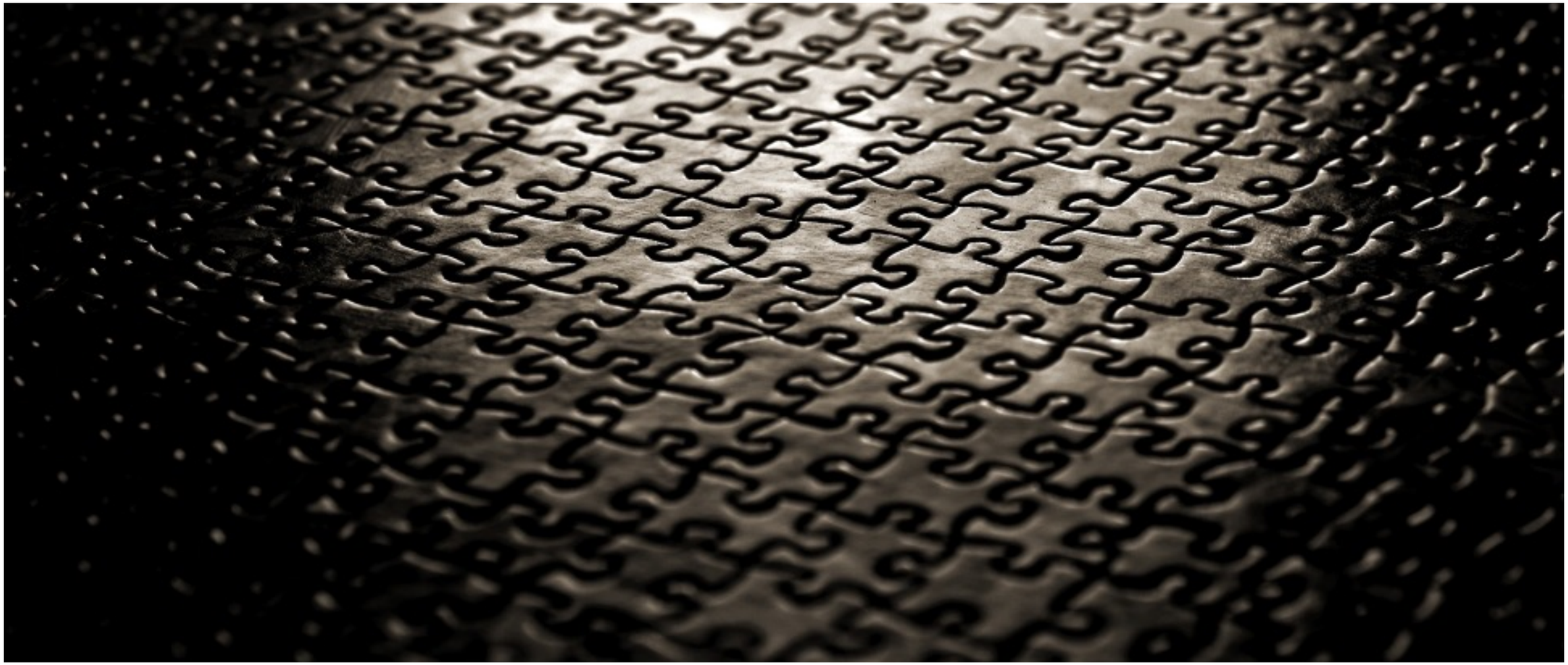


Aligning reads: tools and theory

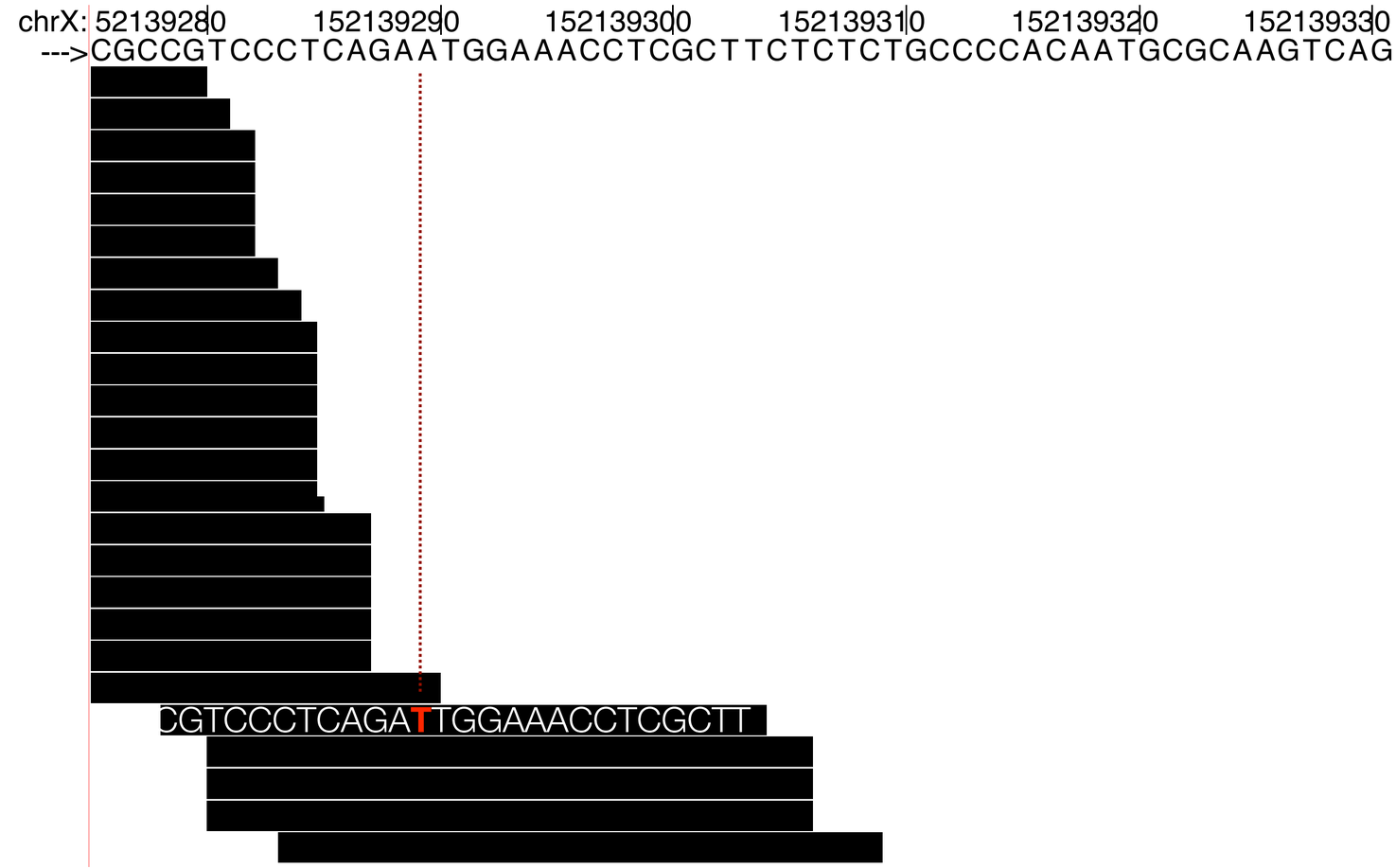
Genome

Sequence read

A simple case of string matching

Genome

Sequence reads



Difficult in practice

- Volume of data: ~3 Gbp
- ~50% of genome is **repeat regions** that cannot be covered by reads
  - Simple repeats, tandem, interspersed
  - Transposons
  - Segmental duplications where mapping is unclear
- **Gaps or unfinished regions**
  - peri-centromere, sub-telomere
  - ~5Mb unique to ethnic groups (e.g., African, Asian)

Challenges:
Human genome is large and complex

- Short reads: 50-150 bp (versus a very long reference)
  - Non-unique alignment
  - Sensitive to sequencing errors
- Massive amount of short reads: one lane produces ≥ 150 million 100 nucleotide reads
- Small insert size: 200-500 bp libraries

Challenges: short read NGS data

Reference	`ATCTCCATAGGACTAGAAGTAG`
Substitution	`ATCTCCATAGCACTAGAAGTAG`
Deletion	`ATCTCCATAGGAC-AGAAGTAG`
Insertion	`ATCTCCATAGGACTAGAAGTTAG`
3bp deletion	`ATCTC--AGGACTAGAAGTAG`

Challenges: non-exact matching

Local alignment vs Global alignment

- **Local alignment** matches the query with a *substring* (k-mer) of the reference
  - Tailored towards finding *regions of highly similar sequence* and aligning around those by working outwards to align the rest

Local Alignment

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
           |||| |||||| ||||||||||||||||
      5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

