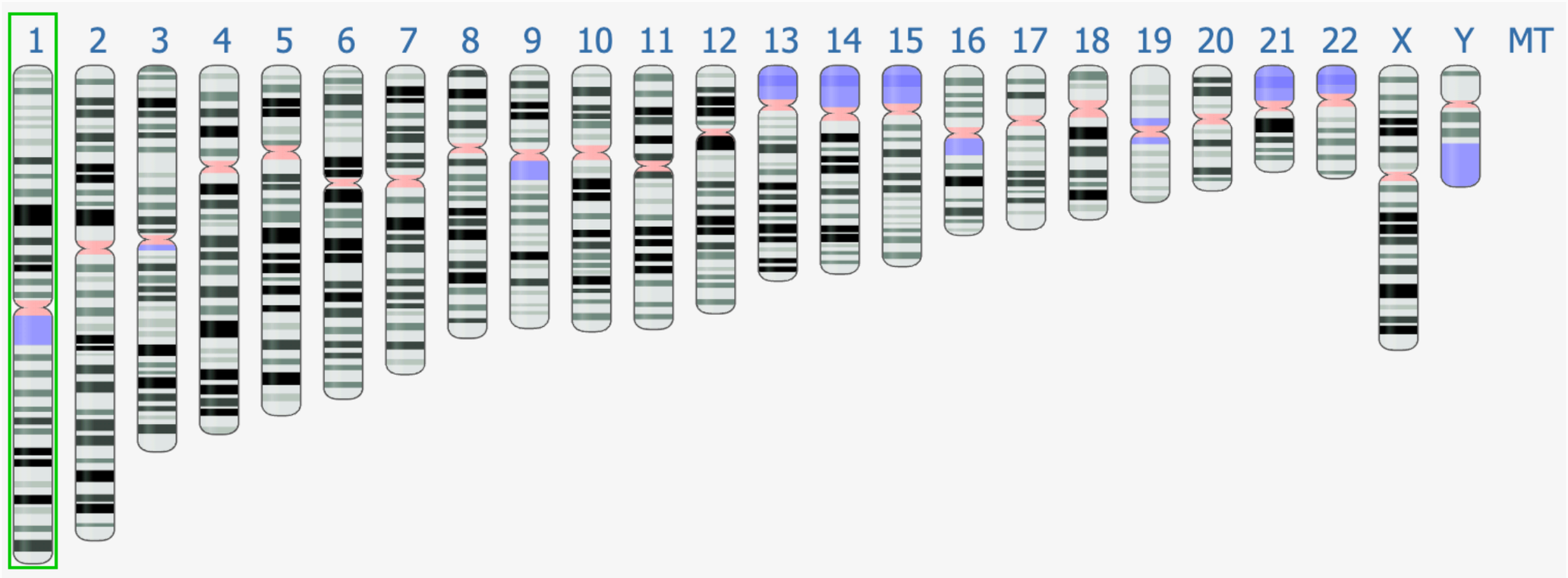


Genomics107

BIOL647
Digital Biology

Rodolfo Aramayo

Genomics107

Mapping and Alignment 103
Mapping with BOWTIE

Open Access

Software

Ultrafast and memory-efficient alignment of short DNA sequences to the human genome

Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg

Address: Center for Bioinformatics and Computational Biology, Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA.

Correspondence: Ben Langmead. Email: langmead@cs.umd.edu

Published: 4 March 2009

Genome ***Biology*** 2009, **10:**R25 (doi:10.1186/gb-2009-10-3-r25)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2009/10/3/R25

Received: 21 October 2008
Revised: 19 December 2008
Accepted: 4 March 2009

Abstract

Bowtie is an ultrafast, memory-efficient alignment program for aligning short DNA sequence reads to large genomes. For the human genome, Burrows-Wheeler indexing allows Bowtie to align more than 25 million reads per CPU hour with a memory footprint of approximately 1.3 gigabytes. Bowtie extends previous Burrows-Wheeler techniques with a novel quality-aware backtracking algorithm that permits mismatches. Multiple processor cores can be used simultaneously to achieve even greater alignment speeds. Bowtie is open source http://bowtie.cbcb.umd.edu.

http://genomebiology.com/2009/10/3/R25

Genomics107

Mapping and Alignment 103
Mapping with BOWTIE - What is BOWTIE?

