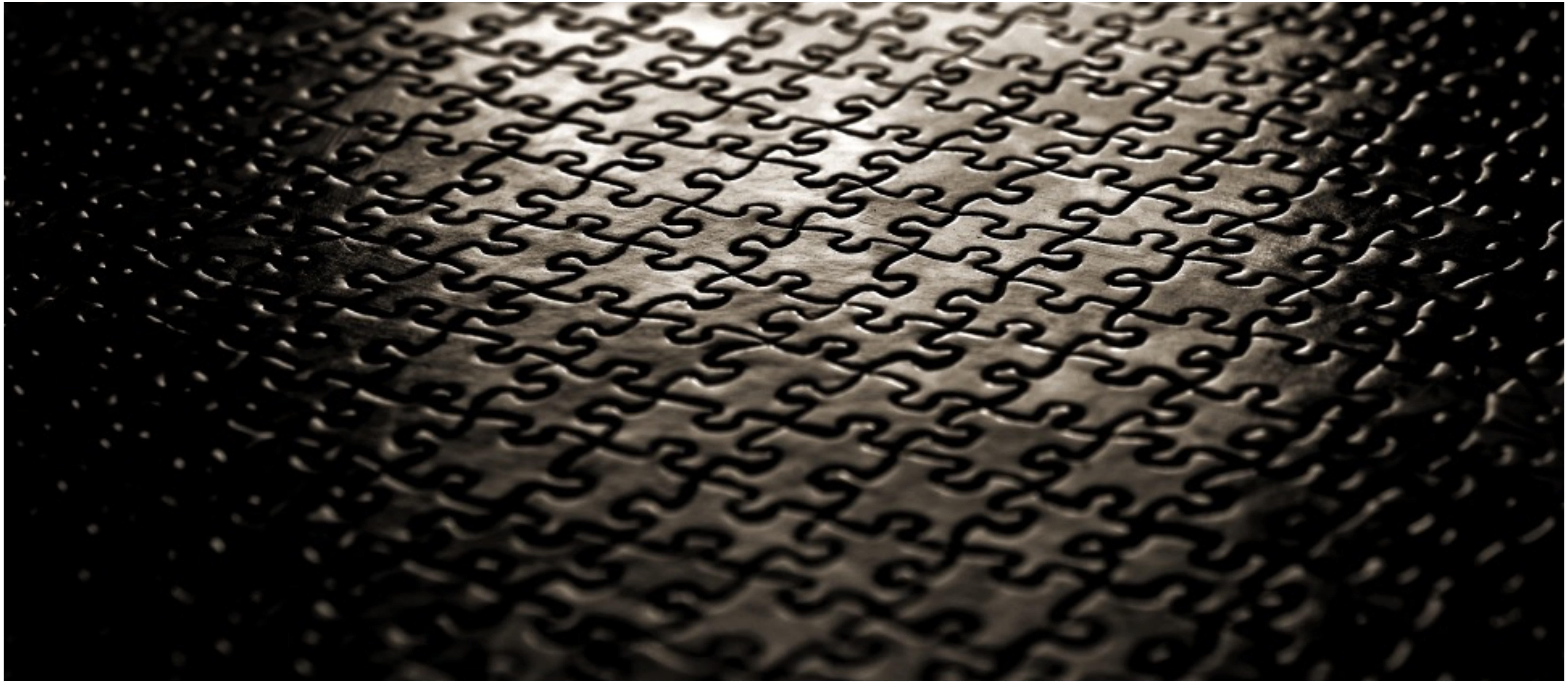
