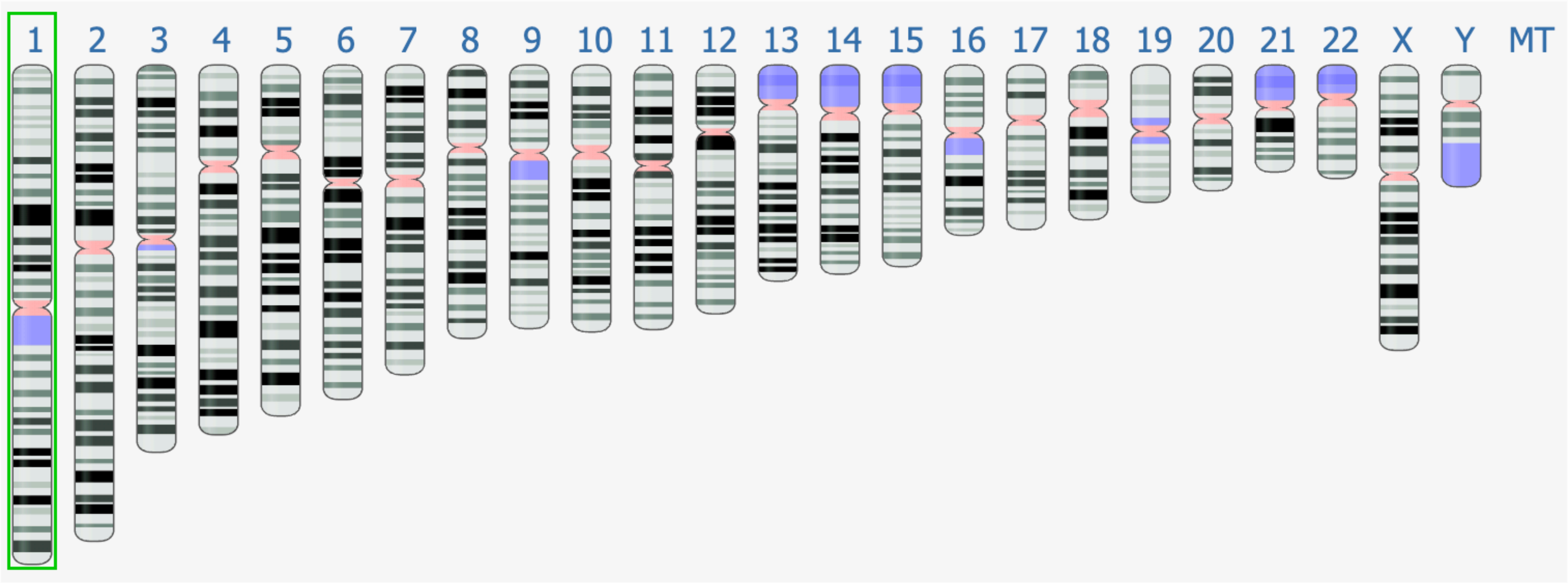


Genomics106

BIOL647
Digital Biology

Rodolfo Aramayo

Genomics106

Mapping and Alignment 102
Introduction

- Mapping small reads a great way to go to understand a highly related genomes

- Traditional sequence alignment algorithms (e.g., blast) cannot be scaled to align millions of reads

- We need to utilize programs that utilize genome indexing algorithms such as Burrows-Wheeler for ultrafast and memory-efficient alignment
  - Years-CPU versus Hours-CPU cost

- The Burrows-Wheeler index – novel approach based on:
  - Mathematics
  - Computer Sciences
- Clean example of theoretical research in the area of data compression applied to computational genomics

- Burrows-Wheeler developed their ideas before small-reads sequencing was available

