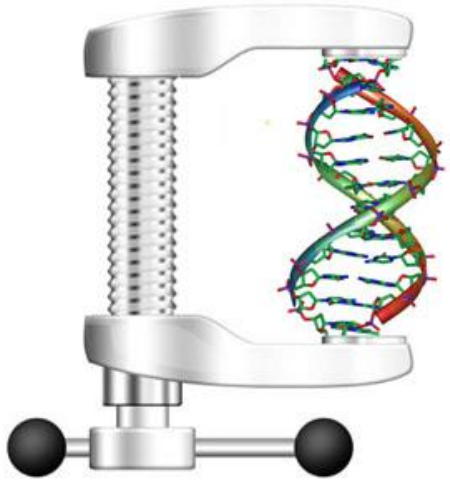
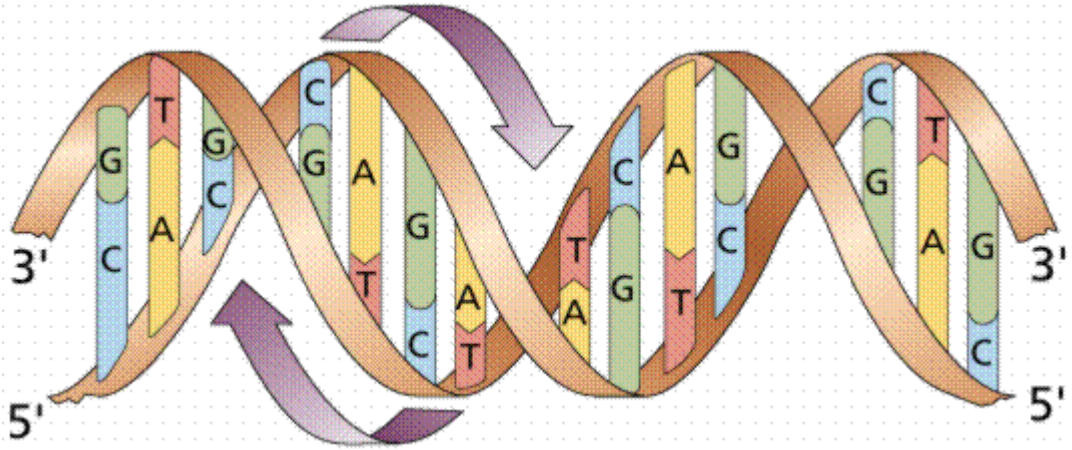
Genome Information Compression



The Human Genome

- 4 possible symbols (**bases**): A,C,G,T
- 3.2 billions base pairs for the human genome
  - 3.2 x 2 bit = 6.4 billion bits / 8 = 800 Mbytes



- So why do we need compression?

Genome sequencing

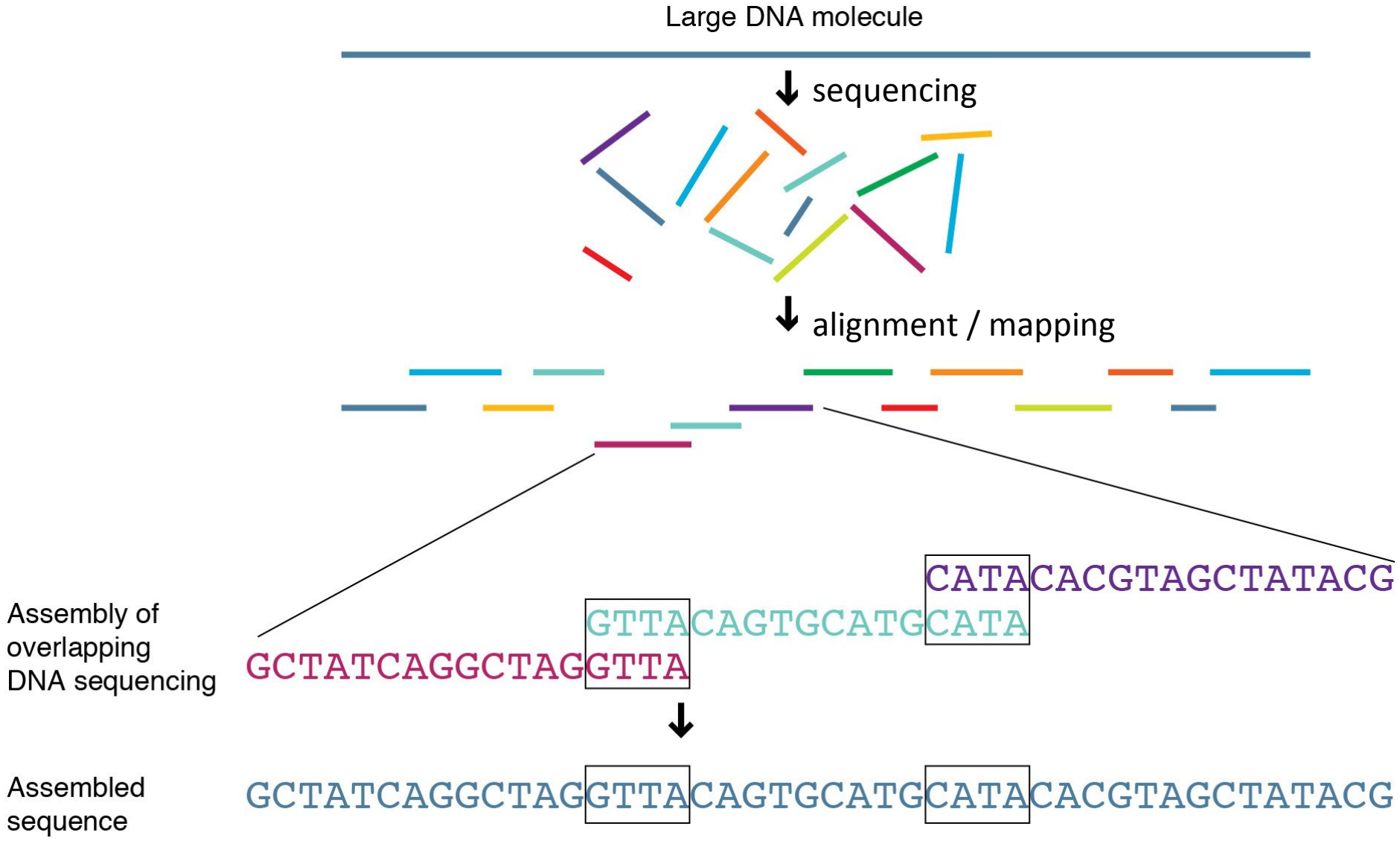
- How do we obtain a genome representation from biological samples?
- Current technology provides random fragments of genome data called “reads”
- This process is called **genome sequencing**

Short and long reads

- Different sequencing technologies provide
  - Read lengths from about 100 to 20,000+ bases
  - Single or coupled (paired) reads
  - Different accuracy levels (from 60% to 99.9%)