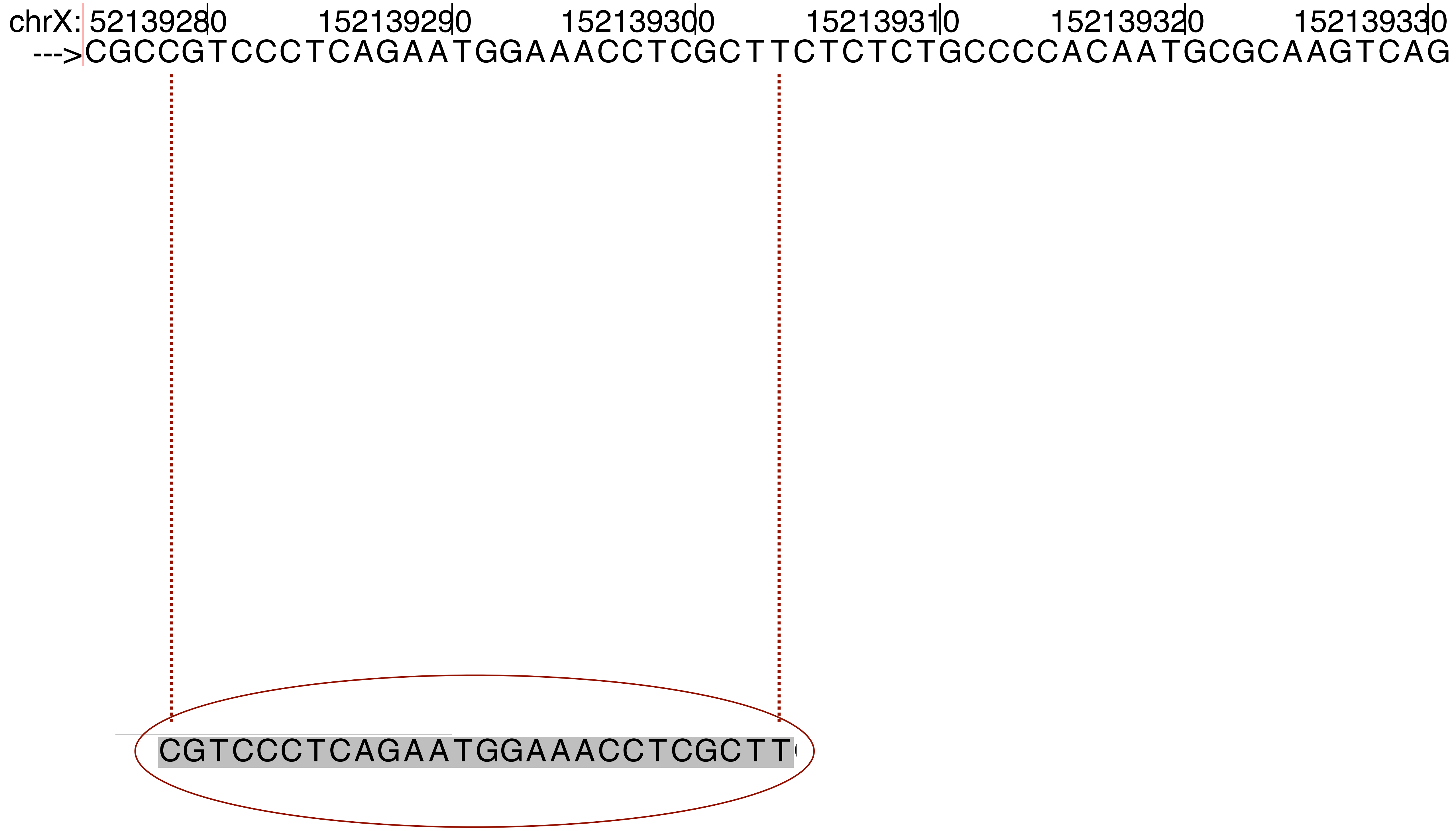


