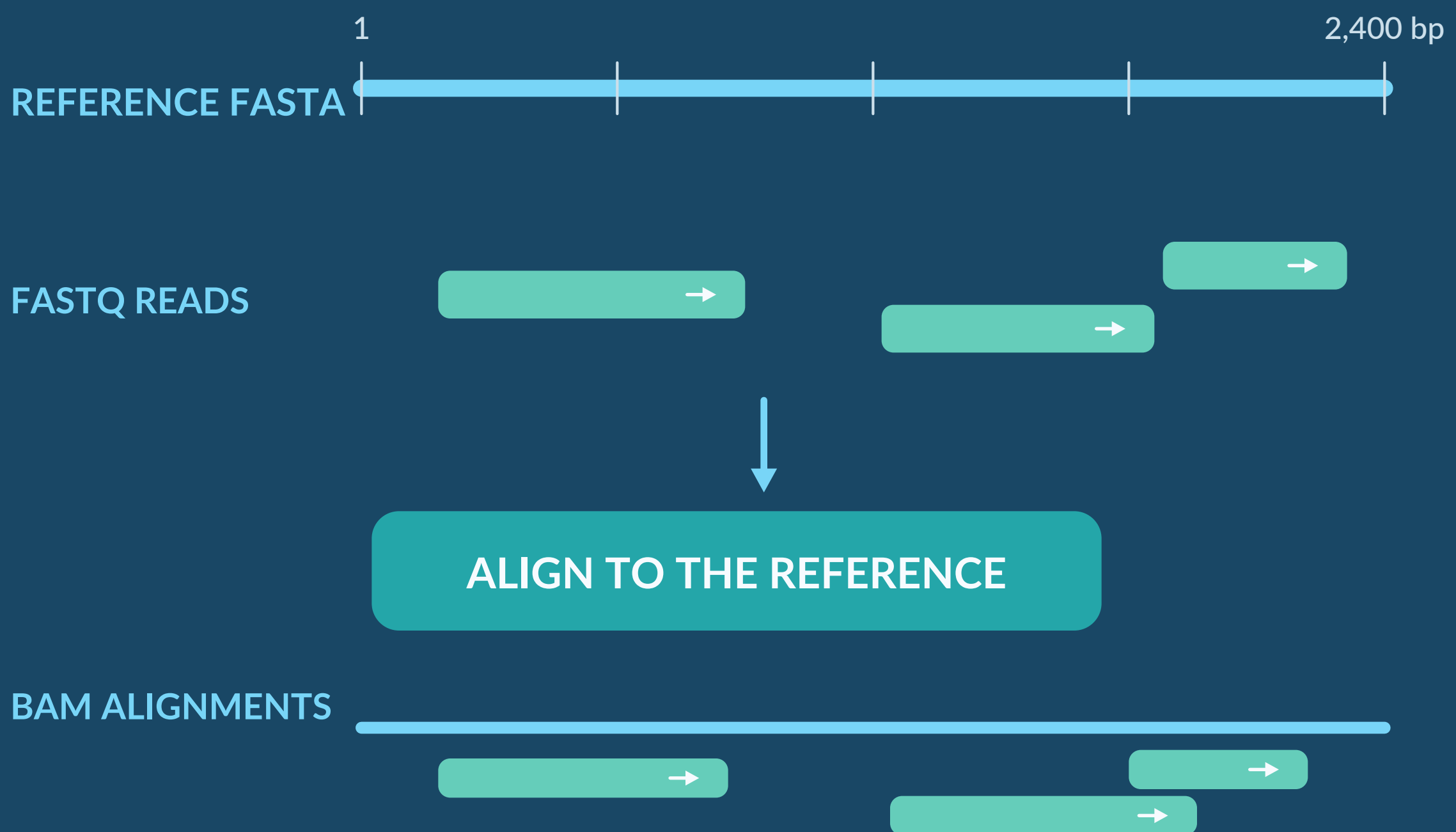


Where do our reads belong?

From individual FASTQ reads to a position on a genome

A chosen reference provides the coordinate system



KEY IDEA

A high-quality read may still be unmapped if it does not match the chosen reference.

Why map the reads?