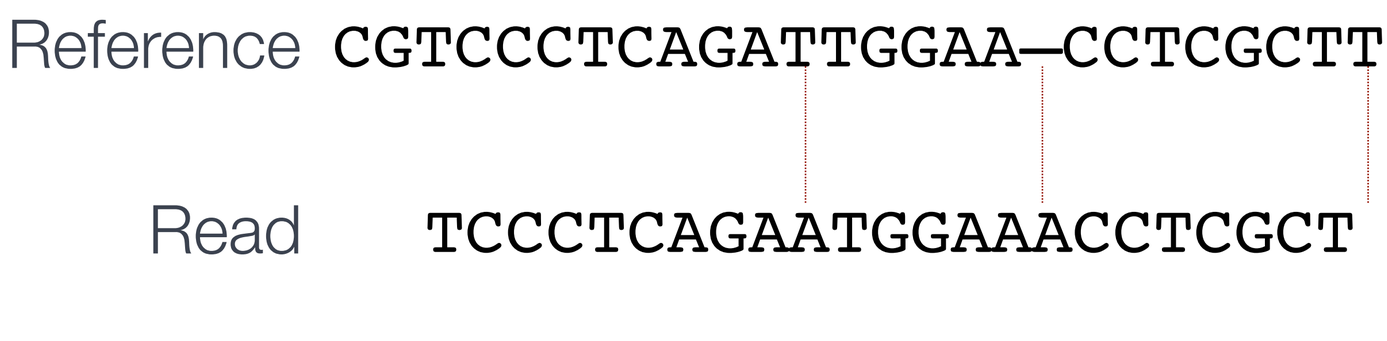
Global Alignment

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   ||||||||||    |||||||  |||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

- A **global alignment** performs end-to-end alignment between the query and the reference

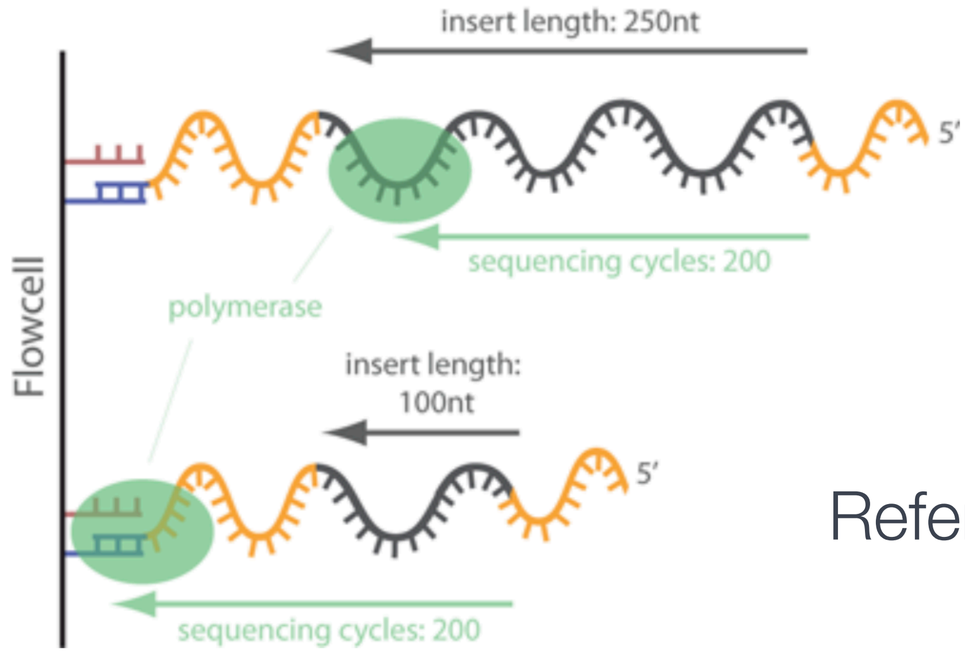
Reference	`ATCTCCATAGGACTAGAAGTAG`
Substitution	`ATCTCCATAGCACTAGAAGTAG`
Deletion	`ATCTCCATAGGAC-AGAAGTAG`
Insertion	`ATCTCCATAGGACTAGAAGTTAG`
3bp deletion	`ATCTC--AGGACTAGAAGTAG`

