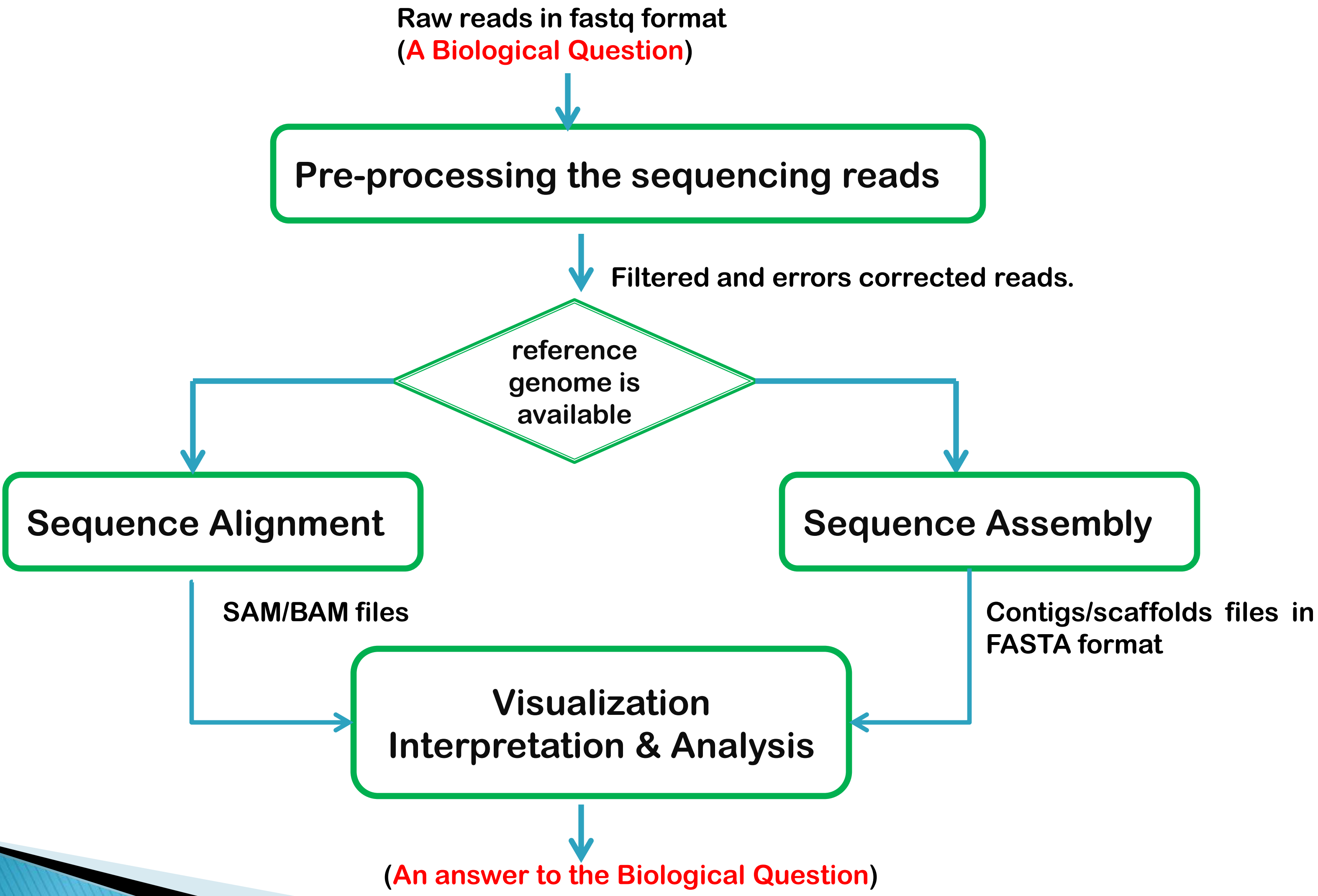
Informatics on High-throughput Sequencing Data

(Summer Course 2020)

Day 13

A typical Data Analysis pipeline



Raw reads in fastq format
(A Biological Question)

Pre-processing the sequencing reads

Filtered and errors corrected reads.

