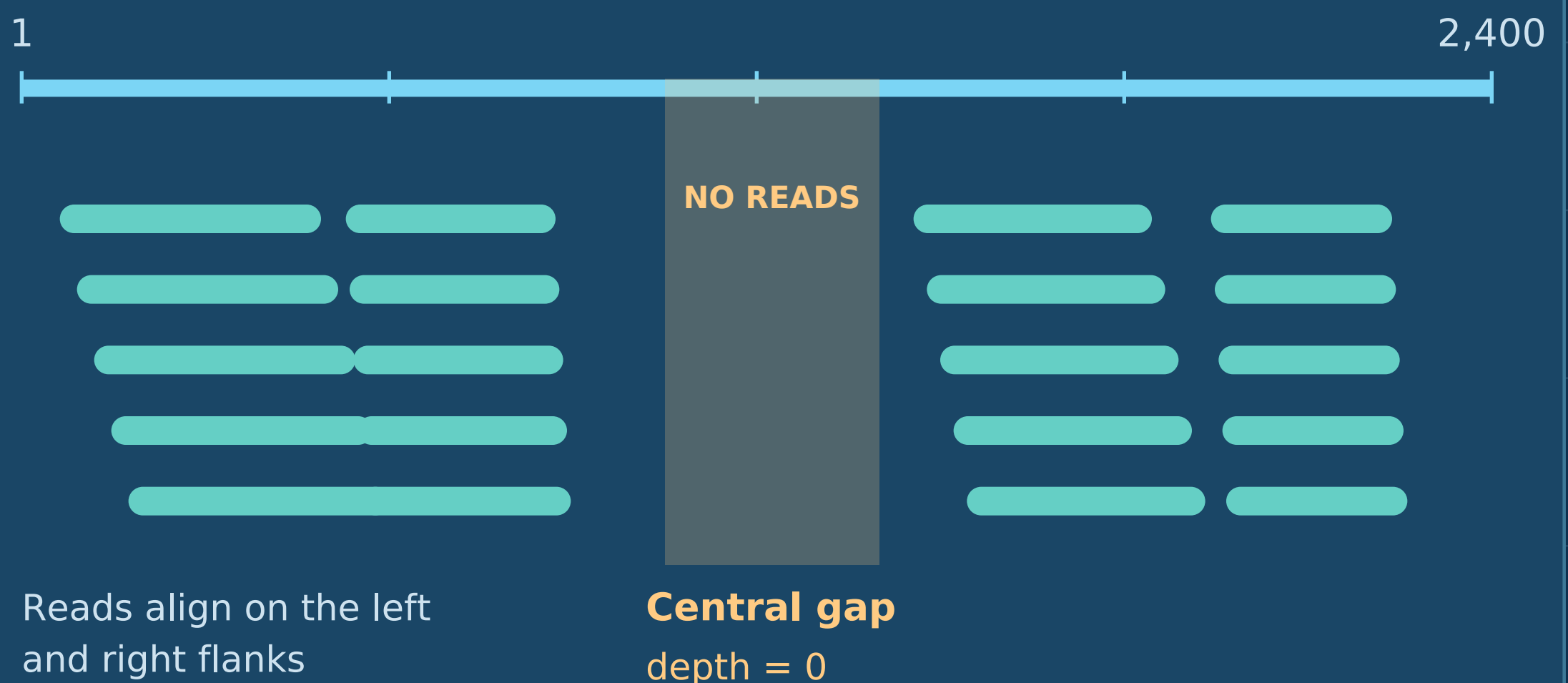


How much of the genome is actually covered?

From mapped reads to evidence across a reference

Synthetic reference: 2,400 bases



WHAT WE NEED TO MEASURE

Mapping

Did the read align?

Depth

How many reads per base?

Breadth

How much reference is covered?

KEY IDEA

Many reads may map while part of a reference has no usable read support.

Mapping, depth, breadth

These measure different parts of the same sequencing dataset.

01

MAPPING RATE

How many reads aligned to the selected reference?

READS are counted.

02

READ DEPTH

How many aligned bases overlap one reference position?

At one BASE.

03

BREADTH

What percentage of the reference meets a depth threshold?

REFERENCE positions are counted.