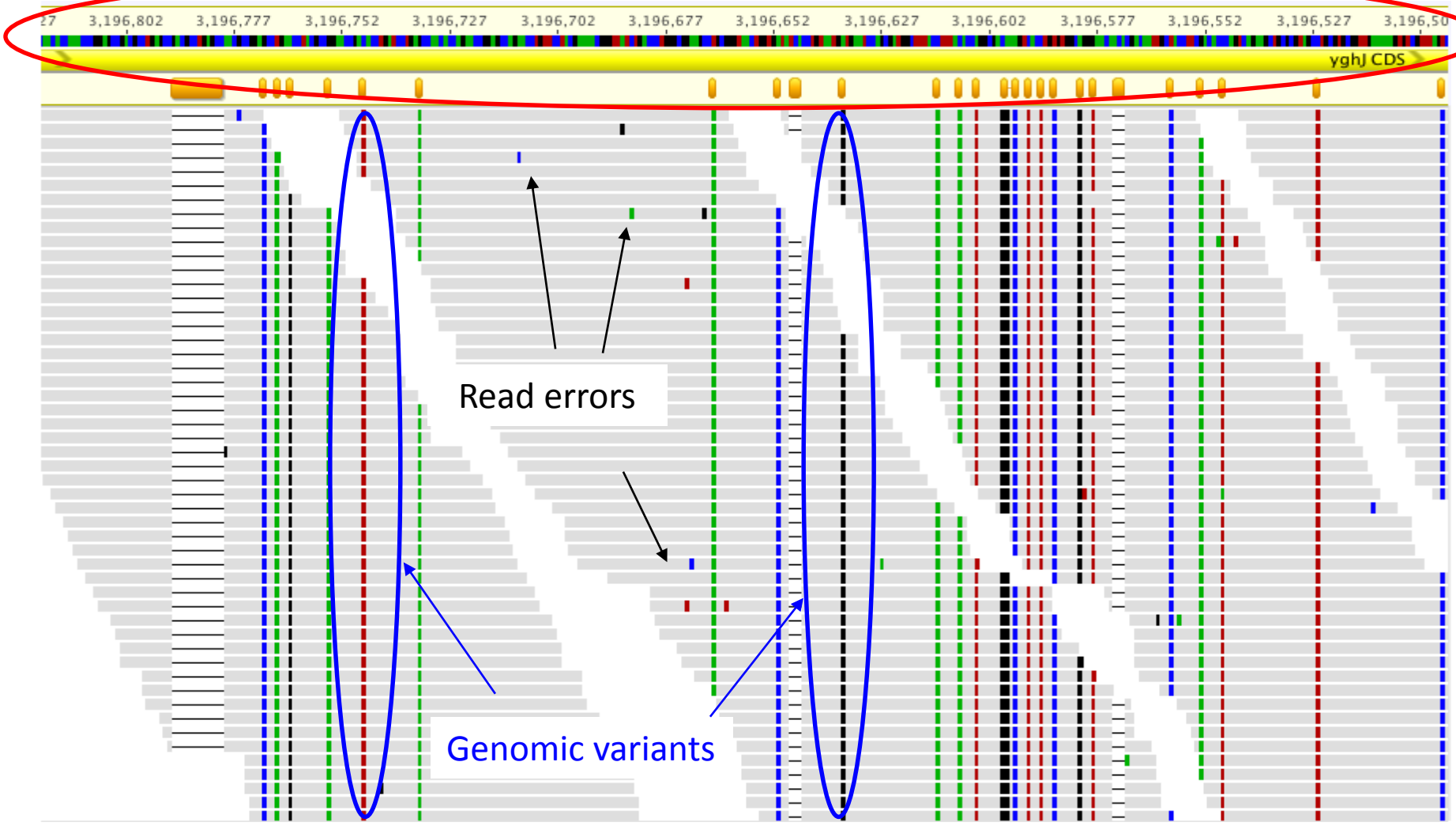
- Shorter reads are
  - more accurate (up to 99.9%)
  - produced in much larger volumes (currently up to 600 billion bases per single run)

From sequencing to assembly



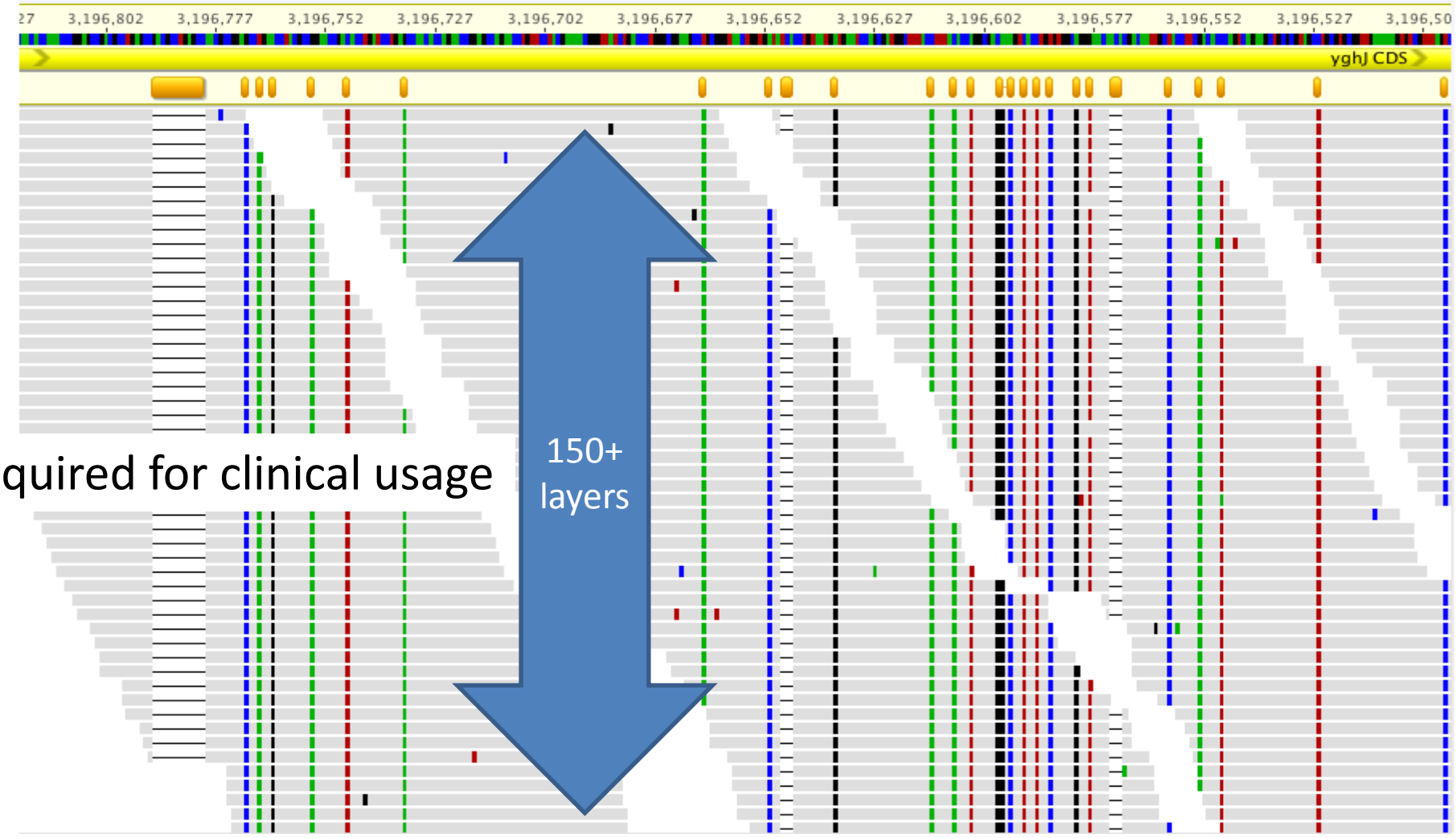
Mapped reads

Reference genome



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Coverage



Required for clinical usage

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

All reads must be preserved

Is this reading noise or a mutation?

