# Short read alignment (using external tools)

Kasper Daniel Hansen <khansen@jhsph.edu>

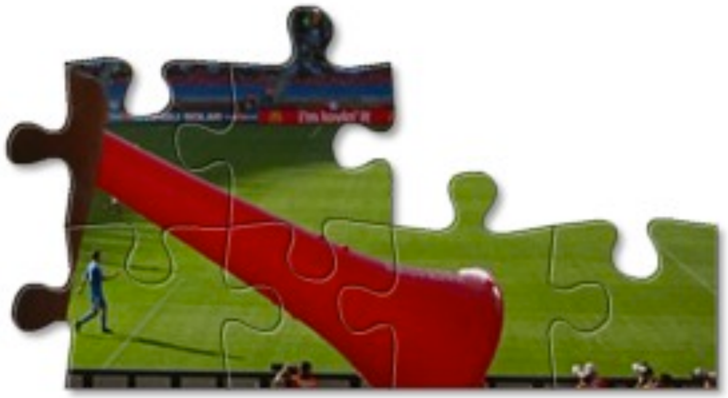
Brixen, 26 June 2011

Many slides are courtesy of
Hector Corrada Bravo and Ben Langmead

# Analyzing reads

```
TATGTCGCAGTATCTT  TATGTCGCAGTATCTG
TATGTCGCAGTATCTT  TATGTCGCAGTATCTG  GTCGCAGTATCTGTCT
GTCGCAGTATCTGTCT
CCGGACACCCTATAT  TATGTCGCAGTATCTT  GTCGCAGTATCTGTCT  GTCGCAGTATCTGTNN
ACACCCTATGTCGCA
GTCGCAGTATCTGTNN
TATGTCGCAGTATCTT  GTCGCAGTATCTGTNN
TATGTCGCAGTATCTG
ACACCCTATGTCGCA  ACACCCTATGTCGCA
CCGGACACCCTATAT
GTCGCAGTATCTGTNN  TATGTCGCAGTATCTG
CCGGACACCCTATAT
CCGGACACCCTATAT  TGTCGCAGTATCTGTC
GTCGCAGTATCTGTNN
TGTCGCAGTATCTGTC
```

↓

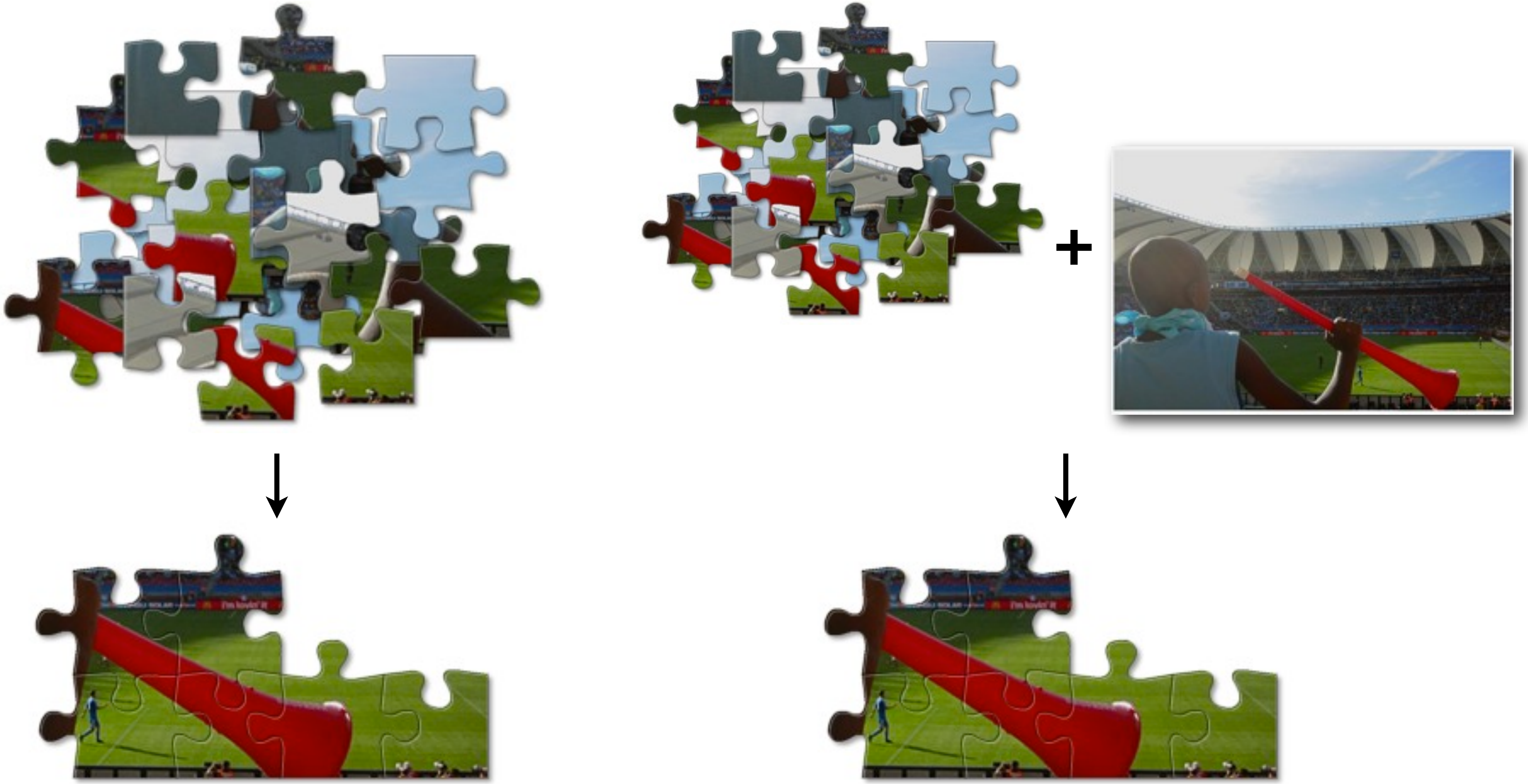


↓

```
CCGGACACCCTATAT
     ||||||||| |
     ACACCCTATGTCGCA
               |||||||
               TGTCGCAGTATCTGTC
                         |||  |||
                         TAT--GTCGCAGTATCTG
```

