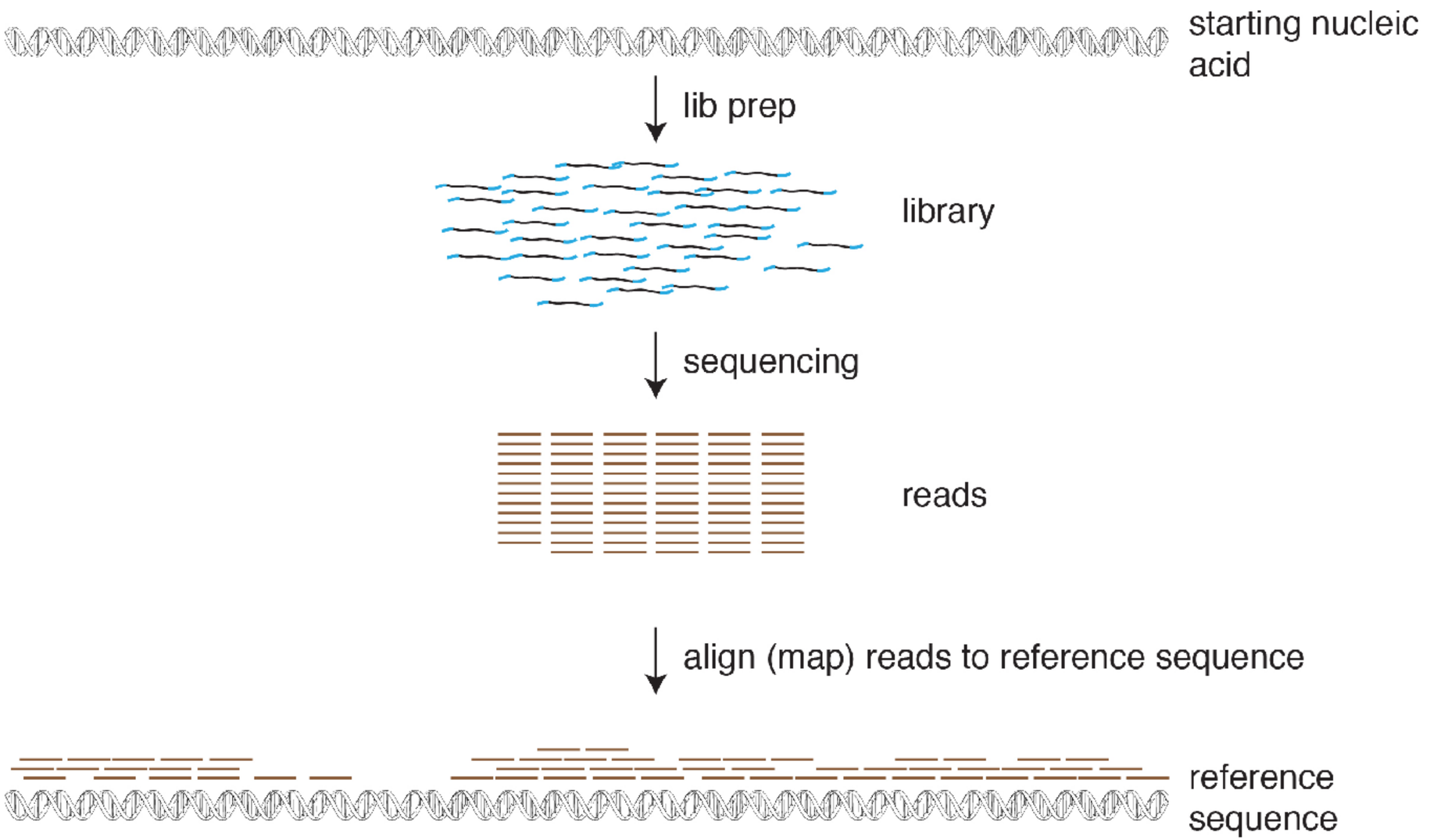
Mapping

Mark Stenglein

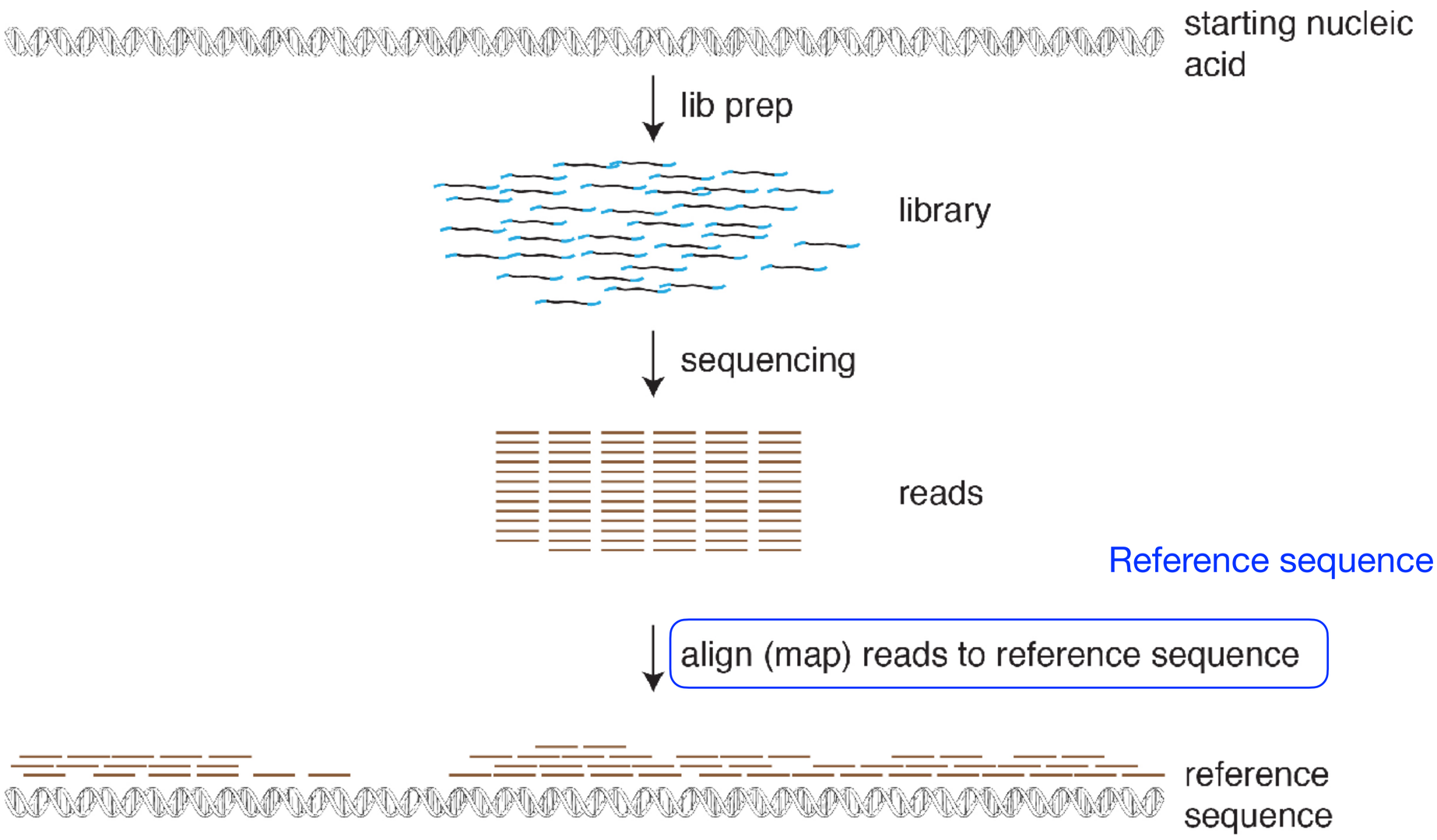
Computational Biology Workshop

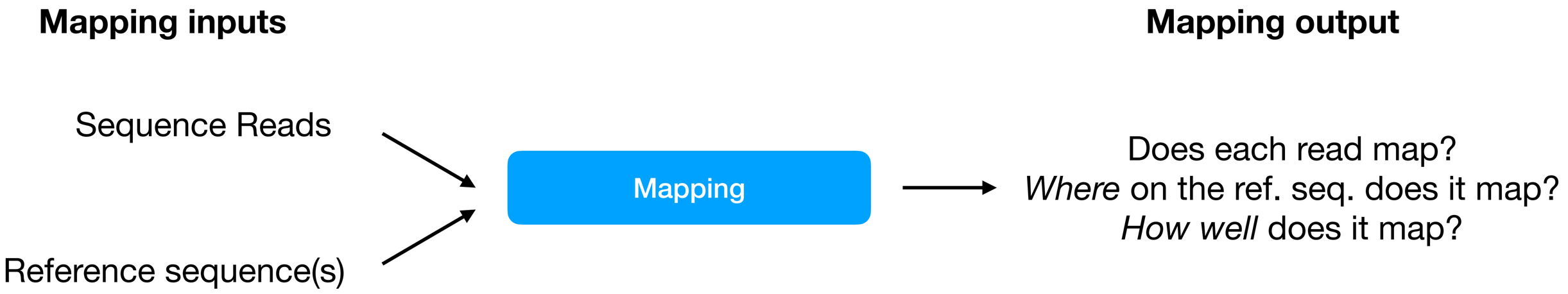
Todos Santos Center
May 9-12, 2022

Mapping is the process by which sequencing reads are aligned to the region of a genome from which they derive.

