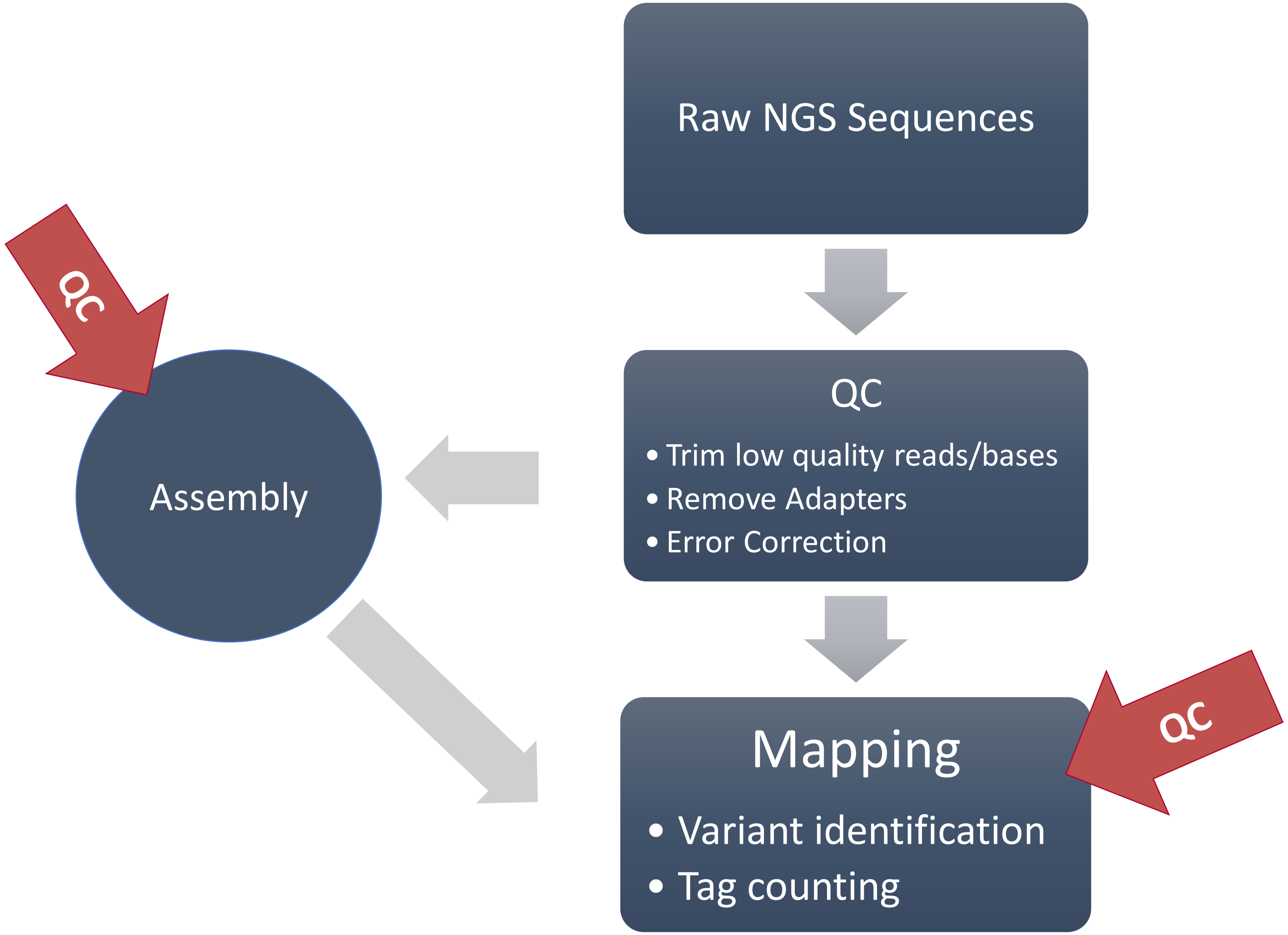
Mapping

Aligning sequencing reads to a reference

The Big Picture



Raw NGS Sequences

QC

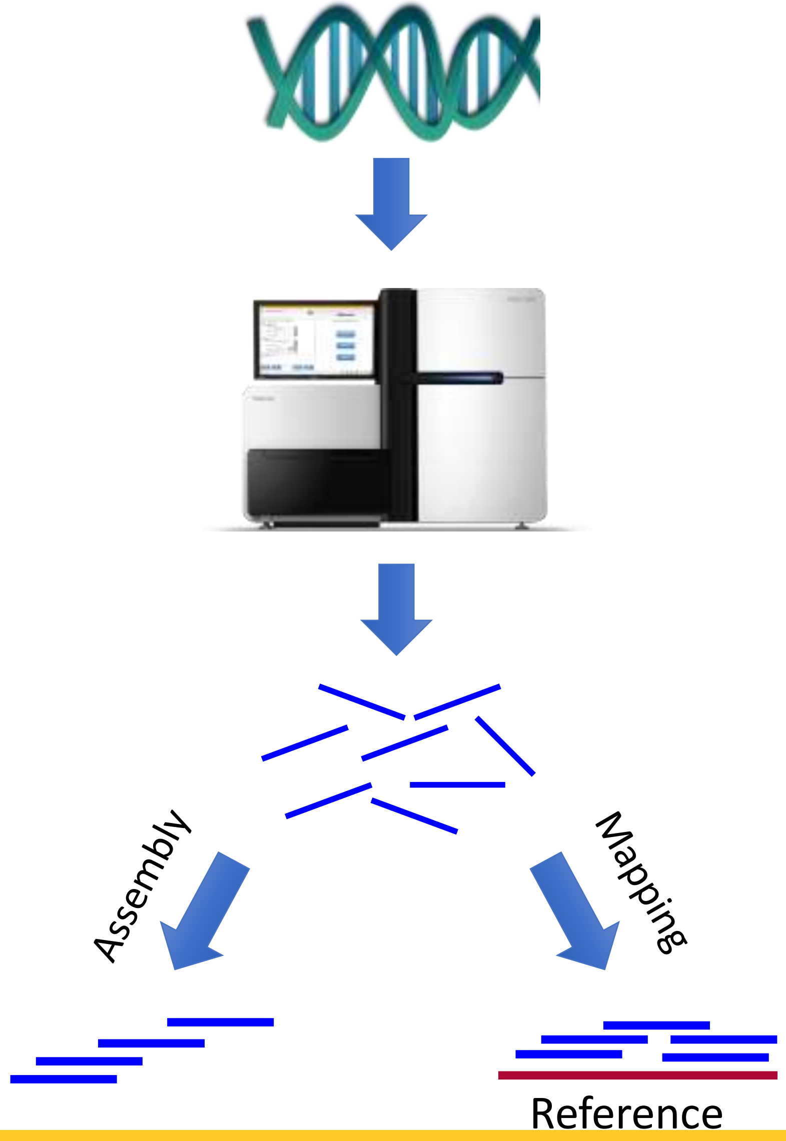
- Trim low quality reads/bases
- Remove Adapters
- Error Correction