Genomics106

Mapping and Alignment 101

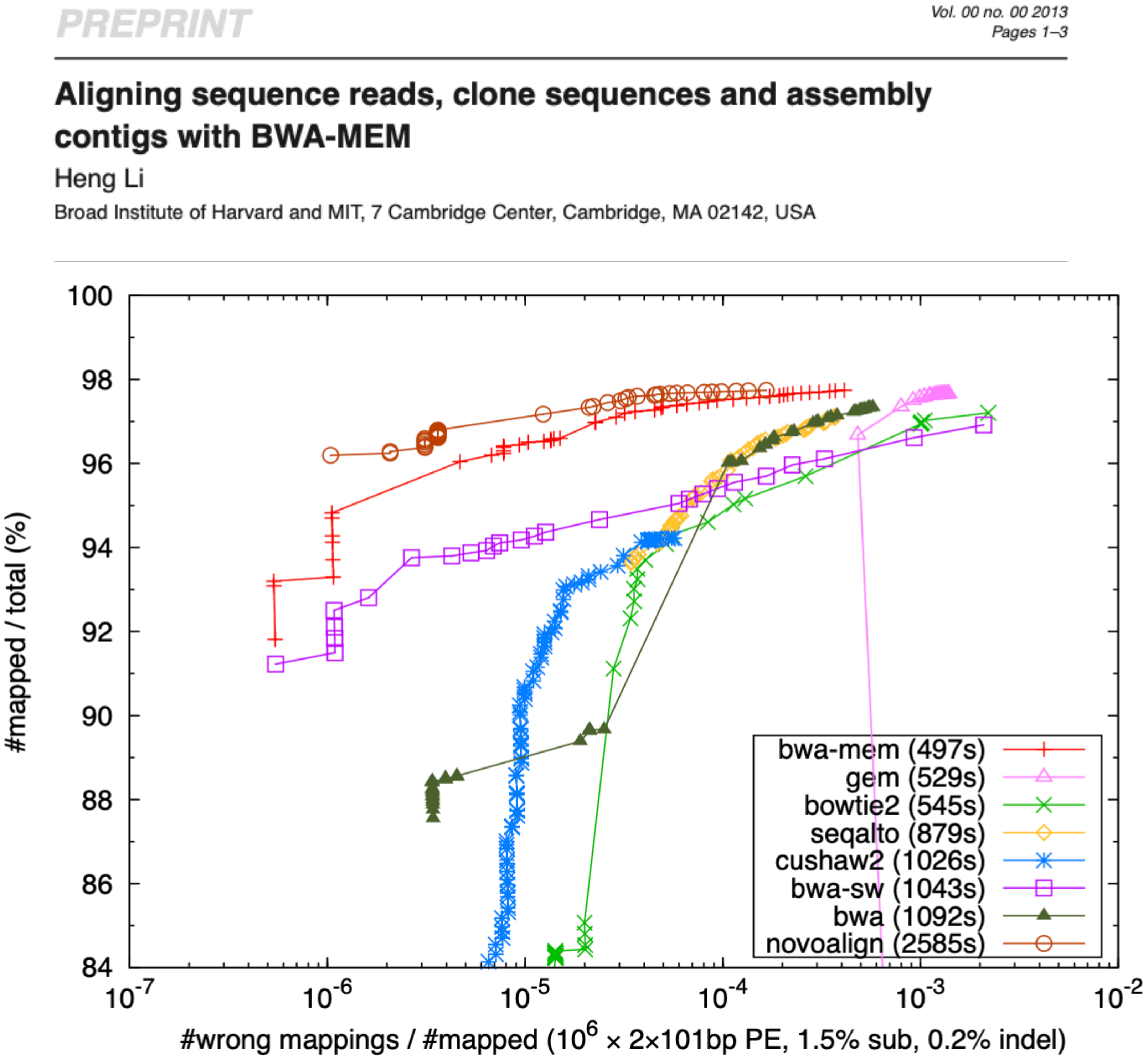
Most Commonly-Used

Sequence Alignment Software

Aligner	Approach	Applications	Availability
BWA-mem	Burrows-Wheeler	DNA, SE, PE, SV	open-source
Bowtie2	Burrows-Wheeler	DNA, SE, PE, SV	open-source
Novoalign	hash-based	DNA, SE, PE	free for academic use
TopHat	Burrows-Wheeler	RNA-seq	open-source
STAR	hash-based (reads)	RNA-seq	open-source
GSNAP	hash-based (reads)	RNA-seq	open-source

Genomics106

Mapping and Alignment 102
Mapping with BWA



PREPRINT

Vol. 00 no. 00 2013
Pages 1–3

Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM

Heng Li

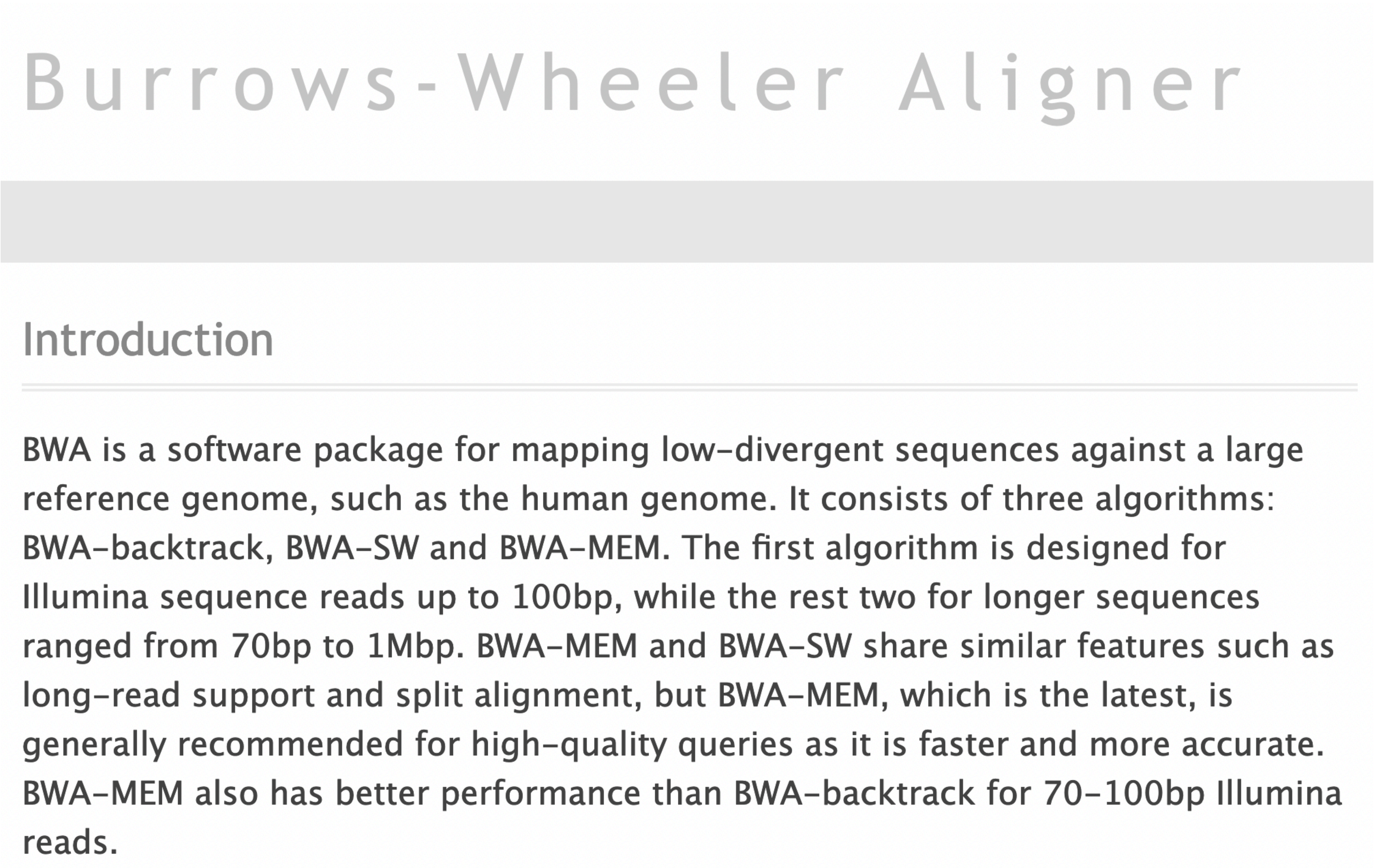
Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA 02142, USA