- **Bowtie is an ultrafast, memory-efficient short read aligner geared toward quickly aligning large sets of short DNA sequences (reads) to large genomes. It aligns 35-base-pair reads to the human genome at a rate of 25 million reads per hour on a typical workstation.**

- **Bowtie indexes the genome with a Burrows-Wheeler index to keep its memory footprint small:**
  - **for the human genome, the index is typically about 2.2 GB (for unpaired alignment) or 2.9 GB (for paired-end alignment).**
  - **Multiple processors can be used simultaneously to achieve greater alignment speed.**

- **Bowtie can also output alignments in the standard SAM format, allowing Bowtie to interoperate with other tools supporting SAM, including the SAMtools consensus, SNP, and indel callers. Bowtie runs on the command line under Windows, Mac OS X, Linux, and Solaris.**

- **Bowtie also forms the basis for other tools, including TopHat: a fast splice junction mapper for RNA-seq reads, Cufflinks: a tool for transcriptome assembly and isoform quantitiation from RNA-seq reads, Crossbow: a cloud-computing software tool for large-scale resequencing data,and Myrna: a cloud computing tool for calculating differential gene expression in large RNA-seq datasets.**

https://github.com/BenLangmead/bowtie/blob/master/MANUAL.markdown

Genomics107

Mapping and Alignment 103
Mapping with BOWTIE - What isn't BOWTIE?

- **Bowtie is not a general-purpose alignment tool like MUMmer, BLAST or Vmatch.**
- **Bowtie works best when aligning short reads to large genomes, though it supports arbitrarily small reference sequences (e.g. amplicons) and reads as long as 1024 bases.**
- **Bowtie is designed to be extremely fast for sets of short reads where**
  - **(a) many of the reads have at least one good, valid alignment,**
  - **(b) many of the reads are relatively high-quality, and**
  - **(c) the number of alignments reported per read is small (close to 1).**
- **Gapped alignments are not currently supported in Bowtie, but they are supported in Bowtie 2.**

https://github.com/BenLangmead/bowtie/blob/master/MANUAL.markdown

Genomics107

Mapping and Alignment 103
BOWTIE Alignment Options

-v (Read Mismatch)

- Allow -v n mismatches in the whole read
  Where n=0,1,2, or 3
- Quality scores are ignored

-n (Seed Mismatch)

- Allow -v n mismatches in the seed region
  Where n=0,1,2, or 3
- Seed length = -l x
  Where n = >5

10mer Seed Region

```
Read:    AATCCCAGAACTTTGGGAGGGTGACATTTGT
         |||||||||||||||||||| ||| ||| ||
Genome:  AATCCCAGAACTTTGGGAGGCTGAGATTAGT
```

Settings: -n 0 -l 10

Read Length

```
Read:    AATGCCAGAACTTTGGGAGGCTGACATTTGT
         ||| |||||||||||||||||| ||| ||
Genome:  AATCCCAGAACTTTGGGAGGCTGAGATTAGT
```