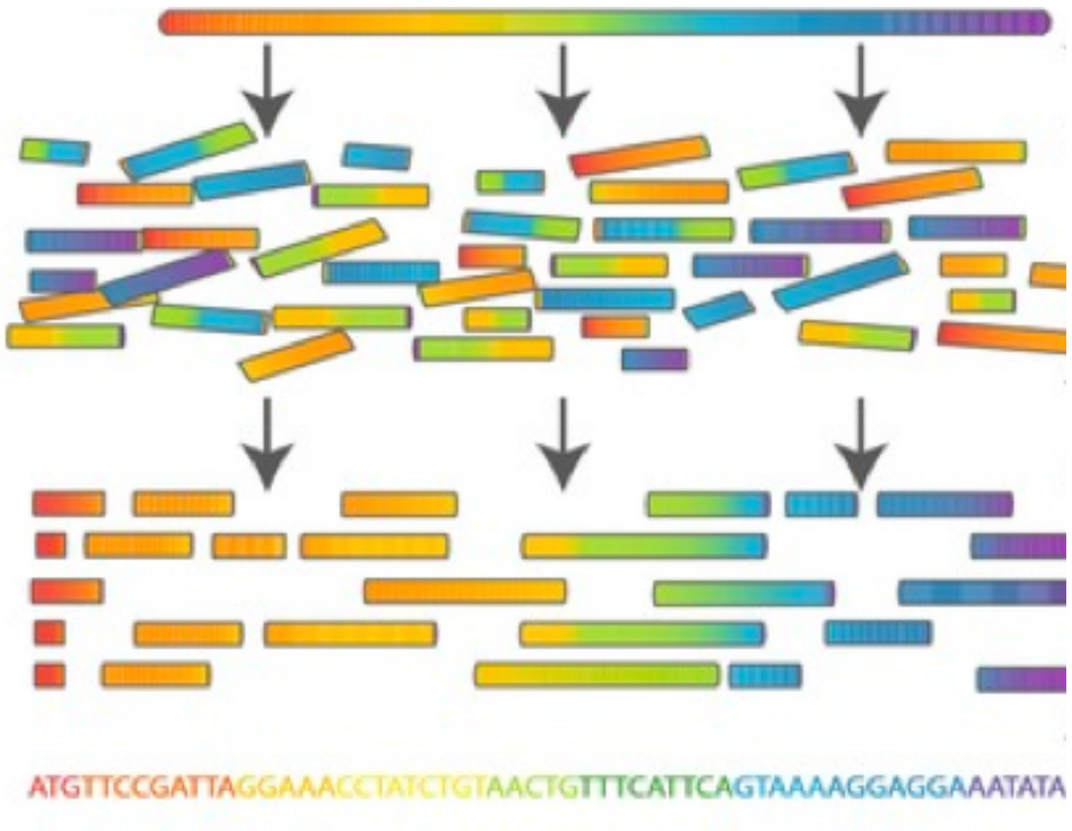


You need to have an existing reference sequence to map to




Mapping inputs
Sequence Reads
Reference sequence(s)
Mapping
Mapping output
Does each read map?
Where on the ref. seq. does it map?
How well does it map?