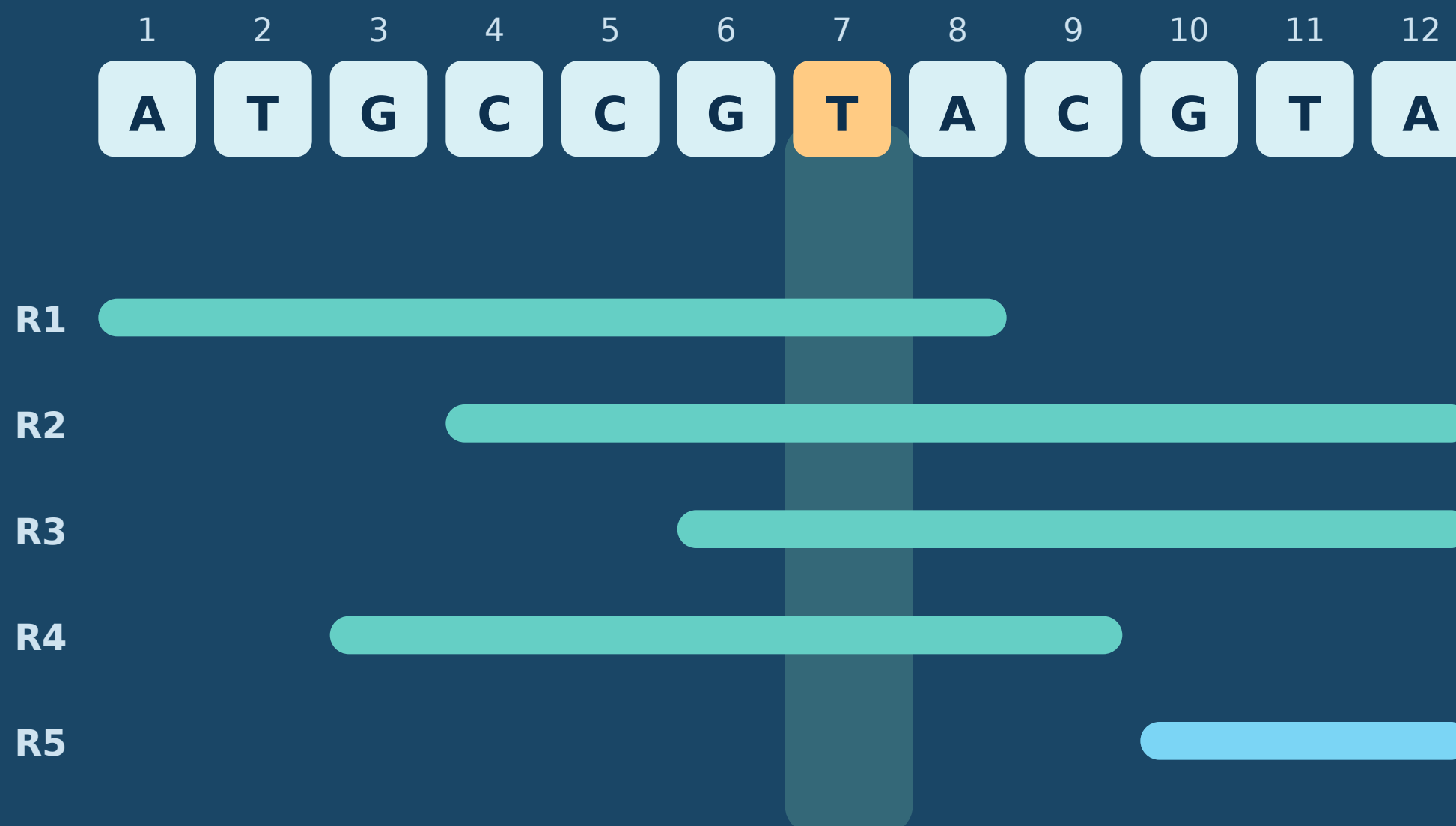
REMEMBER

More mapped reads does not automatically mean more complete genome coverage.

How many reads cover one base?

Depth is counted at each reference position.

Toy reference: 12 positions



Position 7

4 reads cross this position

DEPTH = 4x

A base with no aligned reads has depth 0x.

INTERPRET

A pileup is read evidence at a position, not a guarantee the base call is correct.

How much reference has support?

The answer depends on the minimum depth required.