Settings: -v 2

Read Length

Genomics107

Mapping and Alignment 103
Mapping with BOWTIE

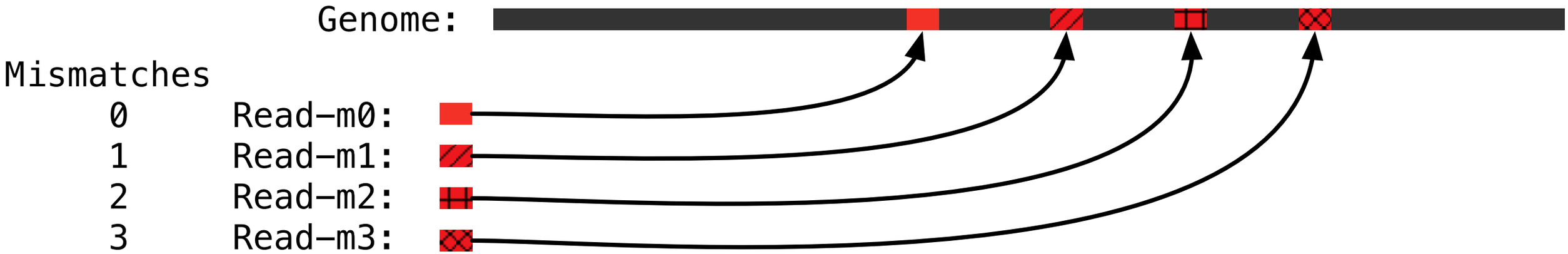
- **Anatomy of an Example Hypothetical Bowtie command:**

```
bowtie                  = Program name
-p 1                    = Use one core
-t                      = Time. Print the amount of wall-clock time taken by each phase
-a                      = Report ALL valid alignments
-v 0                    = No mismatches. In -v mode, alignments may have no more than either
                          0, 1, 2, or 3 mismatches
--best                  = Give me your best alignment
--strata                = Report the total number of alignment for the entire alignment
-m 1                    = -m 1 instructs bowtie to refrain from reporting any alignments
                          for reads having more than 1 reportable alignments
                          Use -1 for no limit
                          This has profound implications for your results
--max                   = Write all reads with a number of valid alignments exceeding the limit
                          set with the -m option to a file (--max)
--un                    = Write all reads that could not be aligned to a file
--al                    = Write aligned reads to a file
--sam                   = Print alignments in SAM format (default)
--sam-nohead            = Suppress the header in the output SAM file
--chunkmbs 1028         = Memory management option
Genome01.fna            = Genome (Indexed)
-q file.fastq           = Query in FastQ format
```

Genomics107

Mapping and Alignment 103

Important Mapping Versus Mismatches Considerations



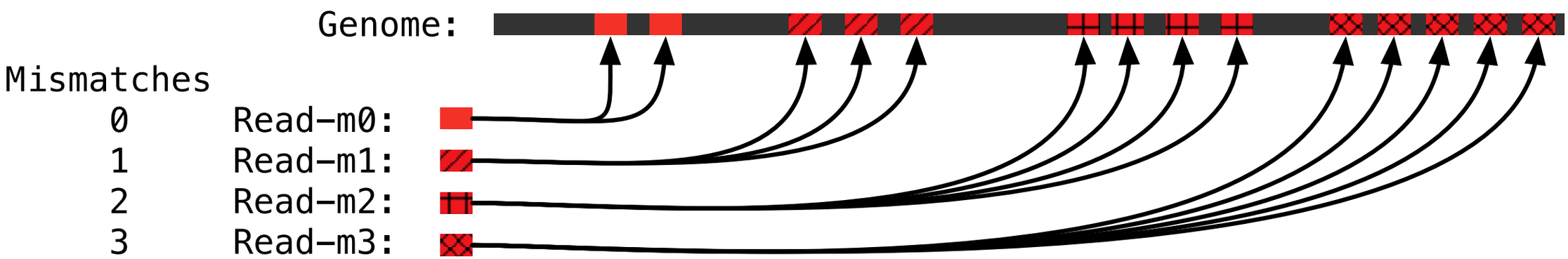

```
Mapping (Alignment Mode: -n or -v)
   0 Mismatches: 01 Hit                    +
   1 Mismatches: 02 Hits                   +     +
   2 Mismatches: 03 Hits                   +     +       +
   3 Mismatches: 04 Hits                   +     +       +       +

Reported Mapping = +
```