Assembly

Mapping

- Variant identification
- Tag counting

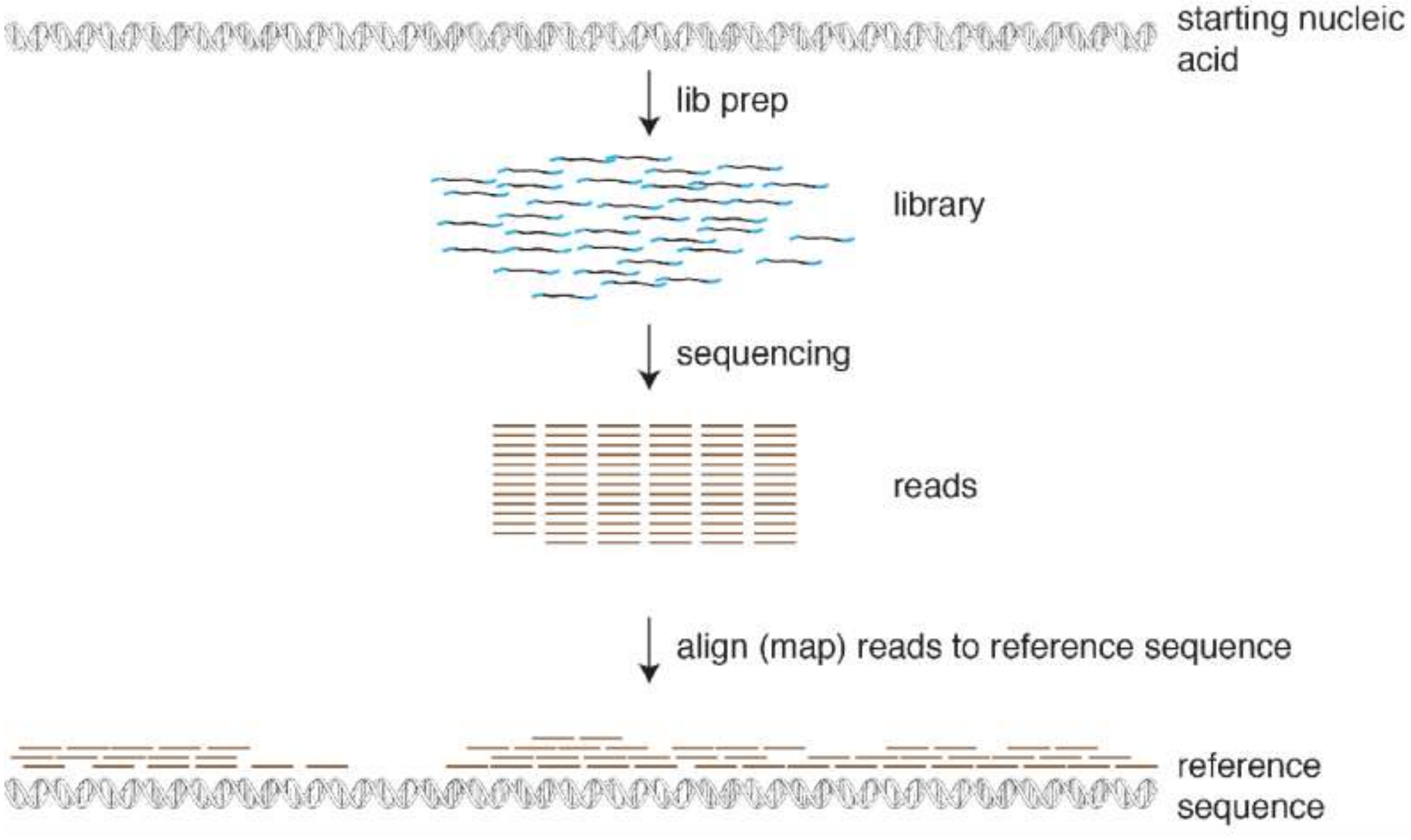
So you got some sequences… now what?



- Quality Control
  - Trim low quality reads/bases
  - Remove Adapters
  - Error Correction
- Mapping
  - Comparison to a reference sequence (genome, transcriptome, etc)
- Assembly
  - Generate a new consensus sequence (genome, transcriptome, etc)

Mapping

- The process by which sequencing reads are aligned (matched) to a region of the genome from which they derive (*reference*)



Why Mapping?

- Identify variants
  - *Substitutions* (fixed difference from reference)
  - *Polymorphisms* (multiple alleles, heterozygous)
    - Single nucleotide polymorphisms = *SNPs*
  - *Structural variants* (insertion-deletion events, duplications, etc)

Variants

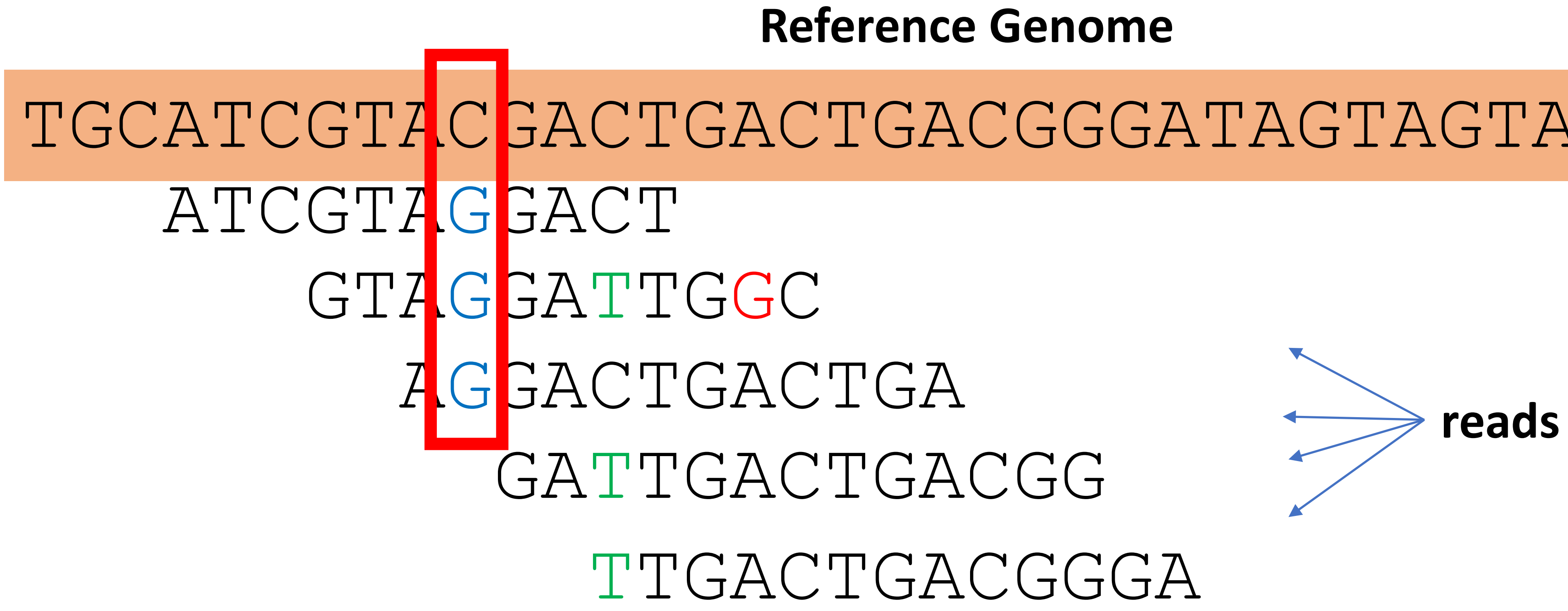
Reference Genome

```
TGCATCGTACGACTGACTGACGGGATAGTAGTAGTCTCTGA
   ATCGTAGGACT
      GTAGGATTGGC
        AGGACTGACTGA
          GATTGACTGACGG
            TTGACTGACGGGA
```

reads

G = substitution   T = polymorphism   G = sequencing error

Variants

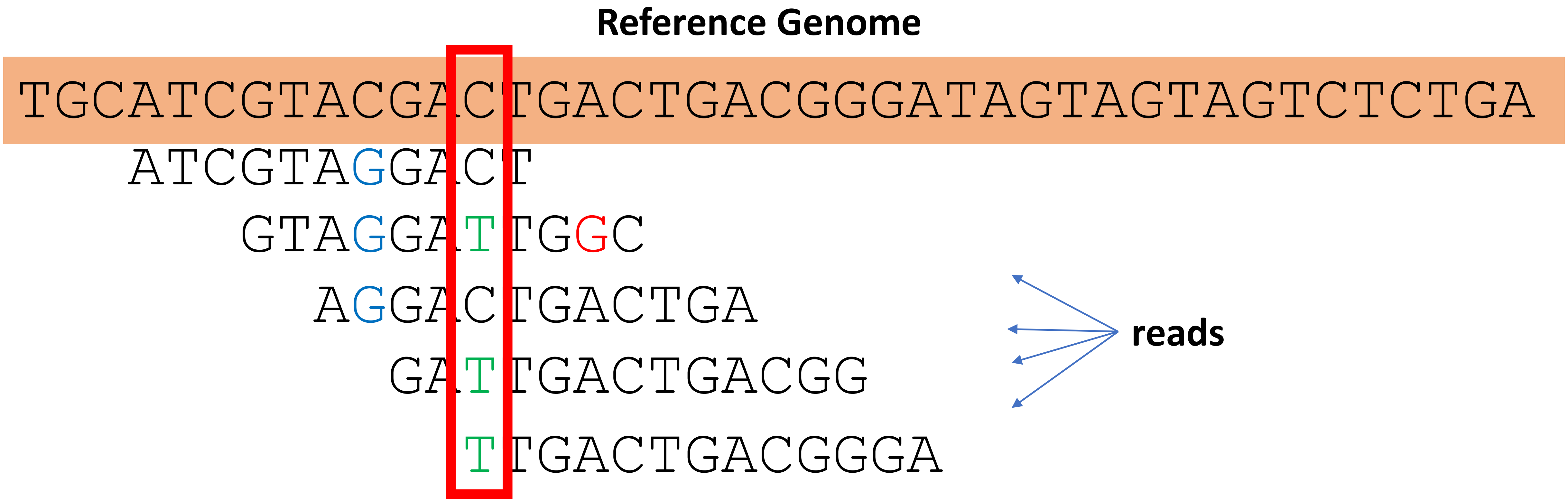


Reference Genome

```
TGCATCGTACGACTGACTGACGGGATAGTAGTAGTCTCTGA
   ATCGTAGGACT
      GTAGGATTGGC
        AGGACTGACTGA
          GATTGACTGACGG
            TTGACTGACGGGA
```

