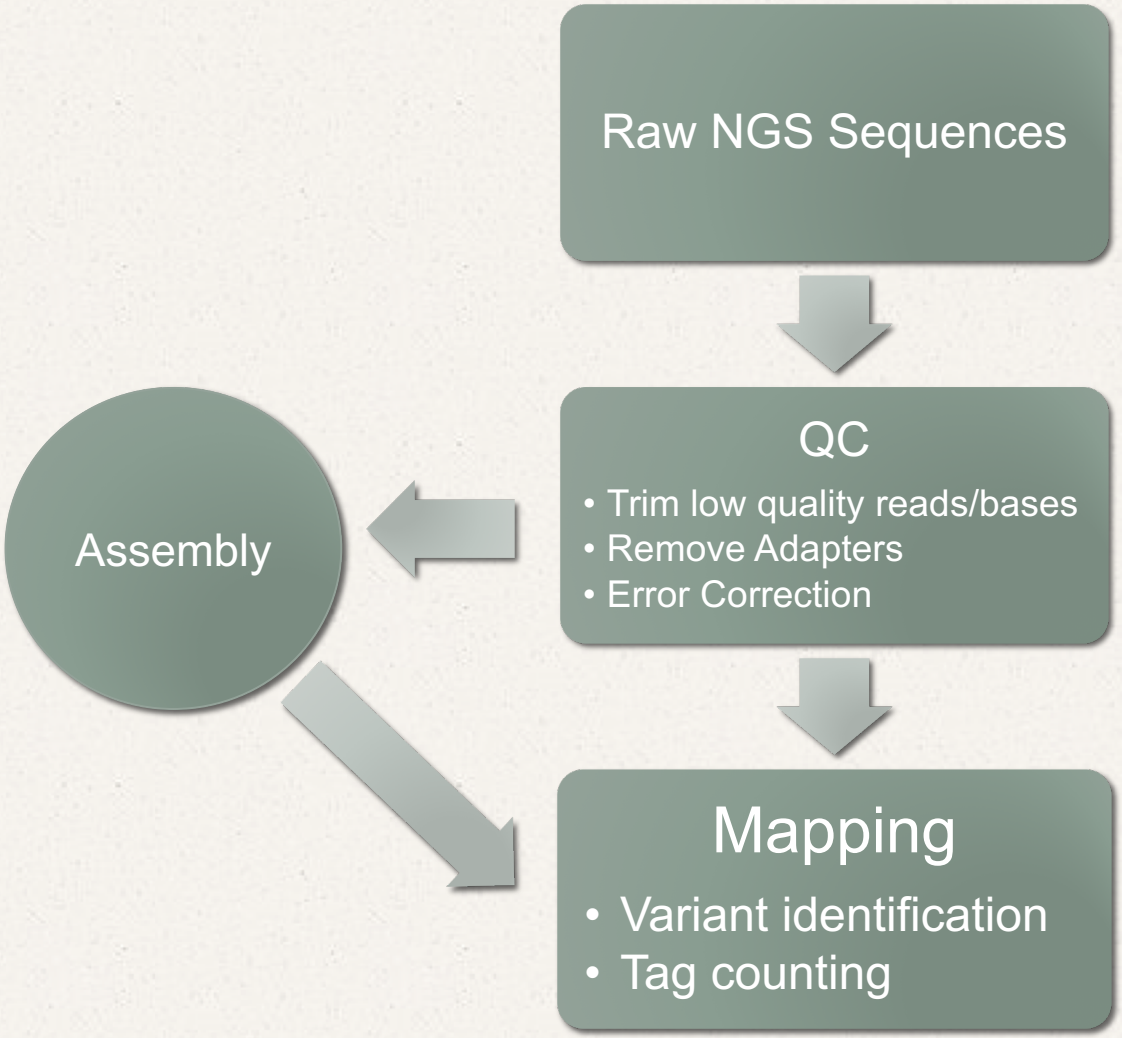
# MAPPING

Aligning sequencing reads to a reference

# Where are we?


Raw NGS Sequences
QC
• Trim low quality reads/bases
• Remove Adapters
• Error Correction
Assembly
Mapping
• Variant identification
• Tag counting

# DIY time!

- Reference:
- “We choose to go to the moon. We choose to go to the moon in this decade and do the other things, not because they are easy, but because they are hard” - JFK

- With a partner….. Map the reads!