Genomics107

Mapping and Alignment 103

Important Mapping Versus Mismatches Considerations




```
Mapping (Alignment Mode: -n or -v) AND Suppress Alignments (-m = -1)
   0 Mismatches: 02 Hits     + +
   1 Mismatches: 05 Hits     + +       + + +
   2 Mismatches: 09 Hits     + +       + + +       + + + +
   3 Mismatches: 15 Hits     + +       + + +       + + + +     + + + + +

Mapping (Alignment Mode: -n or -v) AND Suppress Alignments (-m = 2)
   0 Mismatches: 02 Hits     + +
   1 Mismatches: 05 Hits     + +       + + +
   2 Mismatches: 09 Hits     + +       + + +       + + + +
   3 Mismatches: 15 Hits     + +       + + +       + + + +     + + + + +

Mapping (Alignment Mode: -n or -v) AND Suppress Alignments (-m = 10)
   0 Mismatches: 02 Hits     + +
   1 Mismatches: 05 Hits     + +       + + +
   2 Mismatches: 09 Hits     + +       + + +       + + + +
   3 Mismatches: 15 Hits     + +       + + +       + + + +     + + + + +

Reported Mapping     = +
Not Reported Mapping = +
```

Genomics107

Mapping and Alignment 103
BOWTIE Workflow

Reference genome (FASTA)

```
>chr1
TACCTCCAGGGGGCATCCTCCCCCCCAAT
TCGAAACACAATCGTAGCCCCTGGCACTA
CCTATGTGTGTCAATTCGGAGAGAGAGAG
ATTCACGAAAAAAAAGTCTGGACTCAACT
AGGATACACACATTCGGCTACAGATACCA
AAAAAAAAAAAAAAAAAATTTTCACCATT
GAGGCACCACCTTCTCGTCGCTGCGTCGC
TCTGCTCGCTTCGGCTAAAAATTCGCGCA
ATACATTCGGCTACAGATACCAAA
```

→ **Generate BOWTIE-Index For Reference Sequence**

→ **Align SE-FASTQ Reads To BOWTIE-Index**

Genomics107

Mapping and Alignment 103
BOWTIE Workflow

- **Commands BOWTIE-Index Construction:**

```
$ bowtie-build \
  BRCA2_WildType.fa \
  BRCA2_WildType.fa;
```

Genomics107

Mapping and Alignment 103
BOWTIE Workflow

- **Commands Single-Ends Reads Mapping:**

```
$ bowtie \
      -x BRCA2_WildType.fa \
      -q \
      BRCA2_WildType_PE_Run01_1.fq,BRCA2_WildType_PE_Run01_2.fq \
      -S WildType_x_WildType_Reads.sam;
```

Genomics107

Mapping and Alignment 103
BOWTIE Workflow

- **Commands Paired-Ends Reads Mapping:**

```
$ bowtie \
      -x BRCA2_WildType.fa \
      -q \
      -1 BRCA2_WildType_PE_Run01_1.fq \
      -2 BRCA2_WildType_PE_Run01_2.fq \
      -S WildType_x_WildType_Reads.sam;
```

Genomics107

Mapping and Alignment 103
Mapping with BOWTIE2

Fast gapped-read alignment with Bowtie 2

Ben Langmead[1,2] & Steven L Salzberg[1–3]

As the rate of sequencing increases, greater throughput is demanded from read aligners. The full-text minute index is often used to make alignment very fast and memory-efficient, but the approach is ill-suited to finding longer, gapped alignments. Bowtie 2 combines the strengths of the full-text minute index with the flexibility and speed of hardware-accelerated dynamic programming algorithms to achieve a combination of high speed, sensitivity and accuracy.

https://www.nature.com/articles/nmeth.1923

Genomics107

Mapping and Alignment 103

Mapping with BOWTIE2 - What is BOWTIE2?

- Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences.

- It is particularly good at aligning reads of about 50 up to 100s of characters to relatively long (e.g. mammalian) genomes.