reference genome is available

Sequence Alignment

Sequence Assembly

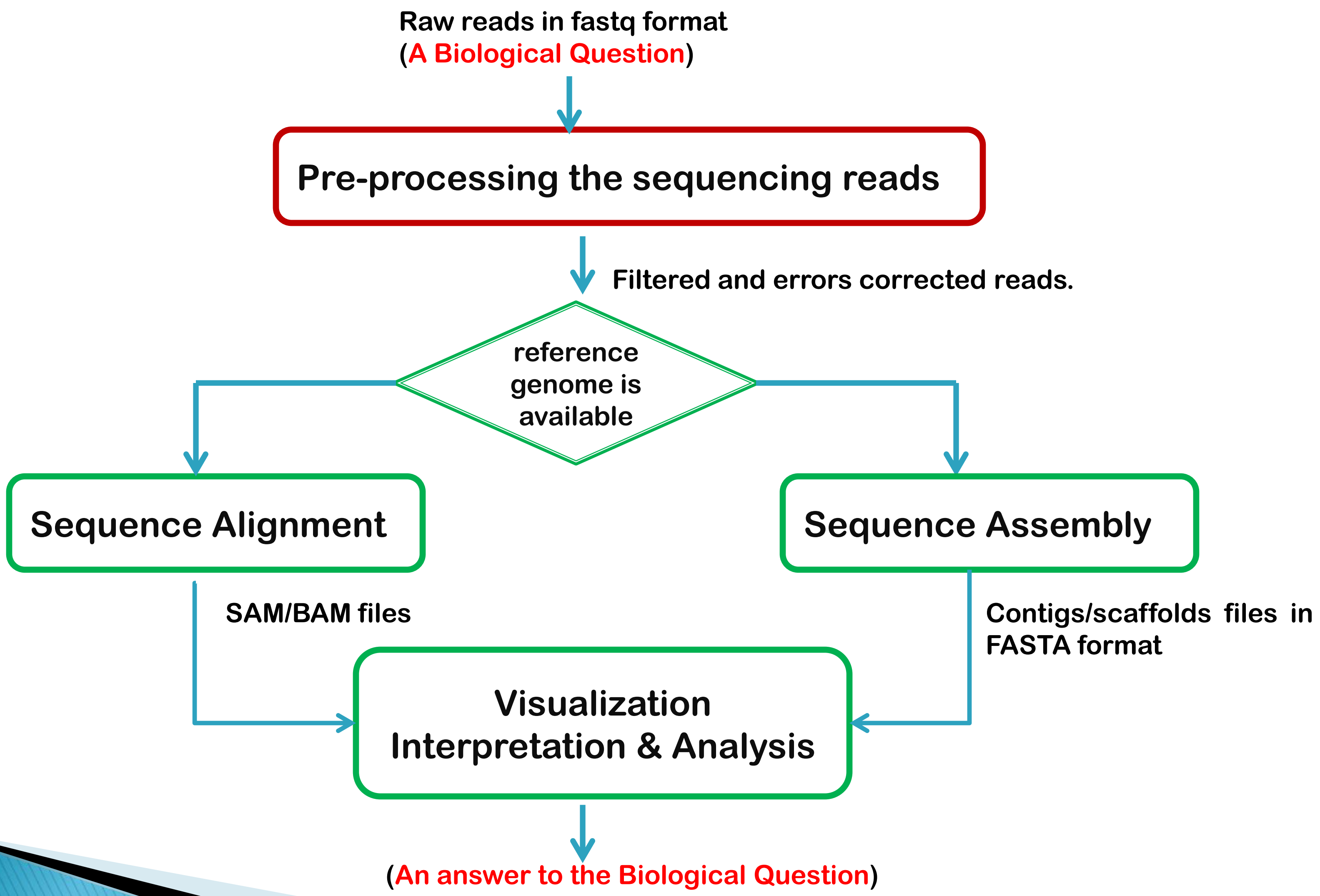
SAM/BAM files

Contigs/scaffolds files in FASTA format

Visualization
Interpretation & Analysis

(An answer to the Biological Question)

A typical Data Analysis pipeline



Raw reads in fastq format
(A Biological Question)

Pre-processing the sequencing reads

Filtered and errors corrected reads.

reference genome is available

Sequence Alignment

Sequence Assembly

SAM/BAM files

Contigs/scaffolds files in FASTA format

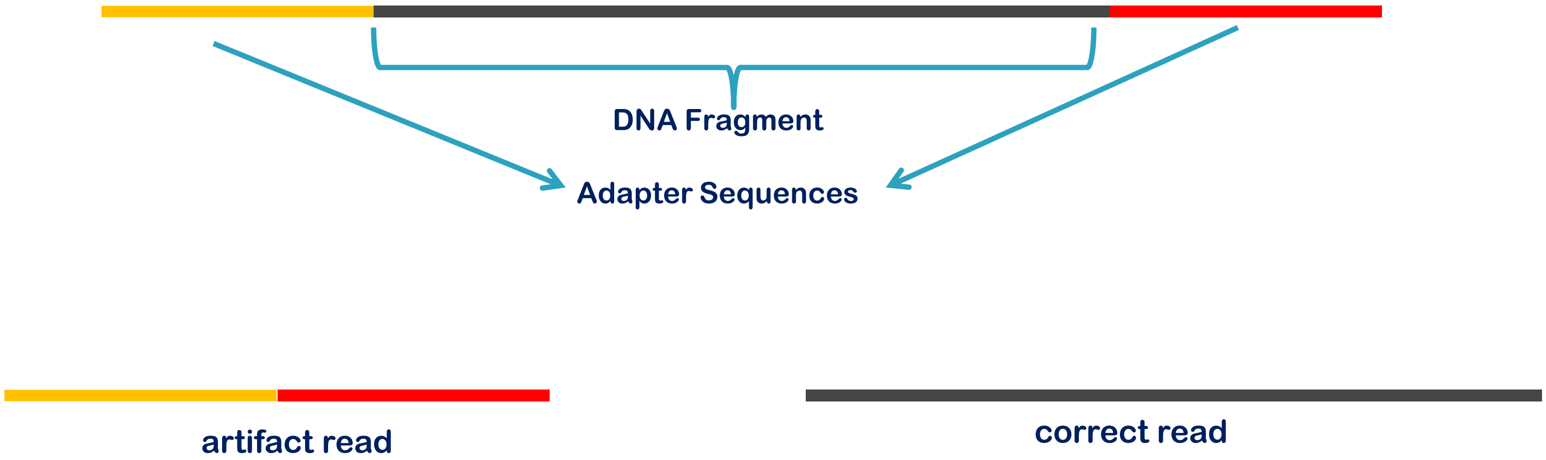
Visualization
Interpretation & Analysis

(An answer to the Biological Question)

Pre-processing reads

- Filter out garbage reads (Reads Trimming)
  - ✓ Reads with low quality base calls.
  - ✓ Reads that are clearly artifacts with chemistry.