https://arxiv.org/pdf/1303.3997v2.pdf

Genomics106

Mapping and Alignment 102
Mapping with BWA

BWA Documentation



Burrows-Wheeler Aligner

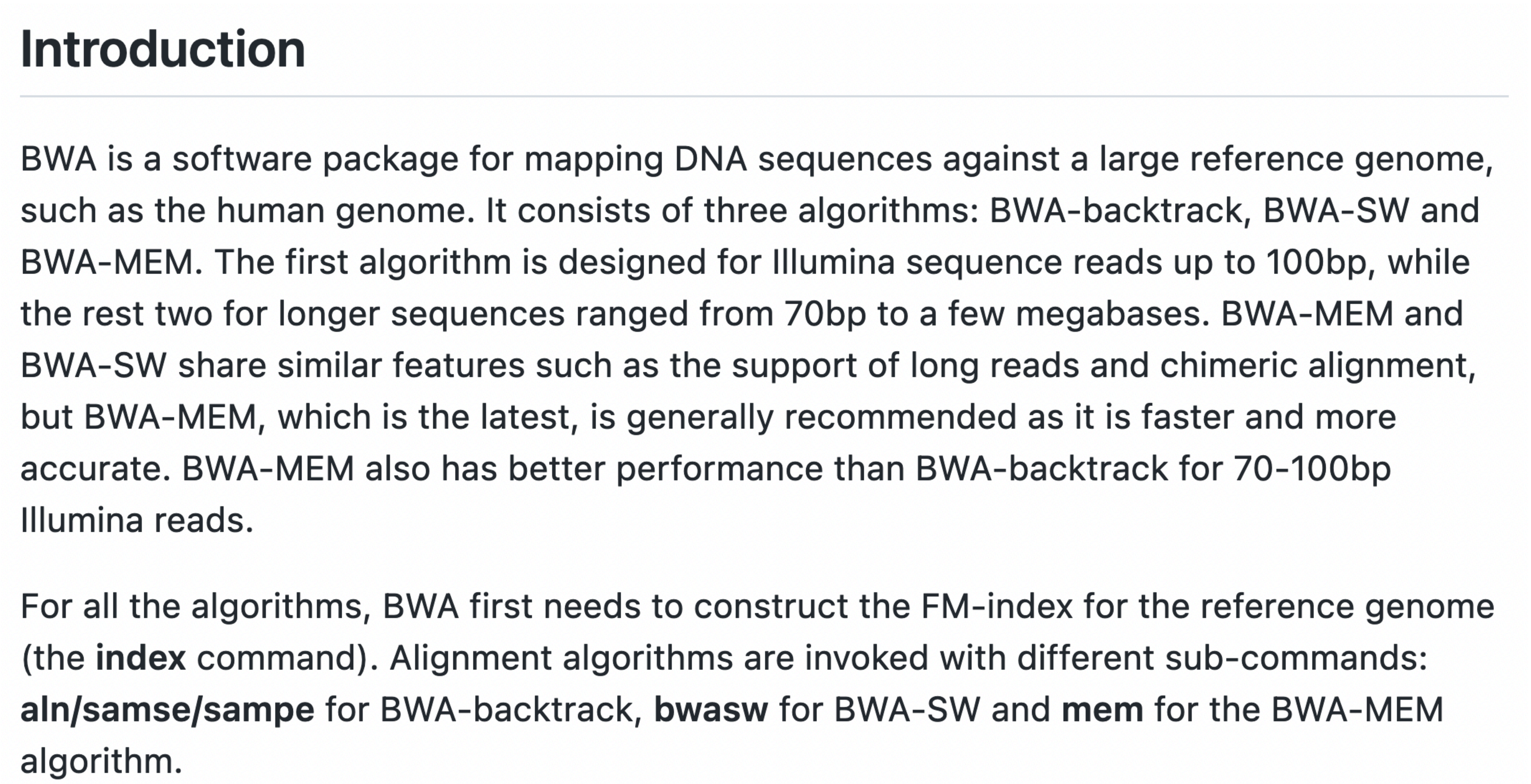
Introduction

BWA is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome. It consists of three algorithms: BWA-backtrack, BWA-SW and BWA-MEM. The first algorithm is designed for Illumina sequence reads up to 100bp, while the rest two for longer sequences ranged from 70bp to 1Mbp. BWA-MEM and BWA-SW share similar features such as long-read support and split alignment, but BWA-MEM, which is the latest, is generally recommended for high-quality queries as it is faster and more accurate. BWA-MEM also has better performance than BWA-backtrack for 70-100bp Illumina reads.

