Mansoura University
Faculty of Computers and Information
Department of Computer Science
Second Semester: 2020-2021

[MED-145] Genomics: Genome Indexing & Reads Mapping

Grade: Third Year (Medical Informatics Program)

Sara El-Metwally, Ph.D.

Faculty of Computers and Information,

Mansoura University,

Egypt.

COURSE OUTLINES

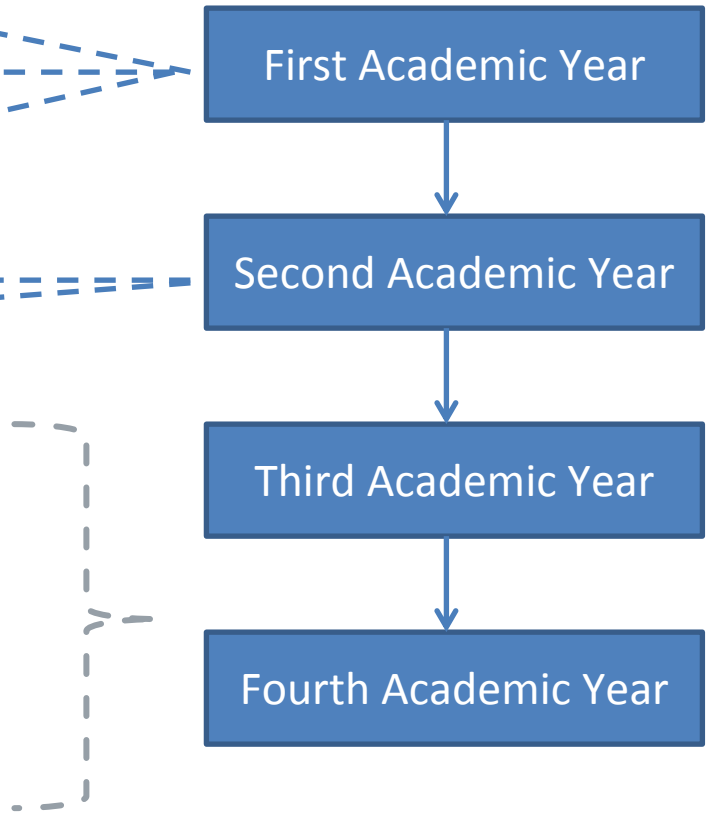
- **Course Meeting Time**: Monday 8:30-10:100, Wednesday 12:20-2.00.
- **Course Instructor**: Sara El-Metwally, PhD
- **Course TAs**: Eng. Nada El-Madah, Eng. Ola Magdy
- **Course Labs**: Unix-based commands, Shell Scripting, Python.
- **Course Grading**:
  - Midterm: 10%
  - Oral: 10%
  - Practical: 10%
  - Project / Paper: 10%
  - Final: 60%

MANSOURA FCIS COURSE DEPENDENCY

Basics of Programming Languages
Basics of Biology
Math & Statistics & Probabilities

Data Structures
Genetics

Bioinformatics, Genomics, Medical Informatics
Artificial Intelligence,
Machine Learning,
Algorithms analysis and design,
Big Data and Data mining techniques,
Distributed Systems,
Cloud Computing,
Pattern Recognition



COURSE PROJECTS

- **Choice No. 1**
  - Teams (up to 5 students).
  - Pick a project from the projects list or propose an idea for your project.
  - Projects outcomes:
    - A Project Proposal that describes the problem, how to solve it, the technologies/tools that you will use, team members and their tasks, etc.
    - A Github page that includes a Readme file that describes your project idea, algorithm, how to run and use the code and any useful links etc. and your project source code with any dependency.
    - Your proposal should be added to your Github page.
    - A Video demo that describes your project (English), the link should be added to your Github page and the video should be uploaded to our course channel on YouTube!
    - Group Photo with a faculty logo.
    - There is a competition among different genomics projects; the top best five projects will awarded a genomics course certificate plus some other prizes in the case of extraordinary projects.
    - Think big drive forward!

COURSE PROJECTS

Genomics Course List Projects 2021	
Project Name	**Example of already existing tools in the field**
Genome/Transcriptome Browsers	UCSC Genome Browser
Genome/Transcriptome Assemblers	Velvet, Canu, LightAssembler, Trinity, SPAdes
Multiple Sequence Aligners (DNA, RNA, Amino Acids)	Clustal Omega, MAFFT, MUSCLE
Local alignment tool	BLAST
Short/long reads aligner to a reference genome	Bowtie, BWA
Phylogenetic Trees Drawing/Analysis Tool	iTOL, phylot, PhyML
Variant Calling Programs	GATK, SAMtools, FreeBayes, DeepVariant
Errors Correction Programs	Bless, Musket, Lighter, Fiona,NanoReviser, MARVEL
Efficient kmers counting tools	Jellyfish, KMC, DSK
Quality control of sequencing reads (short/long)	FastQC, FASTX-Toolkit, LongQC
Assembly Evaluation Software	QUAST
Fastq compression tool	MZPAQ, fqzcomp, Spring
Your Idea?	**Your Program**