A typical read for DNA sequencing process



DNA Fragment

Adapter Sequences

artifact read

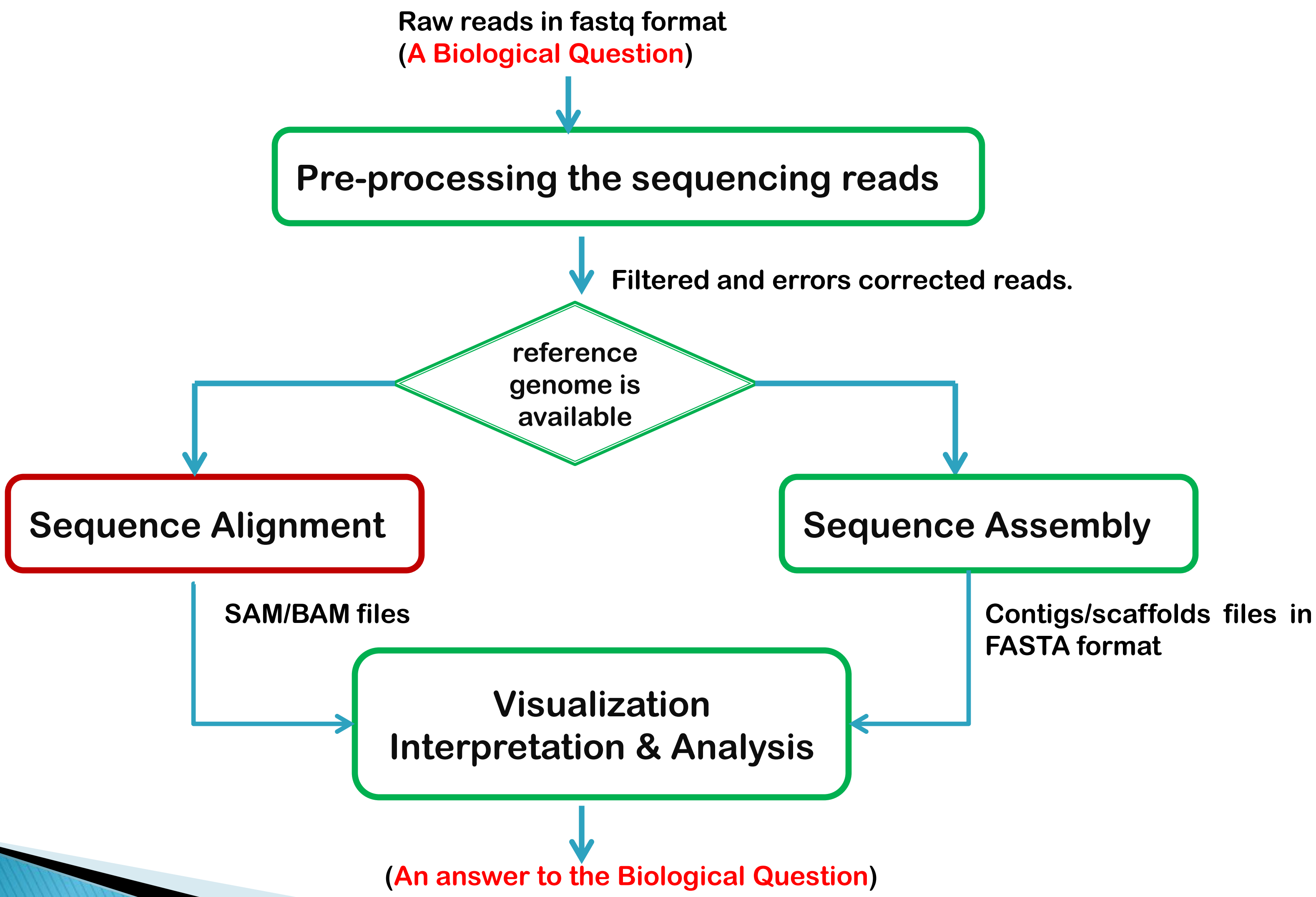
correct read

Pre-processing reads

- Filter out garbage reads (Reads Trimming)
  - ✓ FASTX
  - ✓ AfterQC
  - ✓ Trimmomatic

- Errors detection and correction
  - ✓ Quake
  - ✓ Lighter
  - ✓ Musket

A typical Data Analysis pipeline



Raw reads in fastq format
(A Biological Question)

Pre-processing the sequencing reads

Filtered and errors corrected reads.