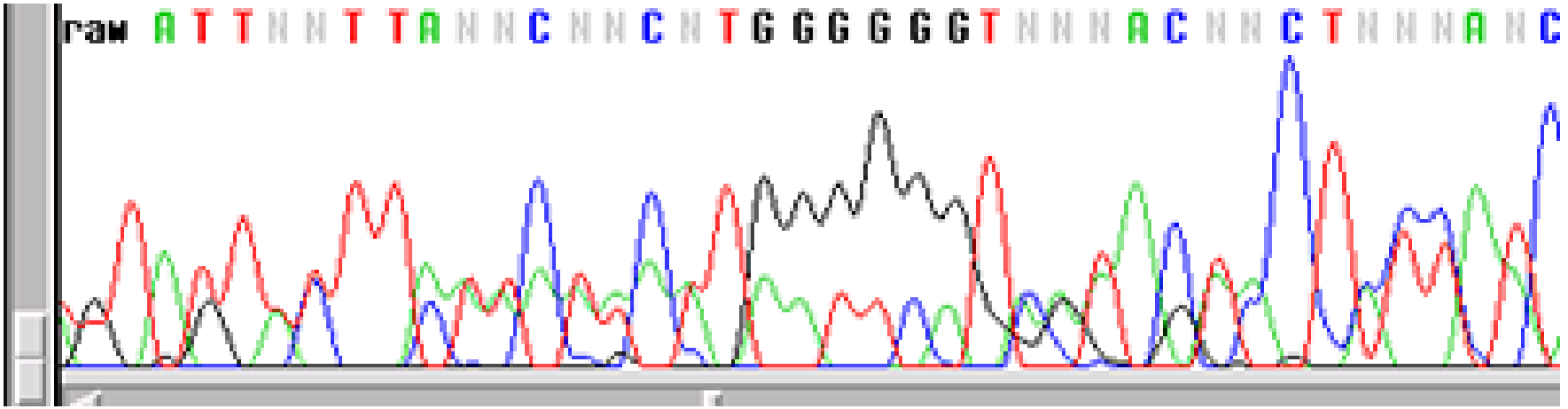
Data volumes

- 3.2 billion per genome
  - X 200+ for clinical usage (to get at least 150 layers)

- Additional information
  - 1 quality score per each base (up to 96 possible levels)

Reads quality scores

- Each base call in a read has a level of confidence



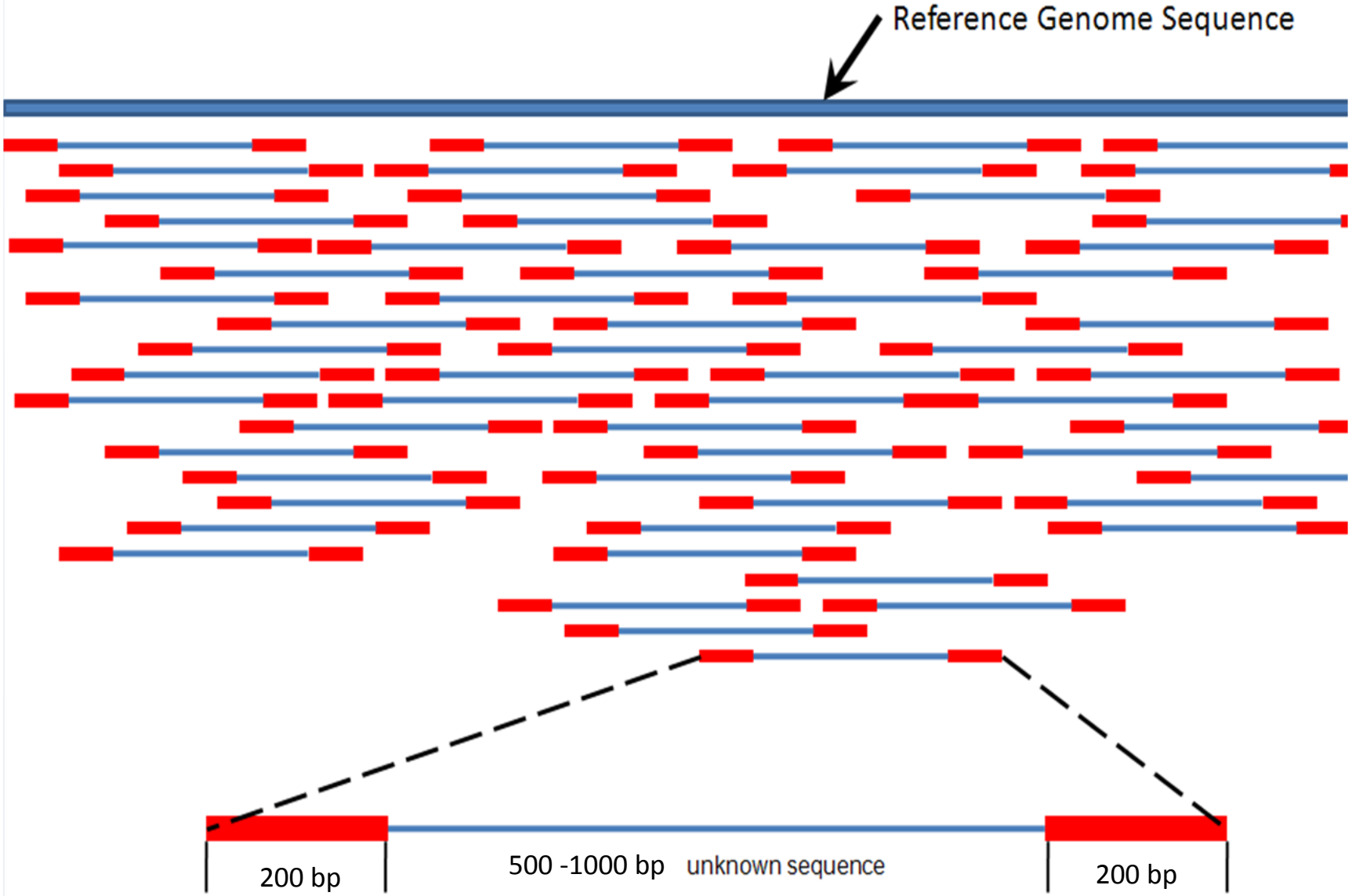
- The level of confidence is expressed as “quality score” in a range that is machine dependent represented as ASCII character.

Data volumes

- 3.2 billion per genome
  - X 200+ for clinical usage (to get at least 150 layers)