Toy reference: 20 positions



Number inside each tile = read depth at that position

at least 1x



15/20 = 75% breadth

at least 5x



10/20 = 50% breadth

Breadth = (reference positions meeting the depth threshold ÷ total reference positions) × 100

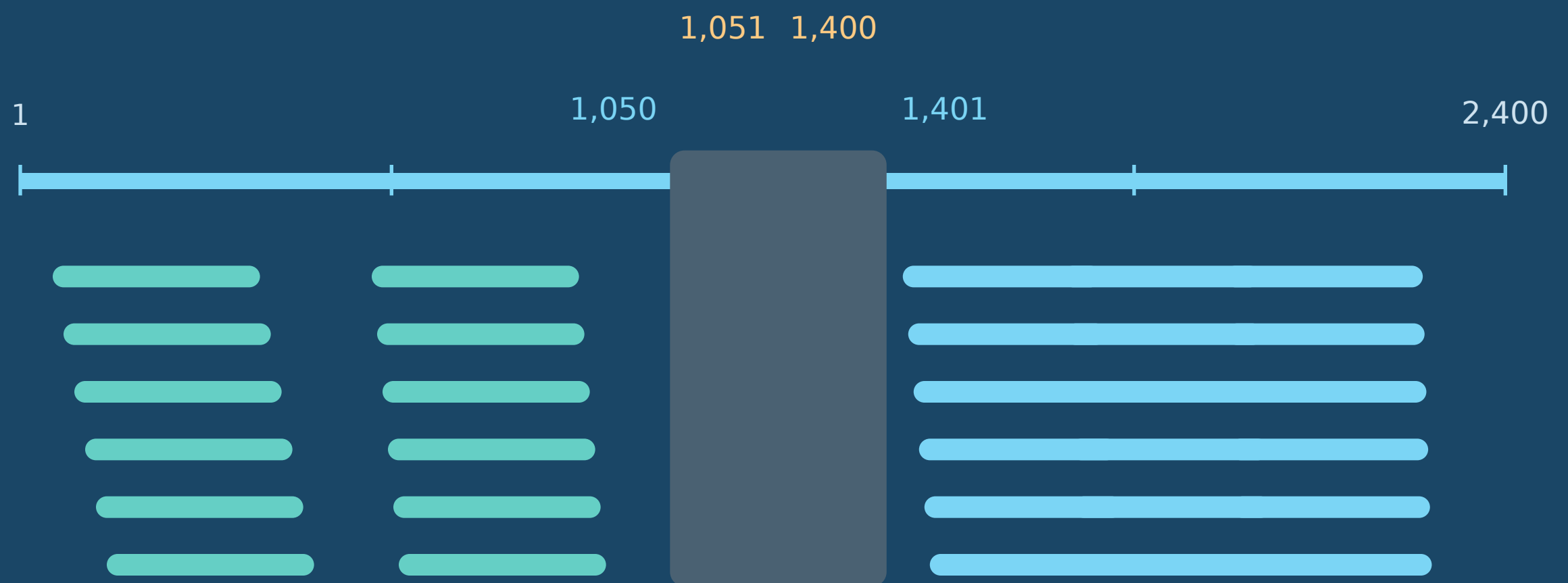
IMPORTANT

A genome may have 75% breadth at 1x but much less breadth at 5x or 10x.

The missing middle matters

Our synthetic reads do not cover positions 1,051 to 1,400.

Reference: 2,400 bases



350 positions with no designed reads

Maximum 1x breadth from this design:

$$2,050 / 2,400 = 85.4\%$$

Design expectation, not a SAMtools measurement

TAKEAWAY

The gap is a feature of these teaching reads. Do not interpret it as a real specimen problem.

Ask SAMtools for the overview

Continue from the mapped BAM you made in Part 7.

1 Return to your Part 7 folder

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

2 Summarize the reference

```
samtools coverage mapped.sorted.bam
```

Look for these columns in the output:

covbases

Number of reference bases with depth $\geq 1x$

coverage

Breadth at 1x, expressed as a percentage

meandepth

Average read depth in the reported region

NOTE

The output describes the chosen reference, not whether every base is equally well supported.

Look at every reference position

Save the depth table, including zero-depth positions.

1 Generate the depth table

```
samtools depth -aa mapped.sorted.bam > depth.tsv
```

2 Peek inside and count its rows

```
head depth.tsv  
wc -l depth.tsv
```

Each row holds three fields:

CHROM

reference name

POS

position

DEPTH

aligned read bases

For this one-reference dataset, expect 2,400 rows.