- Bowtie 2 indexes the genome with an FM Index (based on the Burrows-Wheeler Transform or BWT) to keep its memory footprint small:
  - for the human genome, its memory footprint is typically around 3.2 gigabytes of RAM.

- Bowtie 2 supports gapped, local, and paired-end alignment modes. Multiple processors can be used simultaneously to achieve greater alignment speed.

- Bowtie 2 outputs alignments in SAM format, enabling interoperation with a large number of other tools (e.g. SAMtools, GATK) that use SAM. Bowtie 2 is distributed under the GPLv3 license, and it runs on the command line under Windows, Mac OS X and Linux and BSD.

- Bowtie 2 is often the first step in pipelines for comparative genomics, including for variation calling, ChIP-seq, RNA-seq, BS-seq. Bowtie 2 and Bowtie (also called "Bowtie 1" here) are also tightly integrated into many other tools

https://github.com/BenLangmead/bowtie2/blob/master/MANUAL.markdown

Genomics107

Mapping and Alignment 103

Mapping with BOWTIE2 - What isn't BOWTIE2?

- **Bowtie 2 is geared toward aligning relatively short sequencing reads to long genomes.**
  - **That said, it handles arbitrarily small reference sequences (e.g. amplicons) and very long reads (i.e. upwards of 10s or 100s of kilobases), though it is slower in those settings.**
  - **It is optimized for the read lengths and error modes yielded by typical Illumina sequencers.**

- **Bowtie 2 does not support alignment of colorspace reads. (Bowtie 1 does.)**

https://github.com/BenLangmead/bowtie/blob/master/MANUAL.markdown

Genomics107

Mapping and Alignment 103

Mapping with BOWTIE2 - How is BOWTIE2 different from BOWTIE1?

- **Bowtie 1 was released in 2009 and was geared toward aligning the relatively short sequencing reads (up to 25-50 nucleotides) prevalent at the time.**

- **Since then, technology has improved both sequencing throughput (more nucleotides produced per sequencer per day) and read length (more nucleotides per read).**

https://github.com/BenLangmead/bowtie/blob/master/MANUAL.markdown

Genomics107

Mapping and Alignment 103

Mapping with BOWTIE2 - How is BOWTIE2 different from BOWTIE1?

- **The chief differences between Bowtie 1 and Bowtie 2 are:**

  - **For reads longer than about 50 bp Bowtie 2 is generally faster, more sensitive, and uses less memory than Bowtie 1.**
  - **For relatively short reads (e.g. less than 50 bp) Bowtie 1 is sometimes faster and/or more sensitive.**
  - **Bowtie 2 supports gapped alignment with affine gap penalties. Number of gaps and gap lengths are not restricted, except by way of the configurable scoring scheme.**
  - **Bowtie 1 finds just ungapped alignments.**
  - **Bowtie 2 supports local alignment, which doesn't require reads to align end-to-end. Local alignments might be "trimmed" ("soft clipped") at one or both extremes in a way that optimizes alignment score.**
  - **Bowtie 2 also supports end-to-end alignment which, like Bowtie 1, requires that the read align entirely.**
  - **There is no upper limit on read length in Bowtie 2.**
  - **Bowtie 1 had an upper limit of around 1000 bp.**

https://github.com/BenLangmead/bowtie/blob/master/MANUAL.markdown

Genomics107

Mapping and Alignment 103

Mapping with BOWTIE2 - How is BOWTIE2 different from BOWTIE1?