Image source: http://ngm.nationalgeographic.com/your-shot/jigsaw-puzzles

# Analyzing reads

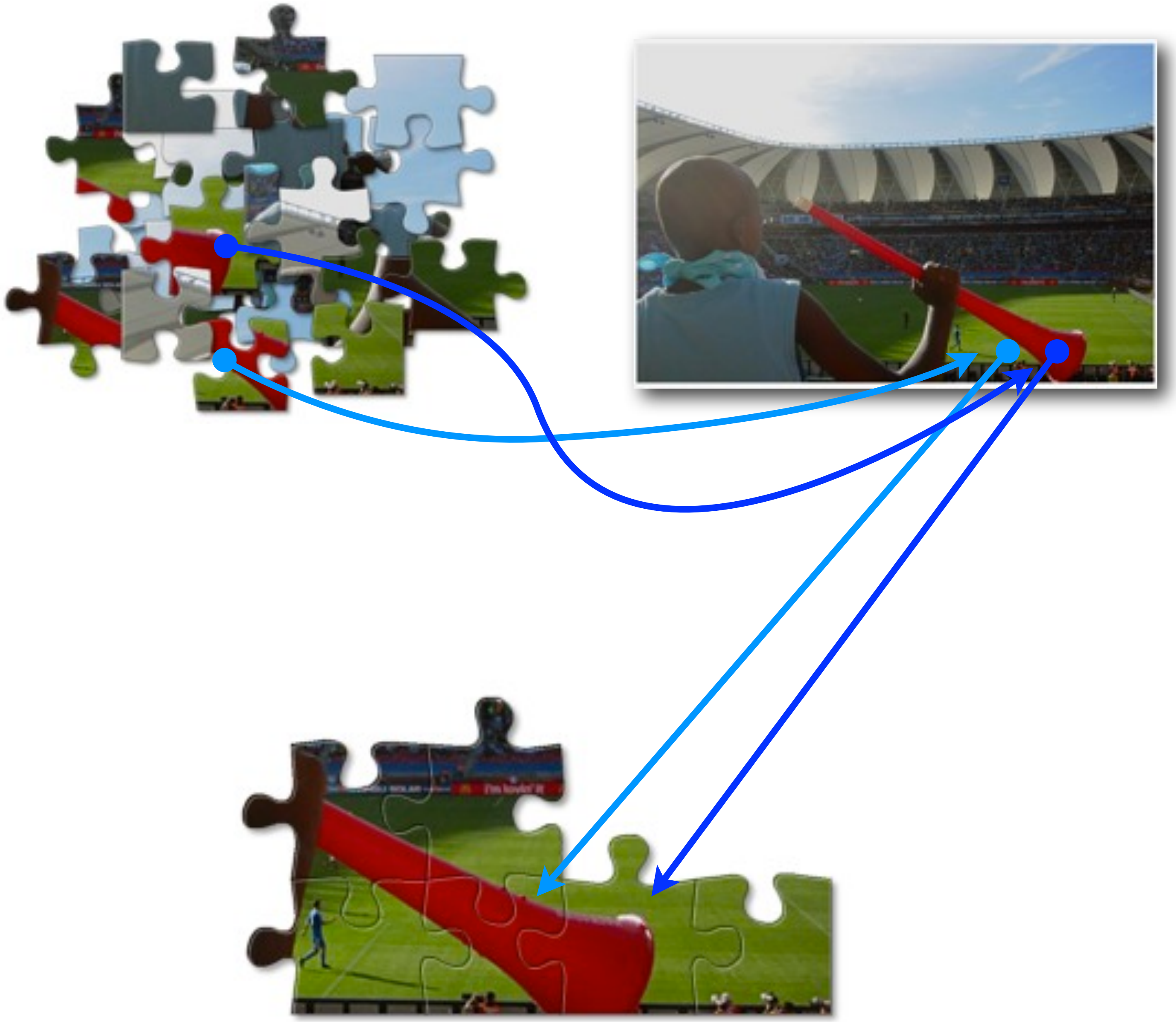


**de novo**

**Comparative**

Image source: http://ngm.nationalgeographic.com/your-shot/jigsaw-puzzles

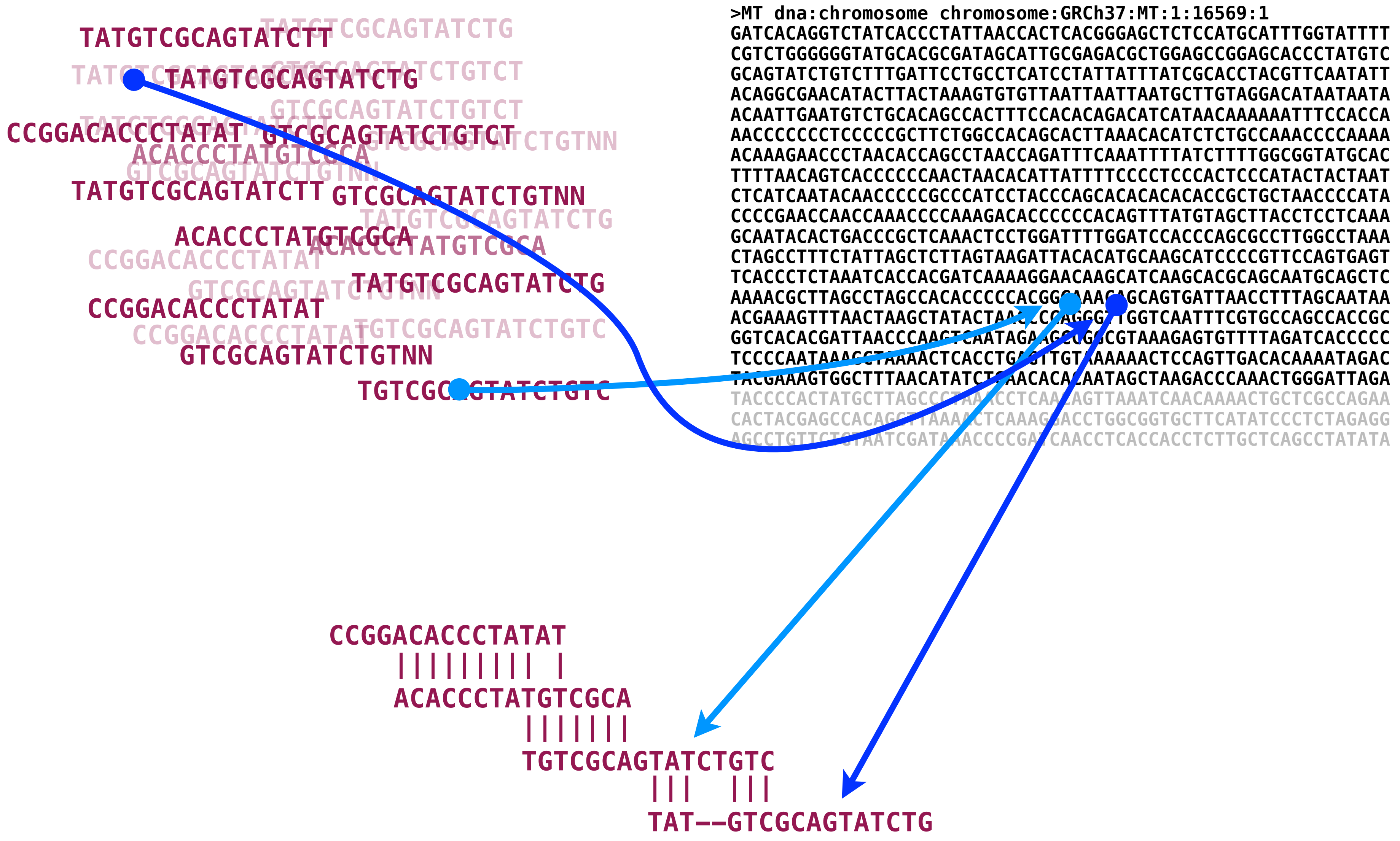
# Comparative



**Comparative**

Image source: http://ngm.nationalgeographic.com/your-shot/jigsaw-puzzles

# Comparative

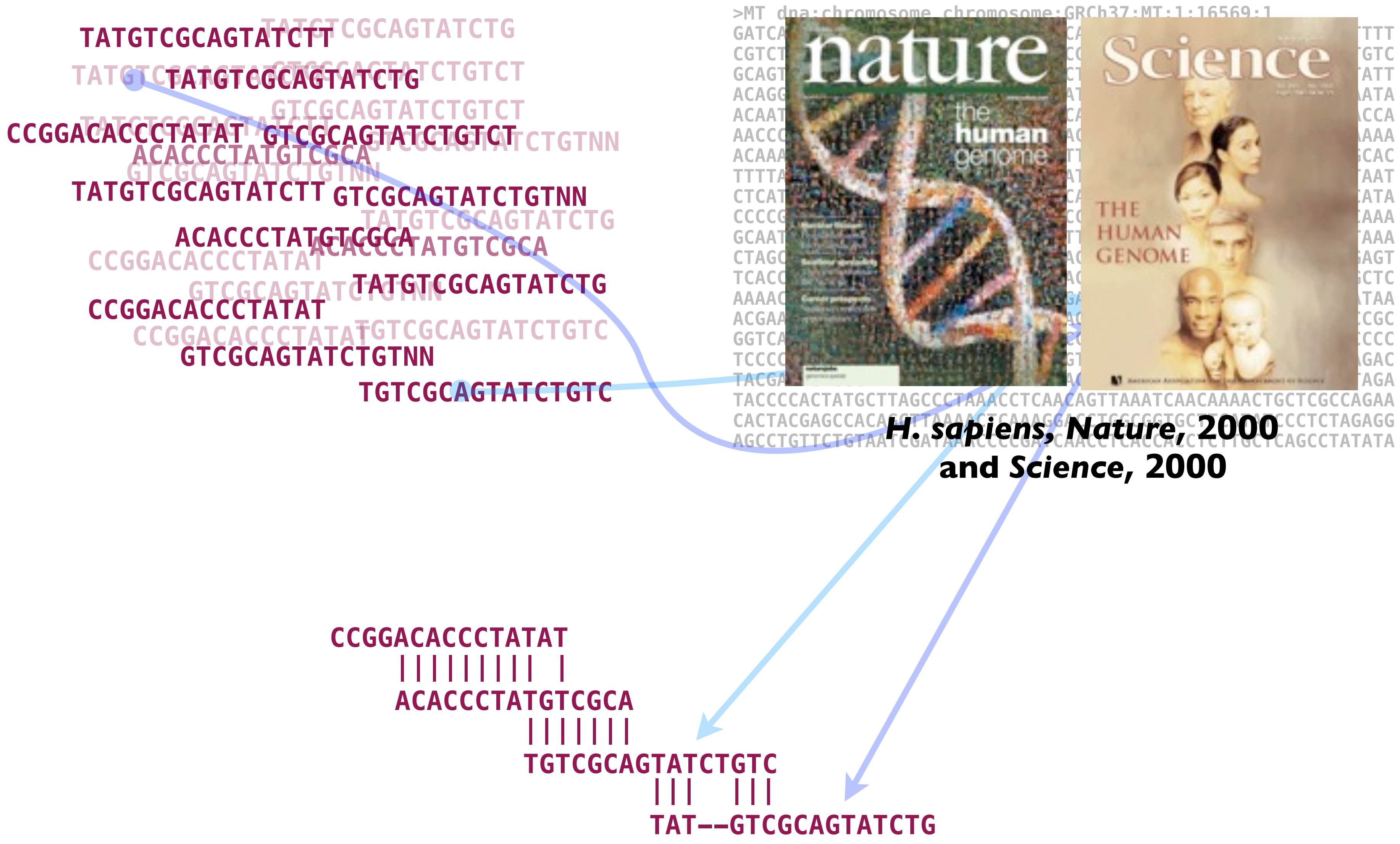


```
>MT dna:chromosome chromosome:GRCh37:MT:1:16569:1
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTT
CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTC
GCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATT
ACAGGCGAACATACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATA
ACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATCATAACAAAAAATTTCCACCA
AACCCCCCCTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAAAA
ACAAAGAACCCTAACACCAGCCTAACCAGATTTCAAATTTTATCTTTTGGCGGTATGCAC
TTTTAACAGTCACCCCCCAACTAACACATTATTTTCCCCTCCCACTCCCATACTACTAAT
CTCATCAATACAACCCCCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATA
CCCCGAACCAACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAA
GCAATACACTGACCCGCTCAAACTCCTGGATTTTGGATCCACCCAGCGCCTTGGCCTAAA
CTAGCCTTTCTATTAGCTCTTAGTAAGATTACACATGCAAGCATCCCCGTTCCAGTGAGT
TCACCCTCTAAATCACCACGATCAAAAGGAACAAGCATCAAGCACGCAGCAATGCAGCTC
AAAACGCTTAGCCTAGCCACACCCCCACGGGAAACAGCAGTGATTAACCTTTAGCAATAA
ACGAAAGTTTAACTAAGCTATACTAACCCCAGGGTTGGTCAATTTCGTGCCAGCCACCGC
GGTCACACGATTAACCCAAGTCAATAGAAGCCGGCGTAAAGAGTGTTTTAGATCACCCCC
TCCCCAATAAAGCTAAAACTCACCTGAGTTGTAAAAAACTCCAGTTGACACAAAATAGAC
TACGAAAGTGGCTTTAACATATCTGAACACACAATAGCTAAGACCCAAACTGGGATTAGA
TACCCCACTATGCTTAGCCCTAAACCTCAACAGTTAAATCAACAAAACTGCTCGCCAGAA
CACTACGAGCCACAGCTTAAAACTCAAAGGACCTGGCGGTGCTTCATATCCCTCTAGAGG
AGCCTGTTCTGTAATCGATAAACCCCGATCAACCTCACCACCTCTTGCTCAGCCTATATA
```

```
CCGGACACCCTATAT
    ||||||||| |
    ACACCCTATGTCGCA
           |||||||
           TGTCGCAGTATCTGTC
                 |||   |||
                 TAT--GTCGCAGTATCTG
```