Aligning reads: tools and theory

Genome



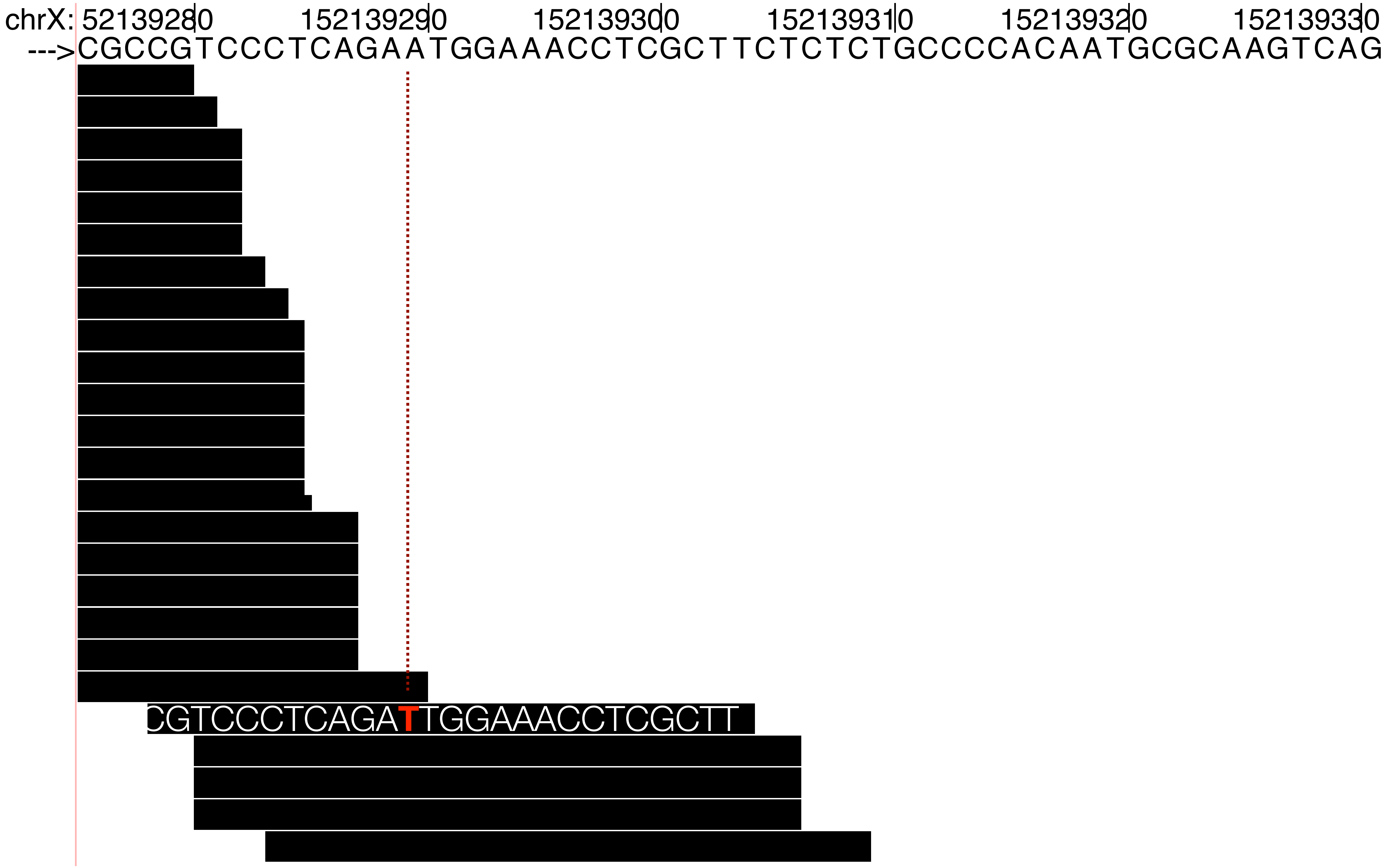
```
chrX: 52139280        152139290        152139300        152139310        152139320        152139330
 --->CGCCGTCCCTCAGAATGGAAACCTCGCTTCTCTCTGCCCCACAATGCGCAAGTCAG
```

Sequence read

```
CGTCCCTCAGAATGGAAACCTCGCTT
```

A simple case of string matching

Genome


chrX: 52139280 152139290 152139300 152139310 152139320 152139330
--->CGCCGTCCCTCAGAATGGAAACCTCGCTTCTCTCTGCCCCACAATGCGCAAGTCAG
CGTCCCTCAGATTGGAAACCTCGCTT

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
- Gap or unfinished regions
  - – peri-centromere, sub-telomere
  - – ~5Mb unique to ethnic groups (e.g., African, Asian)
- Finishing errors(1/10,000bp), miscalled base incorporated

Challenges: Human genome is large and complex

- Genome Reference Consortium: “…working to create assemblies that better represent diversity and provide more robust substrates for genome analysis.”
  - novel assembly algorithm
  - correcting assembly errors (fix patches)
  - addition of new alternate loci (patches)
  - filling in gaps