- **The chief differences between Bowtie 1 and Bowtie 2 are:**

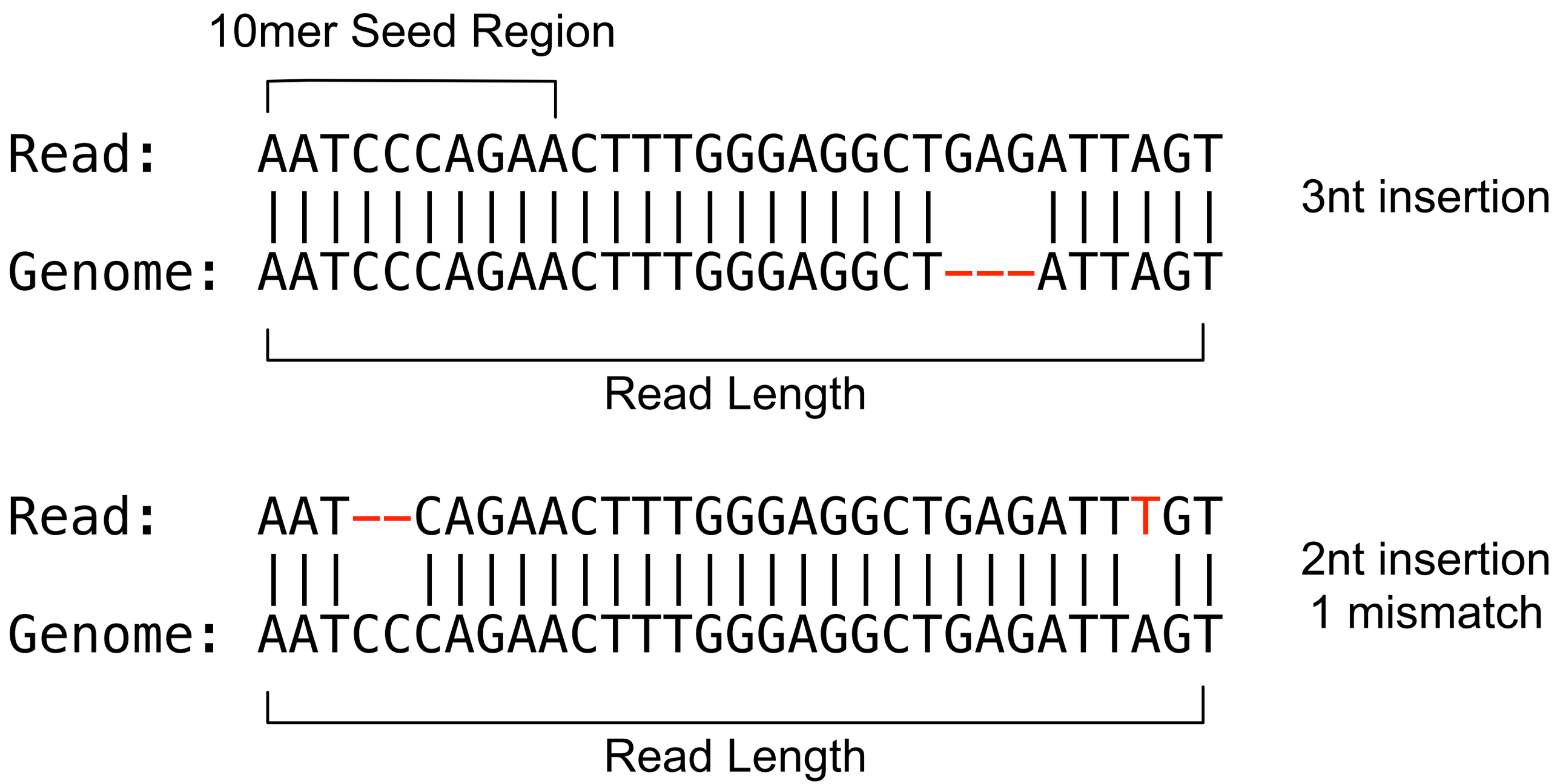
  - **Bowtie 2 allows alignments to overlap ambiguous characters (e.g. Ns) in the reference.**
  - **Bowtie 1 does not.**
  - **Bowtie 2 does away with Bowtie 1's notion of alignment "stratum", and its distinction between "Maq-like" and "end-to-end" modes.**
  - **In Bowtie 2 all alignments lie along a continuous spectrum of alignment scores where the scoring scheme, similar to Needleman-Wunsch and Smith-Waterman.**
  - **Bowtie 2's paired-end alignment is more flexible. E.g. for pairs that do not align in a paired fashion, Bowtie 2 attempts to find unpaired alignments for each mate.**
  - **Bowtie 2 reports a spectrum of mapping qualities, in contrast for Bowtie 1 which reports either 0 or high.**
  - **Bowtie 2 does not align colorspace reads.**
  - **Bowtie 2 is not a "drop-in" replacement for Bowtie 1. Bowtie 2's command-line arguments and genome index format are both different from Bowtie 1's.**