reference genome is available

Sequence Alignment

Sequence Assembly

SAM/BAM files

Contigs/scaffolds files in FASTA format

Visualization
Interpretation & Analysis

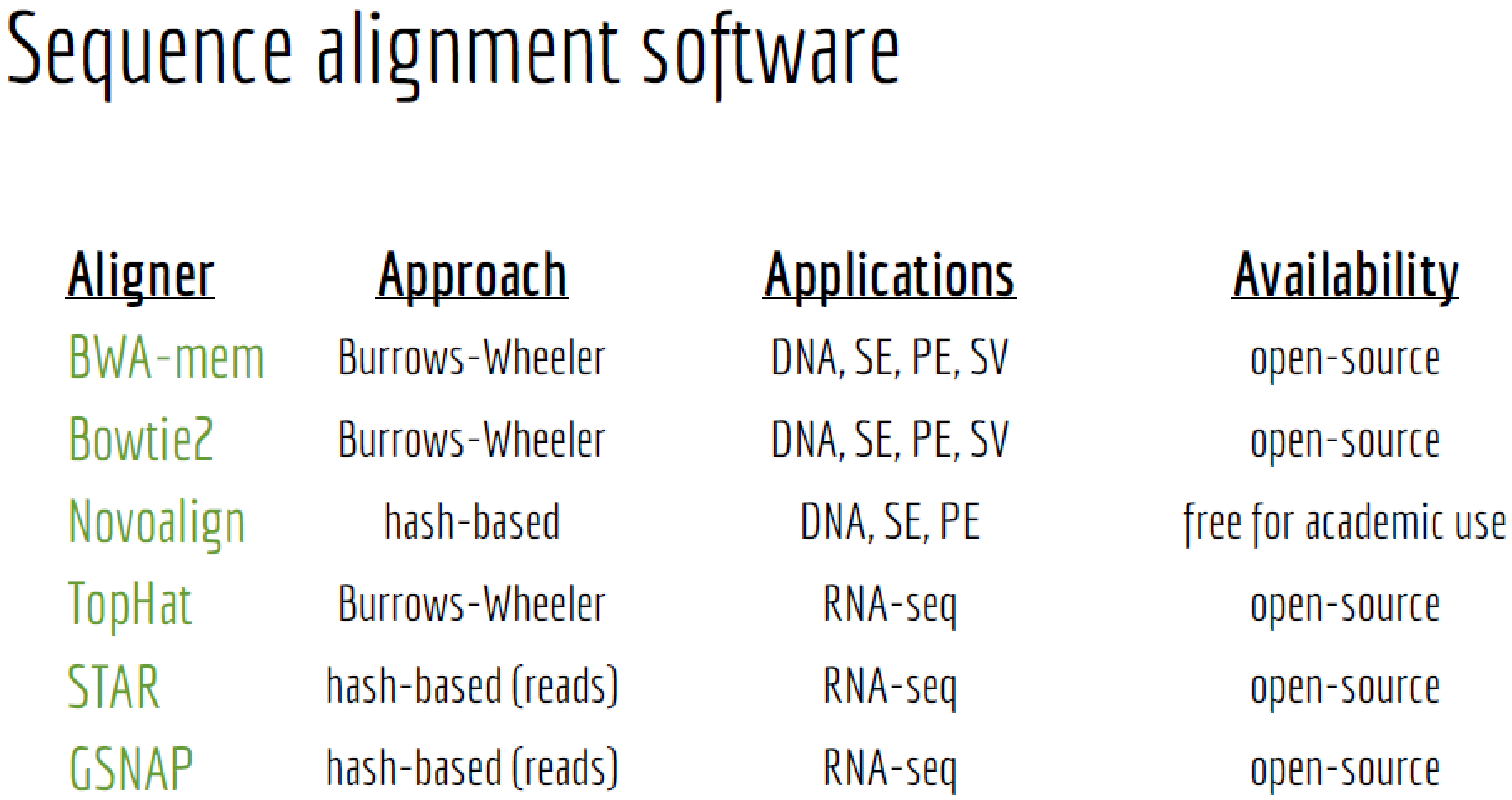
(An answer to the Biological Question)

Sequence alignment

Sequencing Reads

Reference Genome

```
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTT
CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTC
GCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATT
ACAGGCGAACATACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATA
ACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATCATAACAAAAAATTTCCACCA
AACCCCCCCTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAAAA
ACAAAGAACCCTAACACCAGCCTAACCAGATTTCAAATTTTATCTTTTGGCGGTATGCAC
TTTTAACAGTCACCCCCCAACTAACACATTATTTTCCCCTCCCACTCCCATACTACTAAT
CTCATCAATACAACCCCCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATA
CCCCGAACCAACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAA
GCAATACACTGACCCGCTCAAACTCCTGGATTTTGGATCCACCCAGCGCCTTGGCCTAAA
CTAGCCTTTCTATTAGCTCTTAGTAAGATTACACATGCAAGCATCCCCGTTCCAGTGAGT
TCACCCTCTAAATCACCACGATCAAAAGGAACAAGCATCAAGCACGCAGCAATGCAGCTC
AAAACGCTTAGCCTAGCCACACCCCCACGGGAAACAGCAGTGATTAACCTTTAGCAATAA
ACGAAAGTTTAACTAAGCTATACTAACCCCAGGGTTGGTCAATTTCGTGCCAGCCACCGC
GGTCACACGATTAACCCAAGTCAATAGAAGCCGGCGTAAAGAGTGTTTTAGATCACCCCC
TCCCCAATAAAGCTAAAACTCACCTGAGTTGTAAAAAACTCCAGTTGACACAAAATAGAC
TACGAAAGTGGCTTTAACATATCTGAACACACAATAGCTAAGACCCAAACTGGGATTAGA
TACCCCACTATGCTTAGCCCTAAACCTCAACAGTTAAATCAACAAAACTGCTCGCCAGAA
CACTACGAGCCACAGCTTAAAACTCAAAGGACCTGGCGGTGCTTCATATCCCTCTAGAGG
AGCCTGTTCTGTAATCGATAAACCCCGATCAACCTCACCACCTCTTGCTCAGCCTATATA
```