EMBL-EBI

NCBI

wellcome trust sanger institute

THE GENOME INSTITUTE at Washington University

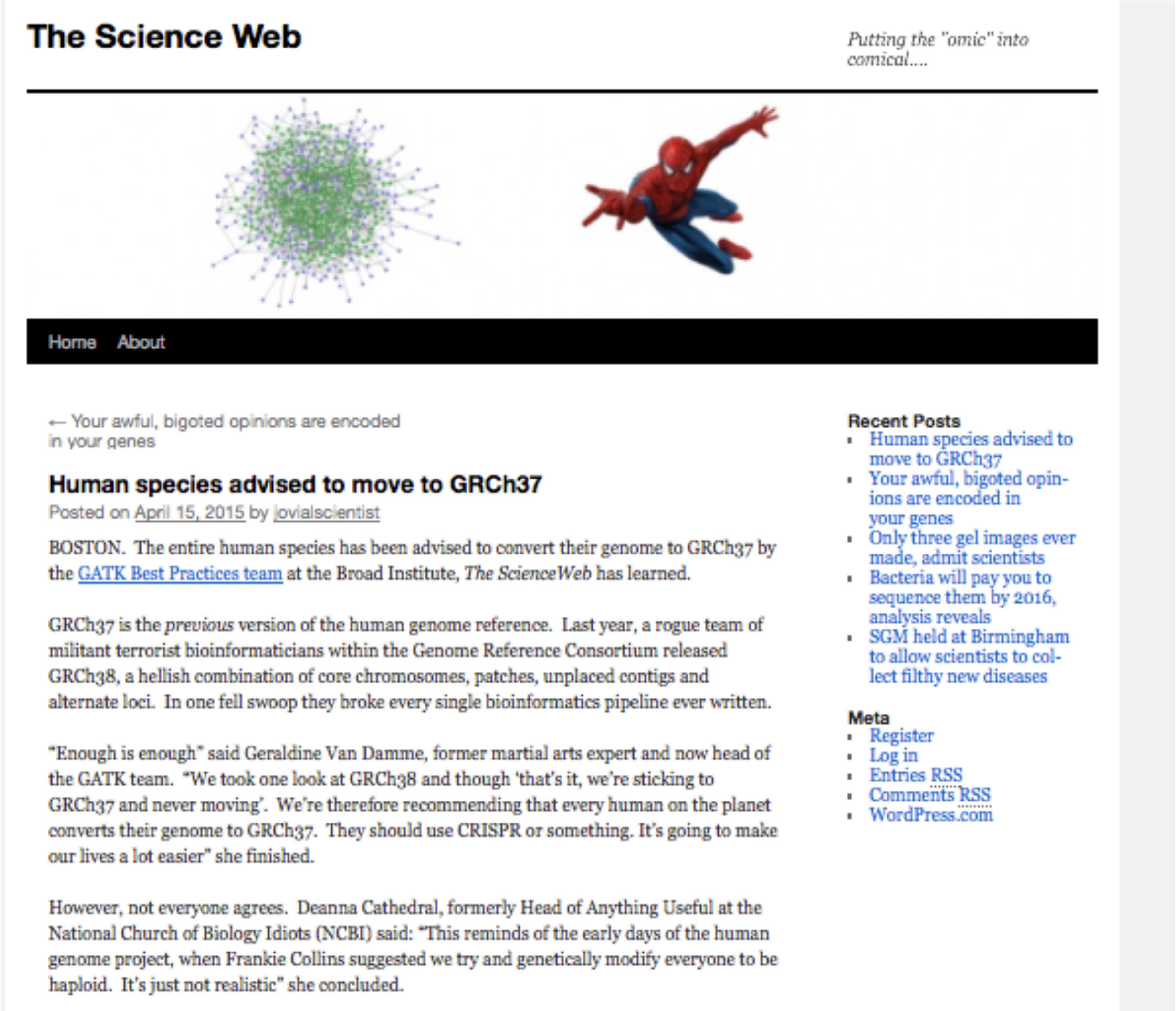
Challenges:
Genome is continuously changing

- Ensembl, UCSC and NCBI all use the same genome assemblies or builds provided by the GRC (i.e GrCh37 = hg19)
- Patches are provided by the GRC, but incorporated as updates by each database at different intervals
- At any point in time, the sequence can vary between databases but coordinates are unchanged
- **Always use the same biological database for all reference data!**

Challenges:
Sources of genome reference sequence

GRCh37 vs. GRCh38

- Closing the gaps; more complete genome information
- 8000 nucleotides altered
- Several misassembled regions corrected
- 261 alternate loci across 178 regions (improved diversity)
- Sequence information for centromeres



The Science Web

Putting the "omic" into comical....

Home About

← Your awful, bigoted opinions are encoded in your genes

Human species advised to move to GRCh37

Posted on April 15, 2015 by jovialscientist

BOSTON. The entire human species has been advised to convert their genome to GRCh37 by the GATK Best Practices team at the Broad Institute, *The ScienceWeb* has learned.

GRCh37 is the *previous* version of the human genome reference. Last year, a rogue team of militant terrorist bioinformaticians within the Genome Reference Consortium released GRCh38, a hellish combination of core chromosomes, patches, unplaced contigs and alternate loci. In one fell swoop they broke every single bioinformatics pipeline ever written.