Genome assembly is the process of trying to reconstruct a genome sequence from reads (making a new reference sequence)



not yet sequenced genome

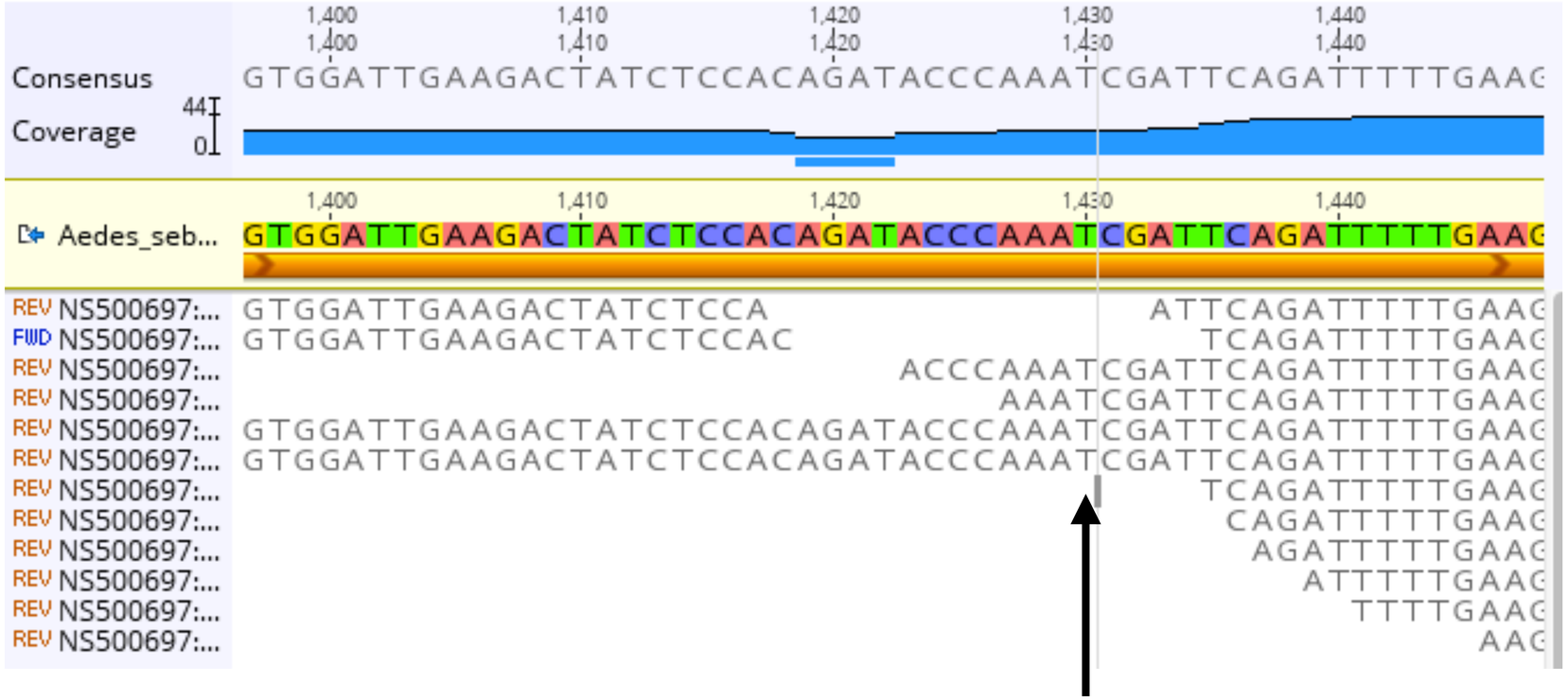
reads

contigs

reconstructed (assembled) genome

Image: Commins, J., Toft, C., Fares, M. A (2009)

Coverage is the number of individual mapped reads that support a particular nucleotide in a reference sequence

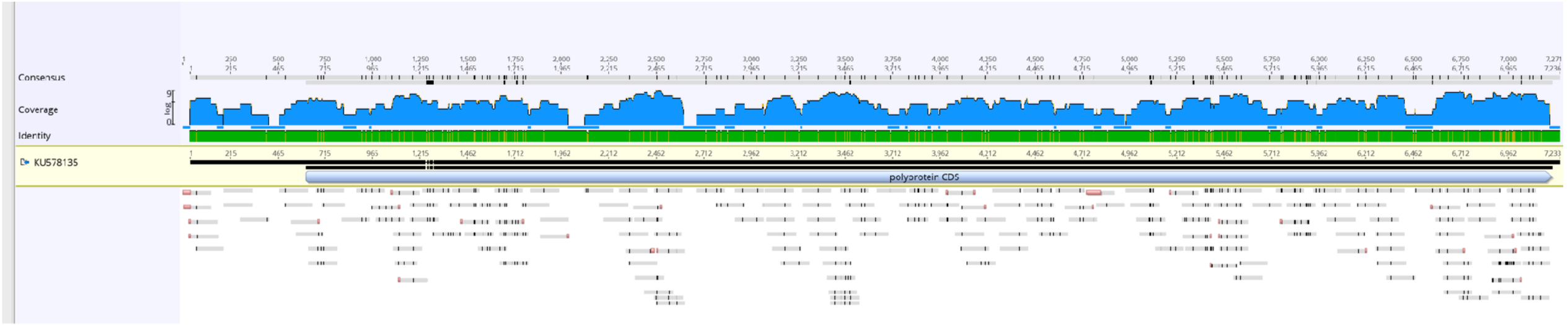


coverage is often referred to as ‘depth’ or ‘depth of coverage’

This T at position 1430 in the reference sequence has 4x coverage

Coverage is also used to describe the fraction of a genome with >0 coverage depth

reads from human oral swab RNA aligned to a coxsackie virus genome



96% genome coverage (96% of bases have >0x coverage)
3.4x average coverage depth (range 0-9x)

hand foot and mouth disease

(Mayo clinic)