G = substitution T = polymorphism G = sequencing error

Variants



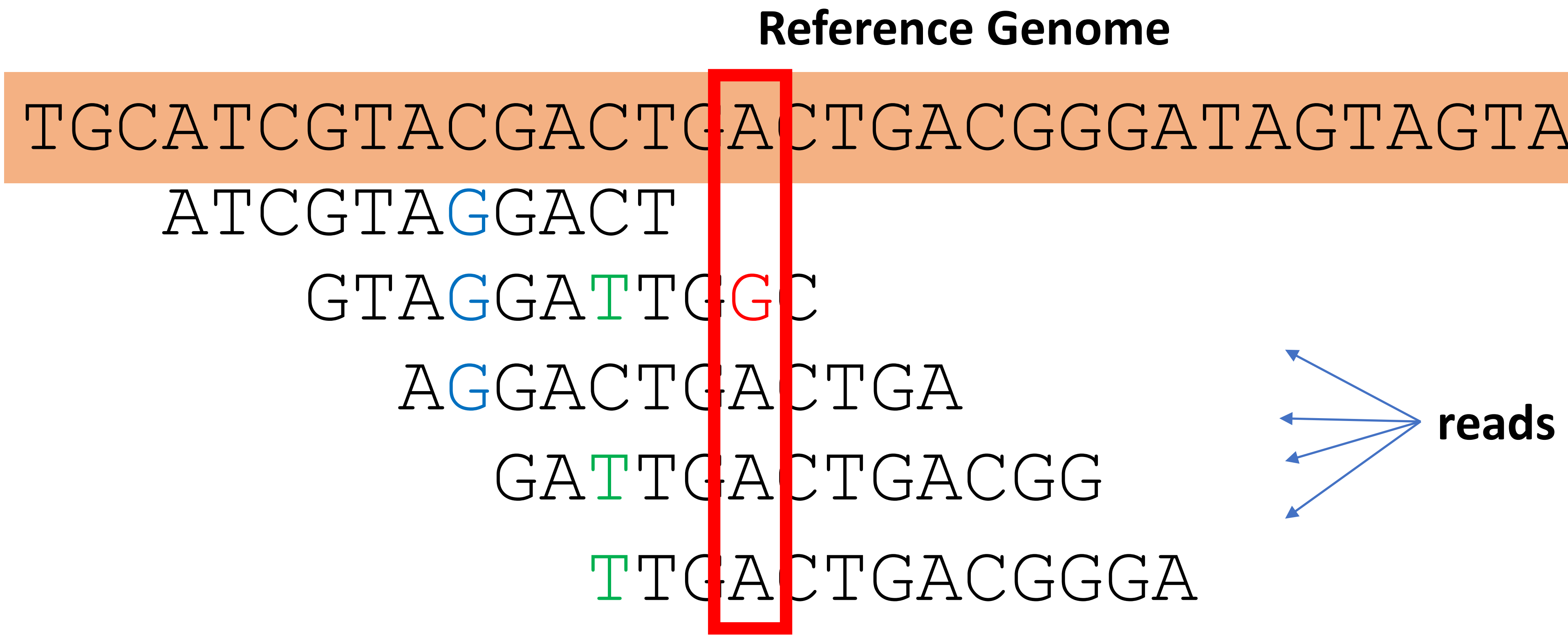
Reference Genome

```
TGCATCGTACGACTGACTGACGGGATAGTAGTAGTCTCTGA
   ATCGTAGGACT
      GTAGGATTGGC
        AGGACTGACTGA
          GATTGACTGACGG
            TTGACTGACGGGA
```

reads

G = substitution   T = polymorphism   G = sequencing error

Variants



Reference Genome

TGCATCGTACGACTGACTGACGGGATAGTAGTAGTCTCTGA

ATCGTAGGACT

GTAGGATTGGC

AGGACTGACTGA

GATTGACTGACGG

TTGACTGACGGGA

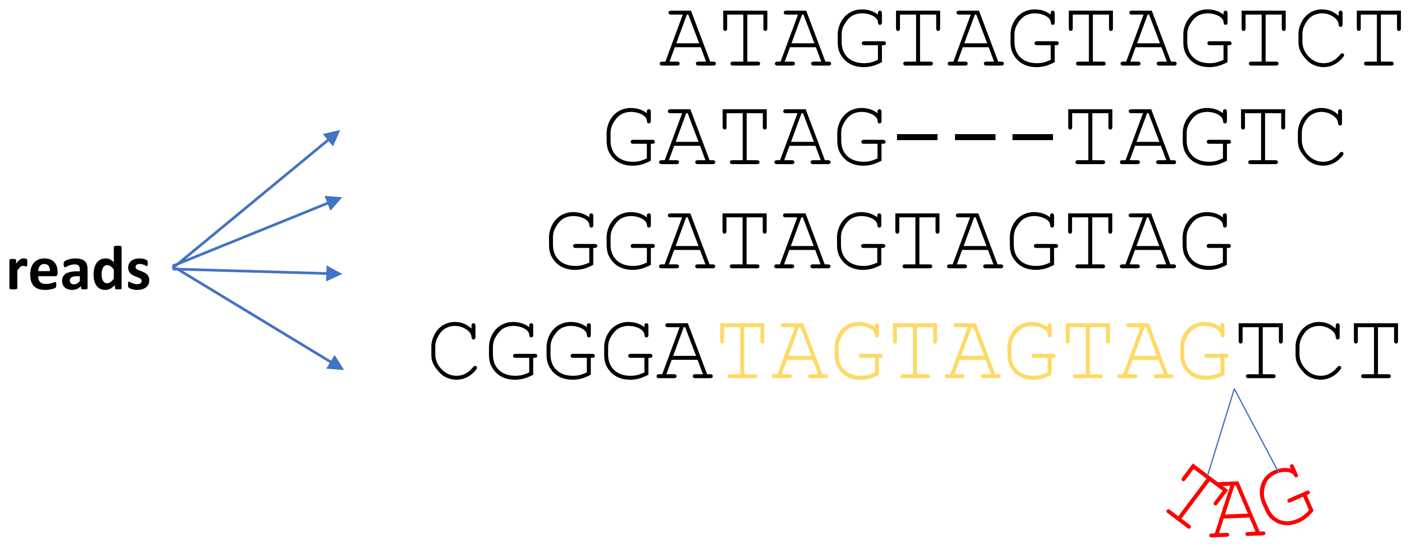
reads

G = substitution T = polymorphism G = sequencing error

Variants

Reference Genome

```
TGCATCGTACGACTGACTGACGGGATAGTAGTAGTCTCTGA
```



reads

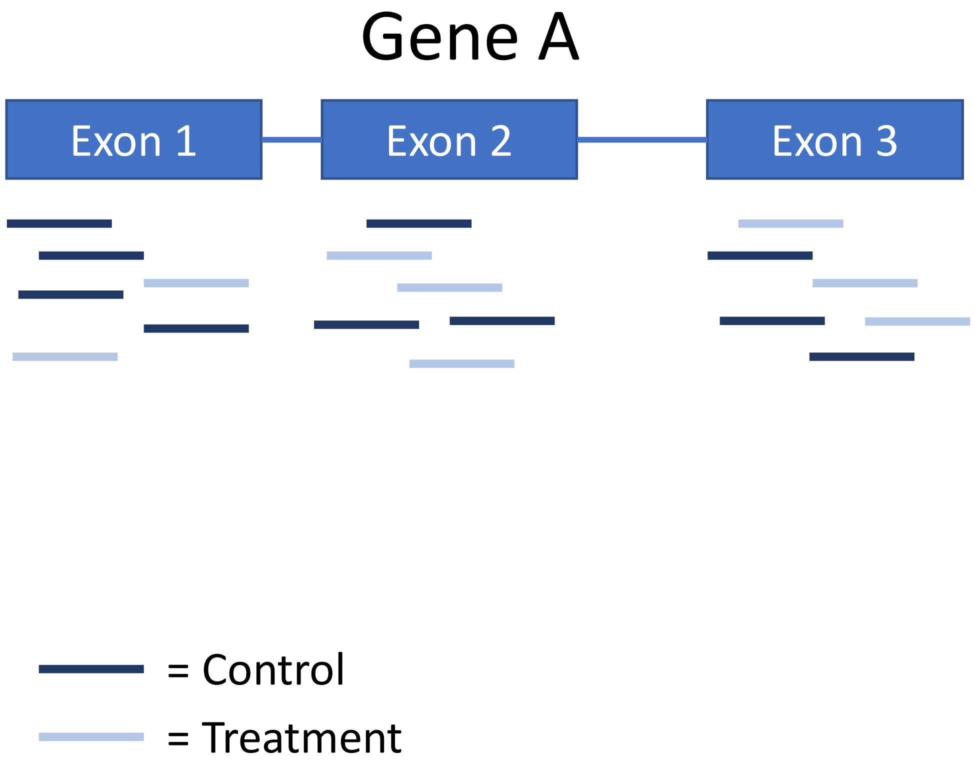
```
                        ATAGTAGTAGTCT
                       GATAG---TAGTC
                      GGATAGTAGTAG
                    CGGGATAGTAGTAGTCT
                                 TAG
```