Connect sequencing data to positions on a selected genome

Two inputs answer one question



reads.fastq.gz

Read sequences + base-call quality



reference.fasta

Chosen genome for comparison

minimap2

SAM / BAM: read positions, strand, and alignment information

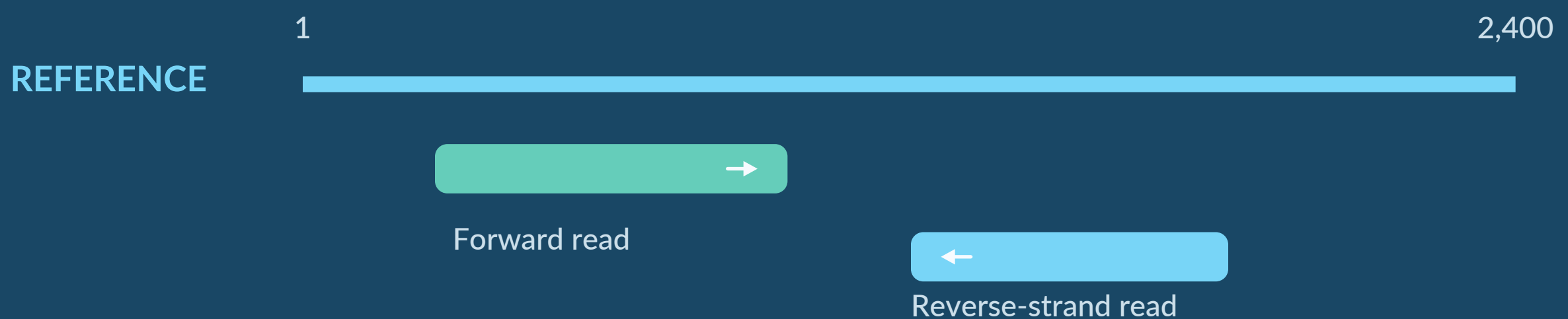
IMPORTANT

Matching this reference does not, by itself, prove the sample organism identity.

One reference. Many reads.

Position, strand, and mismatches all matter

Whole-reference view (schematic)



Reads may align on either strand. A mismatch does not always prevent alignment.

ZOOM IN: a short illustrative sequence window



ONE MISMATCH

Orange base differs from reference. The aligner scores the overall match.

Diagram illustrates the concept; it is not an alignment printed by minimap2.

What files does mapping create?

Follow the information from input to indexed alignment

INPUT



reference.fasta

Reference DNA sequence



reads.fastq.gz

Reads + Phred quality

minimap2

SAM

mapped.sam

Text-based alignment records

samtools sort

BAM

mapped.sorted.bam

Compressed, coordinate-sorted alignments

samtools index

BAI

mapped.sorted.bam.bai

Fast position lookup index for the BAM

Install two tools in Ubuntu

minimap2 maps reads; SAMtools sorts and inspects alignments

minimap2

Align reads to a reference

SAMtools

Sort, index, and inspect results

1. Refresh Ubuntu package information

```
sudo apt update
```

2. Install both tools

```
sudo apt install minimap2 samtools -y
```

3. Confirm both tools are available

```
minimap2 --version  
samtools --version
```

ONE-TIME SETUP

Install once, then reuse both tools in later lessons.

Set up the practice folder

Keep your reference and reads together in one location

1. Create the Part 7 folder

```
mkdir -p ~/bioinformatics_training/part7
```

2. Enter the folder

```
cd ~/bioinformatics_training/part7
```

3. Open it from Ubuntu in Windows Explorer

```
explorer.exe .
```

Copy both practice files into this folder:

```
part7/  
├── reference.fasta  
└── reads.fastq.gz
```

```
ls -lh reference.fasta reads.fastq.gz
```

Create the alignment

Two commands turn reads into a sorted BAM file

1. Map single-end short reads (SAM output)

```
minimap2 -ax sr reference.fasta reads.fastq.gz > mapped.sam
```

2. Sort and compress alignments (BAM output)

```
samtools sort -o mapped.sorted.bam mapped.sam
```

What does coordinate sorting change?

Before sorting

Read aligned at reference position 720

Read aligned at reference position 180

Read aligned at reference position 510



After sorting

Read aligned at reference position 180

Read aligned at reference position 510

Read aligned at reference position 720

REMEMBER

Sorting changes record order by reference position. It does not change the DNA sequence.

Index and inspect the BAM

Check the mapping result, not just whether the command completed

1. Create the position index

```
samtools index mapped.sorted.bam
```

2. Summarize alignment statistics

```
samtools flagstat mapped.sorted.bam
```

3. Count mapped alignment records

```
samtools view -c -F 4 mapped.sorted.bam
```

index

Fast access to positions within BAM

flagstat / view

Counts of alignment records

INTERPRETATION

Mapping rate depends on the reference and settings. It is not a coverage measurement.

Mapping is not coverage

Many reads can align while part of the reference has no data

Toy reference: 2,400 bases



Read pileup (illustrative)



Designed central gap, with no reads across that interval.

Depth at each reference position (schematic)



Mapping rate

Fraction of reads with suitable alignments

Depth and breadth

Reads per position; fraction of reference covered

Run the complete workflow

A short practice checklist before Part 8

- 1 Which file contains the reference sequence?
- 2 Create mapped.sam using minimap2.
- 3 Convert it to a sorted BAM, then create the BAI index.
- 4 Use flagstat to inspect the mapping summary.
- 5 Explain why mapping rate cannot replace coverage analysis.

KEEP: reference.fasta | reads.fastq.gz | mapped.sorted.bam | mapped.sorted.bam.bai

COMING NEXT

Part 8: Measure read depth and breadth of genome coverage.