Applications of mapping

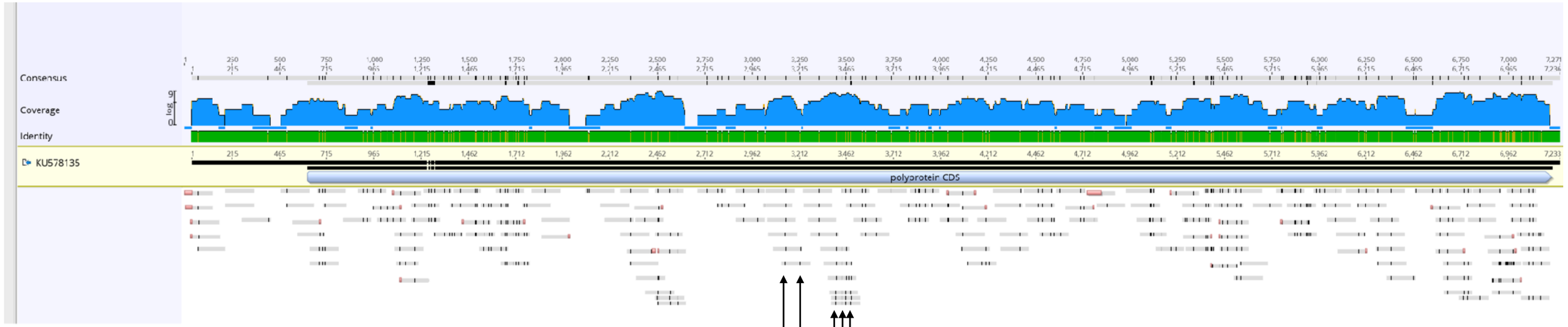
- **Quantification**: using reads for counting: sequence itself not important per se
  - RNA-seq: reads mapping to a particular transcript proportional to its abundance in the sample
  - ChIP-seq: coverage levels proportional to binding of proteins to DNA sequences

- **Variant identification**
  - Single nucleotide variants (SNVs aka SNPs)
  - Structural variants
  - Consensus-changing or sub-consensus

- **Remove sequences** of specific origins
  - Contaminating organisms
  - Plasmid
  - Organellar

There are variants in the reads relative to the co reference sequence: these differences are the basis for ‘variant calling’

reads from human oral swab RNA aligned to a coxsackie virus genome

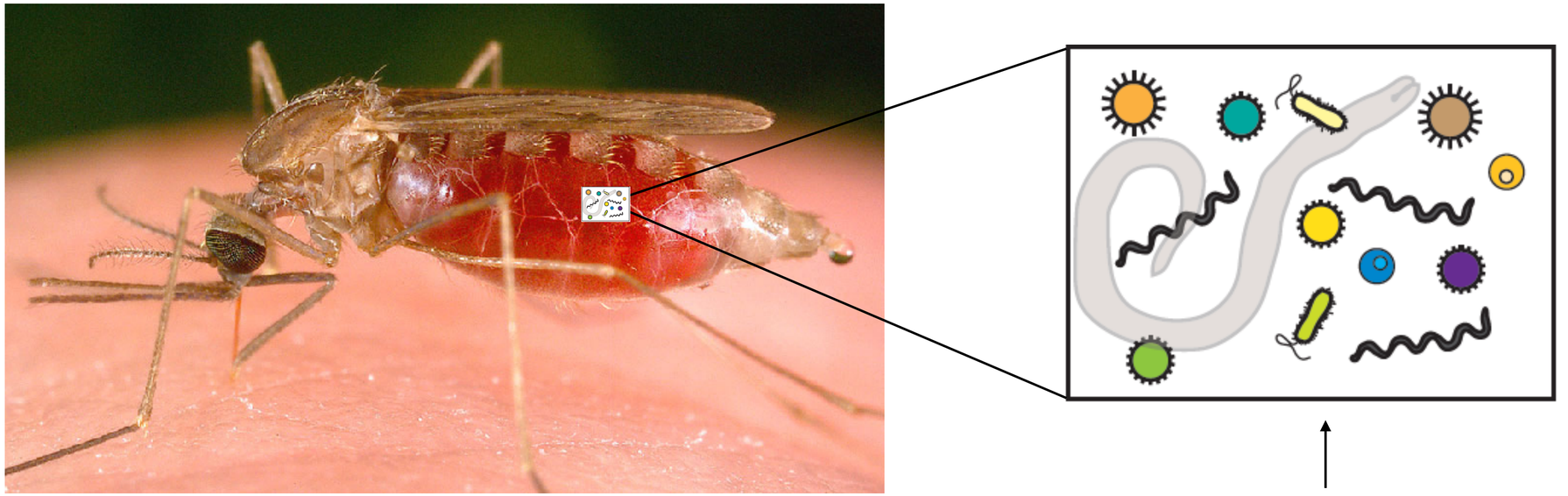


mismatches between reads and the reference sequence

hand foot and mouth disease

(Mayo clinic)

Mapping can be used to remove reads that derive from an organism you don’t care about



mosquito-associated microbiota

You can sequence everything and use a mosquito reference genome to remove all the mosquito reads leaving reads from all the other non-mosquito organisms

Mapping Exercise

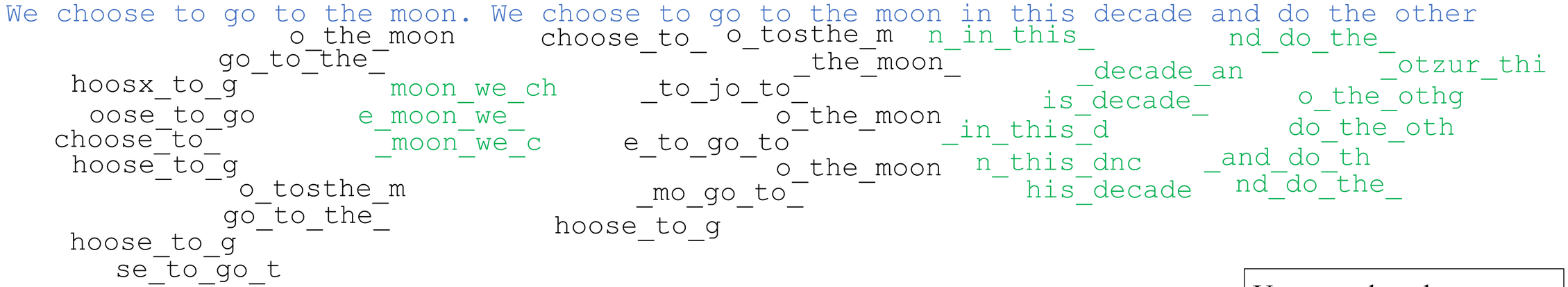
Work in pairs to map reads to the provided ‘genome’