- Things to consider:
  - Error rate
  - Coverage

# DIY time!

- Reference:
- “We choose to go to the moon. We choose to go to the moon in this decade and do the other things, not because they are easy, but because they are hard” - JFK

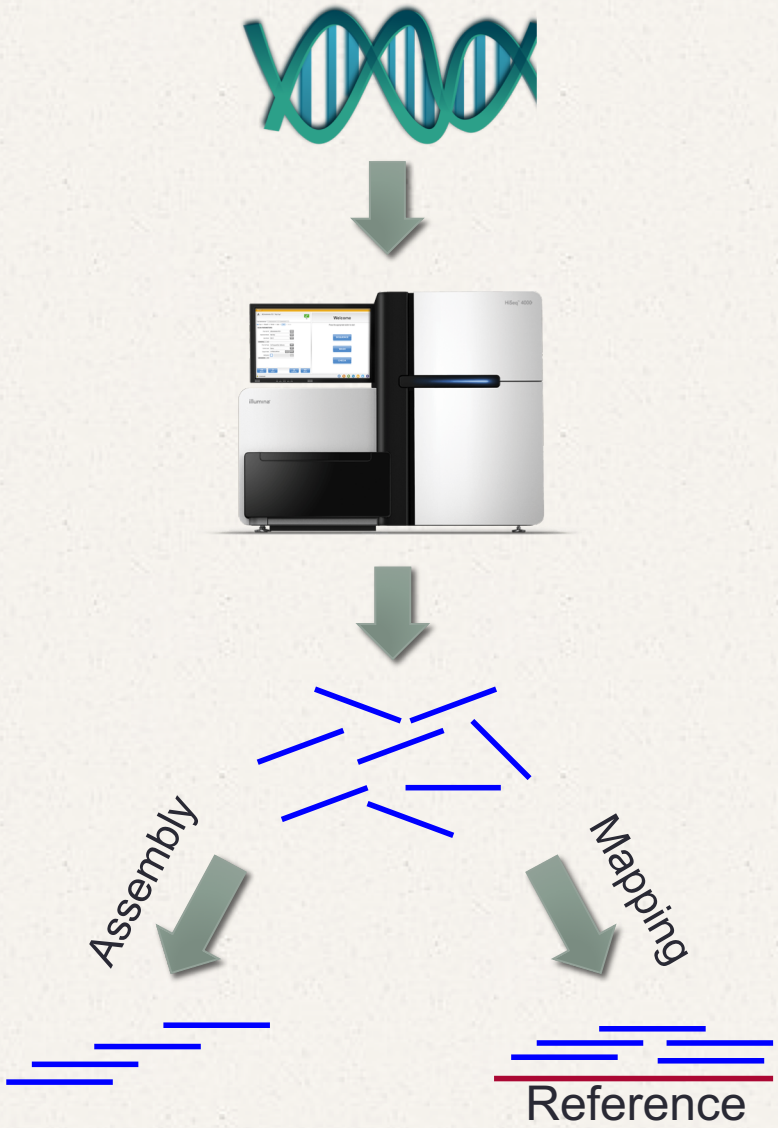
- With a partner….. Map the reads!