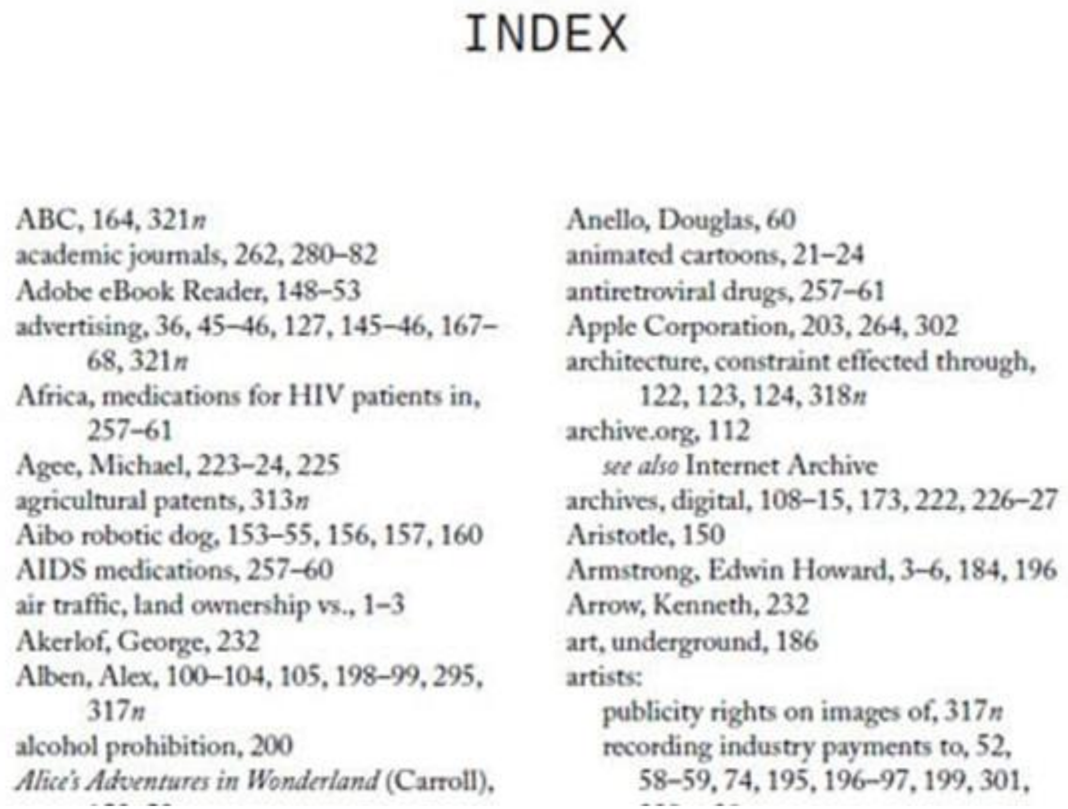


Sequence alignment software

Aligner	Approach	Applications	Availability
BWA-mem	Burrows-Wheeler	DNA, SE, PE, SV	open-source
Bowtie2	Burrows-Wheeler	DNA, SE, PE, SV	open-source
Novoalign	hash-based	DNA, SE, PE	free for academic use
TopHat	Burrows-Wheeler	RNA-seq	open-source
STAR	hash-based (reads)	RNA-seq	open-source
GSNAP	hash-based (reads)	RNA-seq	open-source

Slides curiosity from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

Genome Indexing



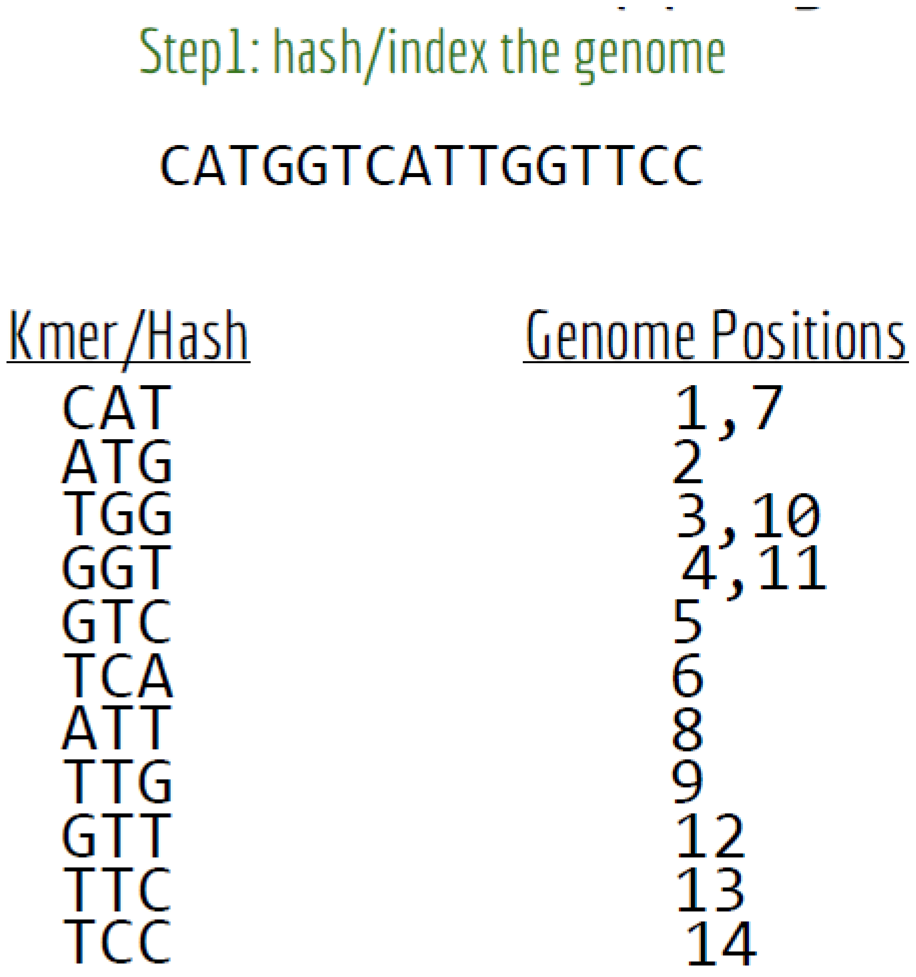
INDEX

ABC, 164, 321*n*
academic journals, 262, 280–82
Adobe eBook Reader, 148–53
advertising, 36, 45–46, 127, 145–46, 167–68, 321*n*
Africa, medications for HIV patients in, 257–61
Agee, Michael, 223–24, 225
agricultural patents, 313*n*
Aibo robotic dog, 153–55, 156, 157, 160
AIDS medications, 257–60
air traffic, land ownership vs., 1–3
Akerlof, George, 232
Alben, Alex, 100–104, 105, 198–99, 295, 317*n*
alcohol prohibition, 200
Alice's Adventures in Wonderland (Carroll),