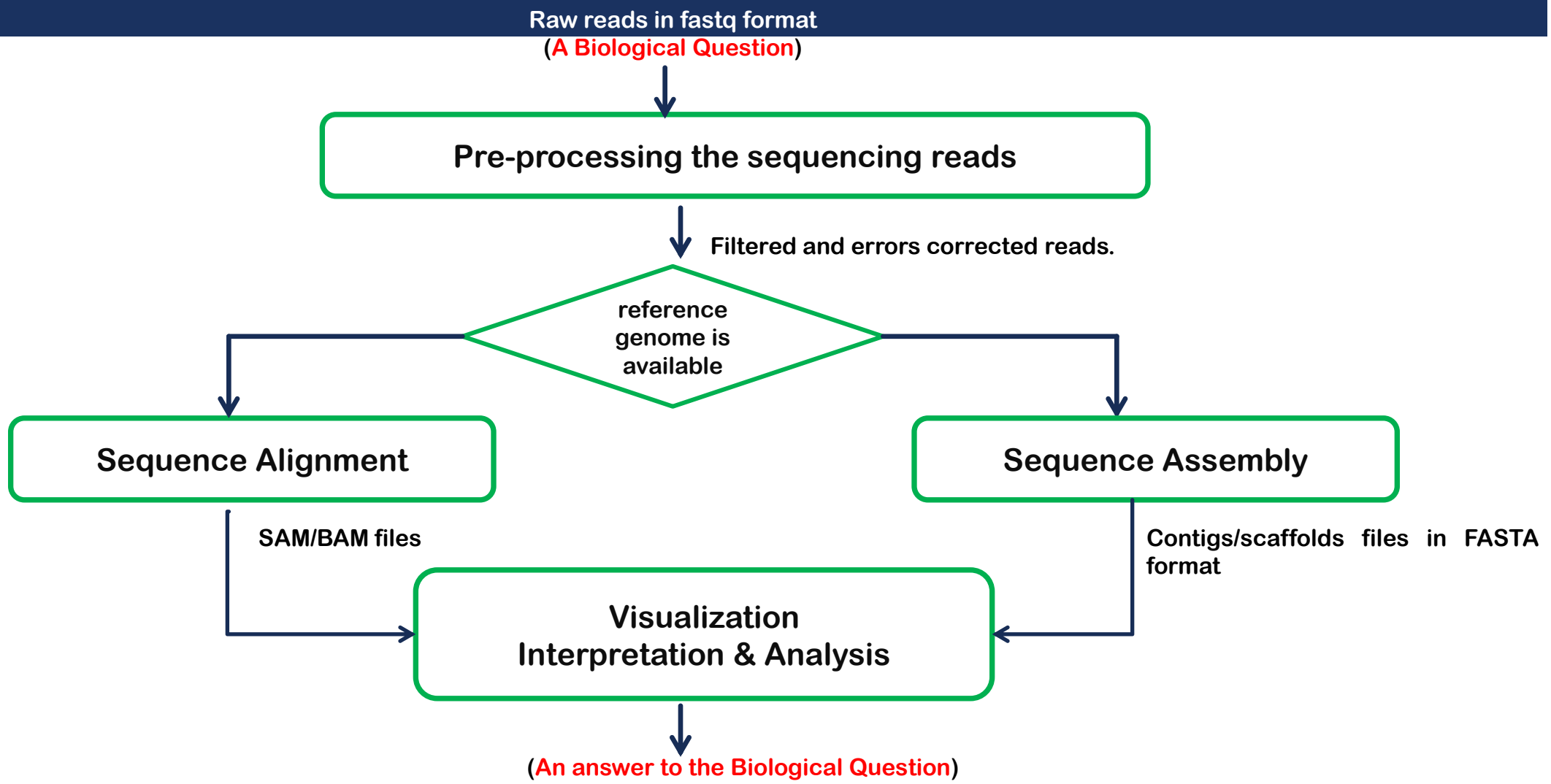
COURSE PROJECTS

- **Choice No. 2**
  - One student.
  - Pick a paper published in 2020/2021 in the genomics field in general or related to SARS-CoV-2 data analysis.
  - Student outcomes:
    - A document that summarizes the paper, the analysis pipeline, the findings and your comment on the paper results (Max. 3 pages).
    - A Video that explains the paper idea using the presentation prepared by you.
    - Student Photo with a faculty logo.
    - A Github page that includes a Readme file that describes the paper idea, data analysis, etc. including the paper reference and your created video along with your photo.

AGENDA

- A Typical Genomics Analysis Workflow.
- Alignment is central to most genomic research.
- Sequence mapping versus alignment.
- Reference based analysis mapping and challenges.
- Mapping Quality.
- Sequence alignment/Mapping software.
- Typical Mapping/Alignment Workflow.
- Hash-based mapping approach.
- Hash-based mapping approach drawbacks.

A TYPICAL GENOMICS DATA ANALYSIS PIPELINE



ALIGNMENT IS CENTRAL TO MOST GENOMIC RESEARCH

One fastq file for single-end reads

FASTQ format → Align to genome → Quality control →

- Find peaks
- Quantify transcripts
- Find SVs
- .
- .
- .
- etc.

→

Analysis, Interpretation, Discovery

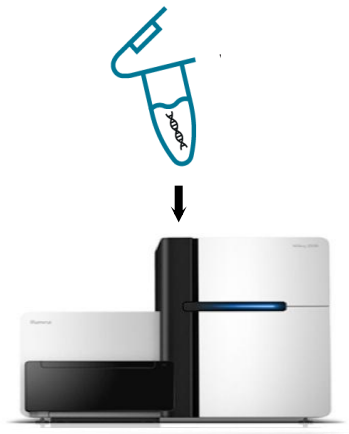
Two fastq files for paired-end reads

1. *Staphylococcus aureus*: **Data download page**
   - Library 1: Fragment
     - Avg Read length: 101bp
     - Insert length: 180bp
     - # of reads: 1,294,104
     - Fastq read file 1
     - Fastq read file 2