- Things to consider:
  - Error rate 20%
  - Coverage 5X

# Just a pairwise alignment, right?

Yes.
x 400 million (or more)

# Mapping

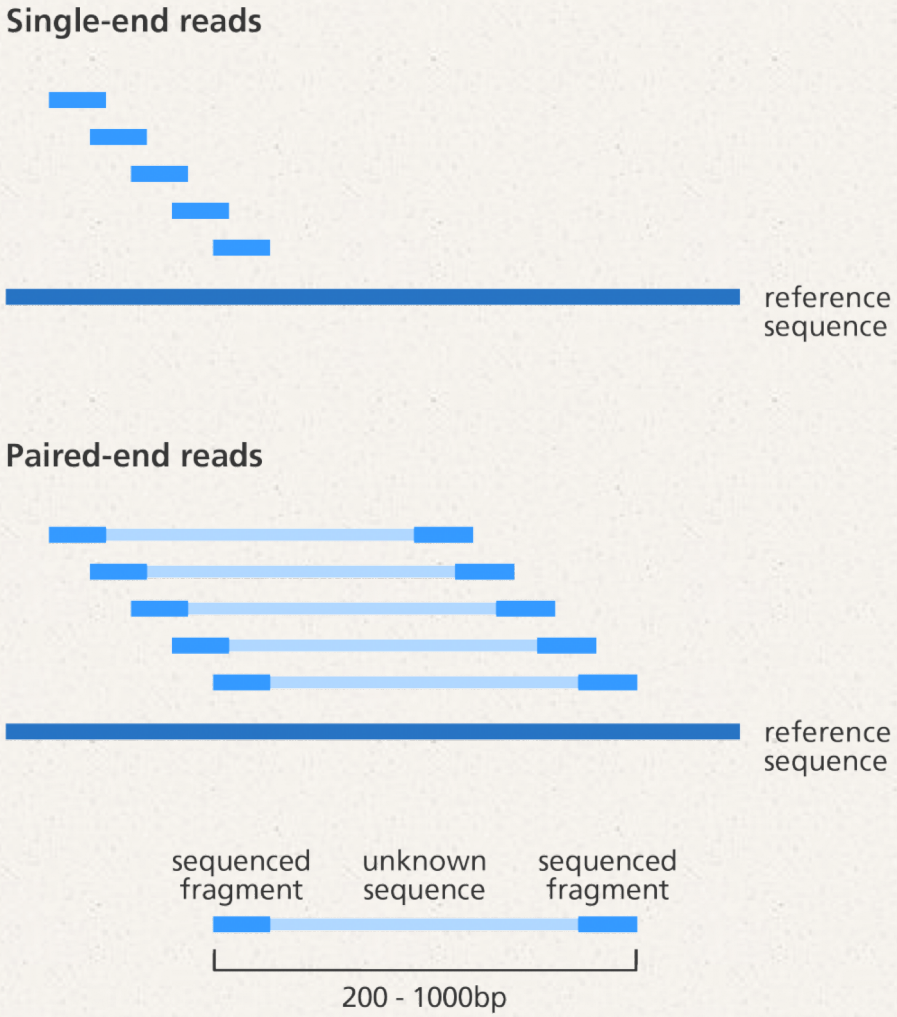



## Challenges

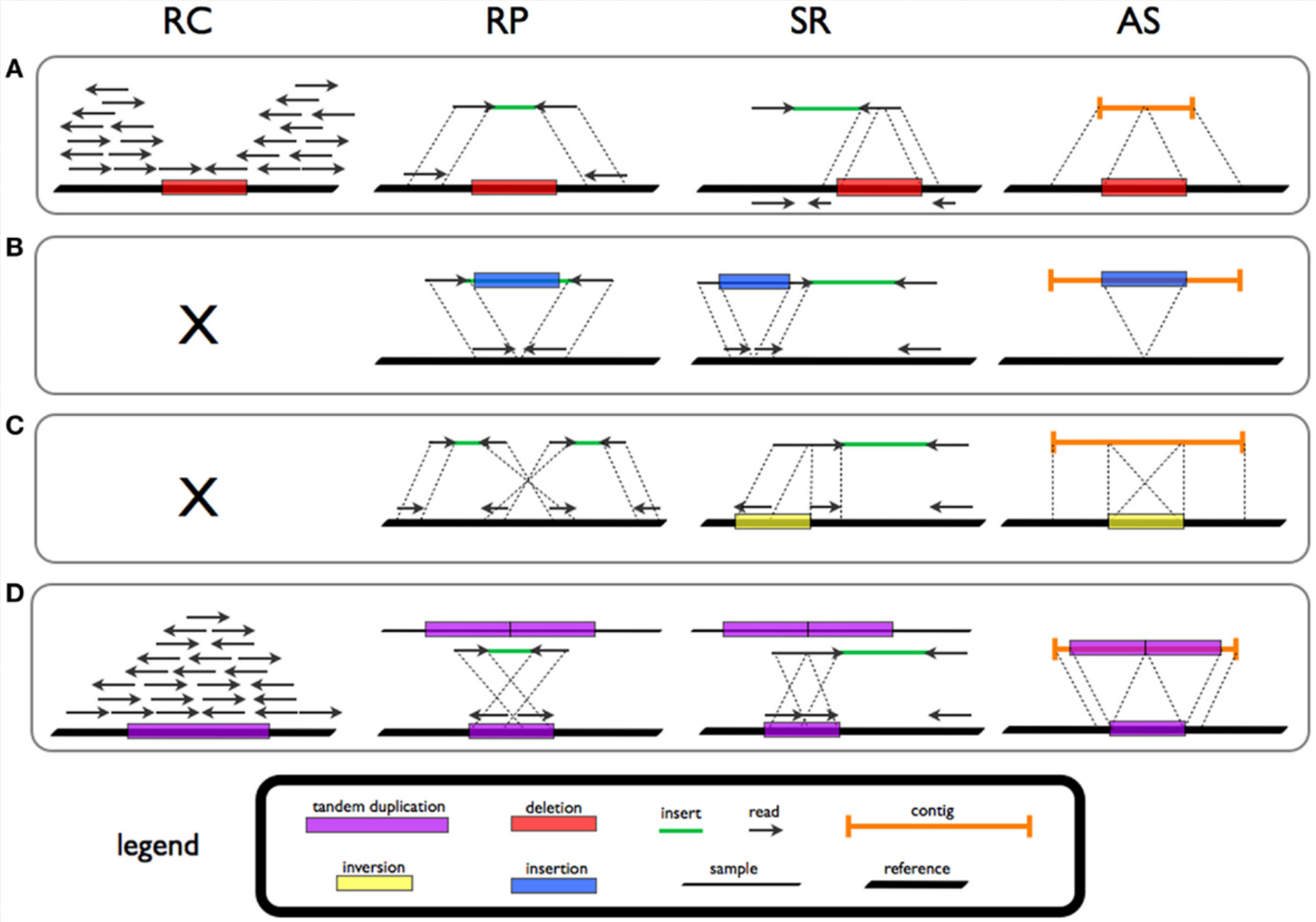
- Large numbers
- Short length
- Sequencing errors
- Repeats
- Indels
- Variants

# What is mapping?

- Which software
  - >70 published programs
  - Input data type
  - Reference
  - Speed vs sensitivity
  - Memory