https://github.com/BenLangmead/bowtie/blob/master/MANUAL.markdown

Genomics107

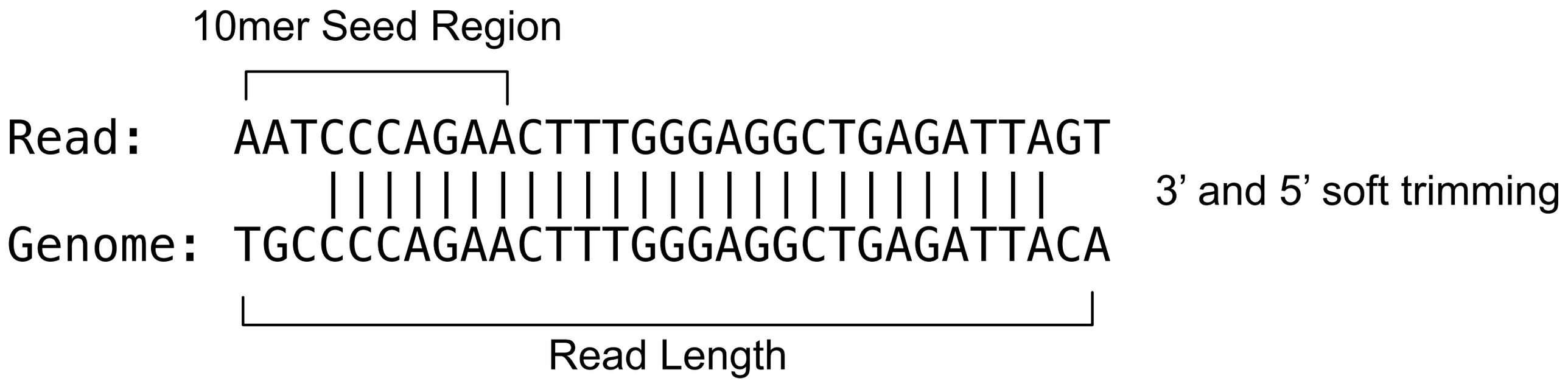
Mapping and Alignment 103

BOWTIE2 Seed Versus Entire read



Genomics107

Mapping and Alignment 103
BOWTIE2 Soft Trimming



Genomics107

Mapping and Alignment 103
BOWTIE2 Workflow

Reference genome (FASTA)

```
>chr1
TACCTCCAGGGGGCATCCTCCCCCCCAAT
TCGAAACACAATCGTAGCCCCTGGCACTA
CCTATGTGTGTCAATTCGGAGAGAGAGAG
ATTCACGAAAAAAAAGTCTGGACTCAACT
AGGATACACACATTCGGCTACAGATACCA
AAAAAAAAAAAAAAAAAATTTTCACCATT
GAGGCACCACCTTCTCGTCGCTGCGTCGC
TCTGCTCGCTTCGGCTAAAAATTCGCGCA
ATACATTCGGCTACAGATACCAAA
```

→ **Generate BOWTIE2-Index For Reference Sequence**

→ **Align SE-FASTQ Reads To BOWTIE2-Index**

Genomics107

Mapping and Alignment 103
BOWTIE2 Workflow

- **Commands BOWTIE2-Index Construction:**

```
$ bowtie2-build \
  BRCA2_WildType.fa \
  BRCA2_WildType.fa;
```