Phase 0 | Phase1 | Phase 2-100

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

BEST CASE SCENARIO: AN ERROR-FREE SEQUENCING TECHNOLOGY



→ ATTCGAAACA
TTCGCGCAAT
CTGGACTCAA

→ Aligner →

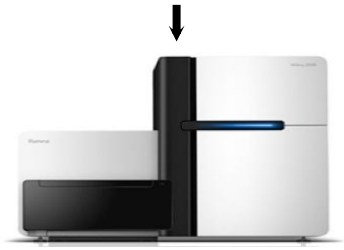
TACCTCCAGGGGGCATCCTCCCCCCCAATTCGAAA
CACAATCGTAGCCCCTGGCACTACCTATGTGTGTC
AATTCGGAGAGAGAGAGATTCACGAAAAAAAAGTC
TGGACTCAACTAGGATACACACATTCGGCTACAGA
TACCAAAAAAAAAAAAAAAAAAATTTTCACCATTG
AGGCACCACCTTCTCGTCGCTGCGTCGCTCTGCTC
GCTTCGGCTAAAAATTCGCGCAATACATTCGGCTA
CAGATACCAAAAAAA

Computers are rather good at finding ***exact*** matches.
Think Google.

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

REALITY CHECK. ERRORS HAPPEN. FREQUENTLY.

ATTCGAAACA



→ ATTTGAAACA → Aligner →

TACCTCCAGGGGGCATCCTCCCCCCCAATTCGAAAC
ACAATCGTAGCCCCTGGCACTACCTATGTGTGTCAA
TTCGGAGAGAGAGAGATTGGAAACAAAAAAGTGCTA
CAGATACCACTAGGATACACACATTCGGCTACAGAT
ACCAAAAAAAAAAAAAAAAAAATTTTCACCATTGAG
GCACCACCTTCTCGTCGCTGCGTCGCTCTGCTCGCG
GCTAAAAAATTAGAAACAACATTCGGCTACAGATAC
CAAAATTT

"Fuzzy" matching is much more computationally expensive.
Think Google's "Did you mean..."

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

SEQUENCE MAPPING VERSUS ALIGNMENT

Mapping: (quickly) find the best possible loci to which a sequence could be aligned.

Alignment: for each locus to which a sequence can be mapped, determine the optimal base by base alignment of the query sequence to the reference sequence.

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

SEQUENCE MAPPING VERSUS ALIGNMENT

ATTCGAAACA

→ ATTGGAAACA → Mapper →

Chr1: 1200

```
>Chr1
TACCTCCAGGGGGCATCCTCCCCCCCAATTGCAAAGACAACG
TAGCCCCTGGCACTACCTATGTGTGTCAA
```

Chr4: 500

```
>Chr 4
TTCGGAGAGAGAGAGATTGGACCCAAAAAAGTGCTACAGATA
CCACTAGGATACACACATTCGGCTACAGATACCAAAAAAAAA
AAAAAAAAAACACCAGAG
```

```
> Chr 10
GCACCACCTTCTCGTCGCTGCGTCGCTCTGCTCGCGGCTAAA
AAATTGGAAACAACATTCGGCTACAGATACCAAA
```

Chr10: 3400

- find possible candidate mapping locations (kmers, subsequences, etc.)

SEQUENCE MAPPING VERSUS ALIGNMENT

ATTCGAAACA

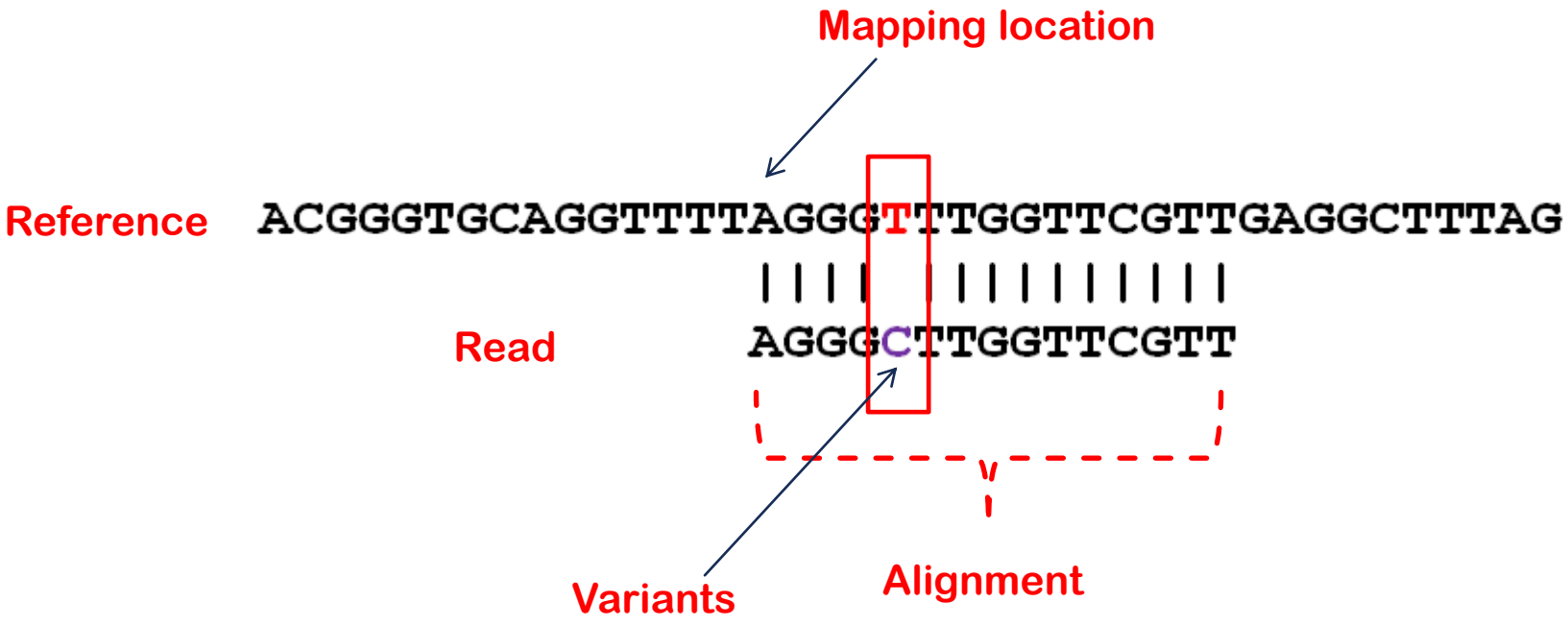
→ ATTGGAAACA → Aligner →

```
>Chr1                                                    # mismatch = 1
TACCTCCAGGGGGCATCCTCCCCCCCAATTGGAAAGACAACG
TAGCCCCTGGCACTACCTATGTGTGTCAA

>Chr 4                                                   # mismatch = 2
TTCGGAGAGAGAGAGATTGGACCCAAAAAAGTGCTACAGATA
CCACTAGGATACACACATTCGGCTACAGATACCAAAAAAAAA
AAAAAAAAAACACCAGAG

> Chr 10
GCACCACCTTCTCGTCGCTGCGTCGCTCTGCTCGCGGCTAAA
AAATTGGAAACAACATTCGGCTACAGATACCAAA
```

