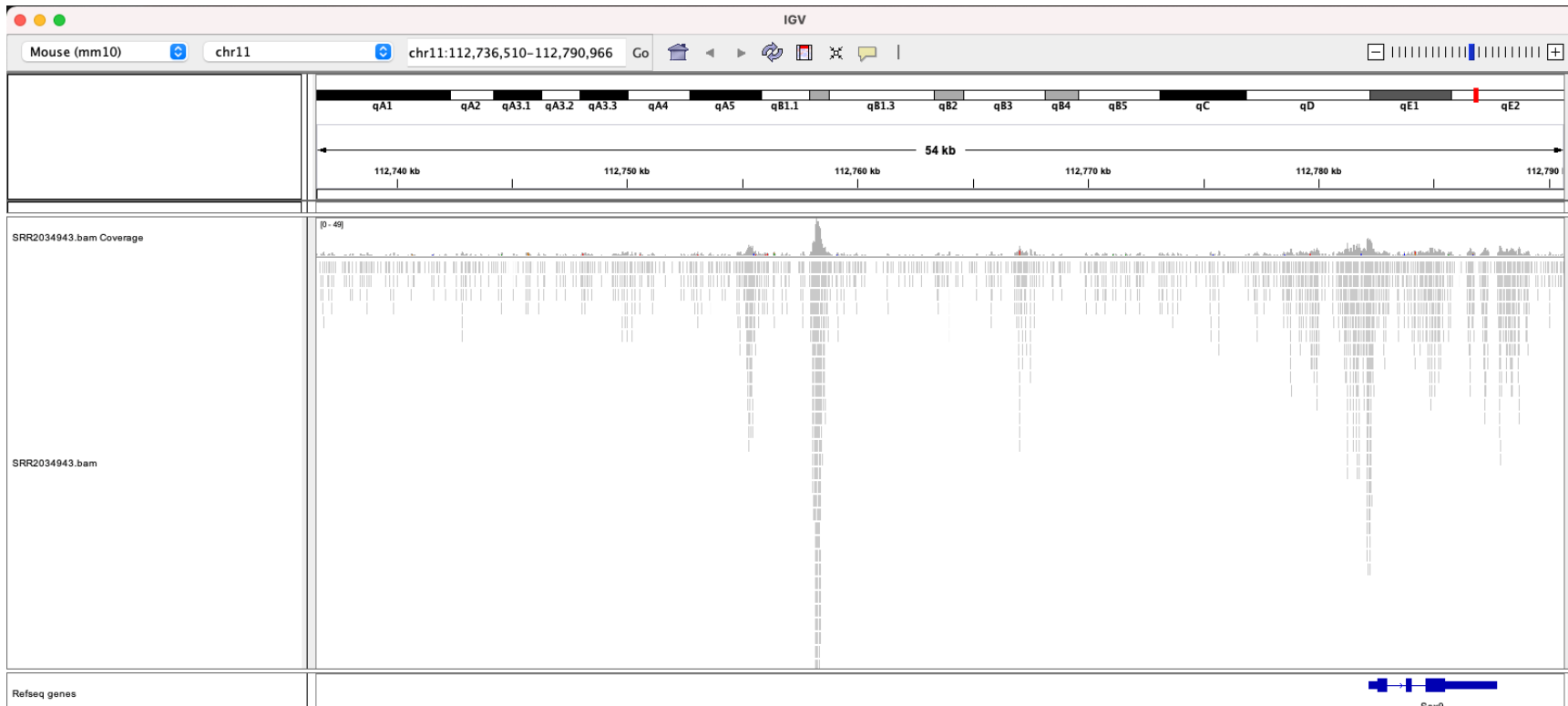
# Visualization using BAM files in IGV

Change to Mm10 →

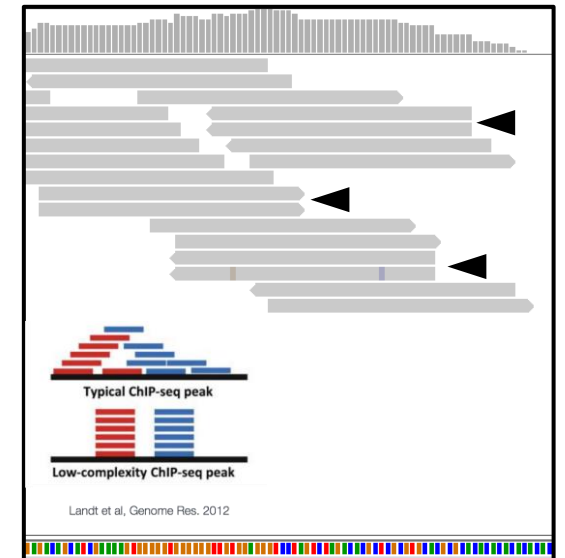
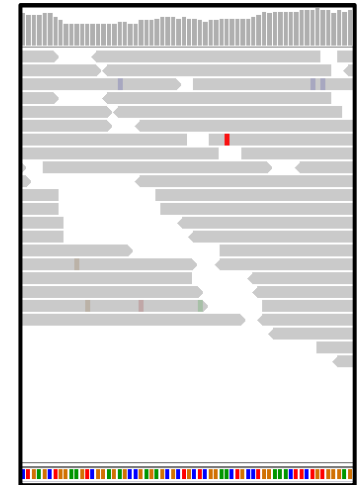