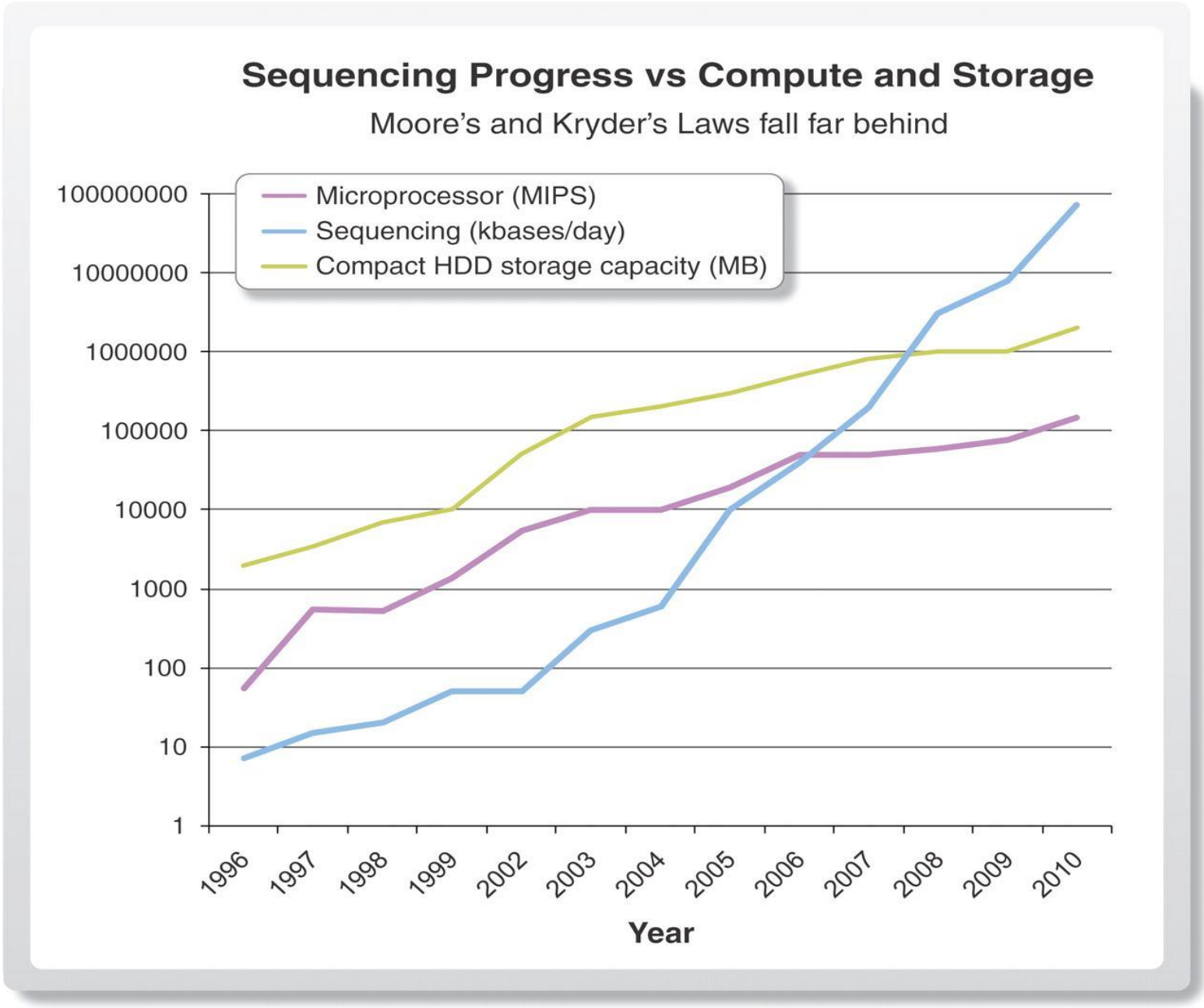
- Additional information
  - 1 quality score (ASCII char) per each base
  - pairing information for coupled reads (labelling)

Paired reads



Data volumes

- 3.2 billion per genome
  - X 200+ for clinical usage (to get at least 150 layers)
- Additional information
  - 1 quality score per each base (up to 96 possible levels)
  - pairing information for coupled reads (labelling)
- Total = 3.2GB x 200 x 2 x labelling ~= 1.5 TB
  - Labelling ~= 1.15


Sequencing Progress vs Compute and Storage
Moore's and Kryder's Laws fall far behind
Microprocessor (MIPS)
Sequencing (kbases/day)
Compact HDD storage capacity (MB)
100000000
10000000
1000000
100000
10000
1000
100
10
1
1996
1997
1998
1999
2002
2003
2004
2005
2006
2007
2008
2009
2010
Year

Raw data format

FASTQ	Field	FASTA
@HWUSI-EAS100R:6:73:941:1973#0/1	*Header (Unique ID plus other information). Only the first character is standard.*	**>HWUSI-EAS100R:6:73:941:1973#0/1**
GATTTGGGGT…..	*Nucleotides sequence*	**GATTTGGGGT……**
+SRR001666.1 071112_SLXA-EAS1_s_7	*Optional description. Only the first character is standard. This is becoming obsolete*	Not present
!''*((((*+)**	*Quality scores*	Not present

Example

One read:

```
HS2000-1240_45:1:1234:6966:12500
AAATATTTTTTAAAATTAGCCAGGTGTGGTGGTGTGTGCCTATAGTTCCAACTAGTTGTAAAGCTGAAACATAAGGACCACTTGGGTACAGGAGTTCCAA
+
CCCFFFFFHHHHHIJIIJIIJIJJFHGGIFHHHIIJJJIJIHIJJFIJJIIJJJHIIIIGIIIIIIFHGGGHFFFFF?BBEECDD?CCCECDD?AC;5@
```

Example

One read:

```
HS2000-1240_45:1:1234:6966:12500/1
AAATATTTTTTAAAATTAGCCAGGTGTGGTGGTGTGTGCCTATAGTTCCAACTAGTTGTAAAGCTGAAACATAAGGACCACTTGGGTACAGGAGTTCCAA
+
CCCFFFFFHHHHHIJIIJIIJIJJFHGGIFHHHIIJJJIJIHIJJFIJJIIJJJHIIIIGIIIIIIFHGGGHFFFFF?BBEECDD?CCCECDD?AC;5@


HS2000-1240_45:1:1234:6966:12500/2
ATCAGATGTATAATTTGCAAATAGTTTCTCTCATTCTTTTTTTTTTTTTTTTAGACAGGGTCTCACTGTATTGCCCAGGCTGGAGTGCAGTGGTGCAATC
+
CCCFFFFFHHHHHJJJJJJJJIIJJJIJJJJIJIJJJIJJJJJJJJHFDD@##################################################
```

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Example

- A read aligned onto a section of the chromosome 1 of the reference genome.

- Chr1 39999220 39999240 39999260 39999280 39999300 39999320 39999340
- . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | .
|
- AATGGTGCAGTCACAGCTGTCTACAAAAATATTTTTTAAAATTAGCCAGGTGTGGTGGTGTGTGCCTATAGTTCCAACTAGTTGTAAAGCTGAAACATAAGGACCACTTGGGTACAGGAGTTCCA AGACTGTGGTGAG
- ..........AAATATTTTTTAAAATTAGCCAGGTGTGGTGGTGTGTGCCTATAGTTCCAACTAGTTGTAAAGCTGAAACATAAGGACCACTTGGGTACAGGAGTTCCAA.....

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Example

- A read aligned onto a section of the chromosome 1 of the reference genome.

```
 Chr1          39999220            39999240            39999260            39999280            39999300            39999320
39999340
                 .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .
|
AATGGTGCAGTCACAGCTGTCTACAAAAATATTTTTTAAAATTAGCCAGGTGTGGTGGTGTGTGCCTATAGTTCCAACTAGTTGTAAAGCTGAAACATAAGGACCACTTGGGTACAGGAGTTCCA AGACTGTGGTGAG
.........................AAATATTTTTTAAAATTAGCCAGGTGTGGTGGTGTGTGCCTATAGTTCCAACTAGTTGTAAAGCTGAAACATAAGGACCACTTGGGTACAGGAGTTCCAA............
```

- A second read, known to be associated with the first read.

```
39999340            39999360            39999380            39999400            39999420            39999440            39999460
  .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |
GTGGTGAGCTGTGATTGCACCACTGCACTCCAGCCTGGGCAATACAGTGAGACCCTGTCTAAAAAAAAAAAAAAAAAAGAATGAGAGAAACTATTTGCAAATTATACATCTGATAAGAGACCTGTAT CTA
............GTTTTGCCCCCCTCACCCCAGCCTGGGGAATAAAGGGGGGCCCTTTCTAAAAAAAAAAAAAAAAAAGAATGAGAGAAACTATTTGCAAATTATACATCTGAT.............
```