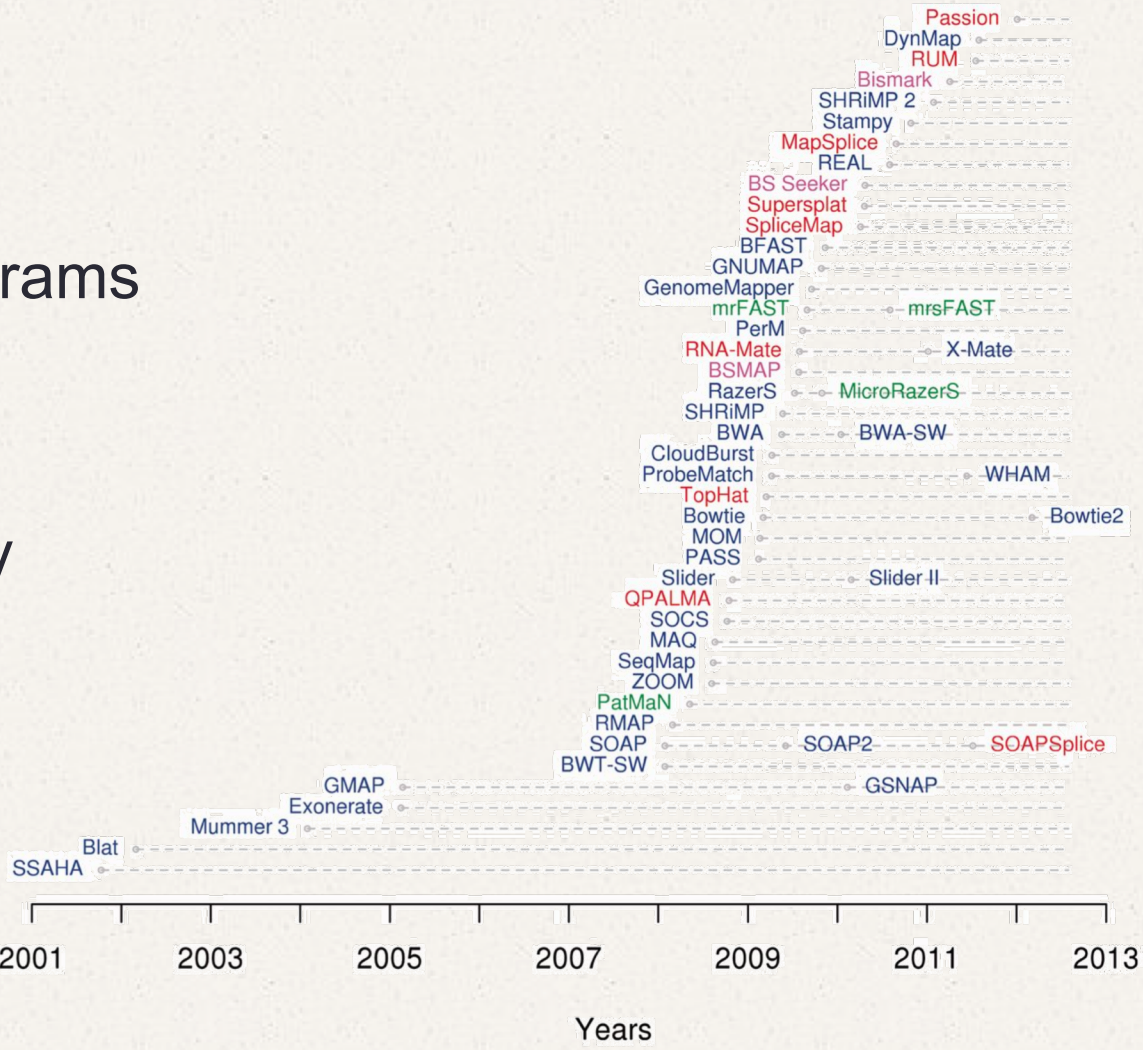

www.yourgenome.org




Tattini et al. (2015) Front. Bioeng. Biotechnol

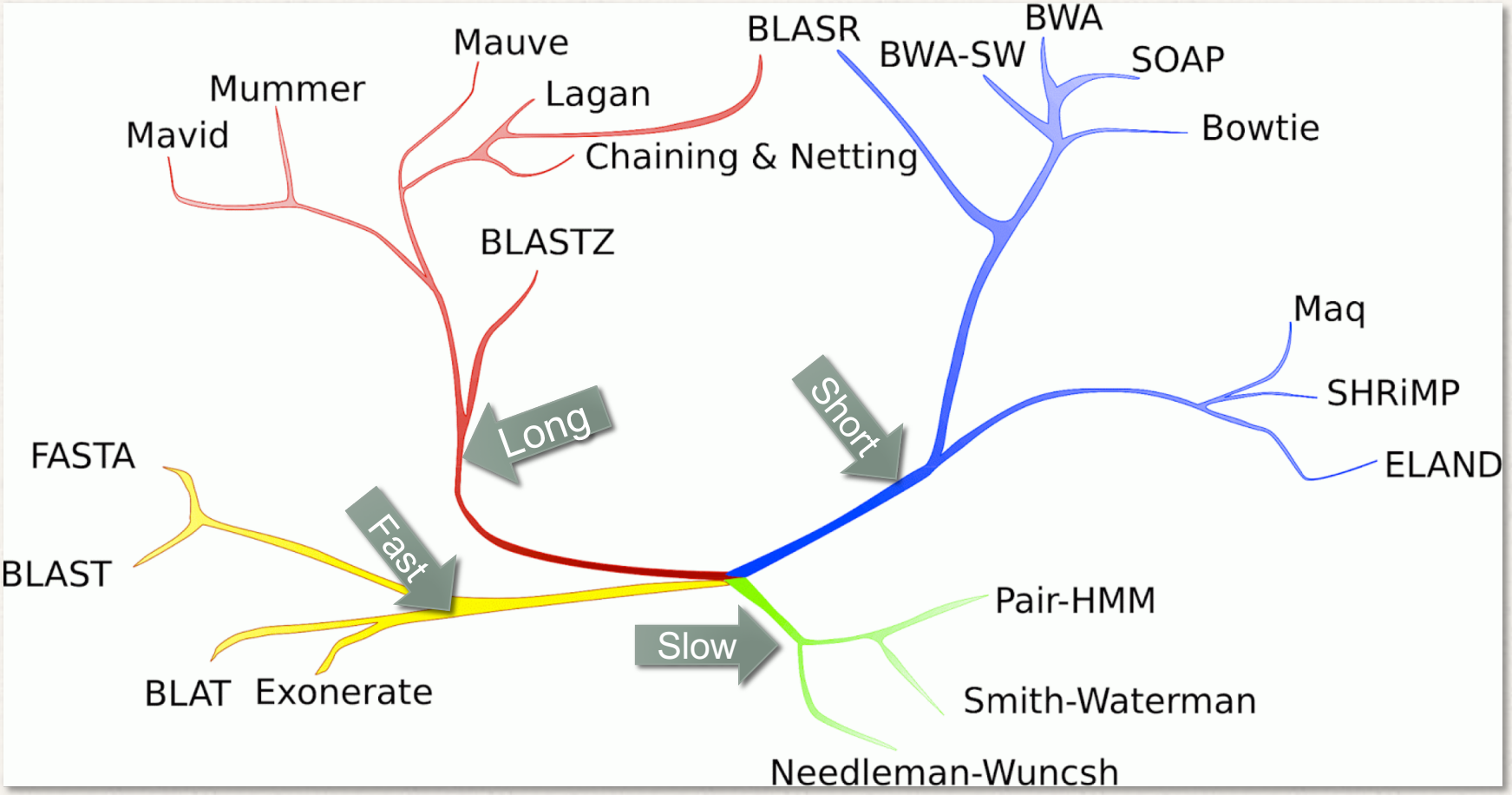
# What is mapping?

- Which software
  - >70 published programs
  - Input data type
  - Reference
  - Speed vs sensitivity
  - Memory