**Comparative**

# Comparative

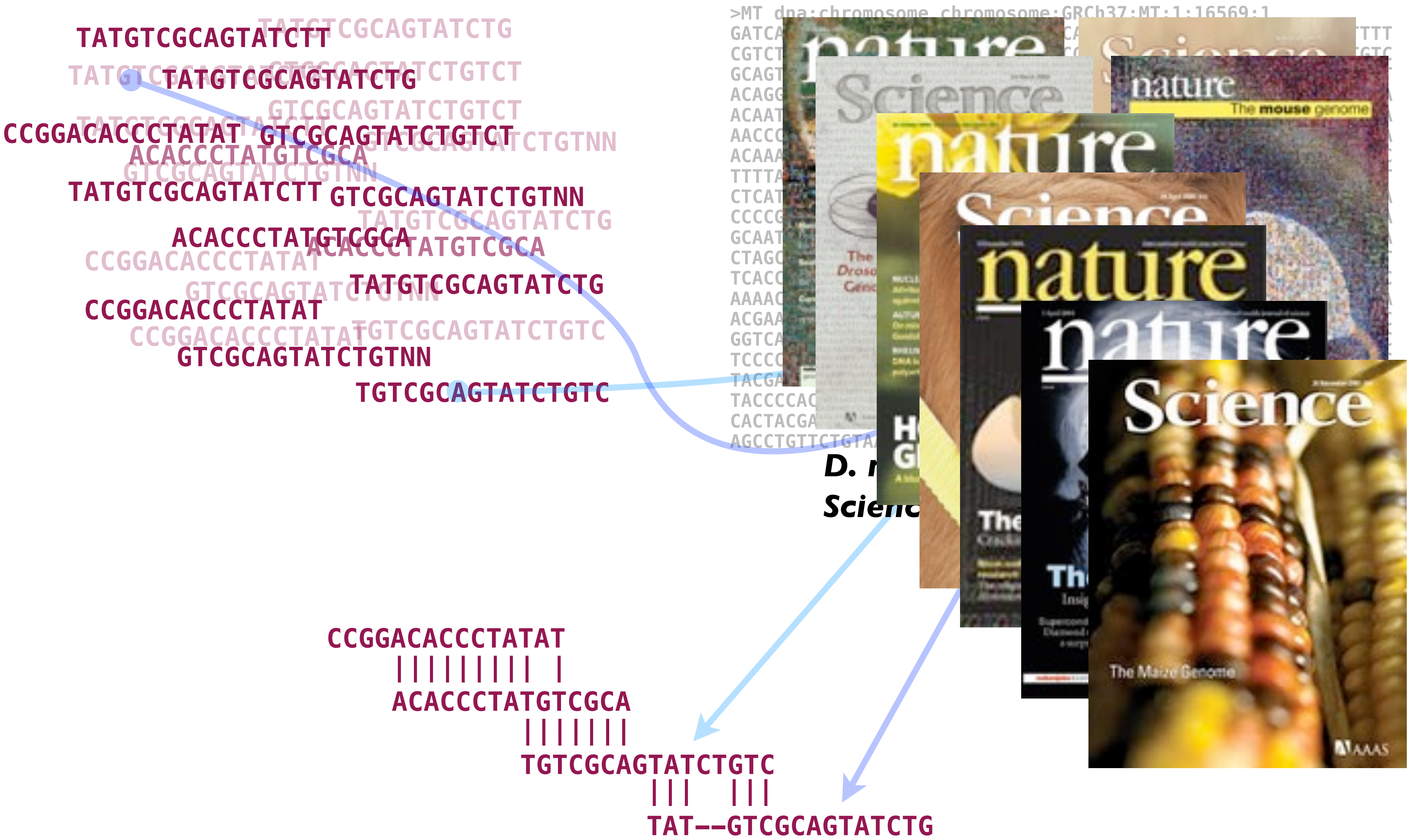


*H. sapiens, Nature*, 2000 and *Science*, 2000

```
CCGGACACCCTATAT
    ||||||||| |
    ACACCCTATGTCGCA
            |||||||
            TGTCGCAGTATCTGTC
                    |||  |||
                    TAT--GTCGCAGTATCTG
```

**Comparative**

# Comparative

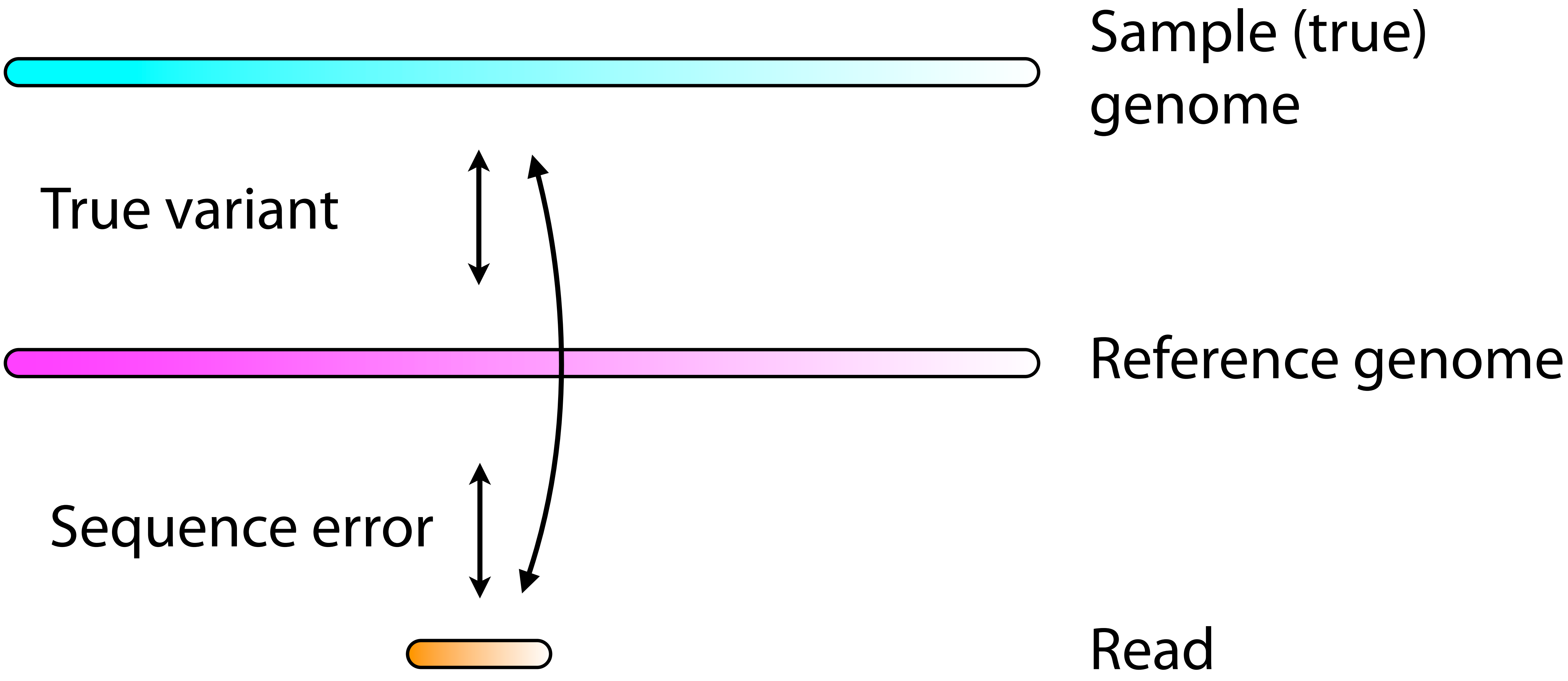


```
TATGTCGCAGTATCTT
TATGTCGCAGTATCTG
CCGGACACCCTATAT GTCGCAGTATCTGTCT
ACACCCTATGTCGCA
TATGTCGCAGTATCTT GTCGCAGTATCTGTNN
ACACCCTATGTCGCA
ACACCCTATGTCGCA
TATGTCGCAGTATCTG
CCGGACACCCTATAT
GTCGCAGTATCTGTNN
TGTCGCAGTATCTGTC
```

```
CCGGACACCCTATAT
    ||||||||| |
    ACACCCTATGTCGCA
            |||||||
            TGTCGCAGTATCTGTC
                    |||  |||
                    TAT--GTCGCAGTATCTG
```