Example

- A read aligned onto a section of the chromosome 1 of the reference genome.

```
 Chr1          39999220            39999240            39999260            39999280            39999300            39999320
39999340
                .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .
|
AATGGTGCAGTCACAGCTGTCTACAAAAATATTTTTTAAAATTAGCCAGGTGTGGTGGTGTGTGCCTATAGTTCCAACTAGTTGTAAAGCTGAAACATAAGGACCACTTGGGTACAGGAGTTCCA AGACTGTGGTGAG
.........................AAATATTTTTTAAAATTAGCCAGGTGTGGTGGTGTGTGCCTATAGTTCCAACTAGTTGTAAAGCTGAAACATAAGGACCACTTGGGTACAGGAGTTCCAA............
```

- A second read, known to be associated with the first read.

```
39999340            39999360            39999380            39999400            39999420            39999440            39999460
  .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |
GTGGTGAGCTGTGATTGCACCACTGCACTCCAGCCTGGGCAATACAGTGAGACCCTGTCTAAAAAAAAAAAAAAAAAAGAATGAGAGAAACTATTTGCAAATTATACATCTGATAAGAGACCTGTAT CTA
............GTTTTGCCCCCCTCACCCCAGCCTGGGGAATAAAGGGGGGCCCTTTCTAAAAAAAAAAAAAAAAAAGAATGAGAGAAACTATTTGCAAATTATACATCTGAT.............
```

HS2000-1240_45:1:1234:6966:12500/2
ATCAGATGTATAATTTGCAAATAGTTTCTCTCATTCTTTTTTTTTTTTTTTTAGACAGGGTCTCACTGTATTGCCCAGGCTGGAGTGCAGTGGTGCAATC

Note that the original sequence of the second read had to be reverse complemented to match the genome.

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Example

- A read aligned onto a section of the chromosome 1 of the reference genome.

```
 Chr1          39999220            39999240            39999260            39999280            39999300            39999320
39999340
                .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .
|
AATGGTGCAGTCACAGCTGTCTACAAAAATATTTTTTAAAATTAGCCAGGTGTGGTGGTGTGTGCCTATAGTTCCAACTAGTTGTAAAGCTGAAACATAAGGACCACTTGGGTACAGGAGTTCCA AGACTGTGGTGAG
..........................>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>> >............
```

- A second read, known to be associated with the first read.

```
39999340            39999360            39999380            39999400            39999420            39999440            39999 460
  .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |
GTGGTGAGCTGTGATTGCACCACTGCACTCCAGCCTGGGCAATACAGTGAGACCCTGTCTAAAAAAAAAAAAAAAAGAATGAGAGAAACTATTTGCAAATTATACATCTGATAAGAGACCTGTAT CTA
............<T<<TGC<<C<CT<<<C<<<<<<<<<<G<<<<A<<G<G<G<<<<T<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<.............. ...
```

Note that the sequence obtained may not reflect exactly the genomic sequence.

Example

- A read aligned onto a section of the chromosome 1 of the reference genome.

```
 Chr1          39999220            39999240            39999260            39999280            39999300            39999320
39999340
                 .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .
|
AATGGTGCAGTCACAGCTGTCTACAAAAATATTTTTTAAAATTAGCCAGGTGTGGTGGTGTGTGCCTATAGTTCCAACTAGTTGTAAAGCTGAAACATAAGGACCACTTGGGTACAGGAGTTCCA AGACTGTGGTGAG
..........................>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
                               CCCFFFFFHHHHHIJIIJIIJIJJFHGGIFHHHIIJJJIJIHIJJFIJJIIJJJJHIIIIGIIIIIIFHGGGHFFFFF?BBEECDD?CCCECDD?AC;5@
```

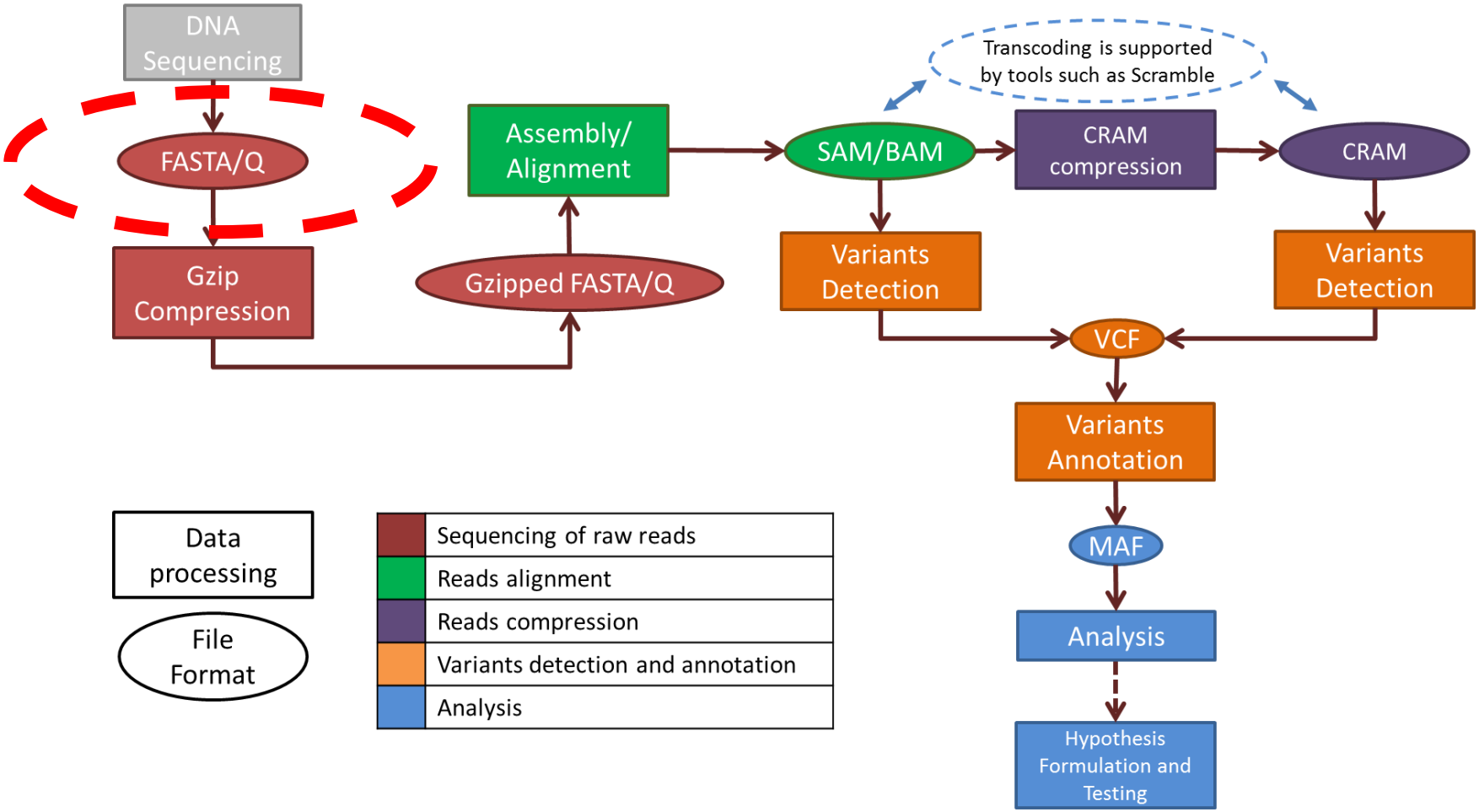
- A second read, known to be associated with the first read.

```
39999340            39999360            39999380            39999400            39999420            39999440            39999460
  .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |
GTGGTGAGCTGTGATTGCACCACTGCACTCCAGCCTGGGCAATACAGTGAGACCCTGTCTAAAAAAAAAAAAAAAAGAATGAGAGAAACTATTTGCAAATTATACATCTGATAAGAGACCTGTAT CTA
............<T<<TGC<<C<CT<<<C<<<<<<<<<<G<<<<A<<G<G<G<<<<T<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<................
            ################################################@DDFHJJJJJJJIJJJIJIJJJJIJJJJIIJJJJJJJHHHHHFFFFFCCC
```