Anello, Douglas, 60
animated cartoons, 21–24
antiretroviral drugs, 257–61
Apple Corporation, 203, 264, 302
architecture, constraint effected through, 122, 123, 124, 318*n*
archive.org, 112
see also Internet Archive
archives, digital, 108–15, 173, 222, 226–27
Aristotle, 150
Armstrong, Edwin Howard, 3–6, 184, 196
Arrow, Kenneth, 232
art, underground, 186
artists:
publicity rights on images of, 317*n*
recording industry payments to, 52, 58–59, 74, 195, 196–97, 199, 301,

https://www.pdfindexgenerator.com/what-is-a-book-index/

Genome Indexing



Step1: hash/index the genome

CATGGTCATTGGTTCC

Kmer/Hash	Genome Positions
CAT	1,7
ATG	2
TGG	3,10
GGT	4,11
GTC	5
TCA	6
ATT	8
TTG	9
GTT	12
TTC	13
TCC	14

Slides curiosity from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

Genome Indexing

Step2: use the index to find reads locations

Toy genome CATGGTCATTGGTTCC

Kmer/Hash	Genome Positions
CAT	1,7
ATG	2
TGG	3,10
GGT	4,11
GTC	5
TCA	6
ATT	8
TTG	9
GTT	12
TTC	13
TCC	14

→ Read TGGTCA

Slides curiosity from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

BWA-MEM

Unaligned
Sample Data
In FASTQ (SE or PE)

```
@seq1
ATTCGAAACA...
+
DDED88(999...
@seq2
CCCCGTTTCA...
+
AAC887BBAC...
```

Reference genome (FASTA)

```
>chr1
TACCTCCAGGGGGCATCCTCCCCCCCAATTCG
AAACACAATCGTAGCCCCTGGCACTACCTATG
TGTGTCAATTCGGAGAGAGAGAGATTCACGAA
AAAAAAGTCTGGACTCAACTAGGATACACACA
TTCGGCTACAGATACCAAAAAAAAAAAAAAAA
AAATTTTCACCATTGAGGCACCACCTTCTCGT
CGCTGCGTCGCTCTGCTCGCTTCGGCTAAAAA
TTCGCGCAATACATTCGGCTACAGATACCAAA
```

BWA MEM

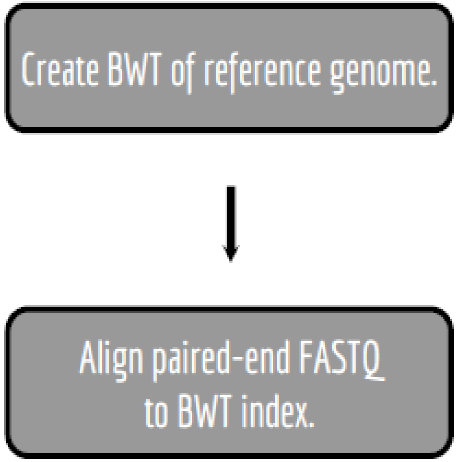
Aligned
Sample Data in
SAM format

```
seq1        99              1
            3666901         60
            149M            =
            3666935         185
            ATTCGAAACA...DDED88(999
            MC:Z:151M       MD:Z:149
            RG:Z:15-0017315_1
            NM:i:0          MQ:i:60
            AS:i:149        XS:i:44
seq2        147             1
            3666935         60
            151M            =
            3666901         -185
            CCCCGTTTCA...AAC887BBAC...
            MC:Z:149M       MD:Z:151
            RG:Z:15-0017315_1
            NM:i:0          MQ:i:60
            AS:i:151        XS:i:59
```

Slides curiosity from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

BWA-MEM workflow

This takes a long time, but you do it once



Create BWT of reference genome.

Align paired-end FASTQ to BWT index.