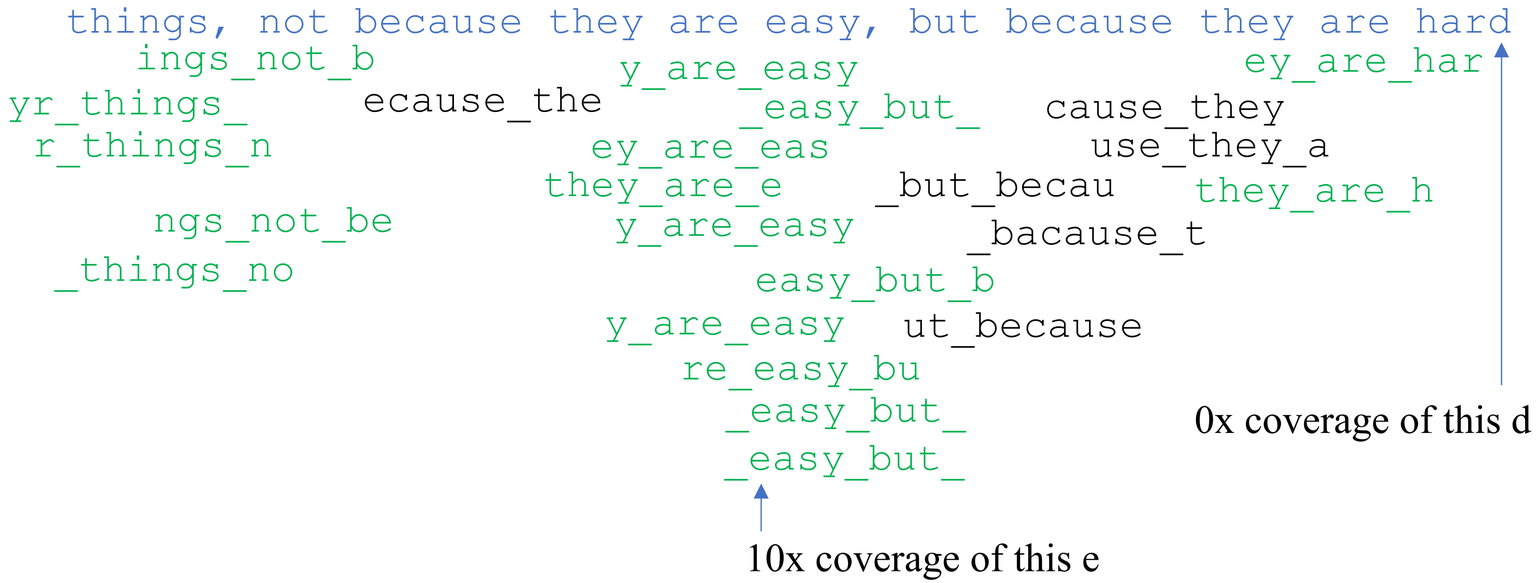
Questions to consider while doing this mapping exercise

Coverage
What was the (approximate) average coverage depth?
What was the maximum coverage depth?
What was the minimum coverage depth?
Was coverage across the ‘genome’ even?
What percent of the genome was covered by at least one read?

Mapping
Were all of your reads mappable?
Where did the unmappable reads come from?
In a real sequencing dataset, why might there be unmappable reads?
What fraction (approximately) of reads mapped unambiguously (uniquely)?
Did you identify any sequencing errors?
Did you identify any variants (SNPs)?

Speed
What was your mapping speed (how many reads per minute did you map)?
How do you think that speed compares to the speed of mapping software like bowtie or bwa?
Could you have mapped faster with more workers in your group?





Unmapped reads:

coronaviru
us_vactine

Uniquely mapped

Ambiguously mapped

59 reads:

- 57 mapped (97%):
  - 32 mapped uniquely (54%)
- 2 unmapped (3%)

Coverage:

- 57 mapped reads x 10 ‘bases’ / read = 570 bases of data
- 570 base / 150 base genome = 3.8x avg coverage

Mapping tools like bowtie2 map fast!

Bowtie2 mapping **1M** 50nt reads to the human genome (3B bp)
1 CPU (1 thread/1 core):

```
mdstengl@cctsi-104:~/analyses/test_human_mapping$ time ./run_bowtie ERR3252925_1_1M.fastq GCA_000001405.15_GRCh38_no_alt_an
alysis_set.fna.bowtie_index
bowtie2 -x GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.bowtie_index -q -U ERR3252925_1_1M.fastq --local --score-min C,1
00,1 --no-unal --threads 1 -S ERR3252925_1_1M.fastq.GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.bowtie_index.sam

real    1m16.427s
user    1m13.836s
sys     0m12.844s
```

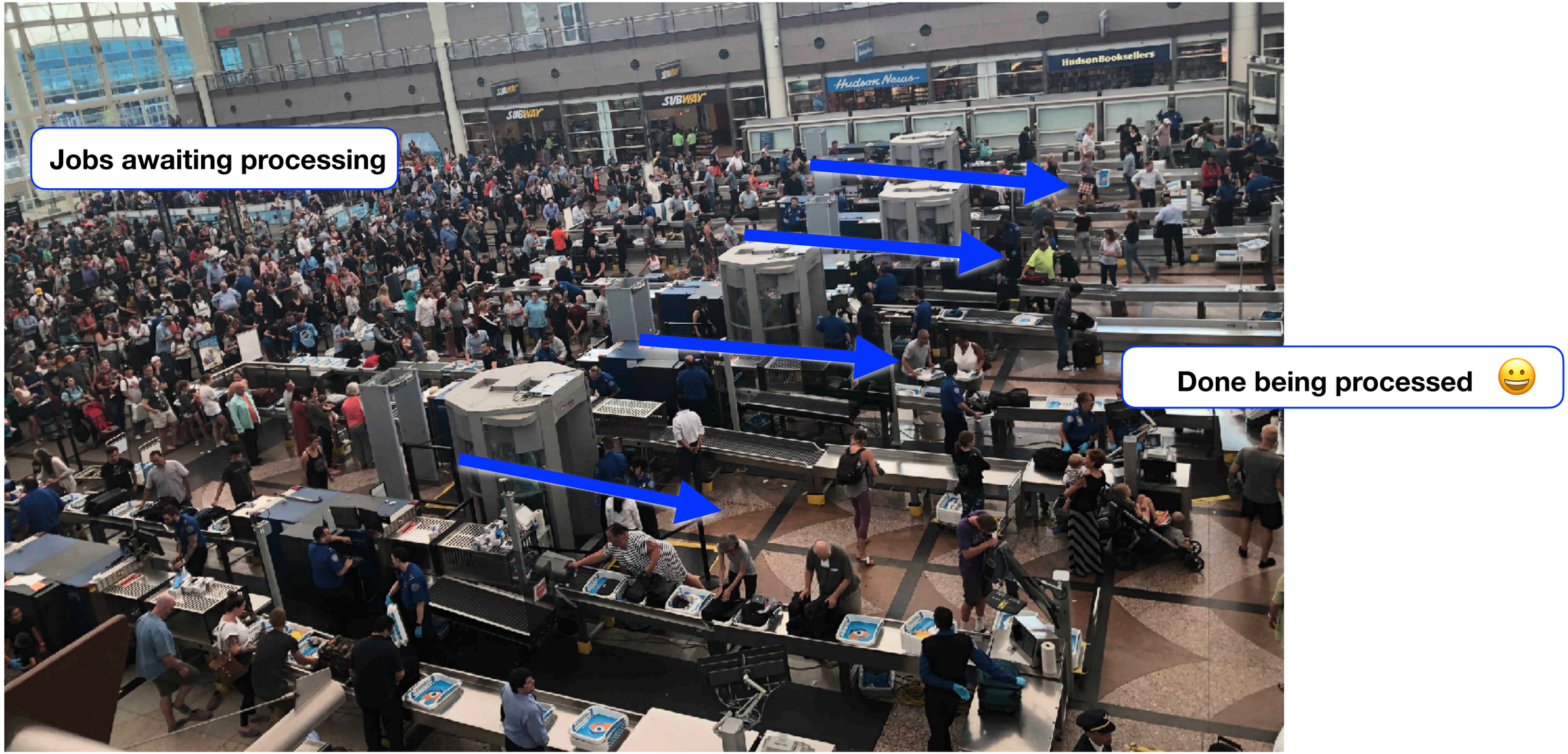
1 minute 16 seconds: 13,000 reads per second