Here, the quality of the sequence obtained, associated to each nucleotide has been flagged as poor in the region where it does not match well the reference genome.

```
@SRR000921.1 E7OAHBI01AQMFC length=41
GGGGGGGTGTAGAAAAGATGTGAGATACCTTTCCTCACCCT
+
::::555GFFGC66;;=GHHHHHHHHHHHFFFFFFGC955C
@SRR000921.3 E7OAHBI01A9JXR length=126
GGGGTGGTATGGAGAGACAATGGGCGATGTTTCTCAGGGCTGCTTCAAGCGGGATTAGGGGCAGCGTGGGAAGCTAGAGTGGGAGAGATTAAGCTGAAGG
+
0777=GFFFFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIFFFFFFFFFFFFCCCCFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@SRR000921.7 E7OAHBI01A9IFH length=173
GGGGGGGGGGGGTTGGGCATGTGGGACCCCCTCCATGTTGACAGTCGAGGCATACGGCCTGCAGGTCCACGCTTACTTGCTGTCCATGCTGTCGGTGTCA
+
//////////////88FFFFFFFFEFEFFFFFEEEFFFGGGGGIIIIFFFFFFFFFFFFFFFFFFFFFFFFFFFFEEEFFFFFFFFFFFFFFFFFFFFFF
@SRR000921.9 E7OAHBI01B333T length=153
GGGGGTTTCCTGGTGGGGTTAGAGGCGGGACGTGAGAAACGTGGCAGGTGGTGGCCAGGCTGTGGGGGAGCCAGGGTATTTCTTGATGGGGTTAGATGGG
+
0000055FFFFFFF;=??EEHGHHHHIIIIIIIIIIIIIIIIIIFFFFFFFFFFFFFFFFFF@@@??GFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@SRR000921.10 E7OAHBI01D6JUW length=87
GGGGGTGCCTCCCAGTTAGGCTGCTCAGGGGTCAGGGGTCAGGGACCCGCTTGAGGAGGCAGTCTGCCCATTCTCAGATCTCCAGCC
+
00000=CFFFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@SRR000921.11 E7OAHBI01C8DW7 length=162
GGGGCAATTGAGTTGGGCATTGATGGCCAAGGAGTTCCGGGGAAACCATGTTGAAGCTGGGAGAGTAGAACACCTGAAATCCTGTGAAATCCATAAAATT
+
BBBBFFFFFFFFFFFFFIIIIHHHHHHIIIIIIIIIIIIIIIIIFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@SRR000921.16 E7OAHBI01EFO7A length=84
GGGGGACAAAATTCAGTCCATAGTACATGCACGATCTCATTTGATCTTCCAGAACCCCATGACATAGTTTCCATTTGAAGATGA
+
////58@DDDDDDDDDDDIIIIIIIIIIIIIIIIIIIIIIIIIIIDDDDDDDCCAAAADDDDDDDDDDDDDDDDDDDDDD
@SRR000921.17 E7OAHBI01BRP08 length=148
GGGGGGGTCAGCCCCCCCGCCCGGCCAGCCGCCCCATCCAGGAGGTGAGGGGCACCTCTGCCCGGCCGCCCCTACTGGGAAGTGAGGAGCCCCTCTGCCC
+
4440000<@@<///////5+++AADDEECCC====CEEEEFFHHICCCCCCCCCCCCCCC===CCCCCCCCCCCGCCCCCCHCCCCCCCCCCCCCCCCC
@SRR000921.18 E7OAHBI01BAECM length=67
GGGGGCGCCTCTTAACACACCAGCTGGTGGACTGTGAATCATTCTGCAGCAGGAAGCGCATGCATGG
+
A@@@@FFFFFFFFFFFFIIIIIIIIIIIIIIIIIIIHHHHHHHGFFFFFFFFFFFFFFFFFFFFF
```