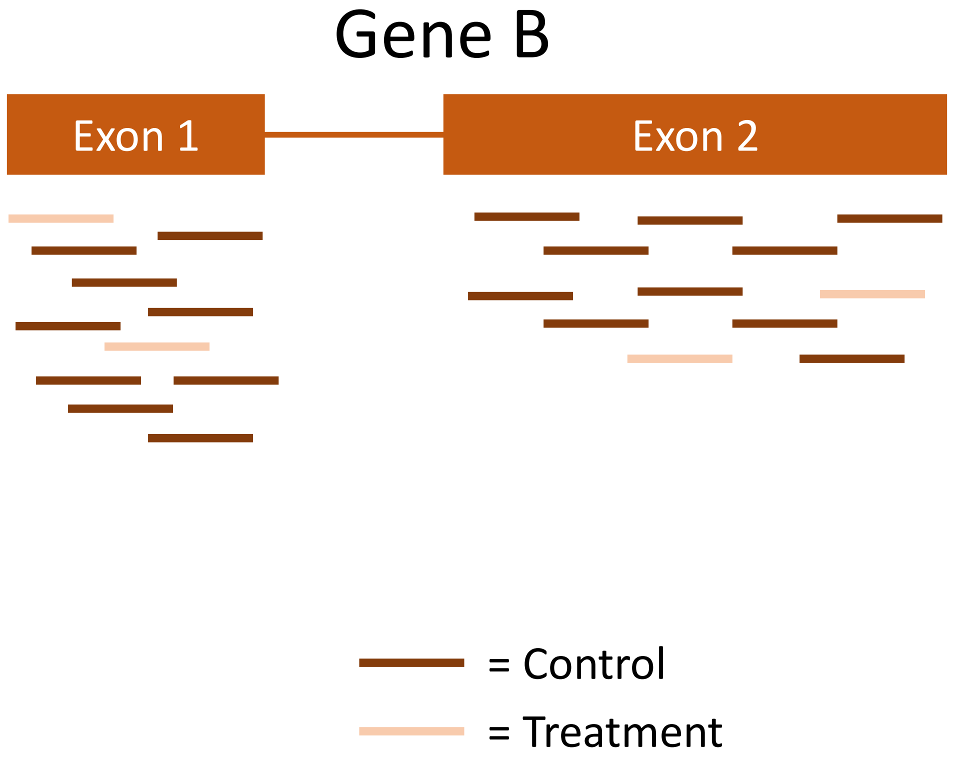
--- = deletion from reference; TAG = insertion relative to reference

Why Mapping?

- Identify variants
  - *Substitutions* (fixed difference from reference)
  - *Polymorphisms* (multiple alleles, heterozygous)
    - Single nucleotide polymorphisms = *SNPs*
  - *Structural variants* (insertion-deletion events, duplications, etc)
- Quantification (counting)
  - Expression level of genes in a transcriptome (RNA-seq)

Quantification

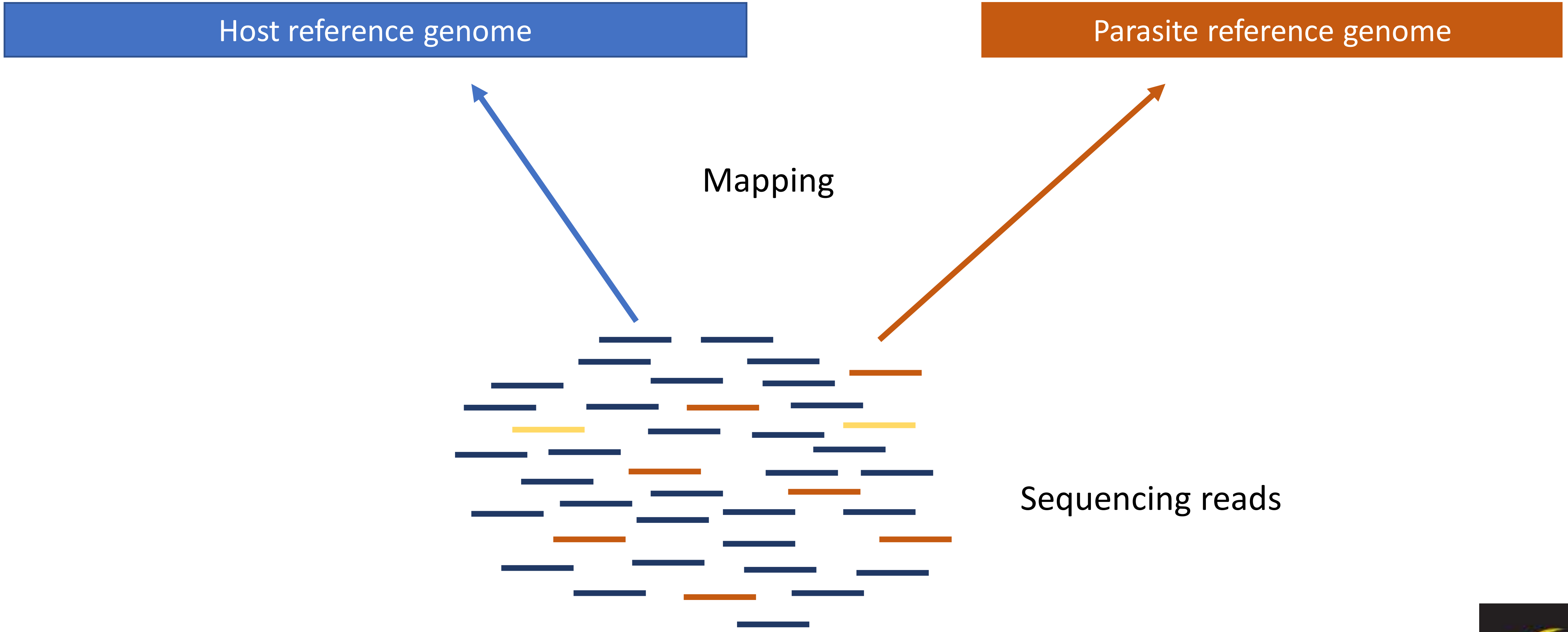




Why Mapping?

- Identify variants
  - *Substitutions* (fixed difference from reference)
  - *Polymorphisms* (multiple alleles, heterozygous)
    - Single nucleotide polymorphisms = *SNPs*
  - *Structural variants* (insertion-deletion events, duplications, etc)
- Quantification
  - Expression level of genes in a transcriptome (RNA-seq)
- Identify or remove sequences of specific origins
  - Contamination
  - Parasites, microbiome, pathogens
  - Organellar DNA (mtDNA, cpDNA)

Identifying/Removing sequences from mixed origin



Host reference genome

Parasite reference genome

Mapping

Sequencing reads

▬ = Other (e.g., contamination, microbiome, etc)

DIY Exercise!!

Map the reads to the reference!