Genomics107

Mapping and Alignment 103
BOWTIE2 Workflow

- **Commands Single-Ends Reads Mapping:**

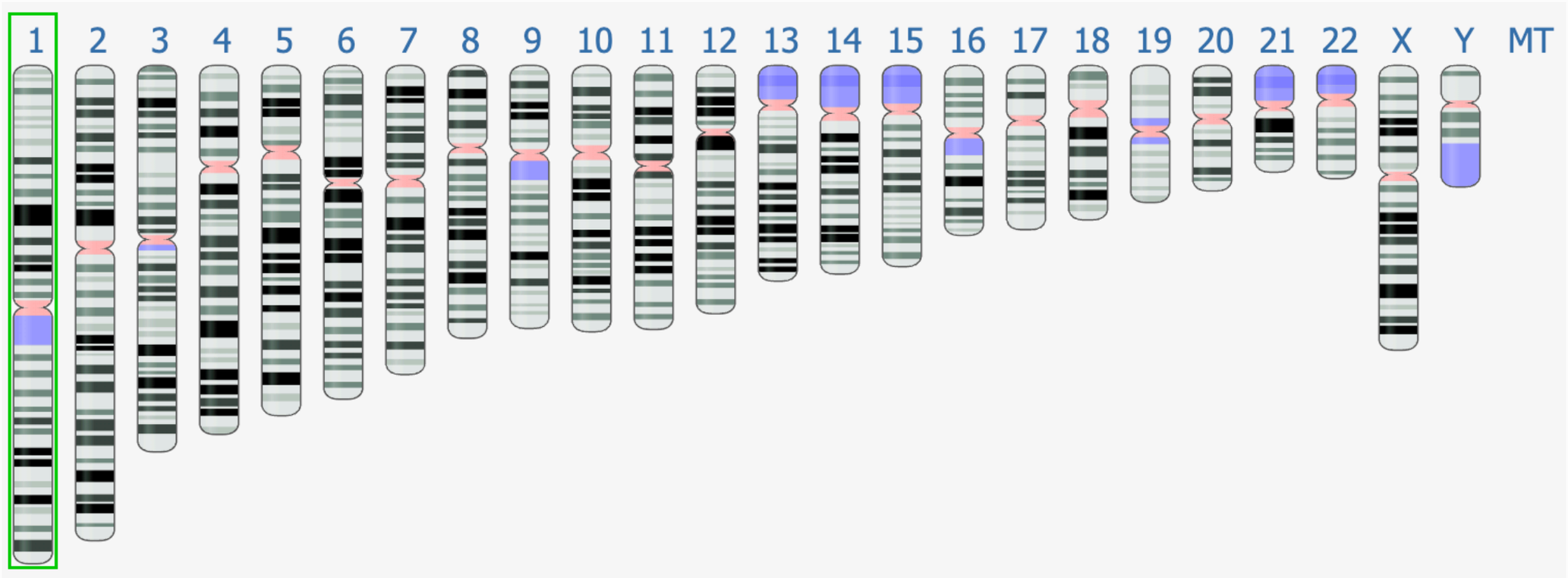
```
$ bowtie2 \
    -x BRCA2_WildType.fa \
    -q \
    BRCA2_WildType_PE_Run01_1.fq,BRCA2_WildType_PE_Run01_2.fq \
    -S WildType_x_WildType_Reads.sam;
```

Genomics107

Mapping and Alignment 103
BOWTIE2 Workflow

- **Commands Paired-Ends Reads Mapping:**

```
$ bowtie2 \
      -x BRCA2_WildType.fa \
      -q \
      -1 BRCA2_WildType_PE_Run01_1.fq \
      -2 BRCA2_WildType_PE_Run01_2.fq \
      -S WildType_x_WildType_Reads.sam;
```



Genomics107

BIOL647
Digital Biology

Rodolfo Aramayo