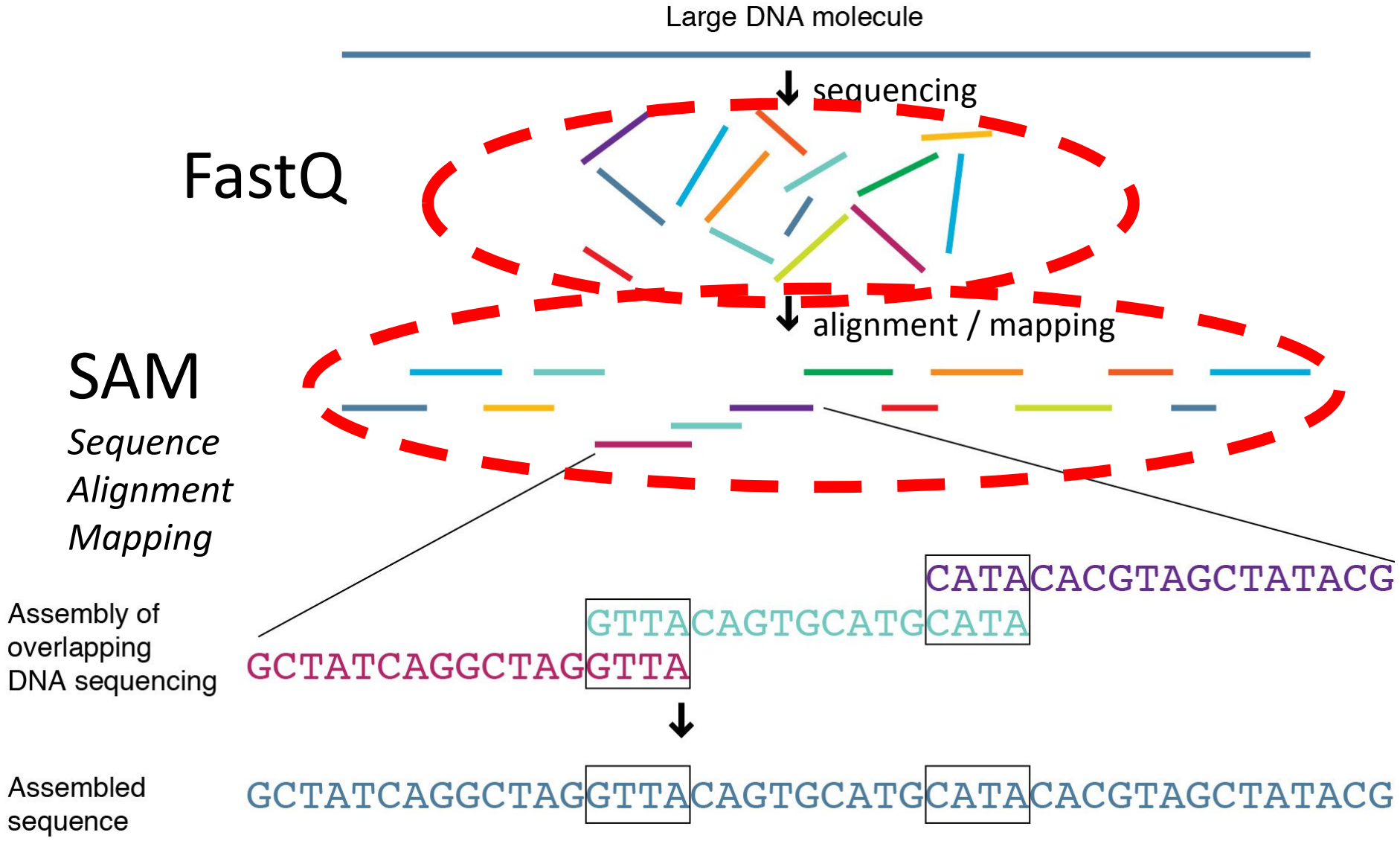
Genome processing pipeline



FastQ compression today

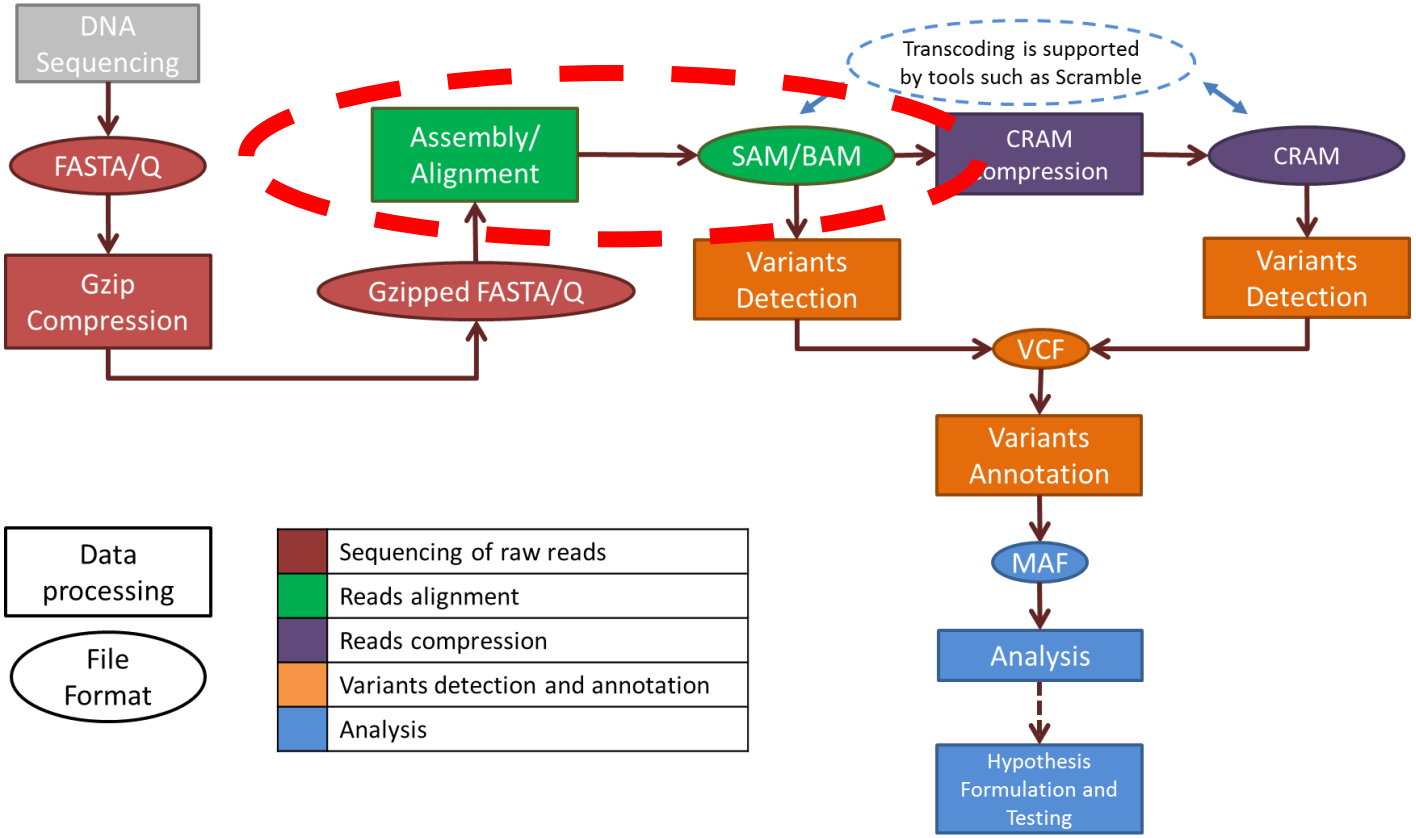
- Gzip of the entire txt file (sometimes split into several files)
- Compression ratio : 3 to 5
- According to the coverage 1 genome can take up to 2 TB