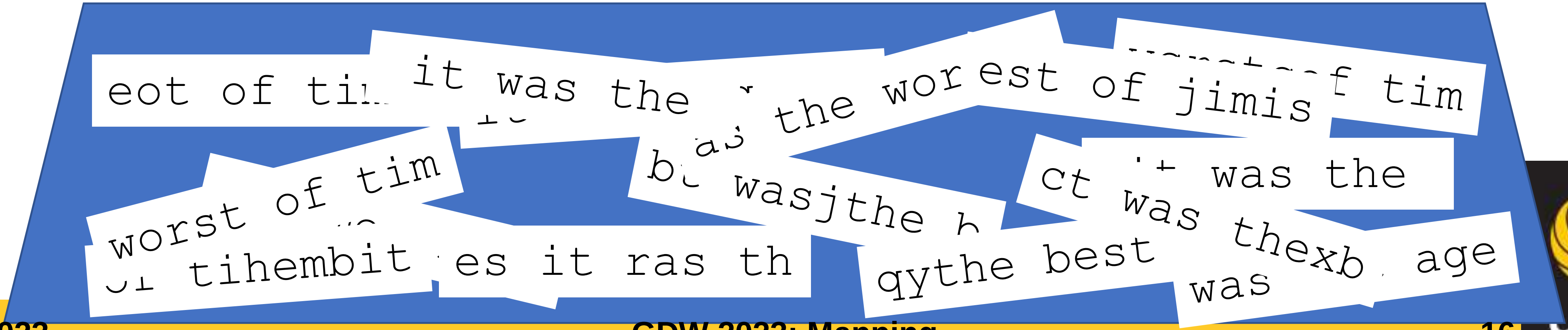
http://ivory.idyll.org/blog/the-assembly-exercise.html

it was the best of times it was the worst of times it was the age

Reference “Genome”

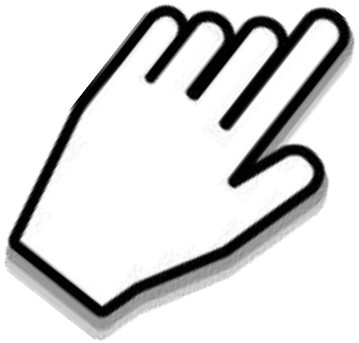
Read “Library”

(shotgun sequenced pieces of the reference genome)

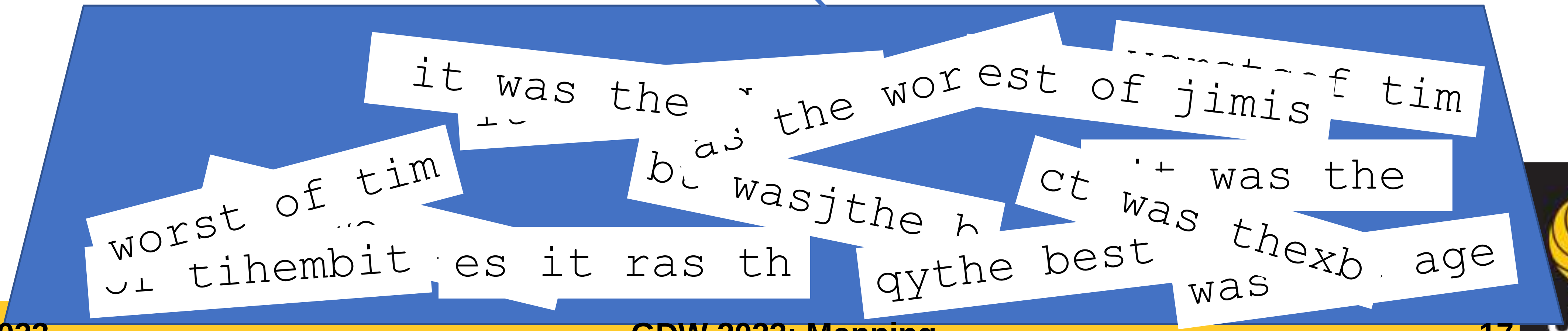


it was the best of times it was the worst of times it was the age

eot of timis



Drag and drop the “reads” from the library and align them to the reference genome.



DIY Exercise!!

- Report the:
  - Coverage
  - Error rate
  - How many variants (SNPs)?
  - Mapping rate (reads/sec)?
- Extra credit: Name the book and author
  - No Googling!

http://ivory.idyll.org/blog/the-assembly-exercise.html

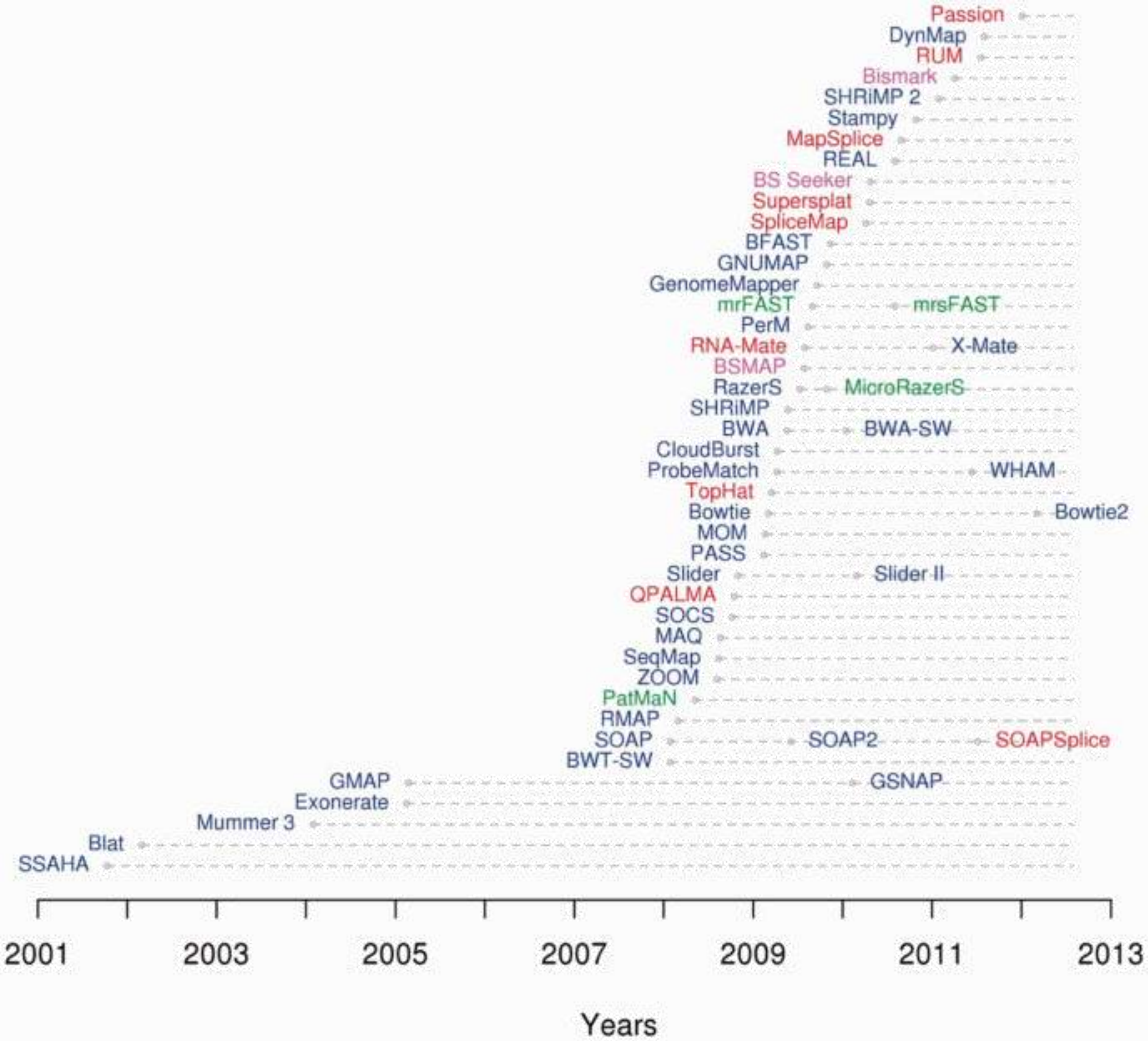
Just a pairwise alignment, right?

Yes.