mismatch = 0

- find alignment score for each candidate alignment.

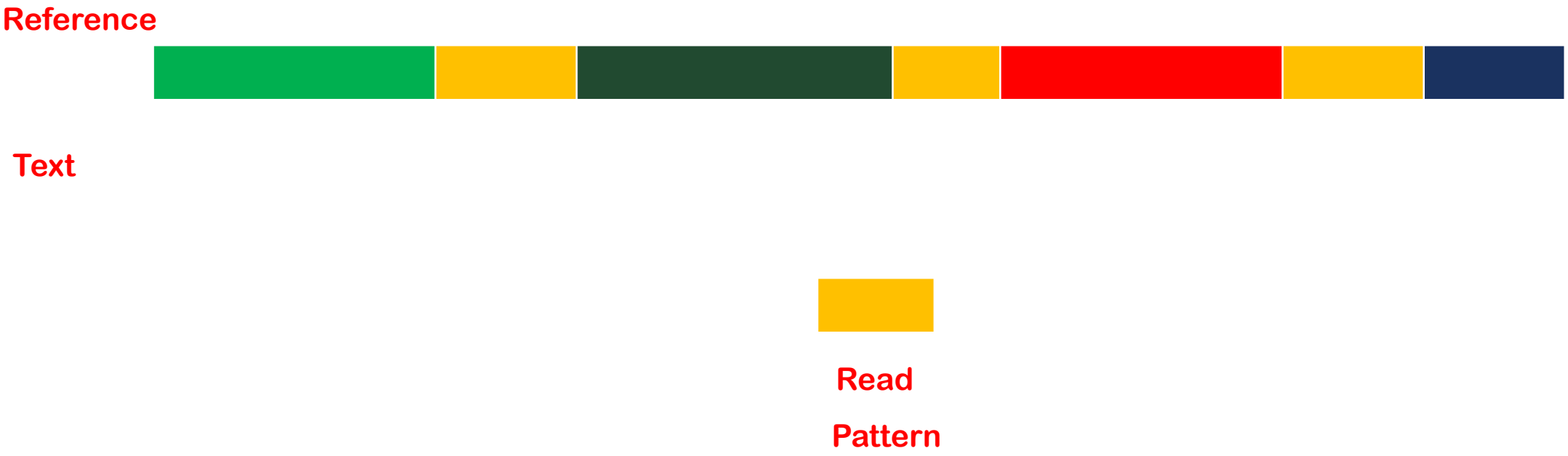
REFERENCE-BASED ANALYSIS



- ✓ Mapping for long reads, aligning for short reads, or used interchangeably.
- ✓Discover genetic variations by comparing reads to a reference genome.
- ✓To do this, the best mapping positions between reads and the reference should be identified (Some Challenges will be here!).

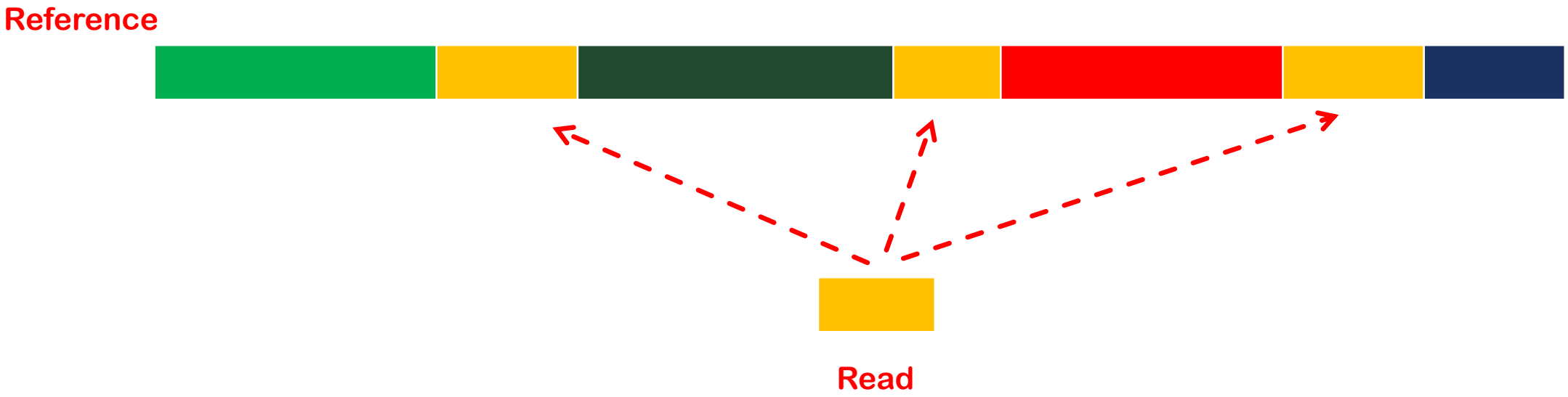
REFERENCE MAPPING CHALLENGES

- Genomes are very large (3 billion bases in human) and have repetitive regions.
- Naïve algorithms would take too much time and memory to map reads to a reference