Bowtie2 mapping **1M** 50nt reads to the human genome (3B bp)
24 CPUs (24 thread/24 core):

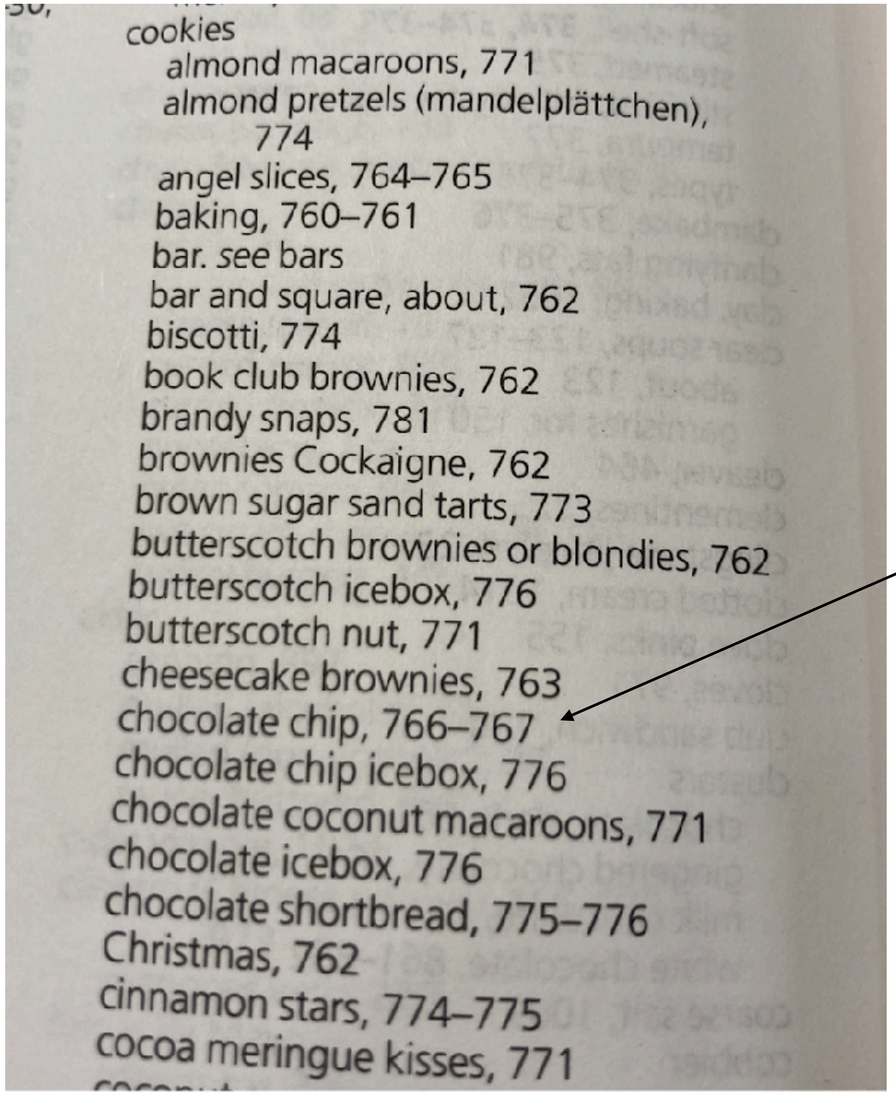
```
mdstengl@cctsi-104:~/analyses/test_human_mapping$ time ./run_bowtie_multiple_threads ERR3252925_1_1M.fastq GCA_000001405.15
_GRCh38_no_alt_analysis_set.fna.bowtie_index
bowtie2 -x GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.bowtie_index -q -U ERR3252925_1_1M.fastq --local --score-min C,1
00,1 --no-unal --threads 24 -S ERR3252925_1_1M.fastq.GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.bowtie_index.sam

real    0m9.641s
user    1m38.696s
sys     0m33.124s
```

9.6 seconds: ~100,000 reads per second

Like airport security, computers can run tasks in parallel to make jobs go faster