General concepts: edit distance



Edit distance =3

General concepts: edit distance



Reference `CGTCCCTCAGATTGGAAACCTCGCTT`

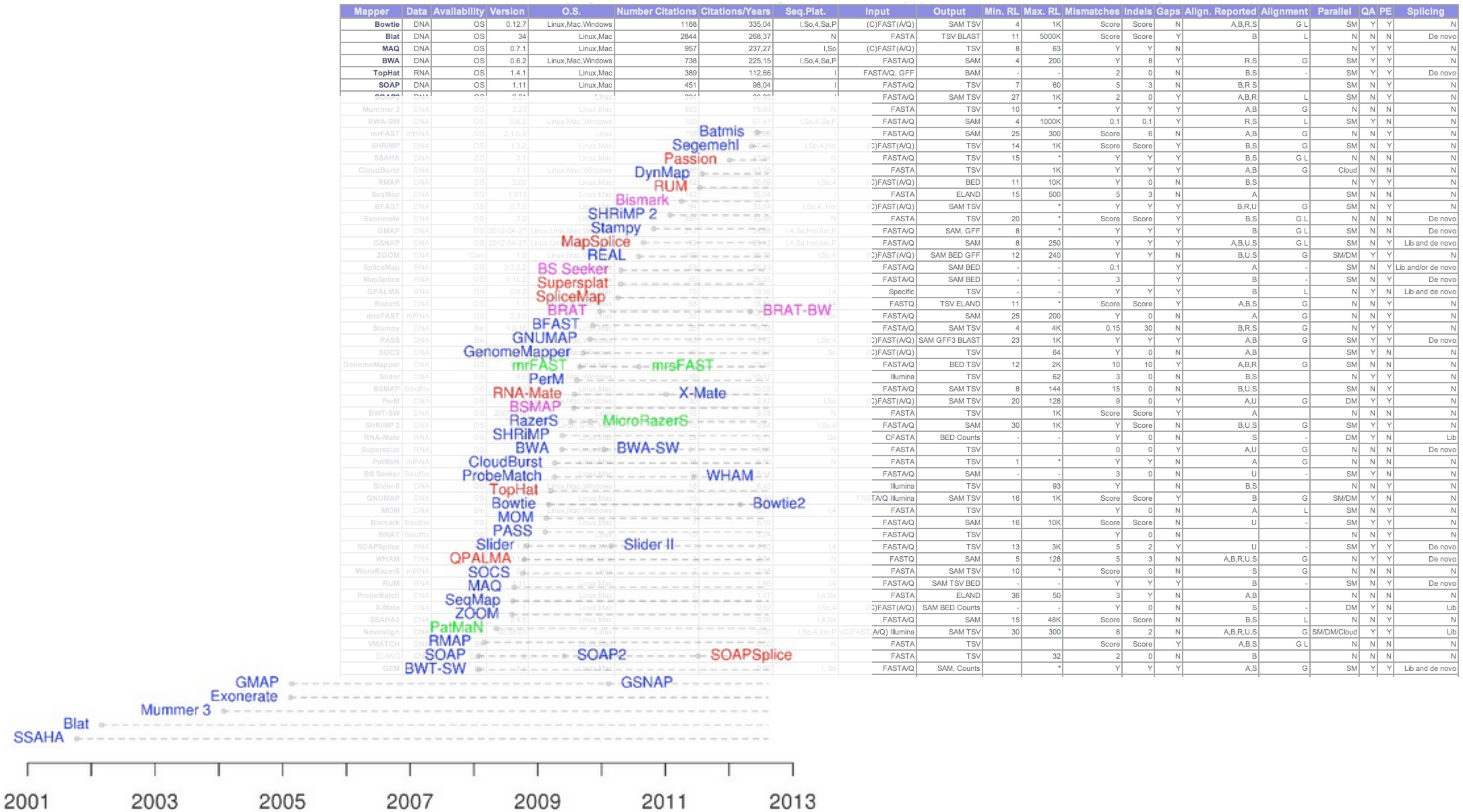
Read `AGGCTACGATTGGAAACCTCGCTT`

General concepts: soft-clipping

Image source: ECSEQ Bioinformatics

Soft-clipping of reads

- Portions of the read that do not match well to the reference genome on either side of the reads are ignored for the alignment
- The procedure can carry a small penalty for each soft-clipped base, but amounts to a significantly smaller penalty than mismatching bases
- Soft-clipped bases are retained in the sequence and simply marked. Trimming methods hard-clip, which deletes the unwanted sequence.