# Visualization using BAM files in IGV

Change to Mm10 →



**PCR duplicates** originate from the library amplification. Performing large numbers of cycles to get enough material to sequence massively increases the number of those duplicates. The expected number of duplicates depends on the method (for instance, a typical run of ATAC-seq has 25-30% of duplicates)





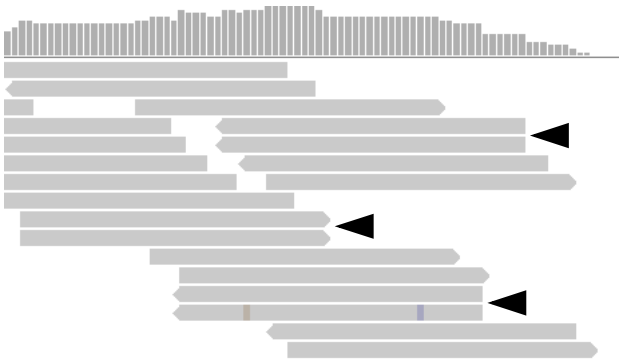
# Optional: removing duplicates using Picard

```
cd Bams
java -jar  picard.jar MarkDuplicates \
  I=bam_file.bam \
  O=bam_file_rmdup.bam \
  M=bam_file_marked_dup_metrics.txt \
  REMOVE_DUPLICATES=True
```

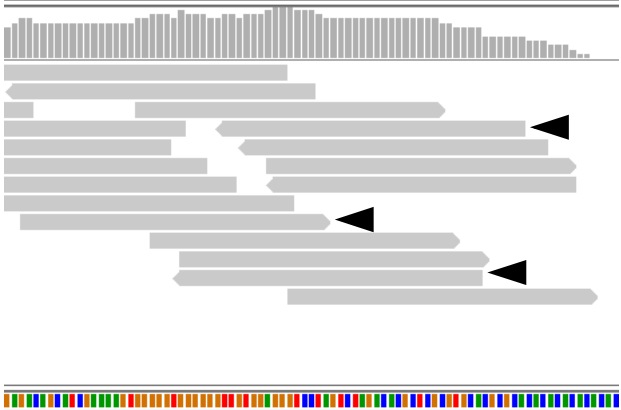
LIBRARY	UNPAIRED READS EXAMINED	UNMAPPED READS	UNPAIRED READ DUPLICATES	PERCENT DUPLICATION
Library	33973845	6553578	6809555	0.200435

Removing PCR duplicates allows a more accurate quantification of the features you are interested in (gene expression, DNA binding, chromatin accessibility, ...).