Fonseca et al. 2012, Bioinformatics

# The 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*

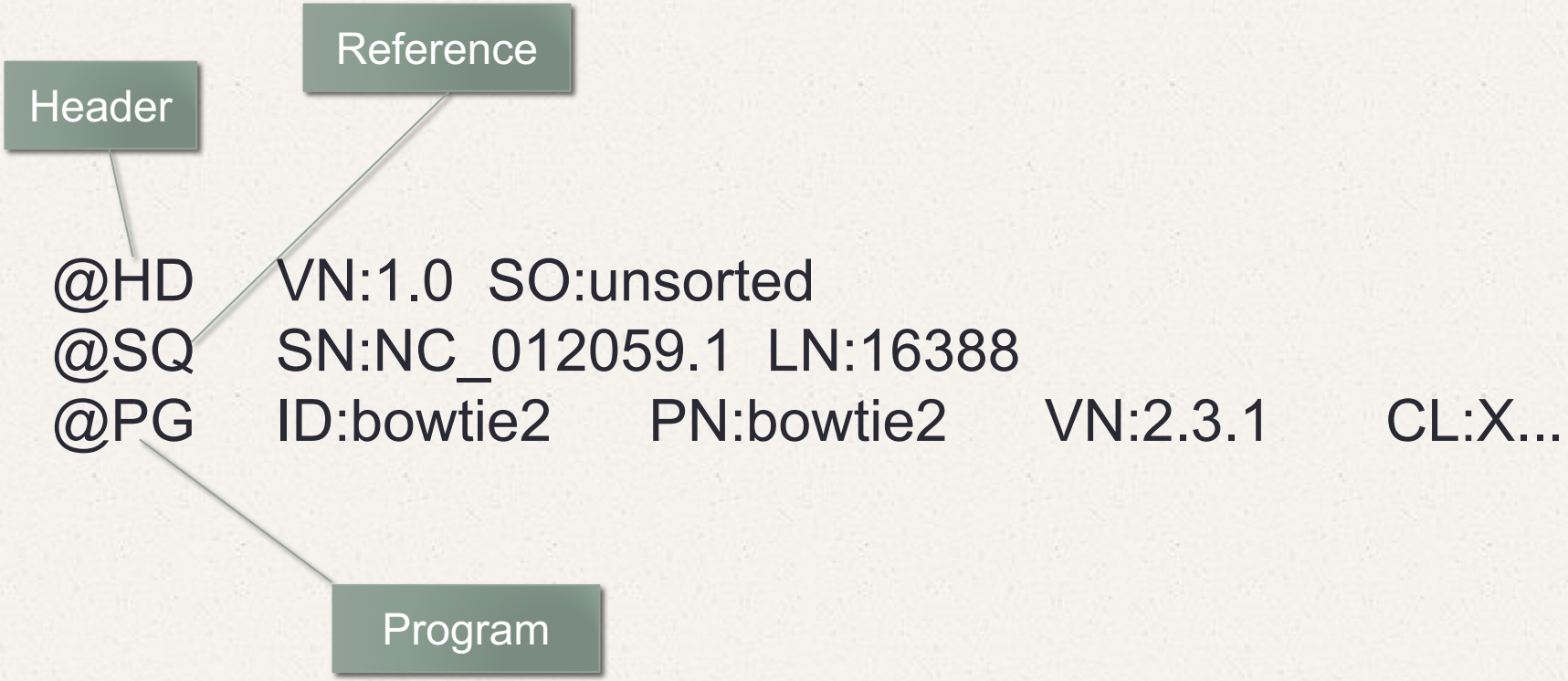
# STORING READ ALIGNMENTS

# 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




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: Alignment Records

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

# SAM: Alignment Records

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

# SAM: Alignment Records

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

# SAM: Alignment Records

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

# SAM: Alignment Records

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

# SAM: Alignment Records

```
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: Alignment Records

```
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: Alignment Records

```
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: Alignment Records

```
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: Alignment Records

```
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): $-10 * \log_{10}(\text{pr[wrongly mapped]})$

# SAM: Alignment Records

```
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

```
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

```
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

```
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

```
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

```
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

```
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

```
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

```
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: Alignment Records

```
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: Alignment Records

```
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)


GENOME MAPPING IS IN...
NO MUTATIONS!
makeameme.org

# NOW WHEN YOU DON'T HAVE A REFERENCE…

Mark Stenglein