Genomics106

Mapping and Alignment 102
Mapping with BWA

BWA GITHUB Repository



Introduction

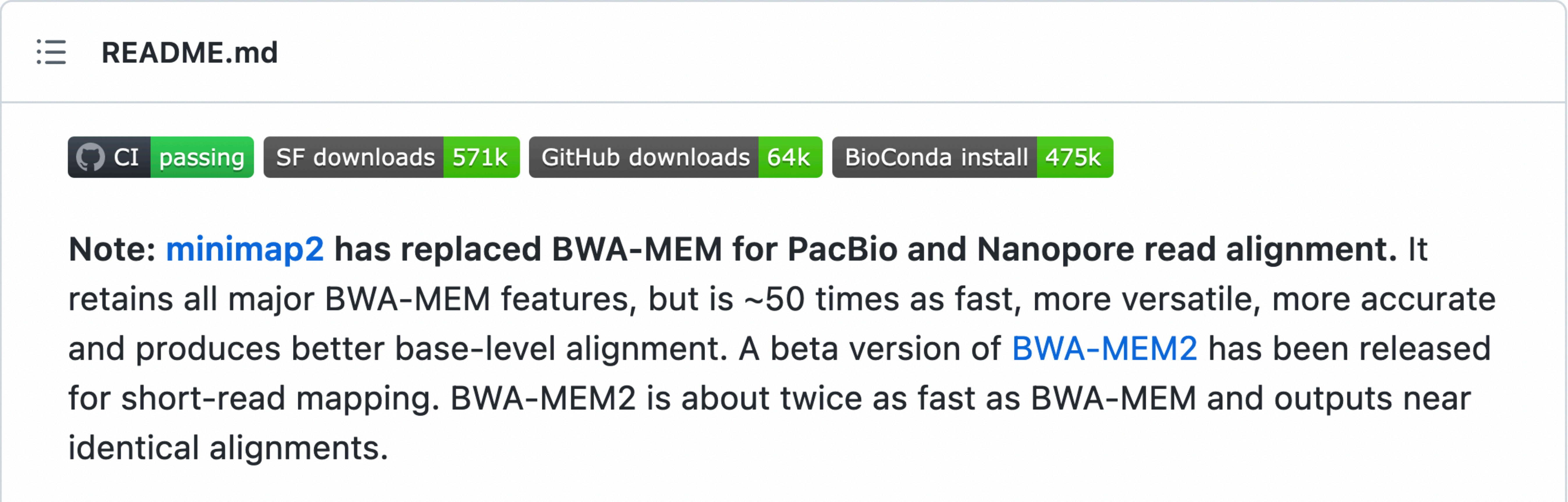
BWA is a software package for mapping DNA sequences against a large reference genome, such as the human genome. It consists of three algorithms: BWA-backtrack, BWA-SW and BWA-MEM. The first algorithm is designed for Illumina sequence reads up to 100bp, while the rest two for longer sequences ranged from 70bp to a few megabases. BWA-MEM and BWA-SW share similar features such as the support of long reads and chimeric alignment, but BWA-MEM, which is the latest, is generally recommended as it is faster and more accurate. BWA-MEM also has better performance than BWA-backtrack for 70-100bp Illumina reads.

For all the algorithms, BWA first needs to construct the FM-index for the reference genome (the **index** command). Alignment algorithms are invoked with different sub-commands: **aln/samse/sampe** for BWA-backtrack, **bwasw** for BWA-SW and **mem** for the BWA-MEM algorithm.