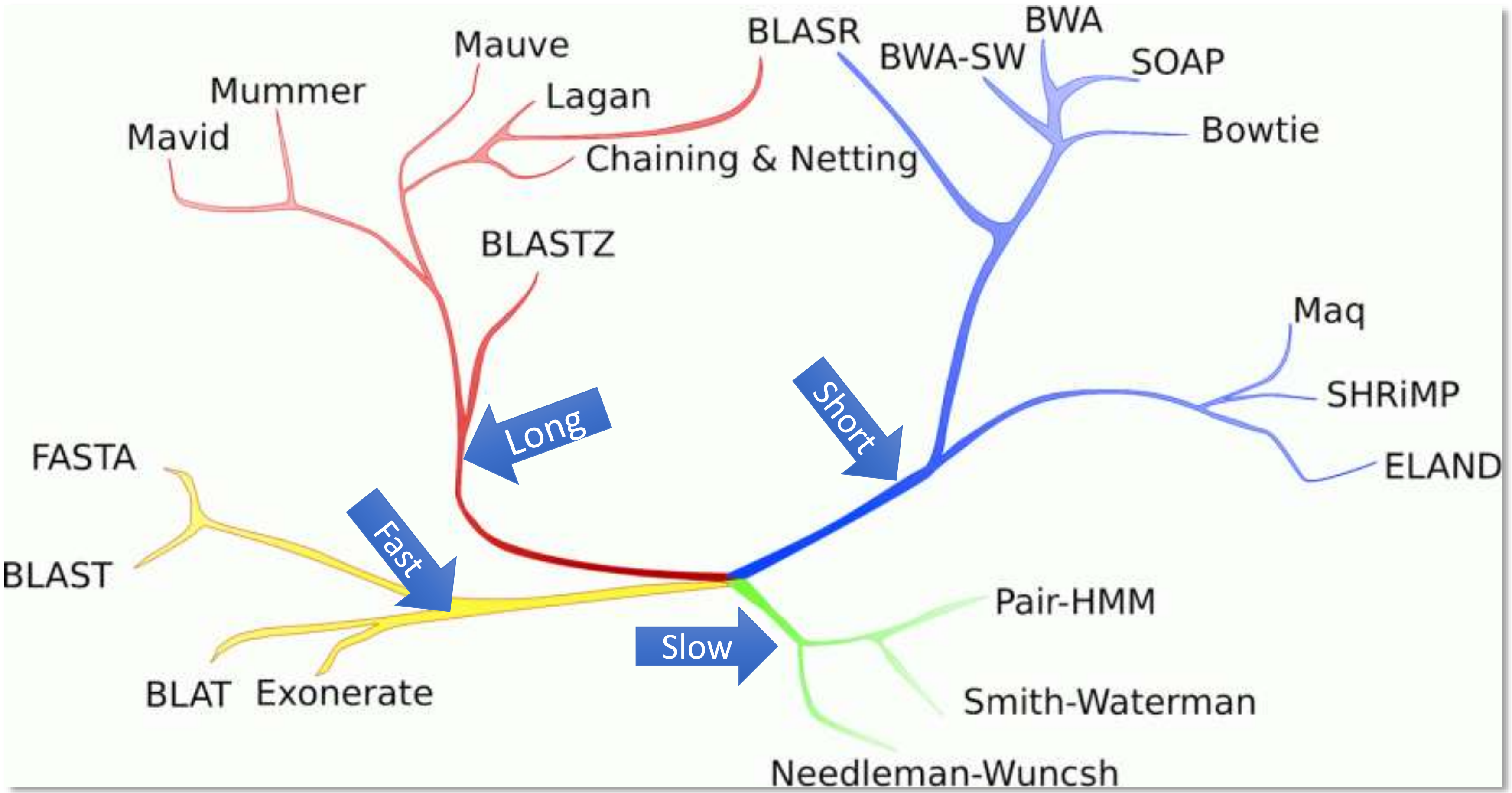
x 400 million (or more)

Which Mapping Soft

- >70 published programs
  - Input data type
  - Reference
  - Speed vs sens
  - Memory



Phylogeny of Pairwise Alignment



Chaisson & Tesler 2012, *BMC Bioinformatics*

Comparison (10 million human reads, 40 bp)

Software	Algorithm	Mismatches	Memory (GB)	Time (min)
BWA	BWT	yes	2.2	73
Bowtie	BWT	yes	7.4	166
BFAST	Spaced seeds	yes	9.7	902
MPScan	Suffix tree	no	2.7	80
PerM	Spaced seeds	yes	13.8	785

Schbath et al. 2012 *J Comput Biol*

Comparison (RNA-seq reads; splice aware)

1 core (cpu)

Software	Read-pairs/hr (millions)	Memory (GB)
STAR	51.5	4.5
STAR sparse	37.9	2.6
TopHat2	1.33	0.68
RUM	0.85	4.5
MapSplice	0.5	0.55
GSNAP	0.3	4.3

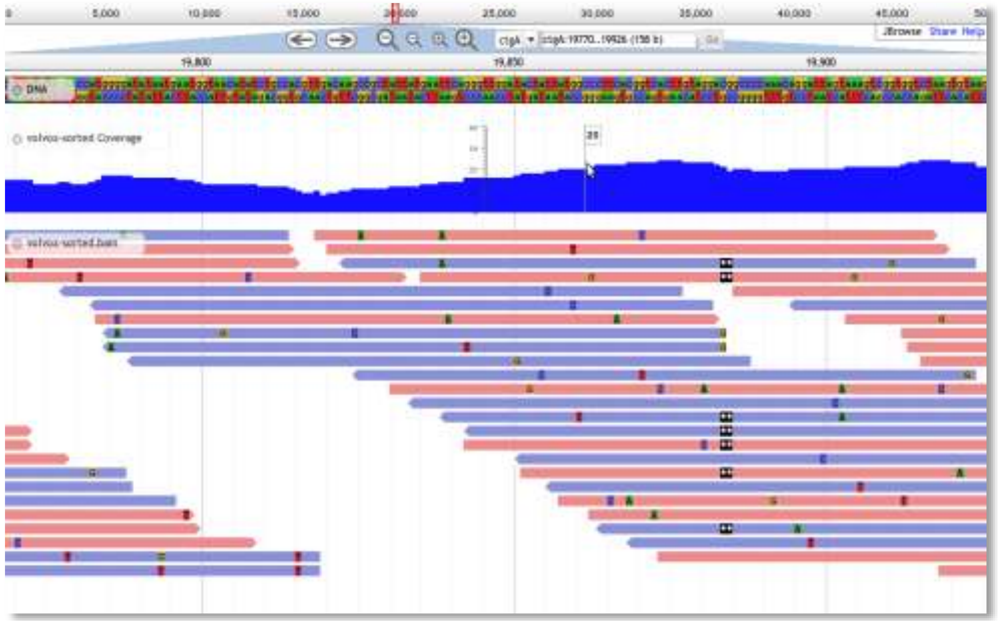
Dobin et al. 2013 *Bioinformatics*

Comparison (RNA-seq reads; splice aware)

	1 core (cpu)		12 cores (cpus)	
Software	**Read-pairs/hr (millions)**	**Memory (GB)**	**Read-pairs/hr (millions)**	**Memory (GB)**
STAR	51.5	4.5	549.9	28.4
STAR sparse	37.9	2.6	423.1	16.0
TopHat2	1.33	0.68	10.1	11.3
RUM	0.85	4.5	7.6	53.3
MapSplice	0.5	0.55	3.1	3.3
GSNAP	0.3	4.3	2.8	27.0