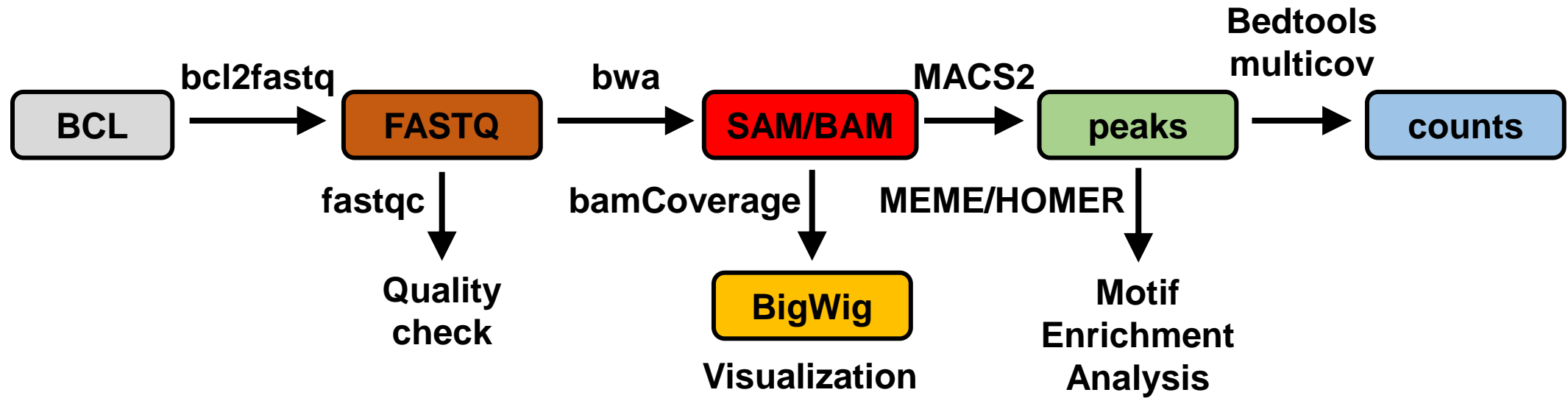
Before



After



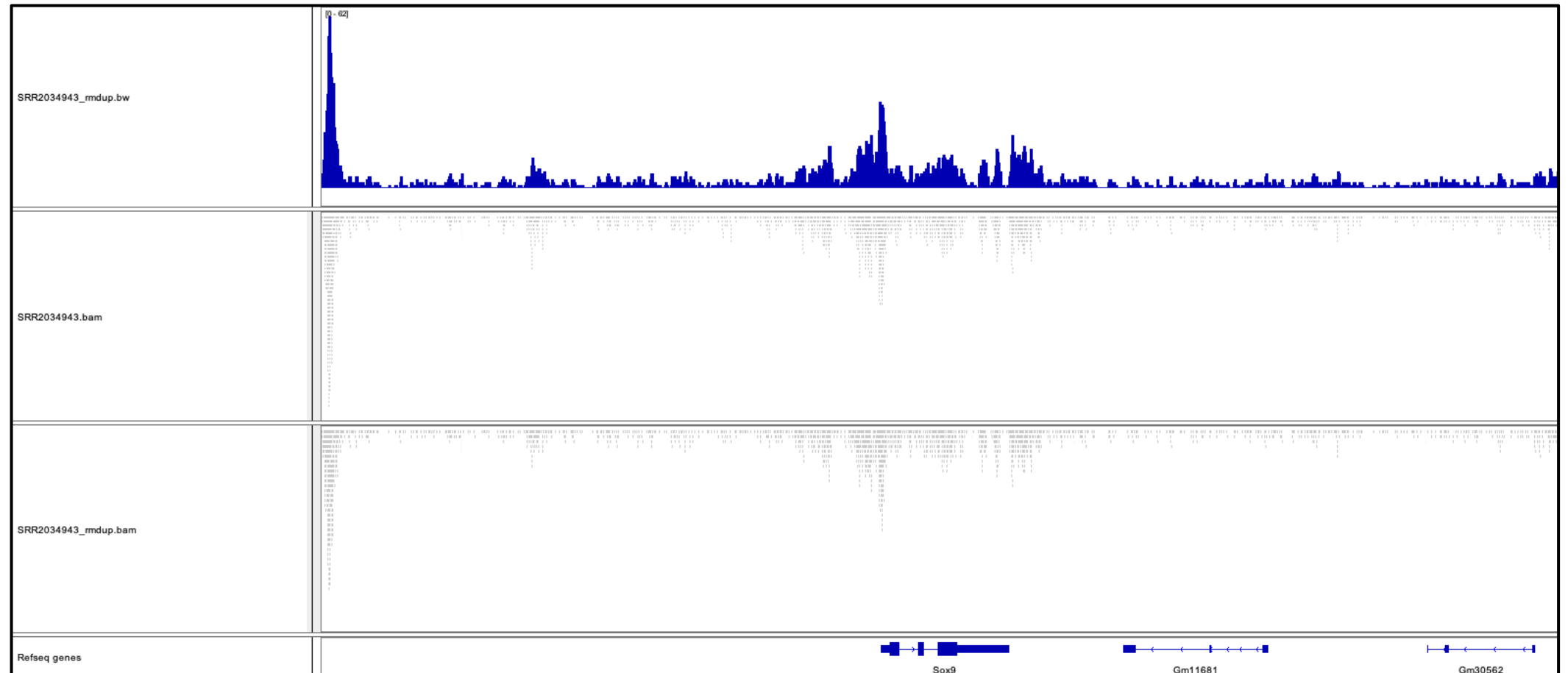
# Full pipeline for ATAC-seq/ChIP-seq



# Visualization using BigWig files

```
sbastide@nobel:/data/personal_folders/sbastide/sequencing_data_tutorial/Bams$ bamCoverage  
usage: An example usage is:$ bamCoverage -b reads.bam -o coverage.bw
```

```
bamCoverage -b ./Bams/SRR2034943_rmdup.bam -o BigWigs/SRR2034943_rmdup.bw
```



# Full pipeline for ATAC-seq/ChIP-seq