Genomics106

Mapping and Alignment 102
Mapping with BWA

BWA GITHUB Repository

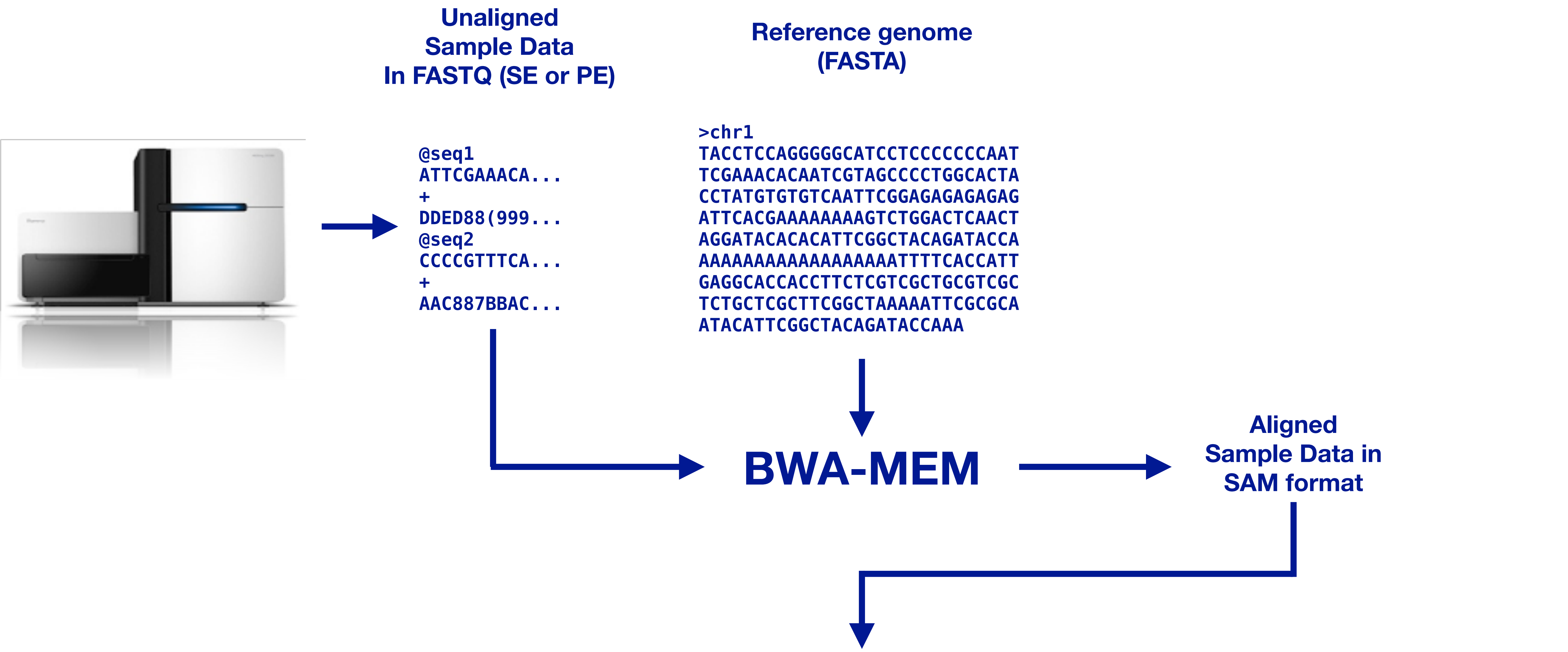


README.md

Note: minimap2 has replaced BWA-MEM for PacBio and Nanopore read alignment. It retains all major BWA-MEM features, but is ~50 times as fast, more versatile, more accurate and produces better base-level alignment. A beta version of BWA-MEM2 has been released for short-read mapping. BWA-MEM2 is about twice as fast as BWA-MEM and outputs near identical alignments.

Genomics106

Mapping and Alignment 102
BWA-MEM



Unaligned
Sample Data
In FASTQ (SE or PE)

```
@seq1
ATTCGAAACA...
+
DDED88(999...
@seq2
CCCCGTTTCA...
+
AAC887BBAC...
```

Reference genome
(FASTA)

```
>chr1
TACCTCCAGGGGGCATCCTCCCCCCCAAT
TCGAAACACAATCGTAGCCCCTGGCACTA
CCTATGTGTGTCAATTCGGAGAGAGAGAG
ATTCACGAAAAAAAAGTCTGGACTCAACT
AGGATACACACATTCGGCTACAGATACCA
AAAAAAAAAAAAAAAAAAATTTTCACCATT
GAGGCACCACCTTCTCGTCGCTGCGTCGC
TCTGCTCGCTTCGGCTAAAAATTCGCGCA
ATACATTCGGCTACAGATACCAAA
```