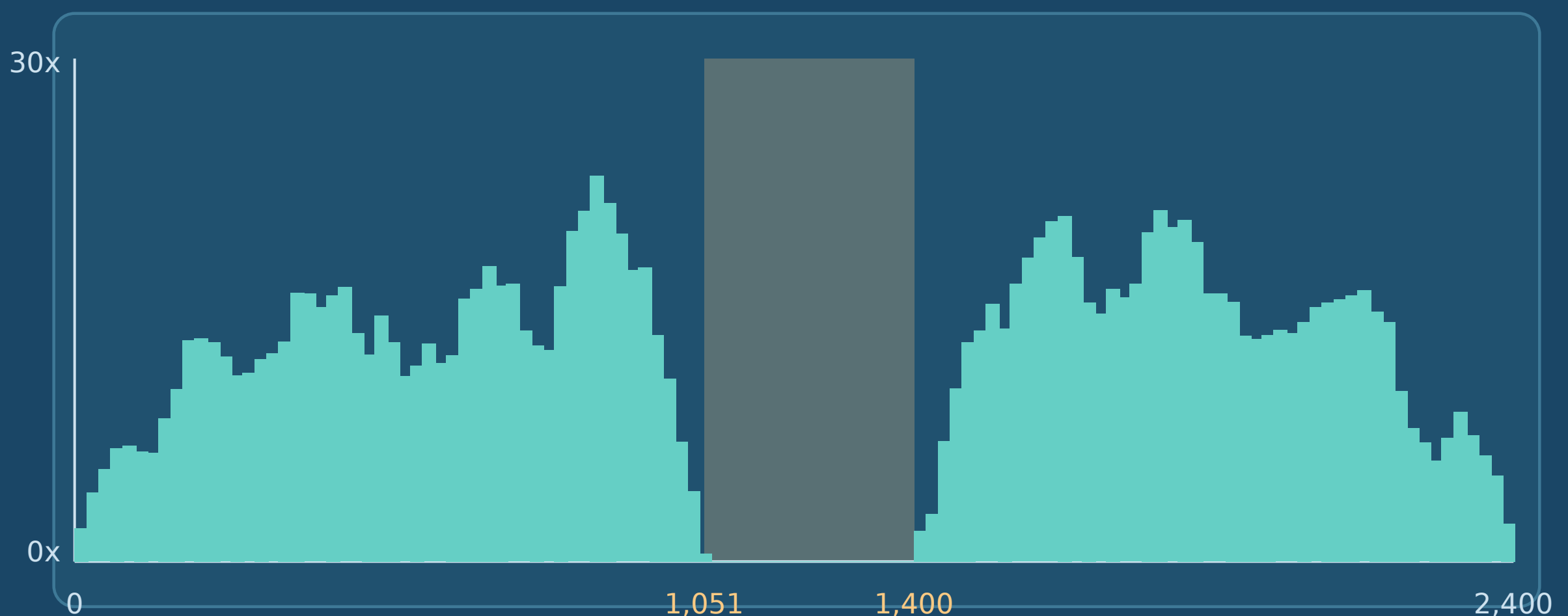
WHY -aa?

It includes positions with depth 0, so the missing region stays visible in your table.

Find the region with no reads

Use the same 2,400-base reference from Part 7.

Designed depth profile (not SAMtools output)



3 Inspect positions near the gap

```
awk '$2>=1048 && $2<=1054' depth.tsv
```

Expect depth 0 at positions 1,051-1,054.

INTERPRET

Coverage plots reveal where evidence is absent; they do not identify the cause on their own.

How much reaches 1x or 10x?

Calculate breadth from the depth table, including zero positions.

Breadth at $\geq 1x$

```
awk '{n++;if($3>0)c++}END{print 100*c/n}' depth.tsv
```

Breadth at $\geq 10x$

```
awk '{n++;if($3>=10)c++}END{print 100*c/n}' depth.tsv
```

Both commands print a percentage, not the number of reads.

$\geq 1x$ at least one aligned base

$\geq 10x$ at least ten aligned bases

CHECK

Higher depth thresholds produce equal or smaller breadth percentages.

A gap is a clue, not a diagnosis

Investigate before concluding why reference positions lack coverage.

01

Sample / library

Input quality, uneven preparation, or amplification dropout?

02

Reference / mapping

Reference divergence or mapping/filter settings?

03

Analysis decision

Do low-depth positions need masking before consensus?

CAUTION

A zero-depth position is not proof the organism or nucleotide is biologically absent.

Interpret your coverage report

Five questions to answer before moving to consensus calling.

1

What is the mean depth reported by samtools coverage?

2

How many reference bases have depth 0?

3

What percentage of the reference has depth $\geq 1x$?

4

What percentage has depth $\geq 10x$?

5

Why can mapped reads still leave a 350-base gap?

COMING NEXT

**PART 9: From mapped reads to a consensus genome.
Which positions can we trust?**