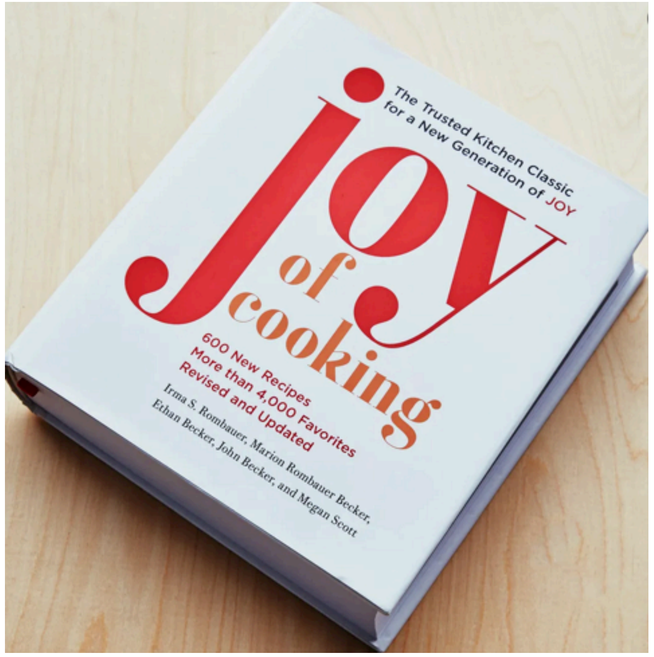


Mapping tools map fast because they **pre-index** reference sequences



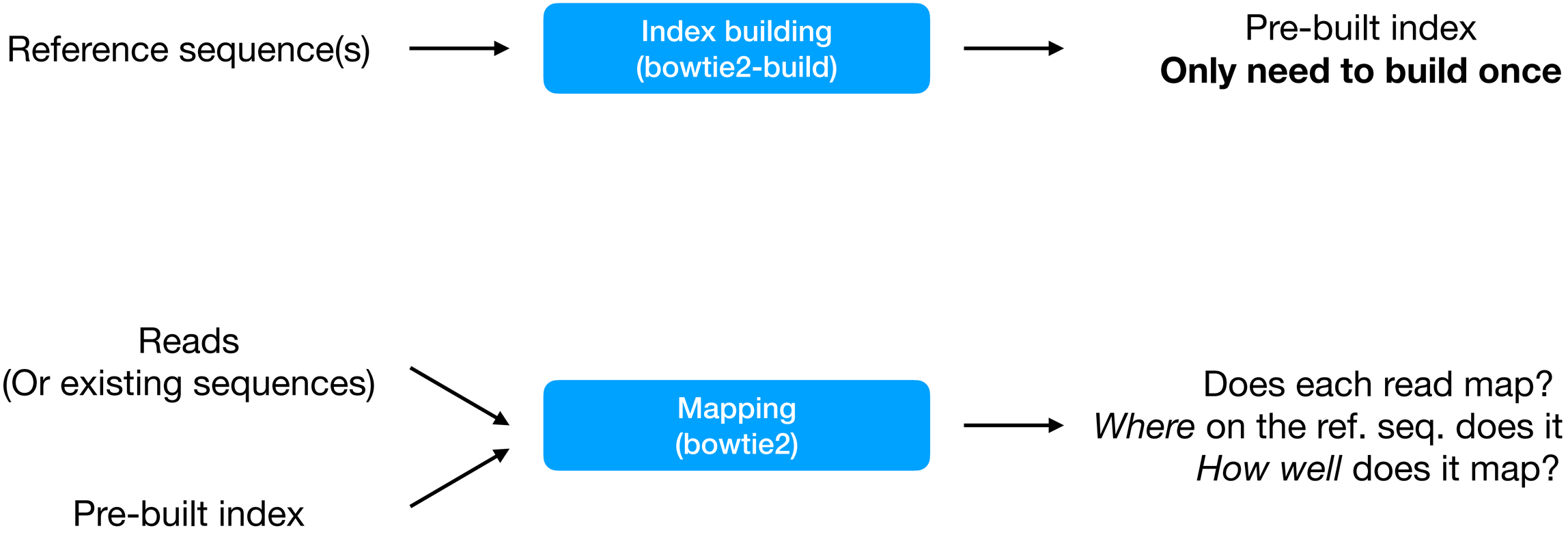
Indexes help you find things faster

chocolate chip cookies on page 766



BLAST databases are another example of pre-built indexes

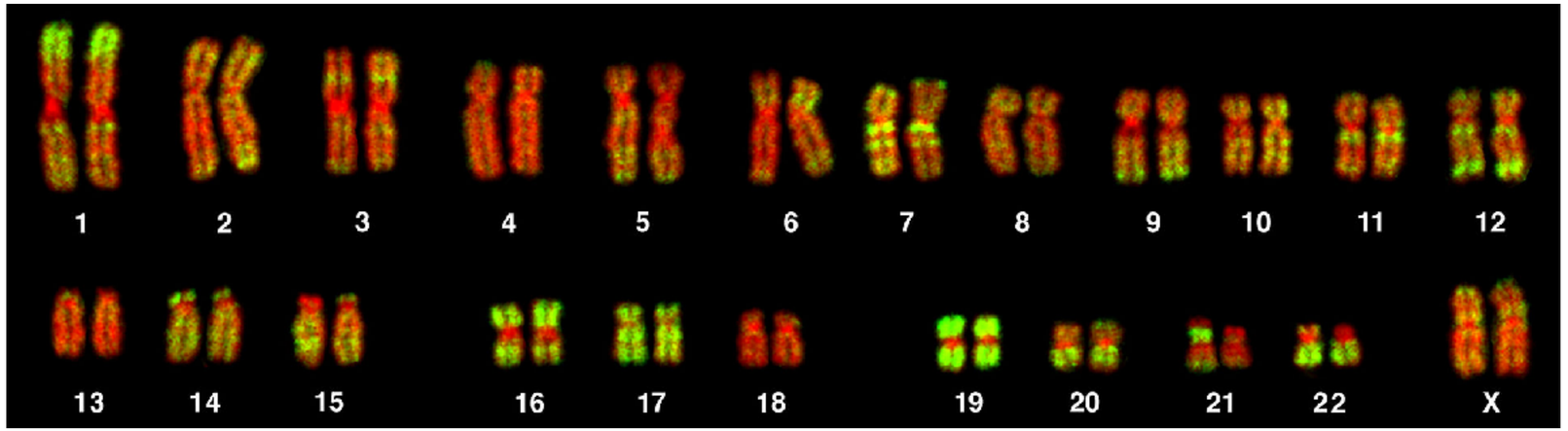
Mapping software includes tools to build the index



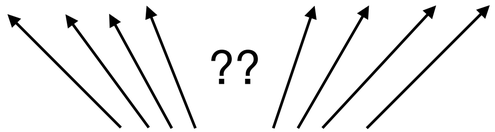
Reads might map to more than one location in a genome

This is especially a problem for reads that derive from repeat elements

Alu sequences in the human genome
1 million copies, ~10% of the mass



Bolzer et al (2005) PLoS Biol



If you had a read derived from an Alu sequence,
which of these million copies should you map it to?