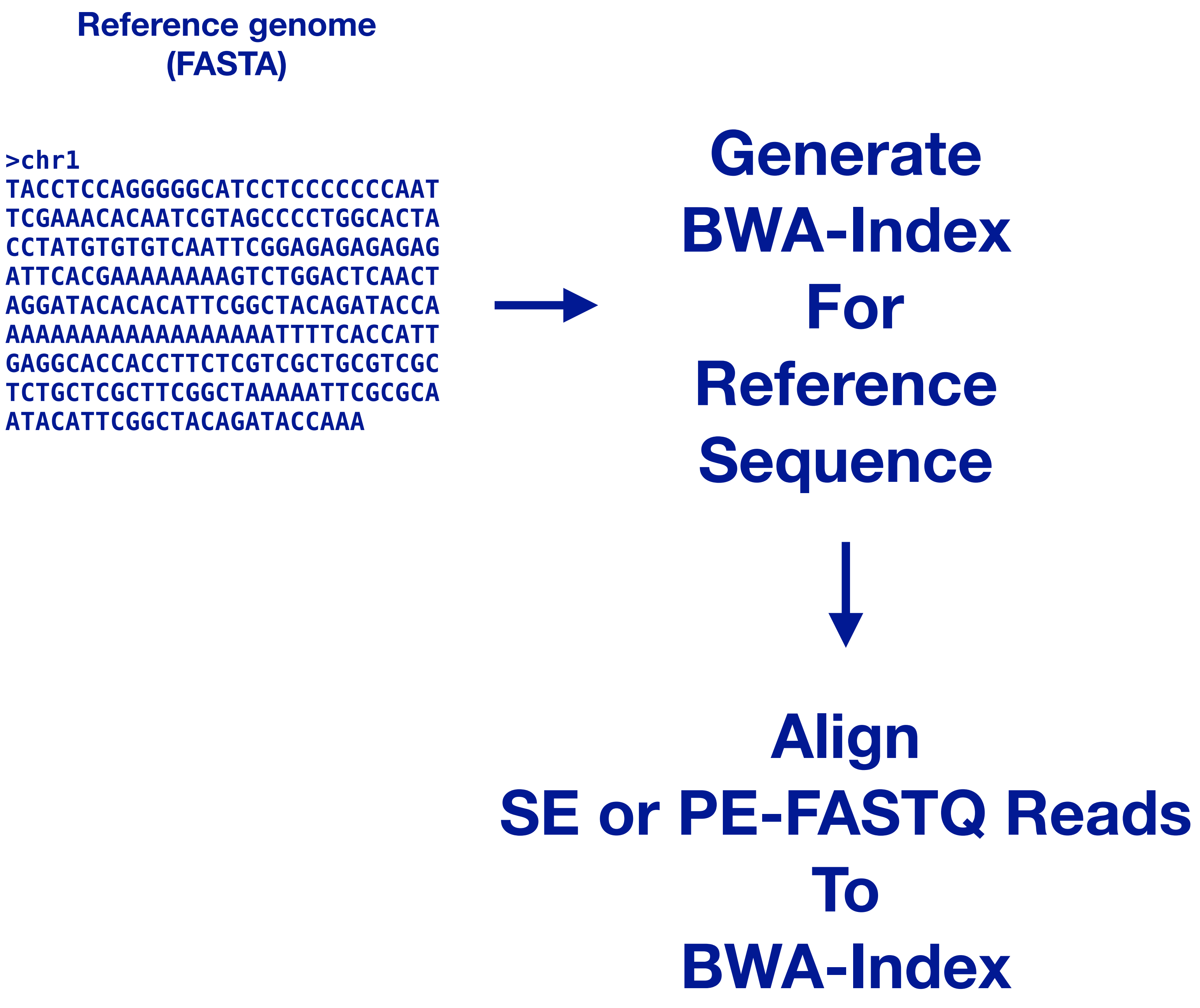
BWA-MEM

Aligned
Sample Data in
SAM format

```
seq1   99  1  3666901  60  149M  =  3666935   185  ATTCGAAACA...  DDED88(999... MC:Z:151M  MD:Z:149  RG:Z:15-0017315_1  NM:i:0  MQ:i:60  AS:i:149  XS:i:44
seq2  147  1  3666935  60  151M  =  3666901  -185  CCCCGTTTCA...  AAC887BBAC... MC:Z:149M  MD:Z:151  RG:Z:15-0017315_1  NM:i:0  MQ:i:60  AS:i:151  XS:i:59
```

Genomics106

Mapping and Alignment 102
BWA-MEM workflow



Reference genome
(FASTA)

```
>chr1
TACCTCCAGGGGGCATCCTCCCCCCCAAT
TCGAAACACAATCGTAGCCCCTGGCACTA
CCTATGTGTGTCAATTCGGAGAGAGAGAG
ATTCACGAAAAAAAAGTCTGGACTCAACT
AGGATACACACATTCGGCTACAGATACCA
AAAAAAAAAAAAAAAAAATTTTCACCATT
GAGGCACCACCTTCTCGTCGCTGCGTCGC
TCTGCTCGCTTCGGCTAAAAATTCGCGCA
ATACATTCGGCTACAGATACCAAA
```

Generate BWA-Index For Reference Sequence

Align SE or PE-FASTQ Reads To BWA-Index

Commands

```
$ bwa \
  index grch38.fa
```

```
$ bwa mem \
  -t 16 \
  grch38.fa \
  1.fq 2.fq \
  > sample.sam
```

Genomics106

Mapping and Alignment 102
Sequence Alignment with BWA

- BWA can map low-divergent sequences against a large reference genome, such as the human genome
  - It consists of three algorithms:
    - BWA-backtrack (Illumina sequence reads up to 100bp)
    - BWA-SW
    - BWA-MEM

- BWA SW and MEM can map longer sequences (70bp to 1Mbp) and share similar features such as long-read support and split alignment, but BWA-MEM, which is the latest, is generally recommended for high-quality queries as it is faster and more accurate

- BWA-MEM also has better performance than BWA-backtrack for 70-100bp Illumina reads

Genomics106

Mapping and Alignment 102
Sequence Alignment with BWA

- **Install BWA**

```
$ cd /vol_b/DB2022_xx/Database

$ mkdir -p Genomics105_bwa_Mapping/data

$ cd Genomics105_bwa_Mapping/data

$ cp -v /vol_b/zzStorage/T16_Data/* ./

$ cd ../

# Soft link the following files to this directory:
# 00_BRCA2_WildType.fa
# 00_BRCA2_WildType_P_050m200_1.fq and 00_BRCA2_WildType_P_050m200_2.fq

# Confirm bwa is installed
$ bwa

# Install bwa in the bioinfosoft environment
$ conda activate bioinfosoft
$ conda install bwa

# Call bwa
$ bwa
```

Genomics106

Mapping and Alignment 102
Sequence Alignment with BWA

- **Getting started - Basic Example Commands**

```
# Generate Index
$ bwa index ref.fa


# Align SE-Reads
$ bwa mem ref.fa read-se.fq.gz | gzip -3 > aln-se.sam.gz


# Align PE-Reads
$ bwa mem ref.fa read1.fq read2.fq | gzip -3 > aln-pe.sam.gz
```