REFERENCE MAPPING CHALLENGES

- Genomes are very large (3 billion bases in human) and have repetitive regions.



REFERENCE MAPPING CHALLENGES

❖ Reads have sequencing errors (substitutions, insertions, and deletions).

```
Reference   ACGGGTGCAGGTTTTAGGGTTTGGTTCGTTGAGGCTTTAG
                             |||| |||||||||||
Read                         AGGGCTTGGTTCGTT
```

```
Reference   ACGGGTGCAGGTTTTAGGGTTTGG----TTGAGGCT
                             |||   ||||     ||
Read                         AGG--TTGGTTCGTT
```

✓Mappers must able to find inexact alignments by tolerating differences.
✓Substitutions are more tolerance than Indels.

REFERENCE MAPPING CHALLENGES

❖ Reads could be mapped to many locations across the genome, which one will be reported?

```
Chr10:1020      AGGGACCGGTTCGTTTAGGGTTTGGTTCGTTGAGGCTTTAG
                                    |||| |||||||||||
                mismatches : 1      AGGGATTGGTTCGTT
```

Start from the base/offset/position no. 2139 in Chr2

Chr2

```
Chr2:2139       AGGGACCGGTTCGTTTAGGGTTTGGTTCGTTGAGGCTTTAG
                |||||   ||||||||
                AGGGATTGGTTCGTT     mismatches : 2
```

REFERENCE MAPPING CHALLENGES

- Reads could be mapped to many locations across the genome, which one will be reported?
  - MQ is an estimation of the probability that a mapping is incorrect (it encodes many factors such as the number of mismatches, type of mismatch, quality scores, etc.)

```
Chr10:1020      AGGGACCGGTTCGTTTAGGGTTTGGTTCGTTGAGGCTTTAG
                                |||| |||||||||||
                Mapping Quality : 10   AGGGATTGGTTCGTT

Chr2:2139       AGGGACCGGTTCGTTTAGGGTTTGGTTCGTTGAGGCTTTAG
                |||||   ||||||||
                AGGGATTGGTTCGTT   Mapping Quality : 1
```

MAPPING QUALITY (MAPQ)

- What is the probability that the sequence should be mapped here and only here?
- MAPQ also uses the Phred (log) scale:

$$MAPQ = -10*\log_{10}(P_{map\_loc\_wrong})$$

$(P_{map\_loc\_wrong})$	$\log_{10}(P_{map\_loc\_wrong})$	MAPQ
1	0	0
0.1	-1	10
0.01	-2	20
0.001	-3	30
0.0001	-4	40

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

REFERENCE MAPPING CHALLENGES

- Reads could be mapped to many locations across the genome, which one will be reported?
  - Low quality mismatches are less important than high quality mismatches.

```
Chr10:1020     AGGGACCGGTTCGTTTAGGGTTTGGTTCGTTGAGGCTTTAG
                              |||| |||||||||||
                              AGGGATTGGTTCGTT
                              Q=15

Chr2:2139      AGGGACCGGTTCGTTTAGGGTTTGGTTCGTTGAGGCTTTAG
               ||||| |||||||||||
               AGGGATCGGTTCGTT
               Q=40
```

REFERENCE MAPPING CHALLENGES

- Reads could be mapped to many locations across the genome, which one will be reported?
  - When two mappings have the same exact alignment, the mapping is ambiguous. Two positions are equal? Which one is correct?

Chr10:1020

```
AGGGTTTGGTTCGTTTTAGGGTTTGGTTCGTTGAGGCTTTAG
|||| |||||||||||
AGGGATTGGTTCGTT
```

Q=40

Chr2:2139

```
AGGGATCGGTTCGTTTAGGGTTTGGTTCGTTGAGGCTTTAG
||||| |||||||||
AGGGAACGGTTCGTT
```

Q=40

REFERENCE MAPPING CHALLENGES

- Reads could be mapped to many locations across the genome, which one will be reported?
  - Choose one mapping position at random, MQ=0