Dobin et al. 2013 *Bioinformatics*

Storing Read Alignments

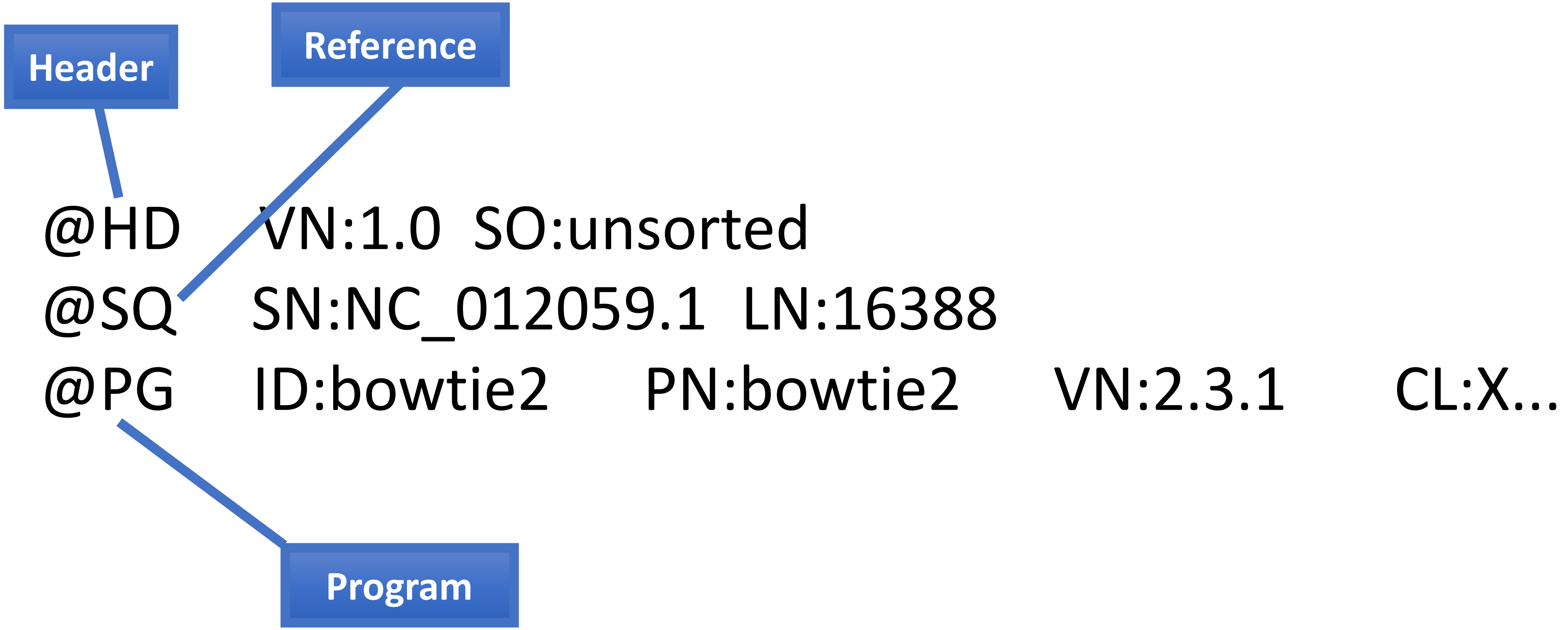


Sequence Alignment (SAM/BAM) Format

- Universal Standard
- SAM (readable)
- BAM (binary, compressed form)
- Specifications:
  - https://samtools.github.io/hts-specs/SAMv1.pdf

- **Structure**
  - Header: programs, version, reference info, sort order, sample info, etc.
  - Read alignment records
    - One record per line

SAM: Header



Header

Reference

```
@HD    VN:1.0  SO:unsorted
@SQ    SN:NC_012059.1  LN:16388
@PG    ID:bowtie2     PN:bowtie2     VN:2.3.1      CL:X...
```

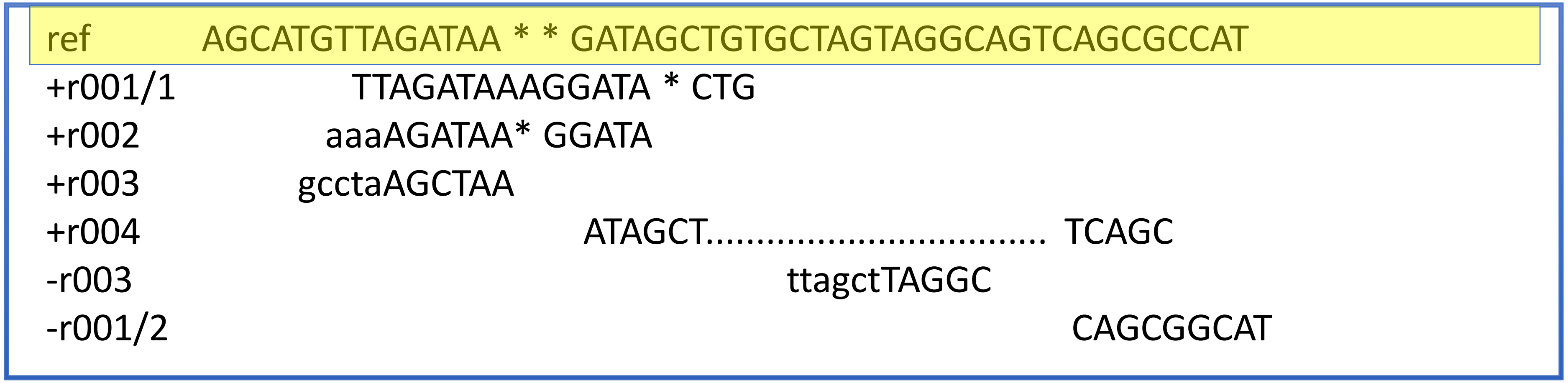
Program

X =bowtie2-align-s --wrapper basic-0 -q --phred33 --very-sensitive -t -p 1 -x NC_012059.1 -1 ERR1938563_1.fq -2 ERR1938563_2.fq

SAM: Alignments

```
ref    AGCATGTTAGATAA * * GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
+r001/1       TTAGATAAAGGATA * CTG
+r002        aaaAGATAA* GGATA
+r003       gcctaAGCTAA
+r004                    ATAGCT..................................  TCAGC
-r003                              ttagctTAGGC
-r001/2                                         CAGCGGCAT
```

SAM: Alignments



```
ref       AGCATGTTAGATAA * * GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
+r001/1          TTAGATAAAGGATA * CTG
+r002           aaaAGATAA* GGATA
+r003         gcctaAGCTAA
+r004                       ATAGCT.................................. TCAGC
-r003                                 ttagctTAGGC
-r001/2                                            CAGCGGCAT
```

SAM: Alignments

```
ref    AGCATGTTAGATAA * * GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
+r001/1       TTAGATAAAGGATA * CTG
+r002       aaaAGATAA* GGATA
+r003      gcctaAGCTAA
+r004                   ATAGCT.......................  TCAGC
-r003                            ttagctTAGGC
-r001/2                                       CAGCGGCAT
```

SAM: Alignments

```
ref       AGCATGTTAGATAA * * GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
+r001/1          TTAGATAAAGGATA * CTG
+r002           aaaAGATAA* GGATA
+r003          gcctaAGCTAA
+r004                       ATAGCT..............................  TCAGC
-r003                                   ttagctTAGGC
-r001/2                                              CAGCGGCAT
```

SAM: Alignments

```
ref      AGCATGTTAGATAA * * GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
+r001/1         TTAGATAAAGGATA * CTG
+r002          aaaAGATAA* GGATA
+r003         gcctaAGCTAA
+r004                      ATAGCT.................................. TCAGC
-r003                                ttagctTAGGC
-r001/2                                           CAGCGGCAT
```

The corresponding SAM format is:

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001    99  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   147  ref  37  30           9M    =   7 -39   CAGCGGCAT * NM:i:1
```

SAM: Alignments

```
ref       AGCATGTTAGATAA * * GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
+r001/1          TTAGATAAAGGATA * CTG
+r002           aaaAGATAA* GGATA
+r003         gcctaAGCTAA
+r004                       ATAGCT................................ TCAGC
-r003                                ttagctTAGGC
-r001/2                                            CAGCGGCAT
```

The corresponding SAM format is:

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001    99  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   147  ref  37  30           9M   =   7 -39   CAGCGGCAT * NM:i:1
```

Read name

SAM: Alignments

```
ref    AGCATGTTAGATAA * * GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
+r001/1      TTAGATAAAGGATA * CTG
+r002       aaaAGATAA* GGATA
+r003      gcctaAGCTAA
+r004                  ATAGCT.................................. TCAGC
-r003                           ttagctTAGGC
-r001/2                                     CAGCGGCAT
```

The corresponding SAM format is:

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001    99 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   147 ref 37  30          9M   =   7 -39  CAGCGGCAT * NM:i:1
```

Flag: pair information, orientation, mapped, etc.

SAM: Alignments

```
ref        AGCATGTTAGATAA * * GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
+r001/1             TTAGATAAAGGATA * CTG
+r002             aaaAGATAA* GGATA
+r003           gcctaAGCTAA
+r004                              ATAGCT.............................. TCAGC
-r003                                              ttagctTAGGC
-r001/2                                                              CAGCGGCAT
```

The corresponding SAM format is:

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001    99  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   147  ref 37  30          9M   =   7 -39  CAGCGGCAT * NM:i:1
```

Reference sequence name & position

SAM: Alignments

```
ref        AGCATGTTAGATAA * * GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
+r001/1           TTAGATAAAGGATA * CTG
+r002            aaaAGATAA* GGATA
+r003          gcctaAGCTAA
+r004                        ATAGCT................................ TCAGC
-r003                                  ttagctTAGGC
-r001/2                                               CAGCGGCAT
```

The corresponding SAM format is:

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001     99  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   147  ref  37  30          9M      =   7 -39   CAGCGGCAT * NM:i:1
```

Mapping Quality (MQ): $-10 * \log_{10}$(pr[wrongly mapped])

SAM: Alignments

```
ref       AGCATGTTAGATAA * * GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
+r001/1          TTAGATAAAGGATA * CTG
+r002           aaaAGATAA* GGATA
+r003         gcctaAGCTAA
+r004                       ATAGCT................................. TCAGC
-r003                                ttagctTAGGC
-r001/2                                            CAGCGGCAT
```

The corresponding SAM format is:

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001    99  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   147  ref  37  30           9M   =   7 -39   CAGCGGCAT * NM:i:1
```