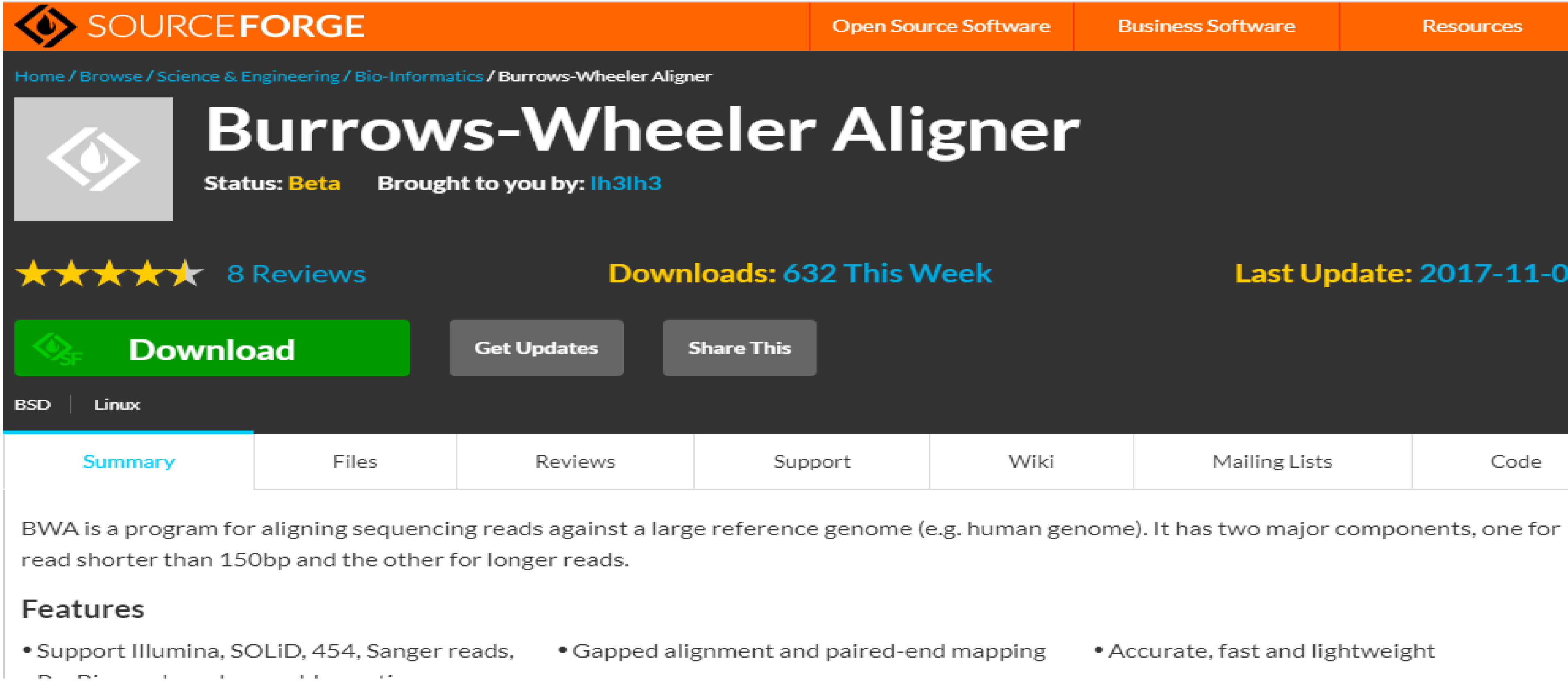
```
$ bwa index grch38.fa
```

Output is in SAM format. Use multiple threads if you have a computer with multiple CPUs.

```
$ bwa mem -t 16 grch38.fa 1.fq 2.fq > sample.sam
```

Slides curiosity from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

BWA-MEM



Burrows-Wheeler Aligner

Status: Beta **Brought to you by:** lh3lh3

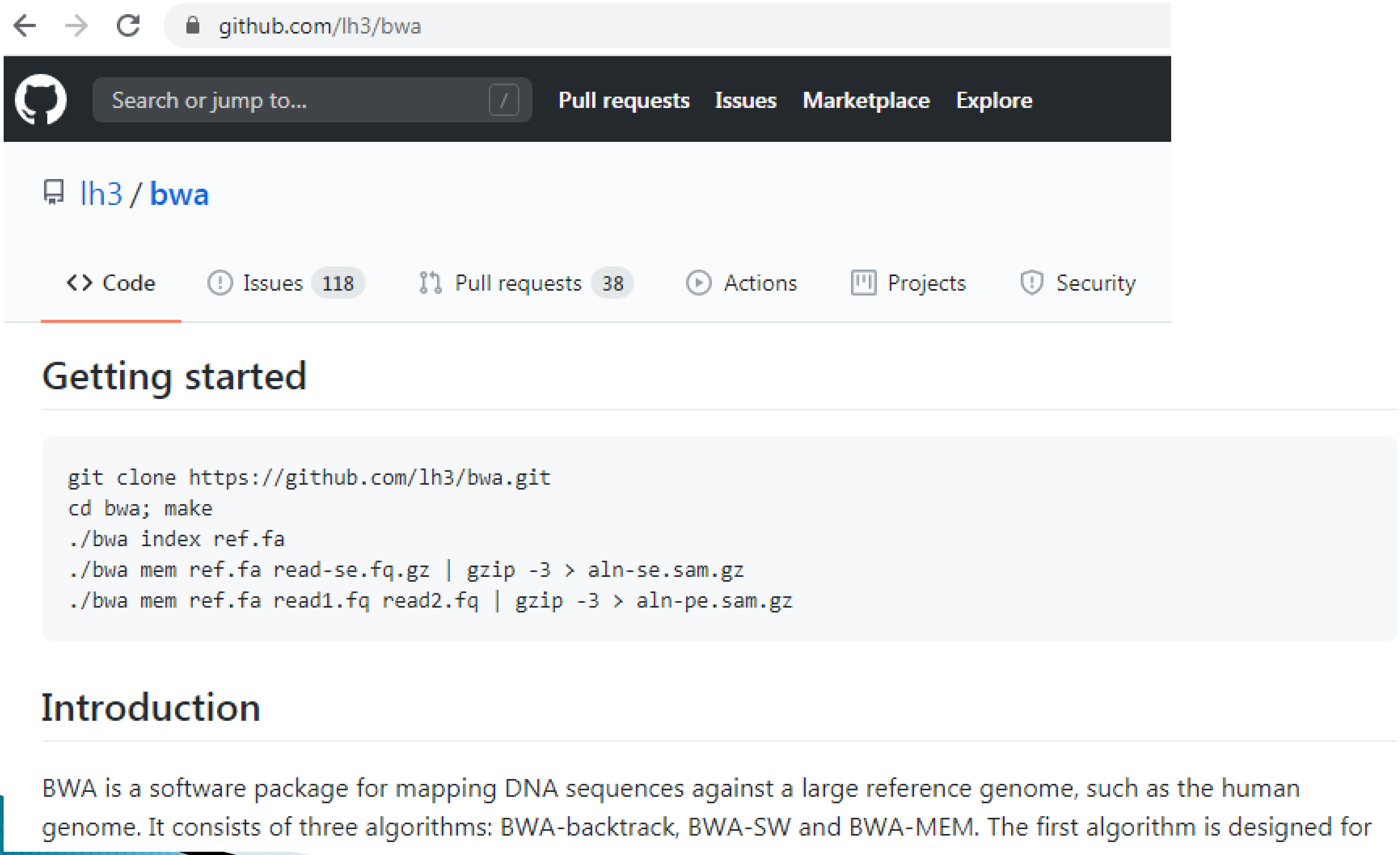
8 Reviews **Downloads:** 632 This Week **Last Update:** 2017-11-0

BWA is a program for aligning sequencing reads against a large reference genome (e.g. human genome). It has two major components, one for read shorter than 150bp and the other for longer reads.