Chr10:1020

```
AGGGTTTGGTTCGTTTTAGGGTTTGGTTCGTTGAGGCTTTAG
|||| |||||||||||
AGGGATTGGTTCGTT
```

Q=40

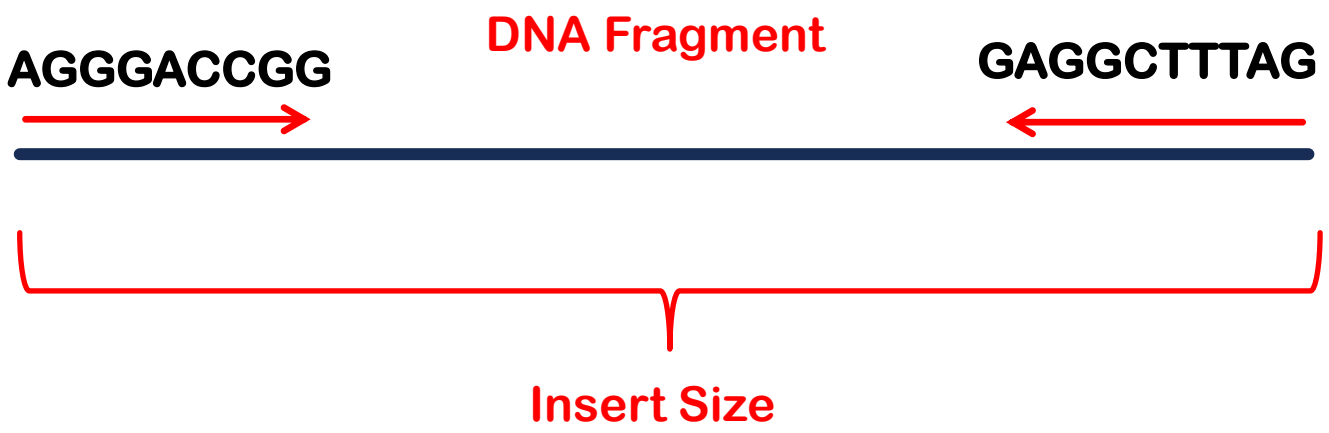
Chr2:2139

```
AGGGATCGGTTCGTTTAGGGTTTGGTTCGTTGAGGCTTTAG
||||| |||||||||
AGGGAACGGTTCGTT
```

Q=40

REFERENCE MAPPING CHALLENGES

- Reads could be mapped to many locations across the genome, which one will be reported?
  - Paired-end reads can help in resolving ambiguous mappings.



REFERENCE MAPPING CHALLENGES

- Reads could be mapped to many locations across the genome, which one will be reported?

Chr10:1020

```
AGGGTTTGGTTCGTTTTAGGGTTTGGTTCGTTGAGGCTTTAG
|||| |||||||||||
AGGGATTGGTTCGTT
```

Chr2:2139

```
AGGGATCGGTTCGTTTAGGGTTTGGTTCGTTGAGGCTTTAG
||||| |||||||||
AGGGAACGGTTCGTT
```

REFERENCE MAPPING CHALLENGES

❖ Reads could be mapped to many locations across the genome, which one will be reported?

Chr10:1020

```
AGGGTTTGGTTCGTTTTAGGGTTTGGTTCGTTGAGGCTTTAG
||||  ||||||||||           ||||||||||  ||||
AGGGATTGGTTCGTT ———————— GGTTCGTTGAAGCTT
```

Chr2:2139

```
AGGGATCGGTTCGTTTAGGGTTTGGTTCGTTGAGGCTTTAG
|||||  |||||||||              ||  |     |  |   ||
AGGGAACGGTTCGTT ———————— TCTTCGTTGATGCAG
```

SEQUENCE ALIGNMENT/MAPPING 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

https://academic.oup.com/bioinformatics/article/28/24/3169/245777

TYPICAL MAPPING/ALIGNMENT WORKFLOW

Reference genome (FASTA)

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

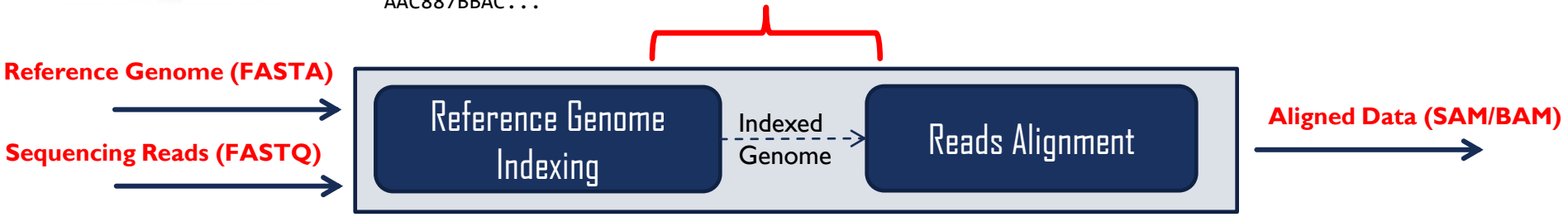
Aligned Sample Data in SAM format

```
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
```

Unaligned Sample Data In FASTQ (SE or PE)

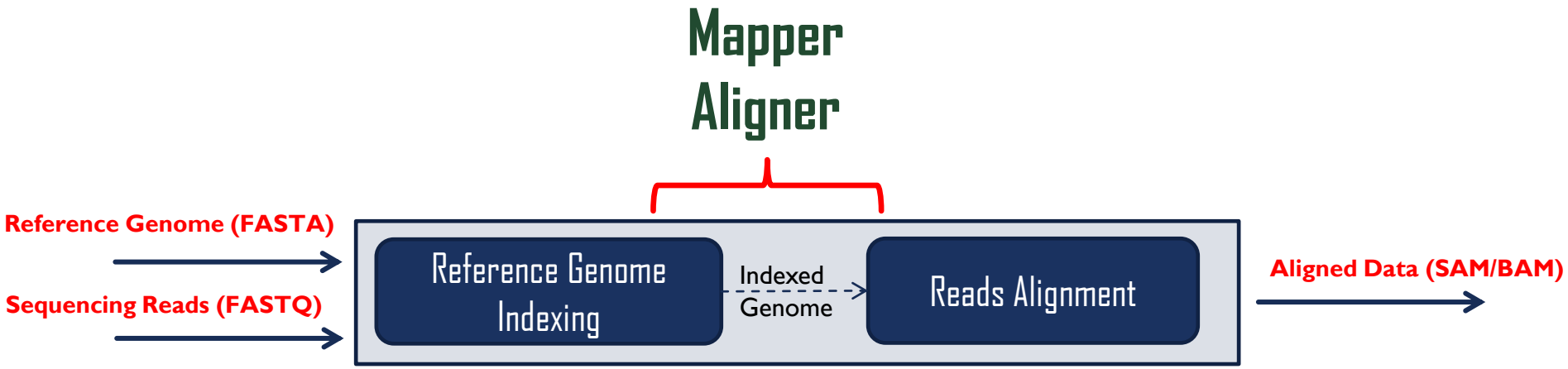
```
@seq1
ATTCGAAACA...
+
DDED88(999...
@seq2
CCCCGTTTCA...
+
AAC887BBAC...
```