CIGAR string

CIGAR String: Compact Idiosyncratic Gapped Alignment Report

```
REF  ACGATACATAC
READ ACGA-ACATAC

CIGAR: 4M1D6M
[4 Matches + 1 Deletion + 6 Matches]
```

```
REF    GACA-AACC
READ atGTCATAACC

CIGAR: 2S4M1I4M
[2 Skipped + 4 Matches + 1 Insertion + 4 Matches]
```

CIGAR String: Compact Idiosyncratic Gapped Alignment Report

```
REF  ACGATACATAC
READ ACGA-ACATAC

CIGAR: 4M1D6M
[4 Matches + 1 Deletion + 6 Matches]
```

```
REF    GACA-AACC
READ atGTCATAACC

CIGAR: 2S4M1I4M
[2 Skipped + 4 Matches + 1 Insertion + 4 Matches]
```

CIGAR String: Compact Idiosyncratic Gapped Alignment Report

```
REF   ACGATACATAC
READ ACGA-ACATAC

CIGAR: 4M1D6M
[4 Matches + 1 Deletion + 6 Matches]
```

```
REF     GACA-AACC
READ atGTCATAACC

CIGAR: 2S4M1I4M
[2 Skipped + 4 Matches + 1 Insertion + 4 Matches]
```

CIGAR String: Compact Idiosyncratic Gapped Alignment Report

```
REF  ACGATACATAC
READ ACGA-ACATAC

CIGAR: 4M1D6M
[4 Matches + 1 Deletion + 6 Matches]
```

```
REF    GACA-AACC
READ atGTCATAACC

CIGAR: 2S4M1I4M
[2 Skipped + 4 Matches + 1 Insertion + 4 Matches]
```

CIGAR String: Compact Idiosyncratic Gapped Alignment Report

```
REF   ACGATACATAC
READ ACGA-ACATAC

CIGAR: 4M1D6M
[4 Matches + 1 Deletion + 6 Matches]
```

```
REF    GACA-AACC
READ atGTCATAACC

CIGAR: 2S4M1I4M
[2 Skipped + 4 Matches + 1 Insertion + 4 Matches]
```

CIGAR String: Compact Idiosyncratic Gapped Alignment Report

```
REF  ACGATACATAC
READ ACGA-ACATAC

CIGAR: 4M1D6M
[4 Matches + 1 Deletion + 6 Matches]
```

```
REF    GACA-AACC
READ atGTCATAACC

CIGAR: 2S4M1I4M
[2 Skipped + 4 Matches + 1 Insertion + 4 Matches]
```

CIGAR String: Compact Idiosyncratic Gapped Alignment Report

```
REF  ACGATACATAC
READ ACGA-ACATAC

CIGAR: 4M1D6M
[4 Matches + 1 Deletion + 6 Matches]
```

```
REF     GACA-AACC
READ atGTCATAACC

CIGAR: 2S4M1I4M
[2 Skipped + 4 Matches + 1 Insertion + 4 Matches]
```

CIGAR String: Compact Idiosyncratic Gapped Alignment Report

```
REF  ACGATACATAC
READ ACGA-ACATAC
```

CIGAR: 4M1D6M

[4 Matches + 1 Deletion + 6 Matches]

```
REF    GACA-AACC
READ atGTCATAACC
```

CIGAR: 2S4M1I4M

[2 Skipped + 4 Matches + 1 Insertion + 4 Matches]

SAM: Alignments

```
ref        AGCATGTTAGATAA  * *GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
+r001/1           TTAGATAAAGGATA * CTG
+r002            aaaAGATAA* GGATA
+r003          gcctaAGCTAA
+r004                        ATAGCT.............................. TCAGC
-r003                                 ttagctTAGGC
-r001/2                                            CAGCGGCAT
```

The corresponding SAM format is:

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001     99  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    147  ref  37   30           9M     =   7 -39  CAGCGGCAT * NM:i:1
```

Mate sequence, location, insert size

SAM: Alignments

```
ref       AGCATGTTAGATAA  * *GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
+r001/1         TTAGATAAAGGATA * CTG
+r002        aaaAGATAA* GGATA
+r003       gcctaAGCTAA
+r004                   ATAGCT..............................  TCAGC
-r003                            ttagctTAGGC
-r001/2                                      CAGCGGCAT
```

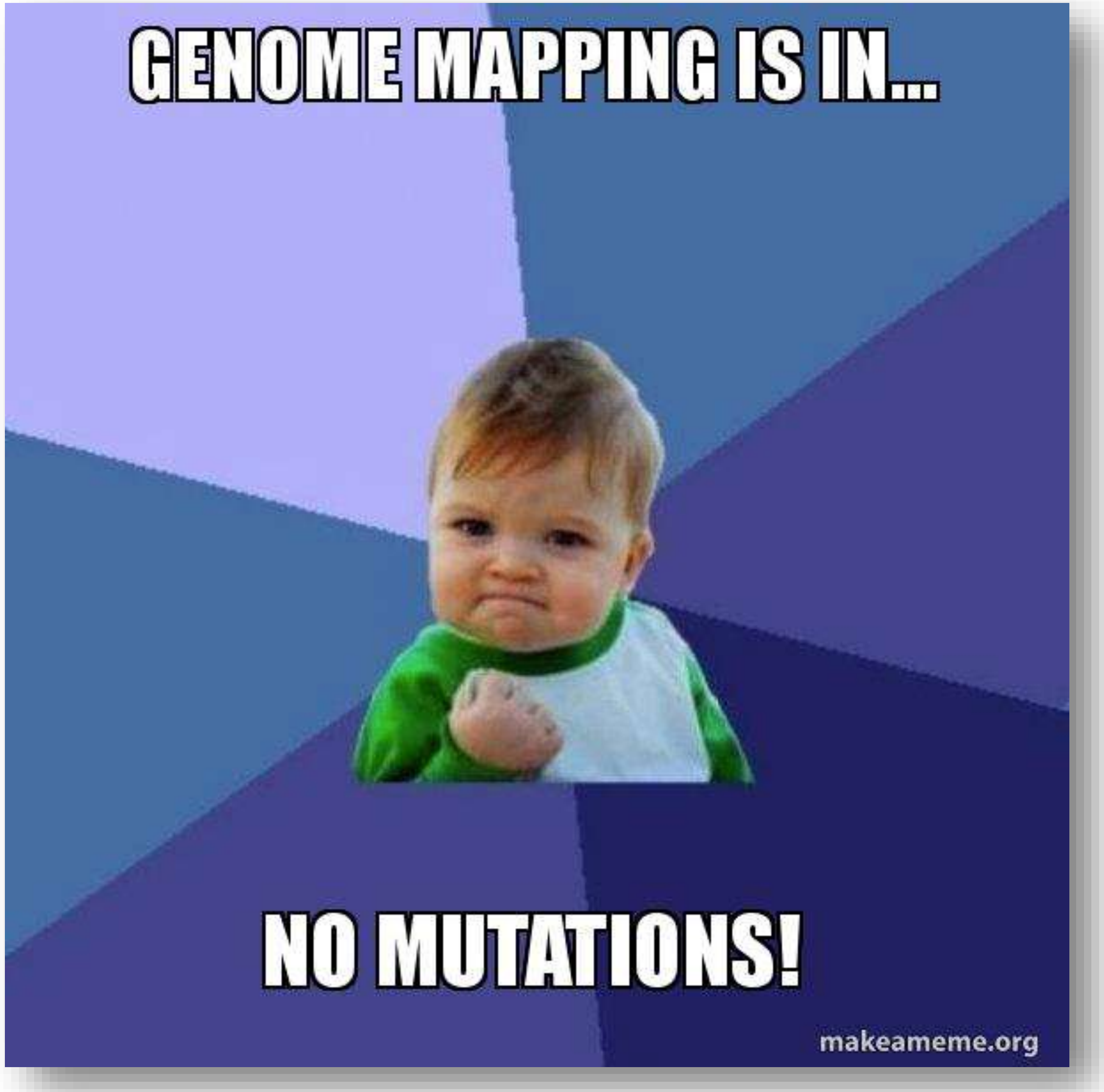
The corresponding SAM format is:

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001    99  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   147  ref  37  30          9M    =   7 -39  CAGCGGCAT * NM:i:1
```

Read sequence & quality (* = no quality stored)

Now for when you DON'T have a reference…

Mark Stenglein



DIY Exercise!!

- Report the:
  - Coverage = 7X
  - Error rate = 10%
  - How many variants (SNPs)? = 2? 3? – tim[i/e]s wa[s/k] ep[o/r]ch
  - Mapping rate (reads/sec)?
- Extra credit: Name the book and author
  - Tale of Two Cities

http://ivory.idyll.org/blog/the-assembly-exercise.html