"Enough is enough" said Geraldine Van Damme, former martial arts expert and now head of the GATK team. "We took one look at GRCh38 and though 'that's it, we're sticking to GRCh37 and never moving'. We're therefore recommending that every human on the planet converts their genome to GRCh37. They should use CRISPR or something. It's going to make our lives a lot easier" she finished.

However, not everyone agrees. Deanna Cathedral, formerly Head of Anything Useful at the National Church of Biology Idiots (NCBI) said: "This reminds of the early days of the human genome project, when Frankie Collins suggested we try and genetically modify everyone to be haploid. It's just not realistic" she concluded.

Recent Posts

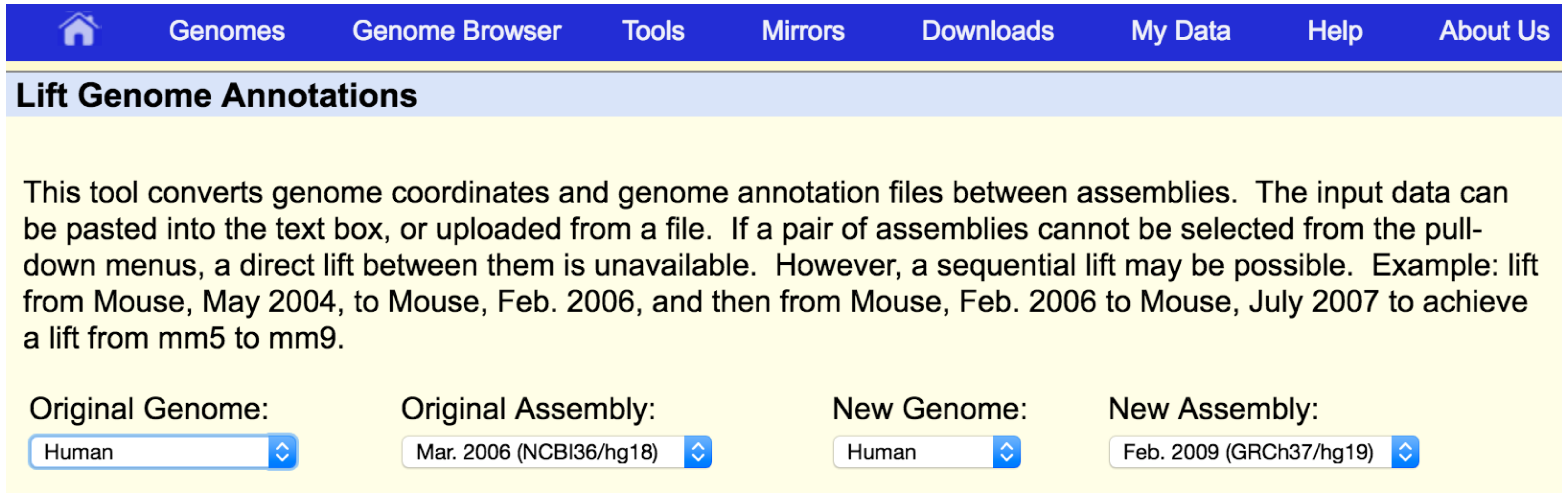
- Human species advised to move to GRCh37
- Your awful, bigoted opinions are encoded in your genes
- Only three gel images ever made, admit scientists
- Bacteria will pay you to sequence them by 2016, analysis reveals
- SGM held at Birmingham to allow scientists to collect filthy new diseases

Meta

- Register
- Log in
- Entries RSS
- Comments RSS
- WordPress.com

LiftOver at UCSC

You can obtain corresponding coordinates of a different genome build, if you have a set of coordinates from a known build using the **LiftOver tool (UCSC)**



Lift Genome Annotations

This tool converts genome coordinates and genome annotation files between assemblies. The input data can be pasted into the text box, or uploaded from a file. If a pair of assemblies cannot be selected from the pull-down menus, a direct lift between them is unavailable. However, a sequential lift may be possible. Example: lift from Mouse, May 2004, to Mouse, Feb. 2006, and then from Mouse, Feb. 2006 to Mouse, July 2007 to achieve a lift from mm5 to mm9.