**Comparative**

# Reference genome



# Smith-Waterman

Aligning two sequences is a classic (and extremely important) problem in computational biology.

An 'efficient' solution is provided by the Smith-Waterman algorithm which produces the 'best' alignment under some statistical model.

It handles insertions and deletions elegantly (the default does not handle base qualities), but is too slow for short reads.

( Biostrings::pairwiseAlignment() )

# Smith-Waterman

Aligning **d** reads of length **m** to reference of length **n** is O(**dmn**)

Say:

**m** = 100 nt<br>
**d** = 2 billion (2 x $10^9$) reads } ≈ 1 week-long run of

**n** = 3 billion (3 x $10^9$) nt ≈ human

**Illumina HiSeq 2000**

Source: http://www.illumina.com/systems/hiseq_2000.ilmn

Total of (6 x $10^{20}$) Smith-Waterman cell updates required

A cluster of 1,000 6 Ghz processors, where each processor computes 1 cell update per clock cycle, would take >3 years

# Alignment

## Take a read:

```
CTCAAACTCCTGACCTTTGGTGATCCACCCGCCTNGGCCTTC
```

## And a reference sequence:

```
>MT dna:chromosome chromosome:GRCh37:MT:1:16569:1
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTT
CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTC
GCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATT
ACAGGCGAACATACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATA
ACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATCATAACAAAAAATTTCCACCA
AACCCCCCCTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAAAA
ACAAAGAACCCTAACACCAGCCTAACCAGATTTCAAATTTTATCTTTTGGCGGTATGCAC
TTTTAACAGTCACCCCCCAACTAACACATTATTTTCCCCTCCCACTCCCATACTACTAAT
CTCATCAATACAACCCCCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATA
CCCCGAACCAACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAA
GCAATACACTGACCCGCTCAAACTCCTGGATTTTGGATCCACCCAGCGCCTTGGCCTAAA
CTAGCCTTTCTATTAGCTCTTAGTAAGATTACACATGCAAGCATCCCCGTTCCAGTGAGT
TCACCCTCTAAATCACCACGATCAAAAGGAACAAGCATCAAGCACGCAGCAATGCAGCTC
AAAACGCTTAGCCTAGCCACACCCCCACGGGAAACAGCAGTGATTAACCTTTAGCAATAA
ACGAAAGTTTAACTAAGCTATACTAACCCCAGGGTTGGTCAATTTCGTGCCAGCCACCGC
GGTCACACGATTAACCCAAGTCAATAGAAGCCGGCGTAAAGAGTGTTTTAGATCACCCCC
TCCCCAATAAAGCTAAAACTCACCTGAGTTGTAAAAAACTCCAGTTGACACAAAATAGAC
TACGAAAGTGGCTTTAACATATCTGAACACACAATAGCTAAGACCCAAACTGGGATTAGA
TACCCCACTATGCTTAGCCCTAAACCTCAACAGTTAAATCAACAAAACTGCTCGCCAGAA
CACTACGAGCCACAGCTTAAAACTCAAAGGACCTGGCGGTGCTTCATATCCCTCTAGAGG
AGCCTGTTCTGTAATCGATAAACCCCGATCAACCTCACCACCTCTTGCTCAGCCTATATA
CCGCCATCTTCAGCAAACCCTGATGAAGGCTACAAAGTAAGCGCAAGTACCCACGTAAAG
ACGTTAGGTCAAGGTGTAGCCCATGAGGTGGCAAGAAATGGGCTACATTTTCTACCCCAG
AAAACTACGATAGCCCTTATGAAACTTAAGGGTCGAAGGTGGATTTAGCAGTAAACTAAG
AGTAGAGTGCTTAGTTGAACAGGGCCCTGAAGCGCGTACACACCGCCCGTCACCCTCCTC
AAGTATACTTCAAAGGACATTTAACTAAAACCCCTACGCATTTATATAGAGGAGACAAGT
CGTAACCTCAAACTCCTGCCTTTGGTGATCCACCCGCCTTGGCCTACCTGCATAATGAAG
AAGCACCCAACTTACACTTAGGAGATTTCAACTTAACTTGACCGCTCTGAGCTAAACCTA
GCCCCAAACCCACTCCACCTTACTACCAGACAACCTTAGCCAAACCATTTACCCAAATAA
AGTATAGGCGATAGAAATTGAAACCTGGCGCAATAGATATAGTACCGCAAGGGAAAGATG
AAAAATTATAACCAAGCATAATATAGCAAGGACTAACCCCTATACCTTCTGCATAATGAA
TTAACTAGAAATAACTTTGCAAGGAGAGCCAAAGCTAAGACCCCCGAAACCAGACGAGCT
ACCTAAGAACAGCTAAAAGAGCACACCCGTCTATGTAGCAAAATAGTGGGAAGATTTATA
GGTAGAGGCGACAAACCTACCGAGCCTGGTGATAGCTGGTTGTCCAAGATAGAATCTTAG
TTCAACTTTAAATTTGCCCACAGAACCCTCTAAATCCCCTTGTAAATTTAACTGTTAGTC
CAAAGAGGAACAGCTCTTTGGACACTAGGAAAAAACCTTGTAGAGAGAGTAAAAAATTTA
ACACCCATAGTAGGCCTAAAAGCAGCCACCAATTAAGAAAGCGTTCAAGCTCAACACCCA
CTACCTAAAAAATCCCAAACATATAACTGAACTCCTCACACCCAATTGGACCAATCTATC
ACCCTATAGAAGAACTAATGTTAGTATAAGTAACATGAAAACATTCTCCTCCGCATAAGC
```

How do we determine the read’s point of origin with respect to the reference?

Match 1:

Read

```
CTCAAAGACCTGACCTTTGGTGATCCACCC-----GCCTNGGCCTTC
||||||  ||||   ||||  |||||||||     |||| |||||
CTCAAACTCCTGGATTTTG--GATCCACCCAGCTGGCCTTGGCCTAA
```

Reference

Match 2:

Read

```
CTCAAACTCCTGACCTTTGGTGATCCACCCGCCTNGGCCTTC
|||||||||||| ||||||||||||||||||||| ||||| |
CTCAAACTCCTG-CCTTTGGTGATCCACCCGCCTTGGCCTAC
```

Reference

Which match is better?

Say match 2 is correct.  Why are there still mismatches and gaps?

JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

# Alignment

## Take a read:

```
CTCAAACTCCTGACCTTTGGTGATCCA
```

## And a reference sequence:

```
>MT dna:chromosome chromosome:GRCh37:MT:1:16569:1
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTT
CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTC
GCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATT
ACAGGCGAACATACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATA
ACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATCATAACAAAAAATTTCCACCA
AACCCCCCCTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAAAA
ACAAAGAACCCTAACACCAGCCTAACCAGATTTCAAATTTTATCTTTTGGCGGTATGCAC
TTTTAACAGTCACCCCCCAACTAACACATTATTTTCCCCTCCCACTCCCATACTACTAAT
CTCATCAATACAACCCCCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATA
CCCCGAACCAACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAA
GCAATACACTGACCCGCTCAAACTCCTGGATTTTGTGATCCACCCAGCGCCTTGGCCTAA
CTAGCCTTTCTATTAGCTCTTAGTAAGATTACACATGCAAGCATCCCCGTTCCAGTGAGT
TCACCCTCTAAATCACCACGATCAAAAGGAACAAGCATCAAGCACGCAGCAATGCAGCTC
AAAACGCTTAGCCTAGCCACACCCCCACGGGAAACAGCAGTGATTAACCTTTAGCAATAA
ACGAAAGTTTAACTAAGCTATACTAACCCCAGGGTTGGTCAATTTCGTGCCAGCCACCGC
GGTCACACGATTAACCCAAGTCAATAGAAGCCGGCGTAAAGAGTGTTTTAGATCACCCCC
TCCCCAATAAAGCTAAAACTCACCTGAGTTGTAAAAAACTCCAGTTGACACAAAATAGAC
TACGAAAGTGGCTTTAACATATCTGAACACACAATAGCTAAGACCCAAACTGGGATTAGA
TACCCCACTATGCTTAGCCCTAAACCTCAACAGTTAAATCAACAAAACTGCTCGCCAGAA
CACTACGAGCCACAGCTTAAAACTCAAAGGACCTGGCGGTGCTTCATATCCCTCTAGAGG
AGCCTGTTCTGTAATCGATAAACCCCGATCAACCTCACCACCTCTTGCTCAGCCTATATA
CCGCCATCTTCAGCAAACCCTGATGAAGGCTACAAAGTAAGCGCAAGTACCCACGTAAAG
ACGTTAGGTCAAGGTGTAGCCCATGAGGTGGCAAGAAATGGGCTACATTTTCTACCCCAG
AAAACTACGATAGCCCTTATGAAACTTAAGGGTCGAAGGTGGATTTAGCAGTAAACTAAG
AGTAGAGTGCTTAGTTGAACAGGGCCCTGAAGCGCGTACACACCGCCCGTCACCCTCCTC
AAGTATACTTCAAAGGACATTTAACTAAAACCCCTACGCATTTATATAGAGGAGACAAGT
CGTAACCTCAAACTCCTGGCCTTTGGTGATCCACCCGCCTTGGCCTACCTGCATAATGAA
AAGCACCCAACTTACACTTAGGAGATTTCAACTTAACTTGACCGCTCTGAGCTAAACCTA
GCCCCAAACCCACTCCACCTTACTACCAGACAACCTTAGCCAAACCATTTACCCAAATAA
AGTATAGGCGATAGAAATTGAAACCTGGCGCAATAGATATAGTACCGCAAGGGAAAGATG
AAAAATTATAACCAAGCATAATATAGCAAGGACTAACCCCTATACCTTCTGCATAATGAA
TTAACTAGAAATAACTTTGCAAGGAGAGCCAAAGCTAAGACCCCCGAAACCAGACGAGCT
ACCTAAGAACAGCTAAAAGAGCACACCCGTCTATGTAGCAAAATAGTGGGAAGATTTATA
GGTAGAGGCGACAAACCTACCGAGCCTGGTGATAGCTGGTTGTCCAAGATAGAATCTTAG
TTCAACTTTAAATTTGCCCACAGAACCCTCTAAATCCCCTTGTAAATTTAACTGTTAGTC
CAAAGAGGAACAGCTCTTTGGACACTAGGAAAAAACCTTGTAGAGAGAGTAAAAAATTTA
ACACCCATAGTAGGCCTAAAAGCAGCCACCAATTAAGAAAGCGTTCAAGCTCAACACCCA
CTACCTAAAAAATCCCAAACATATAACTGAACTCCTCACACCCAATTGGACCAATCTATC
ACCCTATAGAAGAACTAATGTTAGTATAAGTAACATGAAAACATTCTCCTCCGCATAAGC
```

Which match is better?

Match 1:

```
            Read
CTCAAACTCCTGACCTTTGGTGATCCA
|||||||||||| ||||||||||||||
CTCAAACTCCTGCCCTTTGGTGATCCA
          Reference
```

Match 2:

```
            Read

CTCAAACTCCTGACCTTTGGTGATCCA
|||||||||||||||||| ||||||||
CTCAAACTCCTGACCTTTCGTGATCCA

          Reference
```

Is there any way to break the tie?

# Two types of qualities

- **Base (sequence) quality**
  Represents the chance that the sequence machine made an error. Produced by the sequence machine (possibly with some post-processing, "calibration"). The 'Q' in FASTQ files.
- **Alignment quality**
  Represents the chance that the alignment is wrong. Produced by the alignment software.

Does base quality really reflect the chance of a sequence error?

# Alignment

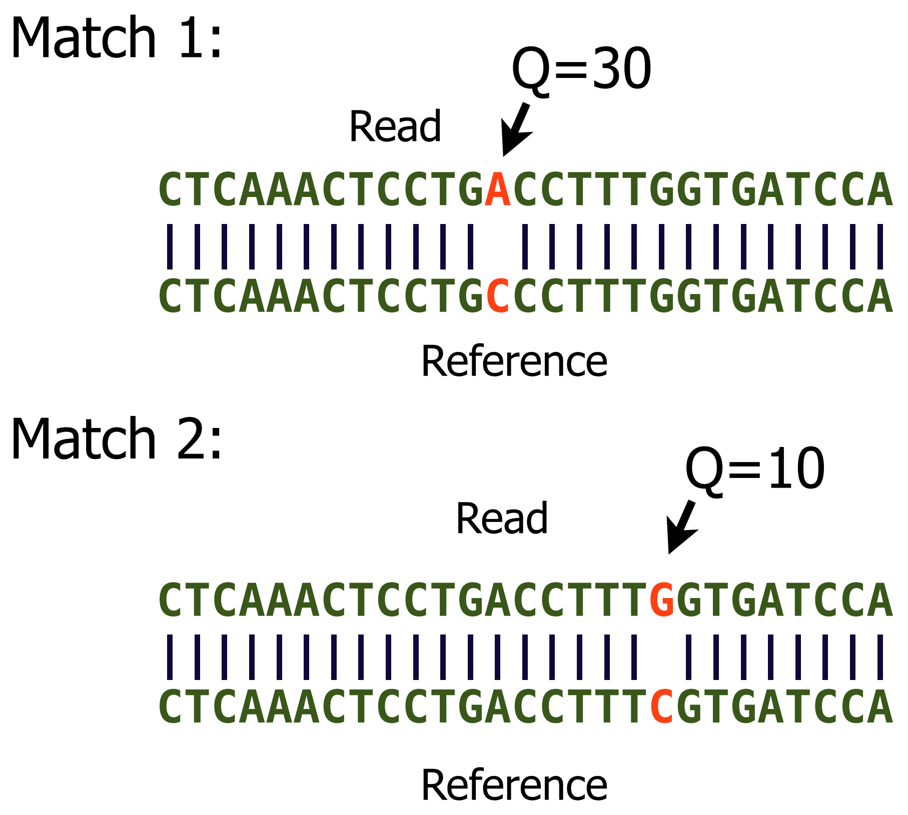
## Take a read:

```
CTCAAACTCCTGACCTTTGGTGATCCA
```

## And a reference sequence:

```
>MT dna:chromosome chromosome:GRCh37:MT:1:16569:1
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTT
CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTC
GCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATT
ACAGGCGAACATACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATA
ACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATCATAACAAAAAATTTCCACCA
AACCCCCCCTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAAAA
ACAAAGAACCCTAACACCAGCCTAACCAGATTTCAAATTTTATCTTTTGGCGGTATGCAC
TTTTAACAGTCACCCCCCAACTAACACATTATTTTCCCCTCCCACTCCCATACTACTAAT
CTCATCAATACAACCCCCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATA
CCCCGAACCAACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAA
GCAATACACTGACCCGCTCAAACTCCTGGATTTTGTGATCCACCCAGCGCCTTGGCCTAA
CTAGCCTTTCTATTAGCTCTTAGTAAGATTACACATGCAAGCATCCCCGTTCCAGTGAGT
TCACCCTCTAAATCACCACGATCAAAAGGAACAAGCATCAAGCACGCAGCAATGCAGCTC
AAAACGCTTAGCCTAGCCACACCCCCACGGGAAACAGCAGTGATTAACCTTTAGCAATAA
ACGAAAGTTTAACTAAGCTATACTAACCCCAGGGTTGGTCAATTTCGTGCCAGCCACCGC
GGTCACACGATTAACCCAAGTCAATAGAAGCCGGCGTAAAGAGTGTTTTAGATCACCCCC
TCCCCAATAAAGCTAAAACTCACCTGAGTTGTAAAAAACTCCAGTTGACACAAAATAGAC
TACGAAAGTGGCTTTAACATATCTGAACACACAATAGCTAAGACCCAAACTGGGATTAGA
TACCCCACTATGCTTAGCCCTAAACCTCAACAGTTAAATCAACAAAACTGCTCGCCAGAA
CACTACGAGCCACAGCTTAAAACTCAAAGGACCTGGCGGTGCTTCATATCCCTCTAGAGG
AGCCTGTTCTGTAATCGATAAACCCCGATCAACCTCACCACCTCTTGCTCAGCCTATATA
CCGCCATCTTCAGCAAACCCTGATGAAGGCTACAAAGTAAGCGCAAGTACCCACGTAAAG
ACGTTAGGTCAAGGTGTAGCCCATGAGGTGGCAAGAAATGGGCTACATTTTCTACCCCAG
AAAACTACGATAGCCCTTATGAAACTTAAGGGTCGAAGGTGGATTTAGCAGTAAACTAAG
AGTAGAGTGCTTAGTTGAACAGGGCCCTGAAGCGCGTACACACCGCCCGTCACCCTCCTC
AAGTATACTTCAAAGGACATTTAACTAAAACCCCTACGCATTTATATAGAGGAGACAAGT
CGTAACCTCAAACTCCTGGCCTTTGGTGATCCACCCGCCTTGGCCTACCTGCATAATGAA
AAGCACCCAACTTACACTTAGGAGATTTCAACTTAACTTGACCGCTCTGAGCTAAACCTA
GCCCCAAACCCACTCCACCTTACTACCAGACAACCTTAGCCAAACCATTTACCCAAATAA
AGTATAGGCGATAGAAATTGAAACCTGGCGCAATAGATATAGTACCGCAAGGGAAAGATG
AAAAATTATAACCAAGCATAATATAGCAAGGACTAACCCCTATACCTTCTGCATAATGAA
TTAACTAGAAATAACTTTGCAAGGAGAGCCAAAGCTAAGACCCCCGAAACCAGACGAGCT
ACCTAAGAACAGCTAAAAGAGCACACCCGTCTATGTAGCAAAATAGTGGGAAGATTTATA
GGTAGAGGCGACAAACCTACCGAGCCTGGTGATAGCTGGTTGTCCAAGATAGAATCTTAG
TTCAACTTTAAATTTGCCCACAGAACCCTCTAAATCCCCTTGTAAATTTAACTGTTAGTC
CAAAGAGGAACAGCTCTTTGGACACTAGGAAAAAACCTTGTAGAGAGAGTAAAAAATTTA
ACACCCATAGTAGGCCTAAAAGCAGCCACCAATTAAGAAAGCGTTCAAGCTCAACACCCA
CTACCTAAAAAATCCCAAACATATAACTGAACTCCTCACACCCAATTGGACCAATCTATC
ACCCTATAGAAGAACTAATGTTAGTATAAGTAACATGAAAACATTCTCCTCCGCATAAGC
```