Genomics106

Mapping and Alignment 102
Sequence Alignment with BWA

- **Create Reference Index:**

```
Usage:   bwa index [options] <in.fasta>

Options: -a STR    BWT construction algorithm: bwtsw, is or rb2 [auto]
         -p STR    prefix of the index [same as fasta name]
         -b INT    block size for the bwtsw algorithm (effective with -a bwtsw) [10000000]
         -6        index files named as <in.fasta>.64.* instead of <in.fasta>.*

Warning: `-a bwtsw' does not work for short genomes, while `-a is' and
         `-a div' do not work not for long genomes
```

```
$ bwa \
  index \
  -a rb2 \
  -p 00_BRCA2_WildType.fa 00_BRCA2_WildType.fa;
```

Genomics106

Mapping and Alignment 102
Sequence Alignment with BWA

- **Align to Reference Genome:**

```
# Using the Index:

                    00_BRCA2_WildType.fa

# Using the following files (PE-Reads):

                    00_BRCA2_WildType_P_050m200_1.fq
                    00_BRCA2_WildType_P_050m200_2.fq
```

```
$ bwa mem \
   -t 2 \
   00_BRCA2_WildType.fa \
   00_BRCA2_WildType_P_050m200_1.fq \
   00_BRCA2_WildType_P_050m200_2.fq  \
     > 00_BRCA2_WildType_P_050m200_ReadsxBRCA2_WildType.sam
```

Genomics106

Mapping and Alignment 102
SAMTOOLS 101

- Install SAMTOOLS

```
$ samtools

$ conda activate bioinfosoft

$ conda install samtools

$ samtools
```

Genomics106

Mapping and Alignment 102
SAMTOOLS 101

- **Viewing alignments**

The samtools view command is the most versatile tool in the samtools package. It’s main function, not surprisingly, is to allow you to convert the binary (i.e., easy for the computer to read and process) alignments in the BAM file view to text-based SAM alignments that are easy for humans to read and process.

```
# All alignments
$ samtools view 00_BRCA2_WildType_P_050m200_ReadsxBRCA2_WildType.sam | more

# Only the first five
$ samtools view 00_BRCA2_WildType_P_050m200_ReadsxBRCA2_WildType.sam | head -n 5

# Counting number of alignments
$ samtools view 00_BRCA2_WildType_P_050m200_ReadsxBRCA2_WildType.sam | wc -l

# View only header
$ samtools view -H 00_BRCA2_WildType_P_050m200_ReadsxBRCA2_WildType.sam

# View header + alignments
$ samtools view -h 00_BRCA2_WildType_P_050m200_ReadsxBRCA2_WildType.sam | more

# Alignment Statistics
$ samtools flagstat 00_BRCA2_WildType_P_050m200_ReadsxBRCA2_WildType.sam
```

Modified from a Tutorial Developed by Aaron Quinlan (November 20, 2013)

Genomics106

Mapping and Alignment 102
SAMTOOLS 101

- **Capturing alignments**

- **The FLAG field in the SAM format encodes several key pieces of information regarding how an alignment aligned to the reference genome**

- **This information can be exploited to isolate specific types of alignments that we want to use in our analysis.**

- **For example, we often want to call variants solely from paired-end sequences that aligned “properly” to the reference genome**

- **To ask the view command to report solely “proper pairs” we use the -f option and ask for alignments where the second bit is true (proper pair is true)**

- **Remember: Decoding SAM Flags**

```
$ samtools view -f 0x2 00_BRCA2_WildType_P_050m200_ReadsxBRCA2_WildType.sam
```

Genomics106

Mapping and Alignment 102
SAMTOOLS 101

- **Capturing alignments**

- **How many properly paired alignments are there?**

```
$ samtools view -f 0x2 00_BRCA2_WildType_P_050m200_ReadsxBRCA2_WildType.sam | wc -l
```

- **Now, let's ask for alignments that are NOT properly paired. To do this, we use the -F option (note the capitalization to denote "opposite")**

```
$ samtools view -F 0x2 00_BRCA2_WildType_P_050m200_ReadsxBRCA2_WildType.sam
```

- **How many improperly paired alignments are there?**

```
$ samtools view -F 0x2 00_BRCA2_WildType_P_050m200_ReadsxBRCA2_WildType.sam | wc -l
```

Genomics106

Mapping and Alignment 102
SAMTOOLS 101

- **Convert SAM to BAM**

```
# Generic Command:
$ samtools \
   view \
   -S \
   -b \
   sample.sam \
      > sample.bam;


$ samtools \
   view \
   -S \
   -b \
   00_BRCA2_WildType_P_050m200_ReadsxBRCA2_WildType.sam \
      > 00_BRCA2_WildType_P_050m200_ReadsxBRCA2_WildType.bam;
```

Genomics106

Mapping and Alignment 102
SAMTOOLS 101

- **Sort BAM File**

```
# Generic Command:
$ samtools \
   sort \
   sample.bam \
   -o sample.sorted.bam;


$ samtools \
   sort \
   00_BRCA2_WildType_P_050m200_ReadsxBRCA2_WildType.bam \
   -o 00_BRCA2_WildType_P_050m200_ReadsxBRCA2_WildType.sorted.bam;
```