Original Genome:	Original Assembly:	New Genome:	New Assembly:
Human	Mar. 2006 (NCBI36/hg18)	Human	Feb. 2009 (GRCh37/hg19)

Challenges: short read NGS data

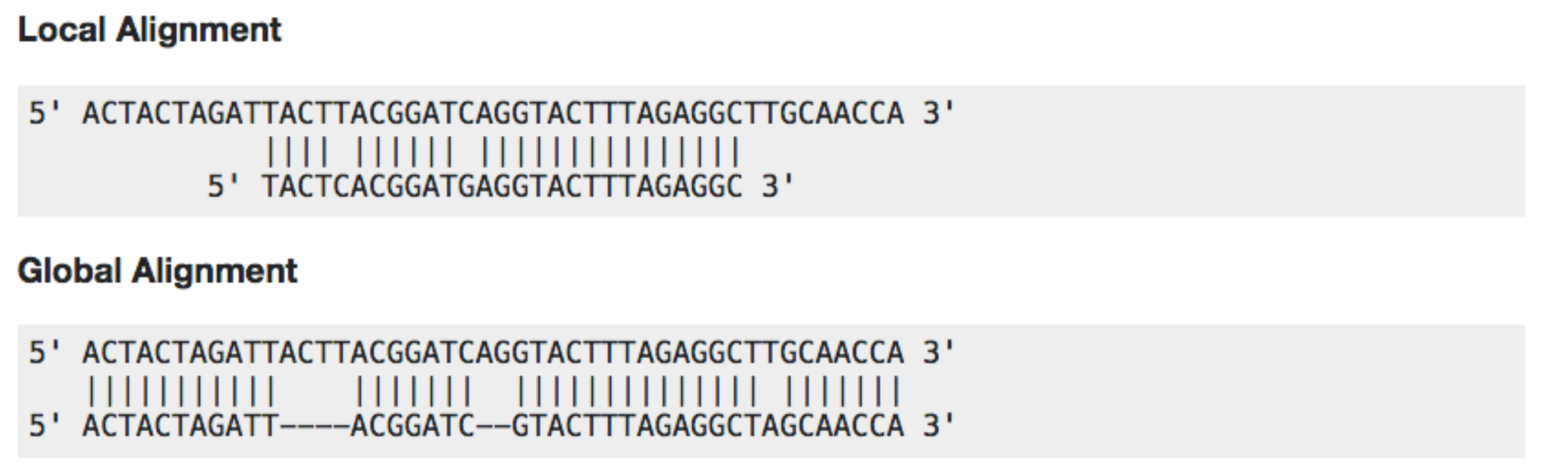
- Short reads: 50-150 bp (versus a very long reference)
  - Non-unique alignment
  - Sensitive to sequencing errors
- Massive amount of short reads: one lane produces ≥ 150 million 100 nucleotide reads
- Small insert size: 200-500 bp libraries

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'
```

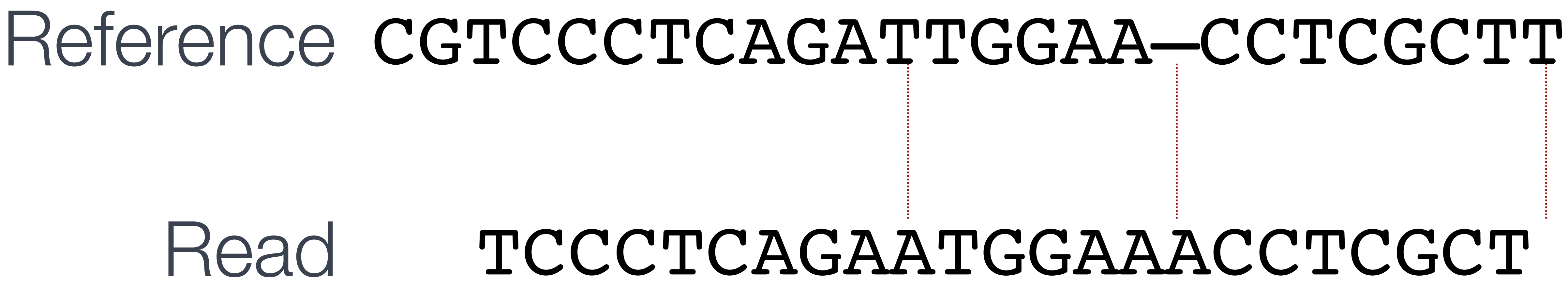
Global Alignment

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

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

General concepts: edit distance

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