Which match is better?



Match 1:

Q=30

Read

```
CTCAAACTCCTGACCTTTGGTGATCCA
|||||||||||| ||||||||||||||
CTCAAACTCCTGCCCTTTGGTGATCCA
```

Reference

Match 2:

Q=10

Read

```
CTCAAACTCCTGACCTTTGGTGATCCA
|||||||||||||||||| ||||||||
CTCAAACTCCTGACCTTTCGTGATCCA
```

Reference

# Alignment

## Read 1:

### Best match:

Read

```
AGCTTATATGCTTTTCAGAGCGATACTAAAACCNAACCTCA
||||||||||||||||||||||||||||||||| |||||||
AGCTTATATGCTTTTCAGAGCGATACTAAAACCTAACCTCA
```

Reference

### Second-best match:

Read

```
AGCTTATATGCTATTTCAGAGCGATACTAAAACCNAACCTTA
|||||||||||| ||||||||||||||||||||| ||||| |
AGCTTATATGCT-TTTCAGAGCGATACTAAAACCTAACCTCA
```

Reference

## Read 2:

### Best match:

Read

```
CTCAAACTCCTGACCTTTGGTGATCCACCCGCCTNGGCCTTC
||||||||||||   ||||||||||||||||||| ||||| |
CTCAAACTCCTG---TTTGGTGATCCACCCGCCTTGGCCTAC
```

Reference

### Second-best match:

Read

```
CTCAAAGACCTGACCTTTGGTGATAAACCC-----GCCTNGGCCTTC
||||    ||||   ||||  |||   |||     |||| |||||
CTCA----CCTGGATTTTG--GATCCGCCCAGCTGGCCTTGGCCTAA
```

Reference

For which read are we more confident that the best match is correct?

# Alignment and Bioconductor

These days, most alignment is done using external tools.

However, it is worth knowing about

matchPDict
matchPWM
pairwiseAlignment

in Biostrings.

# Popular aligners

- Bowtie
- BWA(-SW)
- MAQ
- SOAP2
- Novoalign
- .....

Many programs support more than one alignment 'mode' depending on command line settings.

The choice of settings is often unclear.

# Which aligner is best?

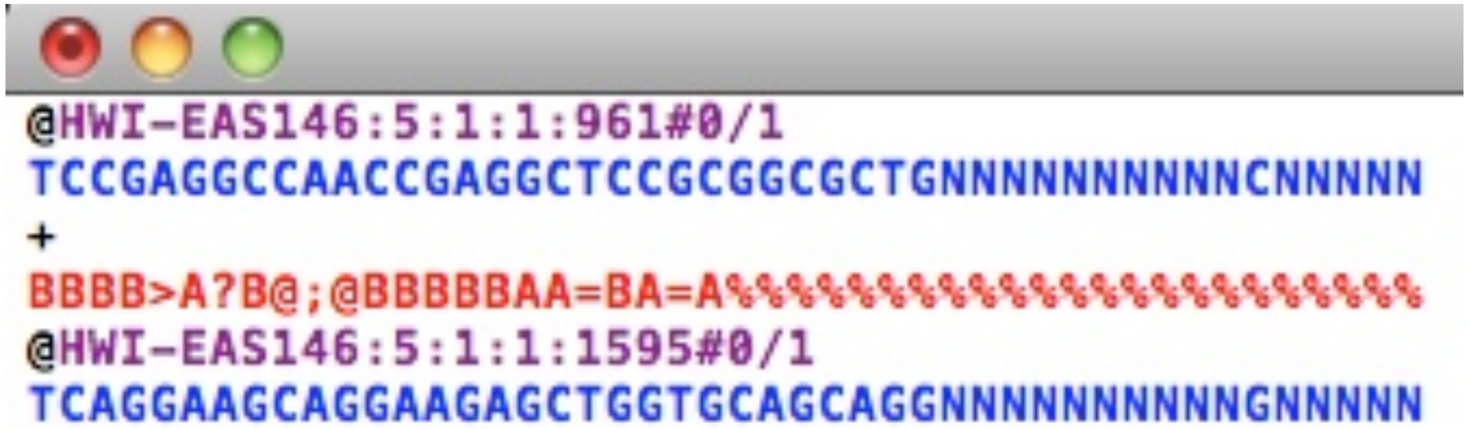
- Two issues: (1) which aligner is the best implementation of a given policy? and (2) which policy is best?
- There has been surprisingly little investigation of which policy is best on real data. It is a hard problem.
- Most aligners have been evaluated in terms of **speed** and **completeness** (% of reads mapped).
- Completeness is probably the wrong metric.
- Some evaluation on simulated data, but we need more.
- Different aligners (policies) produce different end results, sometimes dramatically different.
- Answer also depends on “for what”.

# Fileformats

- Input
  FASTQ, FASTA, QSEQ, SFF
  Vendor specific formats (like CSFASTA+QUAL)
- Output
  BAM/SAM, program-specific format

Tip: Learn the UNIX shell, especially piping

```
gunzip -c INPUT.fastq.gz | \
  bowtie -m 1 -v 2 -p 4 -y --trim3 10 hsapiens_hg19 - | \
  gzip -c > OUTPUT.bwt.gz
```



```
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGGCGCTGNNNNNNNNNNCNNNNN
+
BBBB>A?B@;@BBBBBAA=BA=A%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGCAGCAGGNNNNNNNNNNGNNNNN
```

**name**
**sequence**
**quality scores**

# Interlude

Now for some perspectives on aligning RNA-seq data.