Features

- Support Illumina, SOLiD, 454, Sanger reads,
- Gapped alignment and paired-end mapping
- Accurate, fast and lightweight

BWA-MEM



Getting started

```
git clone https://github.com/lh3/bwa.git
cd bwa; make
./bwa index ref.fa
./bwa mem ref.fa read-se.fq.gz | gzip -3 > aln-se.sam.gz
./bwa mem ref.fa read1.fq read2.fq | gzip -3 > aln-pe.sam.gz
```

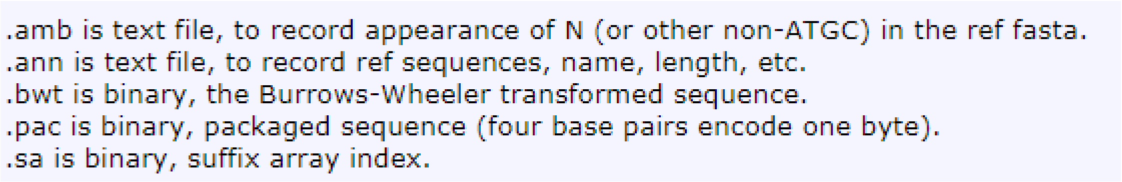
Introduction

BWA is a software package for mapping DNA sequences against a large reference genome, such as the human genome. It consists of three algorithms: BWA-backtrack, BWA-SW and BWA-MEM. The first algorithm is designed for

BWA-MEM

```
bzip2 -d bwa-0.7.17.tar.bz2
tar xvf bwa-0.7.17.tar
make
./bwa index  wu_0.v7.fas
```

- wu_0.v7.fas.amb
- wu_0.v7.fas.ann
- wu_0.v7.fas.bwt
- wu_0.v7.fas.pac
- wu_0.v7.fas.sa

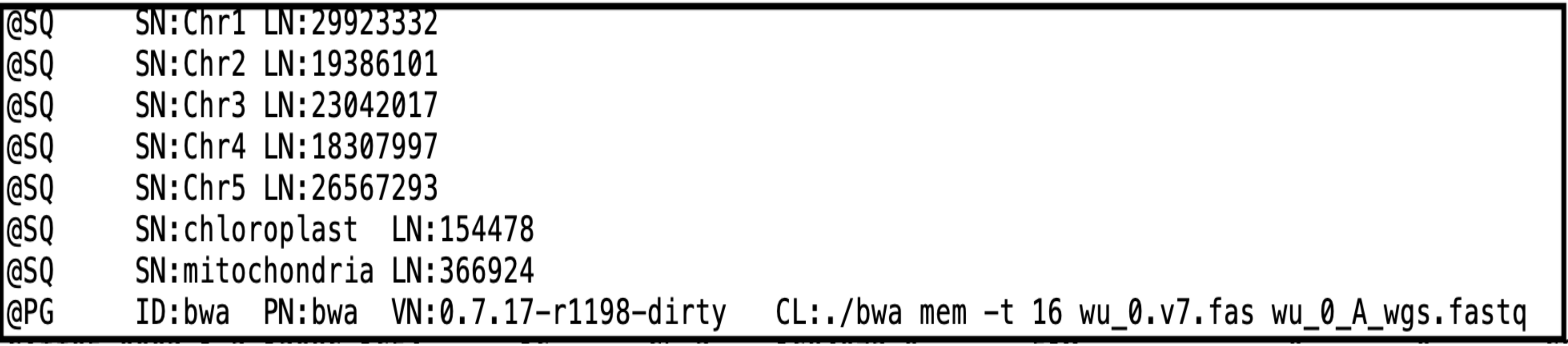


.amb is text file, to record appearance of N (or other non-ATGC) in the ref fasta.
.ann is text file, to record ref sequences, name, length, etc.
.bwt is binary, the Burrows-Wheeler transformed sequence.
.pac is binary, packaged sequence (four base pairs encode one byte).
.sa is binary, suffix array index.

```
./bwa mem -t 16 wu_0.v7.fas wu_0_A_wgs.fastq > results.sam
```

http://seqanswers.com/forums/showthread.php?t=25553

BWA-MEM



```
@SQ     SN:Chr1 LN:29923332
@SQ     SN:Chr2 LN:19386101
@SQ     SN:Chr3 LN:23042017
@SQ     SN:Chr4 LN:18307997
@SQ     SN:Chr5 LN:26567293
@SQ     SN:chloroplast  LN:154478
@SQ     SN:mitochondria LN:366924
@PG     ID:bwa  PN:bwa  VN:0.7.17-r1198-dirty   CL:./bwa mem -t 16 wu_0.v7.fas wu_0_A_wgs.fastq
```

SAM

Field	Meaning
GAII05_0002:1:2:12086:1654	Read ID
16	Flag
Chr2	Chr
1694072	start
0	MAPQ
51M	CIGAR
*	Mate Chr
0	Mate start
0	Mate dis
CCTTGTAAAATCATTATTAATGTTTTTAAACCCCTTTTAAAAATCCTTGTA	read
CCCCCCCCCCCCCCCCCCCBBCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	qual
NM:i:1 MD:Z:20C30 AS:i:46 XS:i:46	Tag-Type-Value

Thanks!

// | ?