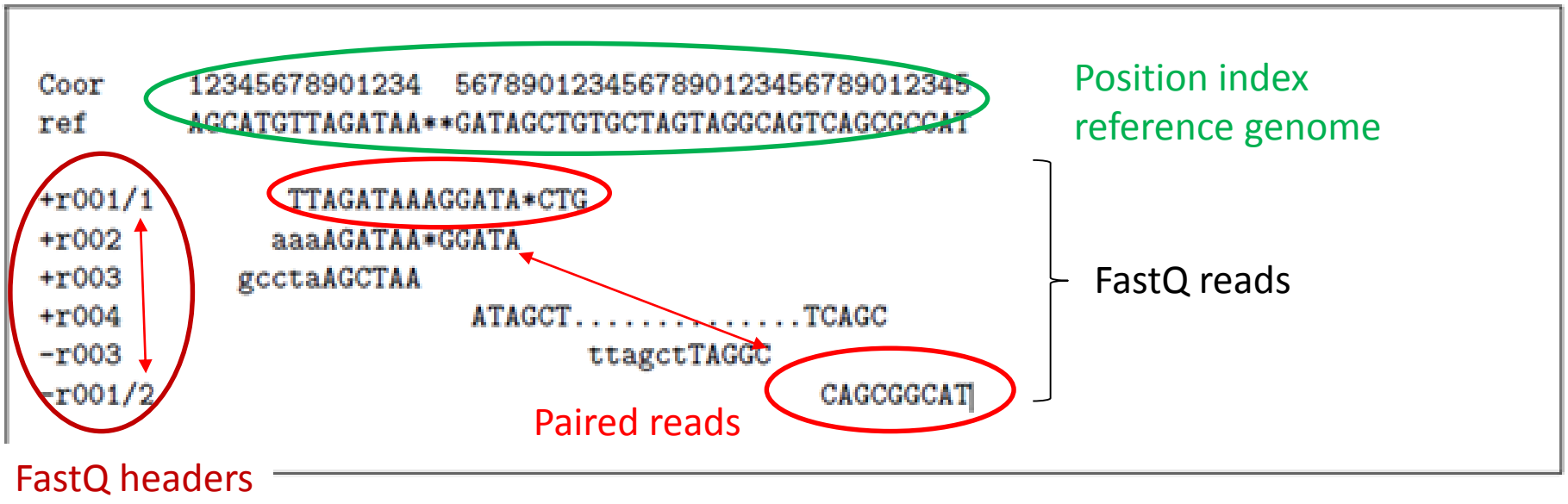
From sequencing to assembly

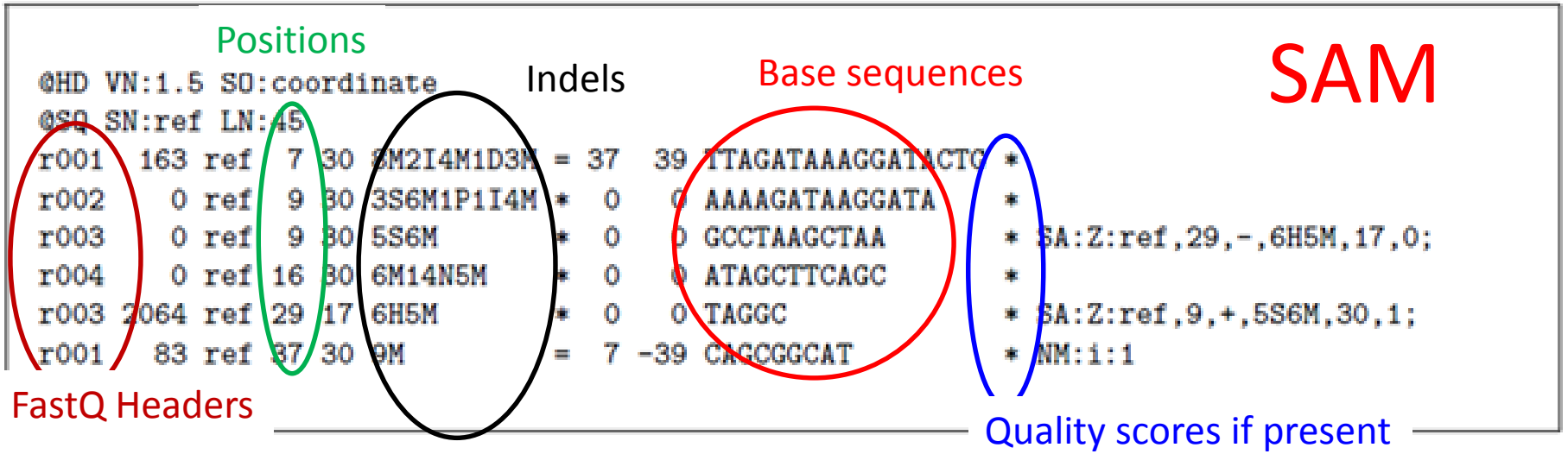


Alignment / mapping

- Raw data + alignment information




Coor 12345678901234 5678901234567890123456789012345
ref AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
+r001/1
+r002
+r003
+r004
-r003
-r001/2
TTAGATAAAGGATA*CTG
aaaAGATAA*GGATA
gcctaAGCTAA
ATAGCT..............TCAGC
ttagctTAGGC
CAGCGGCAT
Position index
reference genome
FastQ reads
Paired reads
FastQ headers


Positions
Indels
Base sequences
SAM
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 83 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
FastQ Headers
Quality scores if present

Compressed SAM = BAM

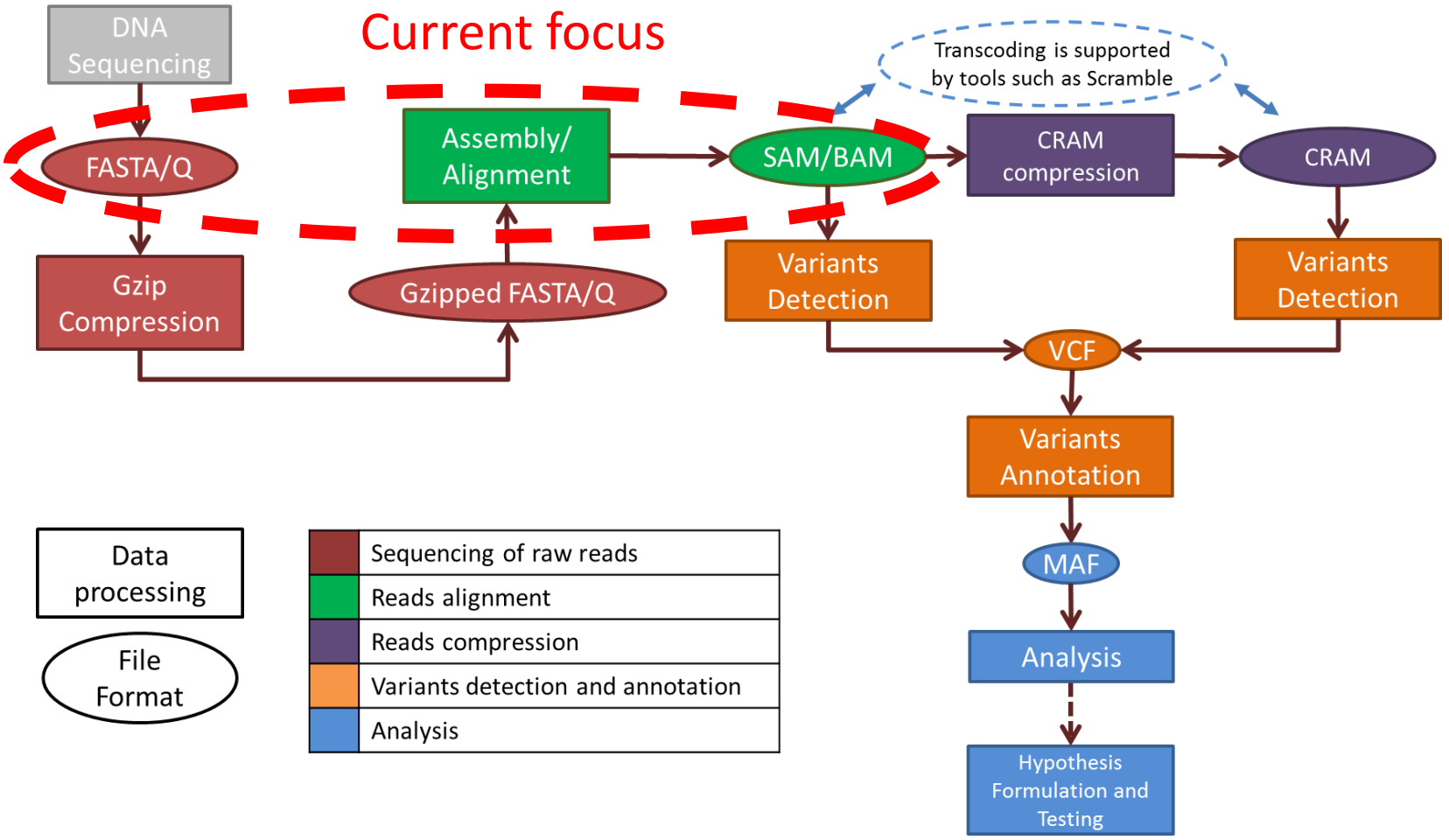
- BAM = Block based zipped SAM
- Indexable for random access
- Compression ratio over textual SAM ~= 4 to 6

- Example (coverage 200x)
  - Fastq = 1.5 TB
    - Fastq.gz = 370 GB
    - Fastq.bzip = 255 GB
    - quip = 205 GB (best compression tool for FastQ)
  - SAM = ~3 TB
  - BAM = 500 GB

SAM/BAM view demo

- Human sample from the MPEG dataset
  - /human/illumina/LowCoverage/NA21144.chrom11
- Chromosome 11
  - Reads length: 100 bases
  - No. of reads: ~10.1 millions
  - Unaligned data: 2.2 GB
  - Aligned SAM: 4.5 GB
  - Compressed BAM: 1 GB

Raw sequence data + Aligned data



Thank you