# Junction reads

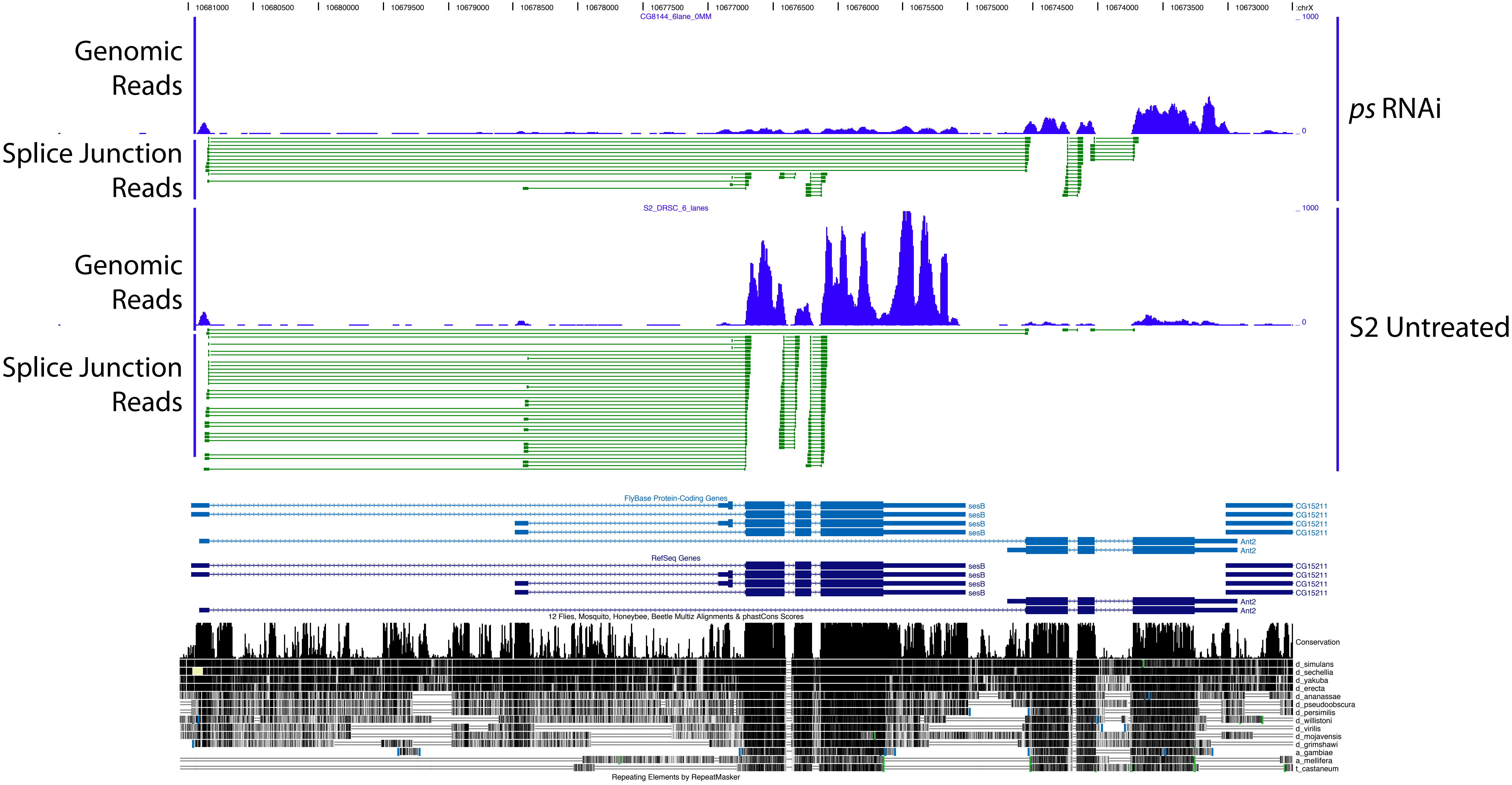


Image from Brenton Gravely

JOHNS HOPKINS BLOOMBERG SCHOOL *of* PUBLIC HEALTH

# Junction reads, zoom

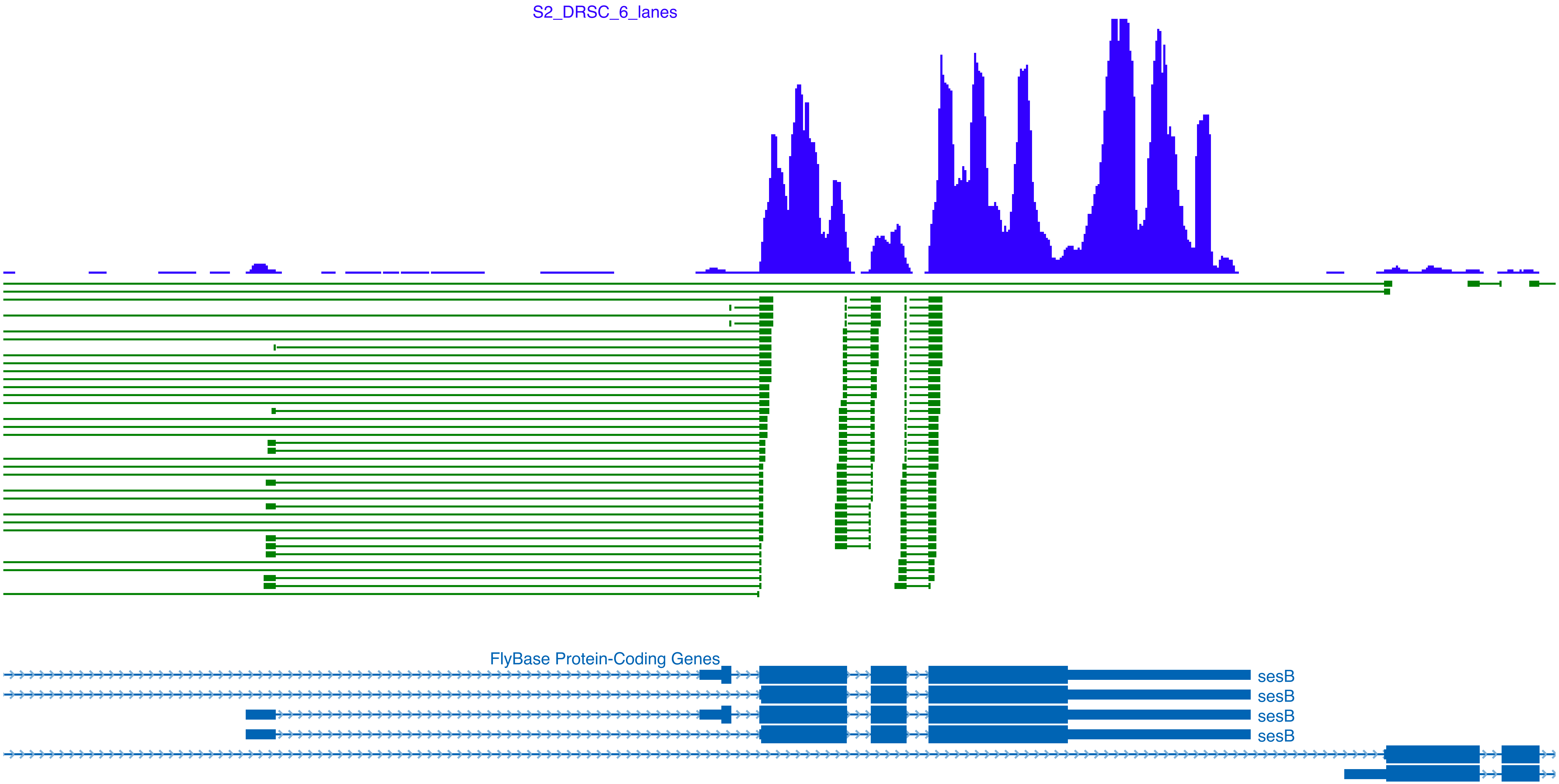
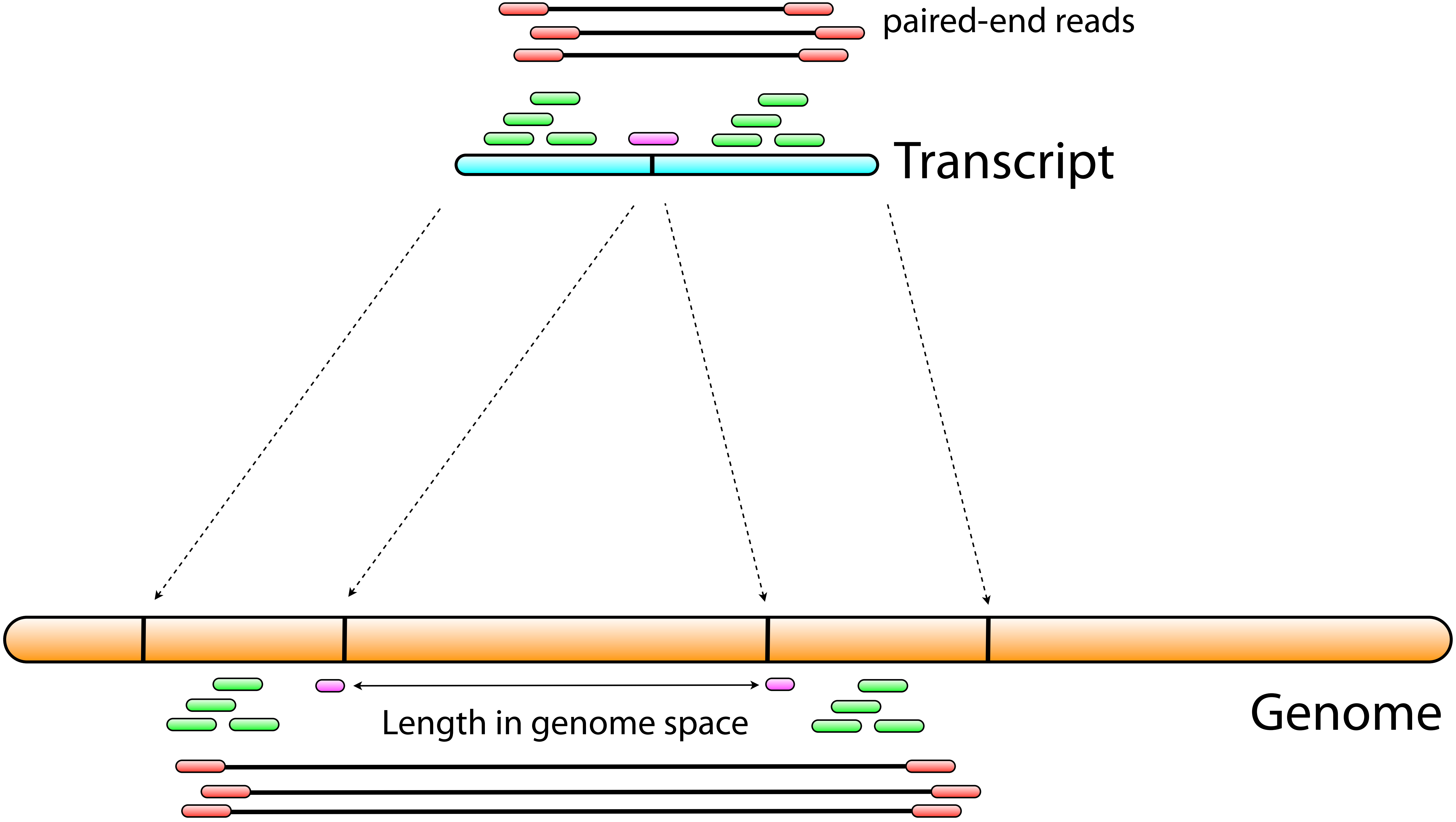