Short-read aligners: choices

http://wwwdev.ebi.ac.uk/fg/hts_mappers/

Building an index

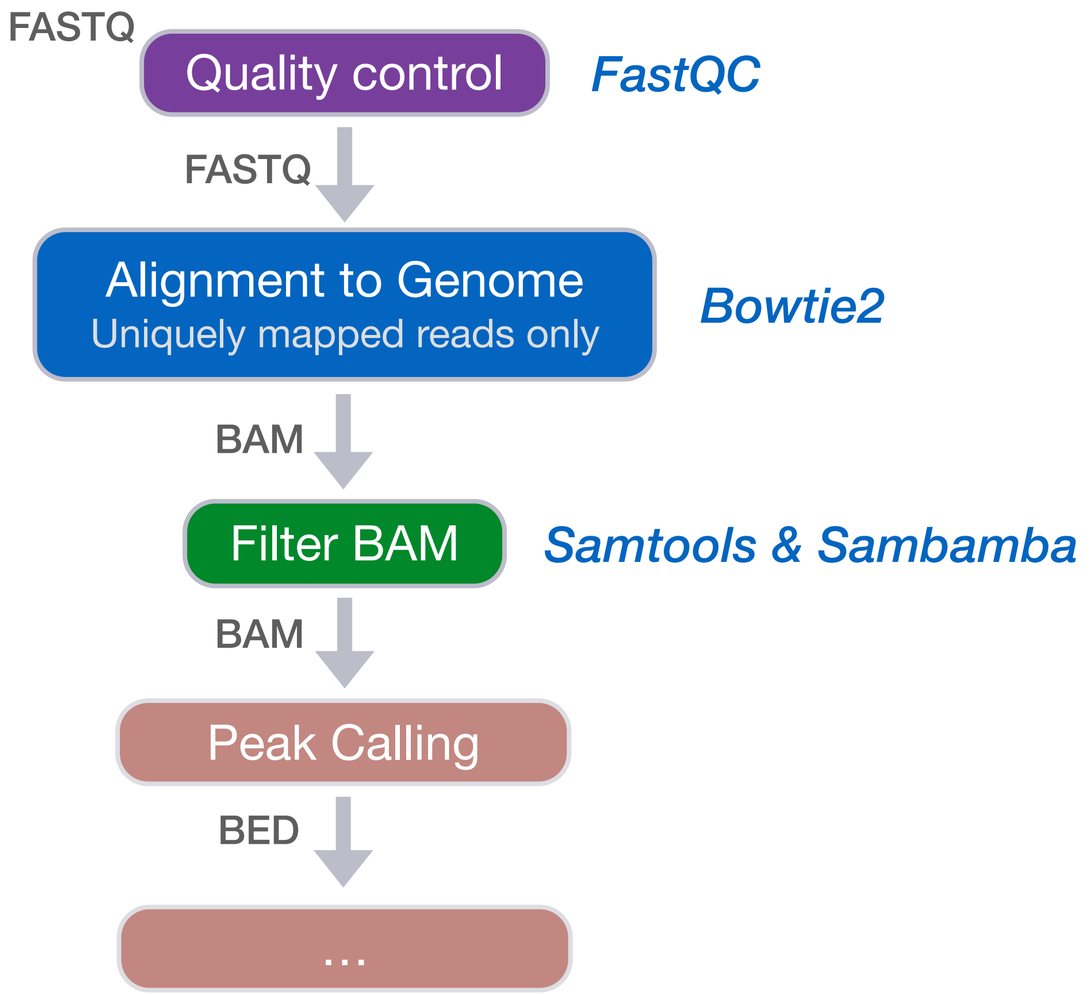
- Having an index of the reference genome provides an efficient way to search
- Once index is built, it can be queried any number of times
- Indexes are genome and tool-specific
- Different types of indices (i.e hash-tables, suffix arrays, Burrows-Wheeler Transform)

- Bowtie2: indexes with an FM Index to keep low memory footprint. Supports gapped, local and paired-end alignment
- BWA: indexes with the Burrows-Wheeler Transform (BWT). Has three algorithms for varying read lengths.
- SOAP2: uses a 2-way BWT for indexing. Fast and accurate alignment of Illumina sequencing reads. Not open source.
- MAQ: first aligns reads to reference sequences and then calls the consensus. Designed for Illumina reads.

Commonly used aligners for ChIP-seq

Alignment considerations for ChIP-seq

- Percentage of mapped reads
    - 75% or higher is good; Percentages vary between organisms
- Duplicates
    - Usually due to over-amplification or short read length
    - ENCODE guidelines suggest < 20% duplication rate with paired-end data; can be more lenient with single-end
- Multi-mappers
    - Reduce by setting allowable mismatches according to sequencing platform
    - For situations when protein binds repetitive DNA, use paired-end sequencing



ChIP-seq: Alignment and Filtering

These materials have been developed by members of the teaching team at the Harvard Chan Bioinformatics Core (HBC). These are open access materials distributed under the terms of the Creative Commons Attribution license (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.