Genomics106

Mapping and Alignment 102
SAMTOOLS 101

- **Create a BAM index file**

```
# Generic Command:
$ samtools \
   index \
   sample.sorted.bam;


$ samtools \
   index \
   00_BRCA2_WildType_P_050m200_ReadsxBRCA2_WildType.sorted.bam;
```

Genomics106

Mapping and Alignment 102
SAMTOOLS 101

- **Visualize BAM Alignment without a Reference Genome**

```
$ samtools \
   tview \
   00_BRCA2_WildType_P_050m200_ReadsxBRCA2_WildType.sorted.bam;
```

- **Visualize BAM Alignment with a Reference Genome**

```
$ samtools \
   tview \
   00_BRCA2_WildType_P_050m200_ReadsxBRCA2_WildType.sorted.bam \
   00_BRCA2_WildType.fa;
```

Genomics106

Mapping and Alignment 102
SAMTOOLS 101

- **View at a Specific Coordinate**

```
# Generic Command:
$ samtools \
   tview \
   00_BRCA2_WildType_P_050m200_ReadsxBRCA2_WildType.sorted.bam \
   00_BRCA2_WildType.fa \
   -p chromosome:coordinate;


$ samtools \
   tview \
   00_BRCA2_WildType_P_050m200_ReadsxBRCA2_WildType.sorted.bam \
   00_BRCA2_WildType.fa \
   -p BRCA2_WildType:10000;


$ samtools \
   tview \
   00_BRCA2_WildType_P_050m200_ReadsxBRCA2_WildType.sorted.bam \
   00_BRCA2_WildType.fa \
   -p BRCA2_WildType:59900;
```

Genomics106

Mapping and Alignment 102

- Factors That Influence Mappability
  - Extent of polymorphism
  - Quality of sequence data
  - Lengths of DNA molecules being sequenced compared with size of reference genome
  - Degree of sequence repetition in regions of the genome to which the reads map

Genomics106

Mapping and Alignment 102

Scripting this Mapping...

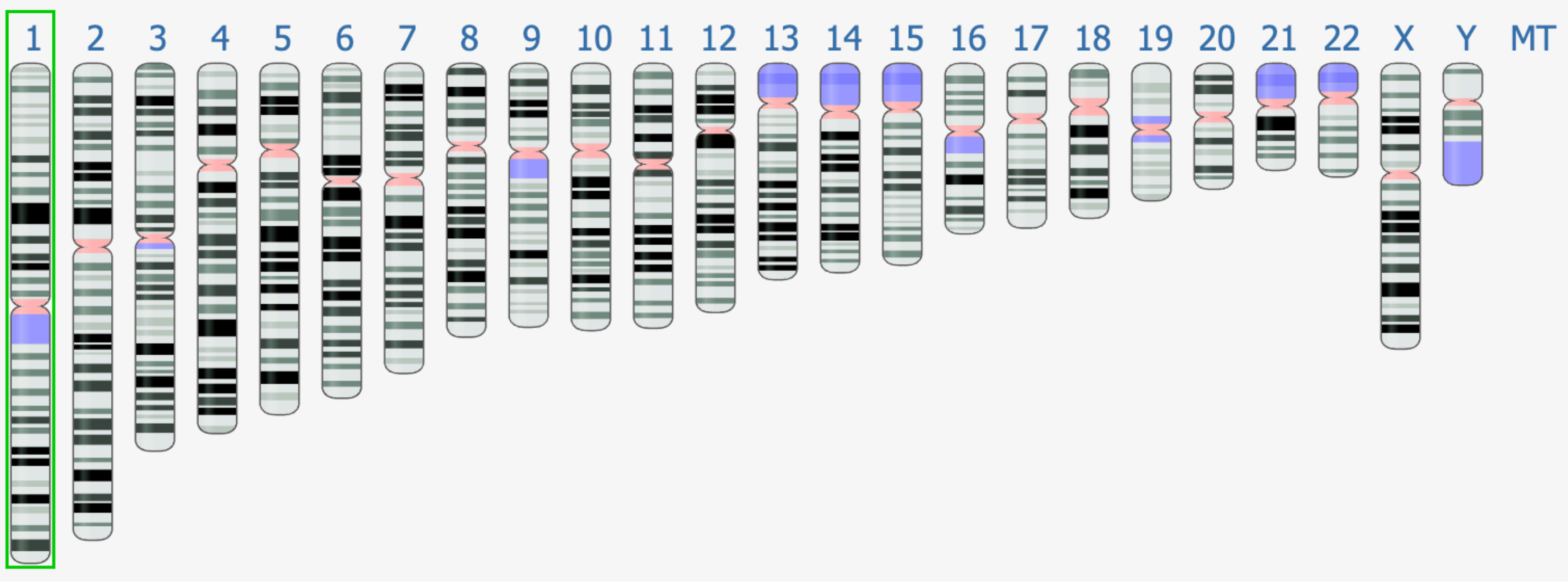
- **In your repositories, on a document entitled:**

12Lecture_BWA_Mapping

- **Outline the logic of a script(s) you need to develop to map reads to genomes using BWA**

- **What logical questions you need to answer to be able to execute the script?**

- **What code you need to have in order to be able to answer those logical questions?**



Genomics106

BIOL647
Digital Biology

Rodolfo Aramayo