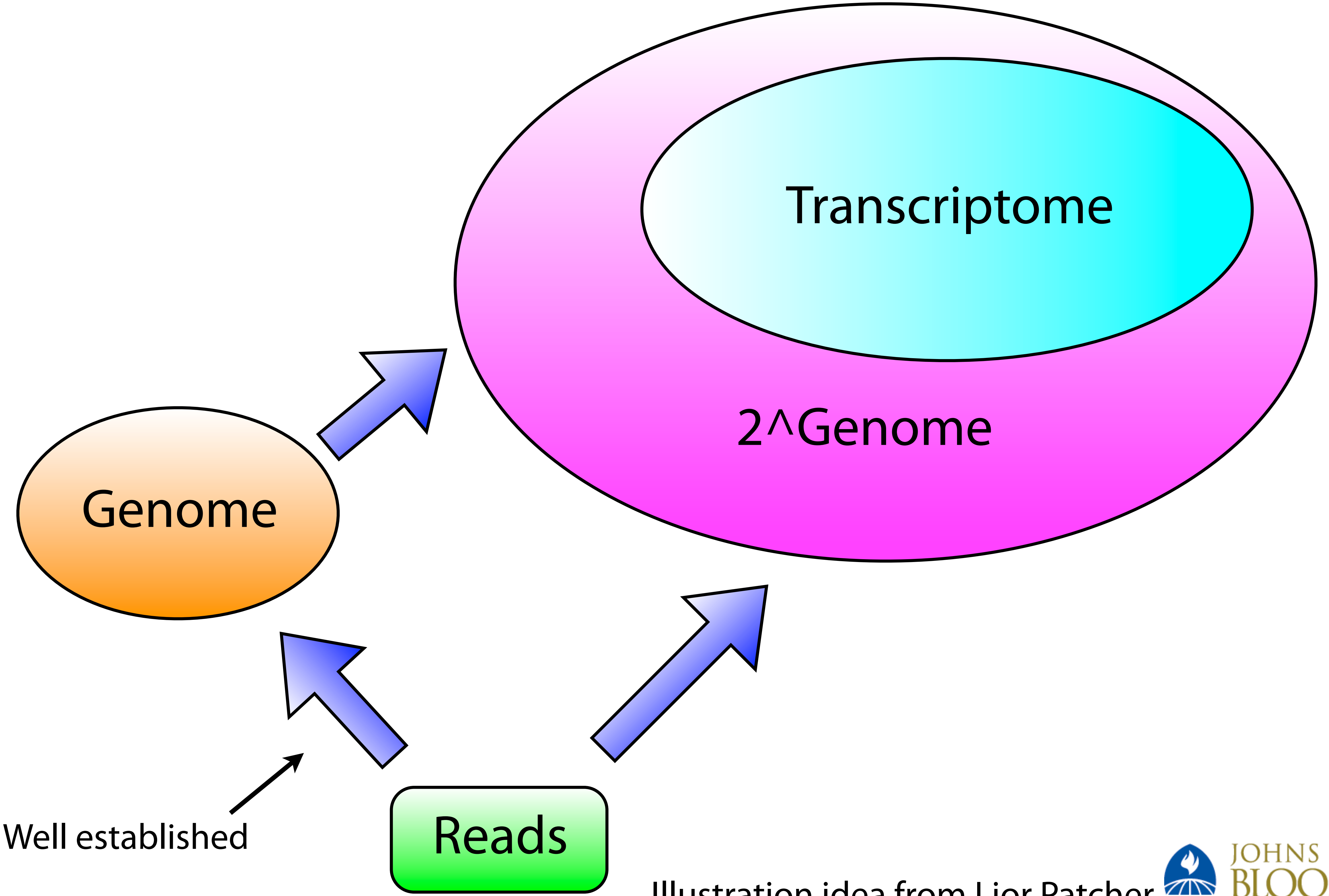
Image from Brenton Gravely

JOHNS HOPKINS BLOOMBERG SCHOOL *of* PUBLIC HEALTH

# Mapping transcripts



# Mapping reads to the transcriptome



Transcriptome

2^Genome

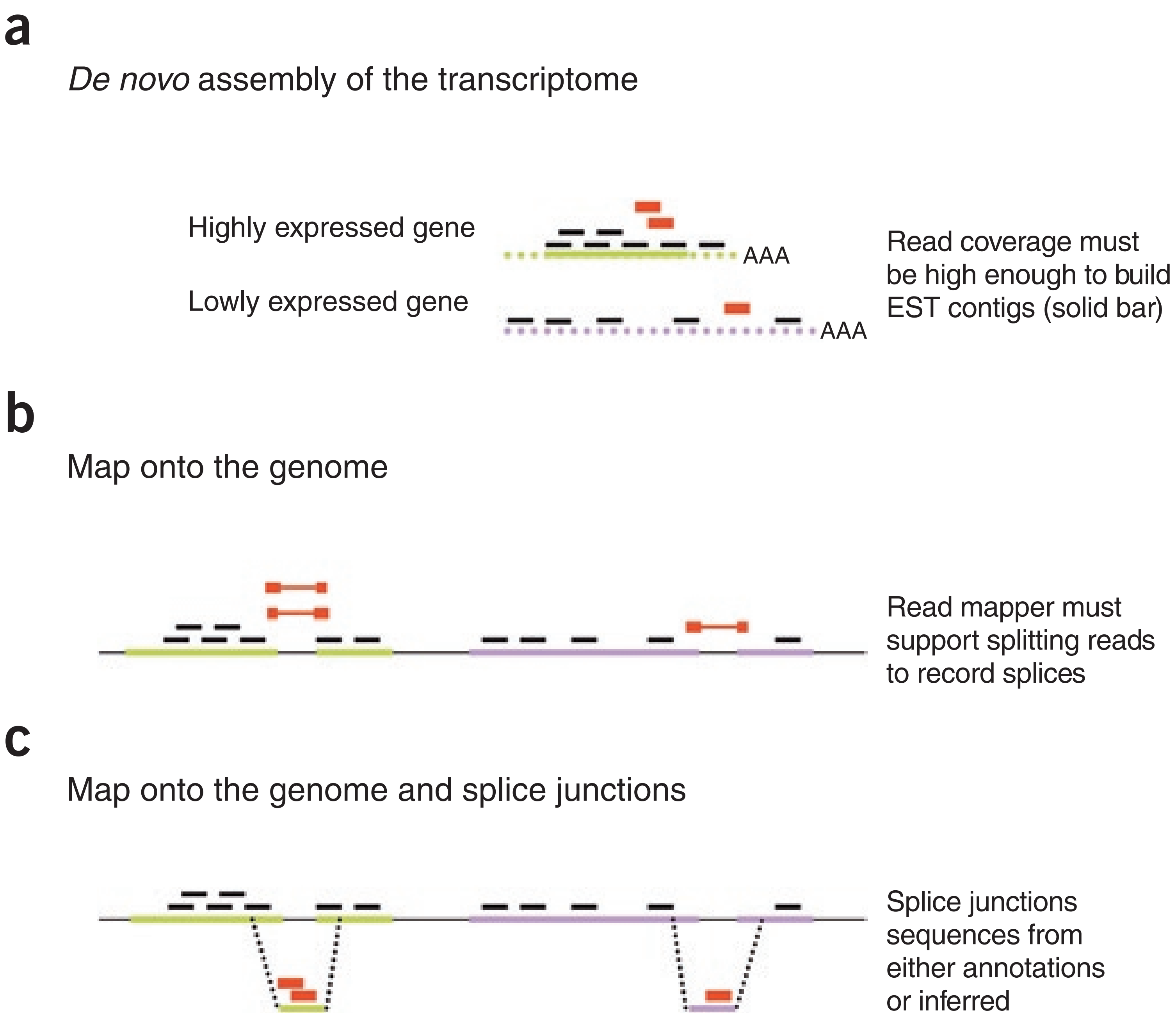
Genome

Reads

Well established

Illustration idea from Lior Patcher

# The basic approaches



**a**

*De novo* assembly of the transcriptome

Highly expressed gene

Lowly expressed gene

Read coverage must be high enough to build EST contigs (solid bar)

**b**

Map onto the genome

Read mapper must support splitting reads to record splices

**c**

Map onto the genome and splice junctions

Splice junctions sequences from either annotations or inferred

From Pepke (2009 Nat Methods)

Popular tools: Tophat/Cufflinks, GSNAP