Mapping tools like bowtie2 have options that define how they will deal with ambiguously mapping reads and provide information about whether a read mapped uniquely or not

Mapping quality measures whether a read maps uniquely or not

Mapping quality: higher = more unique

The aligner cannot always assign a read to its point of origin with high confidence. For instance, a read that originated inside a repeat element might align equally well to many occurrences of the element throughout the genome, leaving the aligner with no basis for preferring one over the others.

Aligners characterize their degree of confidence in the point of origin by reporting a mapping quality: a non-negative integer Q = -10 log10 p, where p is an estimate of the probability that the alignment does not correspond to the read's true point of origin. Mapping quality is sometimes abbreviated MAPQ, and is recorded in the SAM `MAPQ` field.

Mapping quality is related to "uniqueness." We say an alignment is unique if it has a much higher alignment score than all the other possible alignments. The bigger the gap between the best alignment's score and the second-best alignment's score, the more unique the best alignment, and the higher its mapping quality should be.

Accurate mapping qualities are useful for downstream tools like variant callers. For instance, a variant caller might choose to ignore evidence from alignments with mapping quality less than, say, 10. A mapping quality of 10 or less indicates that there is at least a 1 in 10 chance that the read truly originated elsewhere.

From: http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml

Mapping quality scores are like basecall quality scores in FASTQ files

Quality score = -10 $\log_{10}$ (p)

Basecall Q score = -10 $\log_{10}$ (**probability baseball is incorrect**)

Mapping Q score = -10 $\log_{10}$ (**probability that the read is not mapped to its true location**)

Q score	P
10	0.1 = 1/10
20	0.01 = 1/100
30	0.001 = 1/1,000
40	0.0001 = 1/10,000

bowtie2 mapped reads from *Drosophila melanogaster* to the *D. melanogaster* reference genome

```
bowtie2 -x /home/databases/fly/fly_genome -q -U dros_pool_R1_fu.fastq --local --score-min C,120,1 --no-unal --time --al dros_pool_R1_fu.
astq.fly_genome.hits.fastq --threads 12 -S dros_pool_R1_fu.fastq.fly_genome.sam
Time loading reference: 00:00:00
Time loading forward index: 00:00:01
Time loading mirror index: 00:00:00
Multiseed full-index search: 00:00:08
186708 reads; of these:
  186708 (100.00%) were unpaired; of these:
    20301 (10.87%) aligned 0 times
    87911 (47.08%) aligned exactly 1 time
    78496 (42.04%) aligned >1 times
89.13% overall alignment rate
Time searching: 00:00:09
Overall time: 00:00:09
```

10% of reads didn't map, what are these?

42% of reads mapped non-uniquely

47% of reads mapped uniquely

The output of mapping software is SAM format files

Header lines, start with @, provide info about the reference sequences and the mapping software used

```
@HD     VN:1.0  SO:unsorted
@SQ     SN:KP714088_Chaq        LN:1488
@SQ     SN:KP757930_Spock_RdRp  LN:1710
@SQ     SN:KP714099_Galbut_1    LN:1413
@SQ     SN:KP714100_Galbut_2    LN:1589
@PG     ID:bowtie2      PN:bowtie2      VN:2.3.2        CL:"/home/apps/bin/bowtie2-align-s --wrapper basic-0 -x galbut -q --local --score-min C,120,1 --time --threads 24 --pa
ssthrough -U F4-17_R1_fuh.fastq"
NS500697:120:HGWL5AFXY:1:11102:12345:6629       16      KP714100_Galbut_2       188     44      150M    *       0       0       CAACCGGTTTTTGAGTATGAGCGTGCAATGGAACGCCTCAACCGGC
AACAACAACCGCGTCCAGTCCAGAAAAACGCGGGCAGCAGACCAAAGAAGTCTAGCGATAGAGGCCCAGAGCCTAAACCGACAACAGTCGAGCAGGGGCAGCGA      AEEEAA<AAEEAAEAEEEAAAEEEAEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEE
AEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEAAAA  AS:i:300        XN:i:0  XM:i:0  XO:i:0      XG:i:0  NM:i:0  MD:Z:150        YT:Z:UU
NS500697:120:HGWL5AFXY:1:11208:16336:12687      16      KP757930_Spock_RdRp     861     36      150M    *       0       0       TTTTTGTTCACTGACTATACGGCGTTTGACAGTTCAGTGCCCTCAT
GGTTAATCCGTGAGTGTTTCAAAATCGTAATGGATTGCTTCTACAAAAACCTAGGCCCCGGCGACTATCAAGTCTTCAGTAAGATAGTCAACTACTTTATTAAC      AEEEEEA<EEEEAEEEAEEEAEEEEEEEEEEEEAEEAEEEEEEEEEEEAEEEEEEAAEEEEE/E
EEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEAAAA  AS:i:225        XN:i:0  XM:i:11 XO:i:0      XG:i:0  NM:i:11 MD:Z:29T8A9C13C5G2A24T10T2
8G0T2T9       YT:Z:UU
NS500697:120:HGWL5AFXY:1:11210:23050:9458       16      KP757930_Spock_RdRp     857     36      140M10S *       0       0       CAAATTTTTGTTCACTGACTATACGGCGTTTGACAGTTCAGTGCCC
TCATGGTTAATCCGTGAGTGTTTCAAAATCGTAATGGATTGCTTCTACAAAAACCTAGGCCCCGGCGACTATCAAGTCTTCAGTAAGATAGTCAACTACTTTAT      EEEEEEEEEEEEEAEEEAEEEEEEEAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAAAA  AS:i:224        XN:i:0  XM:i:8  XO:i:0      XG:i:0  NM:i:8  MD:Z:33T8A9C13C5G2A24T10T2
8     YT:Z:UU
```

After header lines, one line per mapped read, with 11 columns separated by tabs

SAM column	Info
1	The read's name
3	The mapped-to reference sequence name
4	Position in the reference sequence where the read maps
5	Mapping quality score
6	Whether there are mismatches to the reference sequence
10	The read's sequence
11	The read's basecall quality scores

SAM files are used as input to downstream tools that use mapping data