```
Reference  CGTCCCTCAGATTGGAA—CCTCGCTT

Read         TCCCTCAGAATGGAAACCTCGCT
```

Edit distance =3

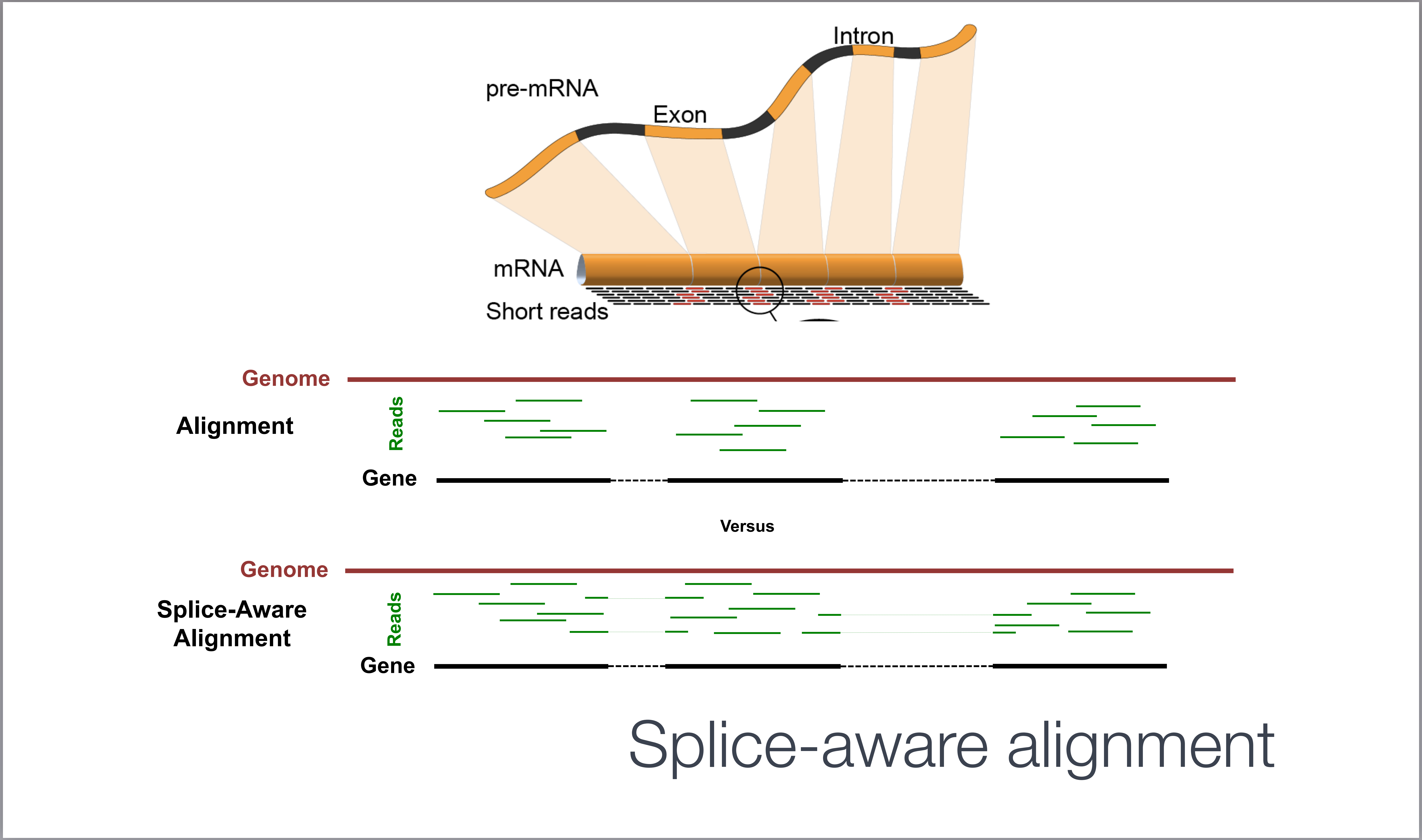
General concepts: edit distance

Building an index

- For each read we need to scan the entire corpus as fast as possible
- 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

Alignment tools can be grouped based on indexing method

- Some examples include:
    - Hash-based
    - Suffix arrays
    - Burrows-Wheeler Transform



Splice-aware alignment

Splice-aware alignment tools:

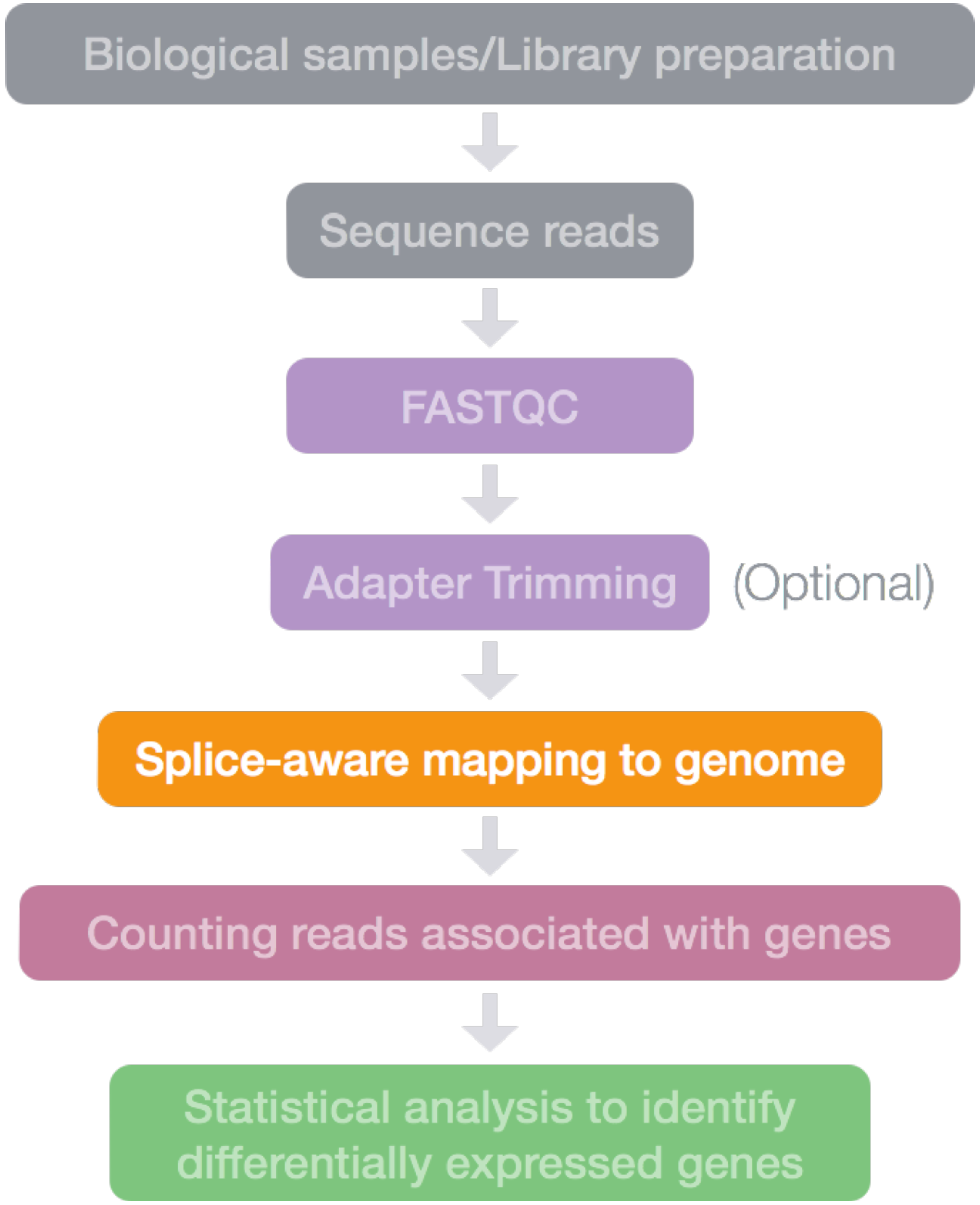
HISAT2, STAR, MapSplice, SOAPSplice, Passion, SpliceMap, RUM, ABMapper, CRAC, GMAP-GSNAP, HMMSplicer, Olego, BLAT

There are excellent aligners available that are not splice-aware. These are useful for aligning directly to genes. However, you will lose isoform information.

Bowtie2, BWA, Novoalign (not free), SOAPaligner

Splice-aware alignment

- Use the genome and GTF from the same source (i.e. Ensembl, NCBI, UCSC)
- Choose an aligner that can allow for a read to be "split" across distant regions to account for splice events
- Evaluate your computational resources and use an aligner that would work best within the confines of the available memory and CPU



Biological samples/Library preparation

Sequence reads

FASTQC

Adapter Trimming (Optional)

Splice-aware mapping to genome

Counting reads associated with genes

Statistical analysis to identify differentially expressed genes

Alignment for RNA-seq