Mapper Aligner



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

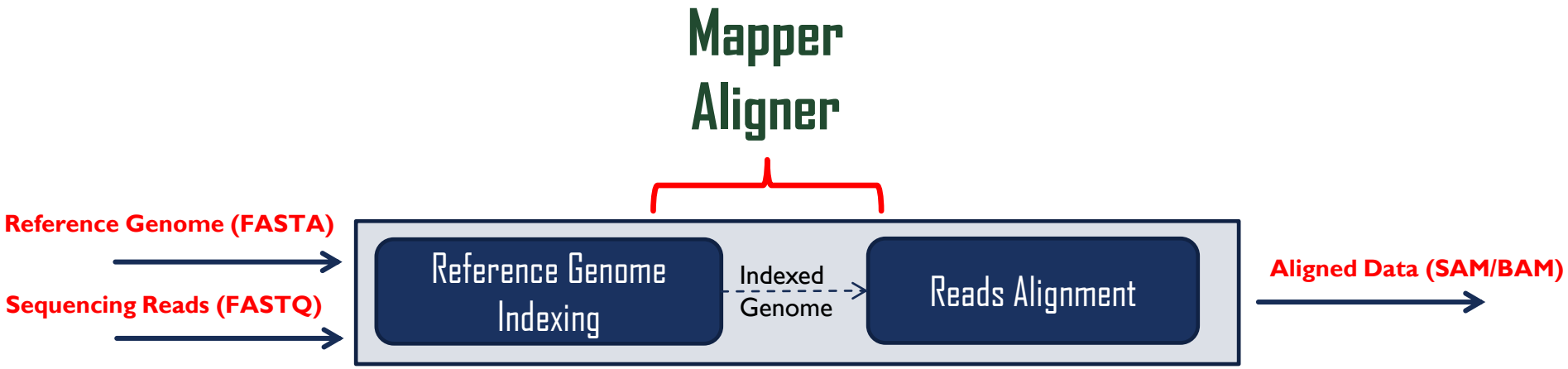
TYPICAL MAPPING/ALIGNMENT WORKFLOW



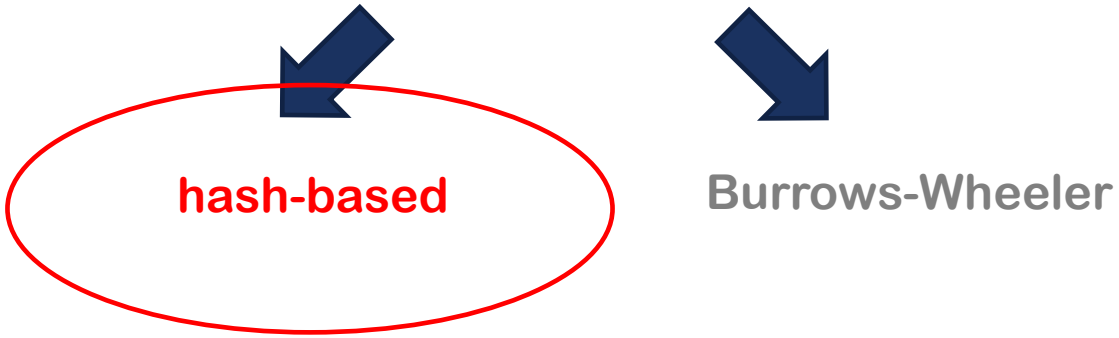
Genome Indexing and Mapping Approaches

hash-based **Burrows-Wheeler**

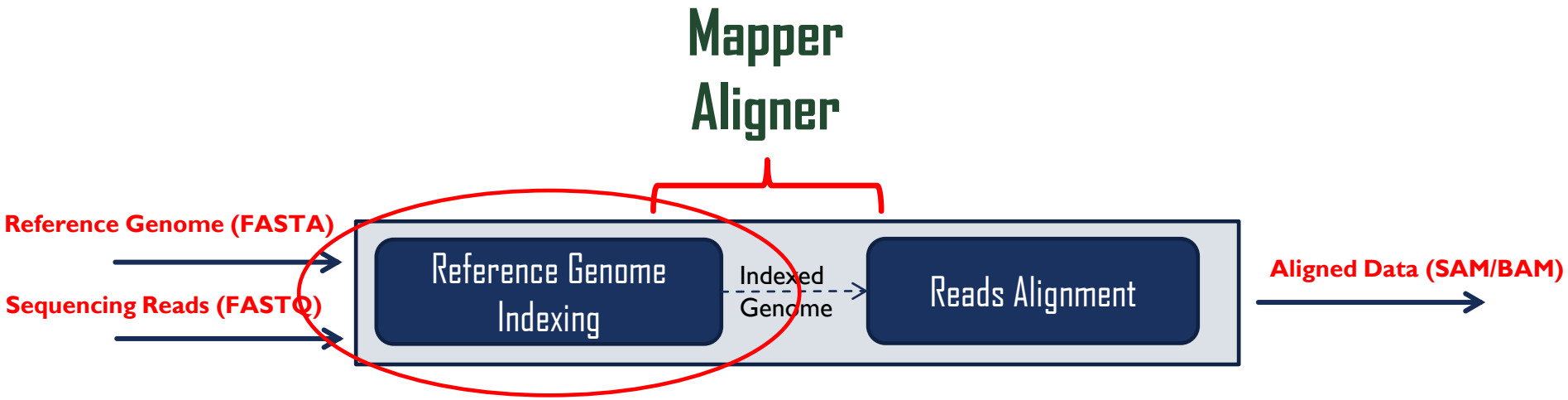
TYPICAL MAPPING/ALIGNMENT WORKFLOW



Genome Indexing and Mapping Approaches



TYPICAL MAPPING/ALIGNMENT WORKFLOW



Genome Indexing and Mapping Approaches

HASH-BASED MAPPING:

Step1: hash/index the genome

Toy genome (16 bp) CATGGTCATTGGTTCC

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

HASH-BASED MAPPING:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3	Kmer/Hash	Genome Positions
	CAT	1

Could be zero-based indexing

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

HASH-BASED MAPPING:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3

Kmer/Hash	Genome Positions
CAT	1
ATG	2

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

HASH-BASED MAPPING:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3

Kmer/Hash	Genome Positions
CAT	1
ATG	2
TGG	3

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

HASH-BASED MAPPING:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3

Kmer/Hash	Genome Positions
CAT	1
ATG	2
TGG	3
GGT	4

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

HASH-BASED MAPPING:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3

Kmer/Hash	Genome Positions
CAT	1
ATG	2
TGG	3
GGT	4
GTC	5

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

HASH-BASED MAPPING:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3

Kmer/Hash	Genome Positions
CAT	1
ATG	2
TGG	3
GGT	4
GTC	5
TCA	6

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

HASH-BASED MAPPING:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3

Kmer/Hash	Genome Positions
CAT	1,7
ATG	2
TGG	3
GGT	4
GTC	5
TCA	6

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

HASH-BASED MAPPING:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
C	A	T	G	G	T	C	A	T	T	G	G	T	T	C	C

Step1: hash/index the genome

CATGGTCATTGGTTCC

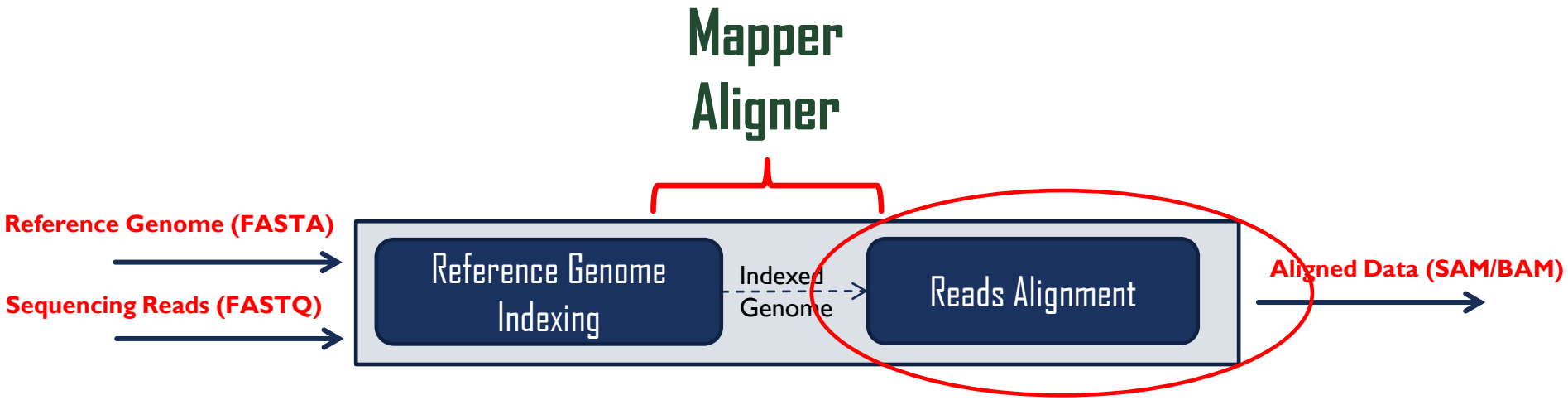
Complete hash/kmer index of our toy genome (forward strand only), k=3

Kmer/hash	Genomic position
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

Genome Index

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

TYPICAL MAPPING/ALIGNMENT WORKFLOW



Genome Indexing and Mapping Approaches

HASH-BASED MAPPING:

Step2: use the index to map (i.e., find alignment locations) reads

→ Read TGGTCA

Genome Index

Kmer/hash	Genomic position
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

HASH-BASED MAPPING:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
C	A	T	G	G	T	C	A	T	T	G	G	T	T	C	C

Step2: use the index to map (i.e., find alignment locations) reads.

Read TGGTCA

Hash match

TGG	3,10

Genome Index

Kmer/hash	Genomic position
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

HASH-BASED MAPPING:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
C	A	T	G	G	T	C	A	T	T	G	G	T	T	C	C

Step2: use the index to map (i.e., find alignment locations) reads.

Read TGGTCA

Hash match

TGG	3,10
TCA	6

Genome Index

Kmer/hash	Genomic position
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

HASH-BASED MAPPING:

Step2: use the index to map (i.e., find alignment locations) reads.

→ Read TGGTCA

TGG	3,10
TCA	6

3+ # skipped kmers + 1 =
3+2+1=6

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
C	A	T	G	G	T	C	A	T	T	G	G	T	T	C	C

HASH-BASED MAPPING:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
C	A	T	G	G	T	C	A	T	T	G	G	T	T	C	C

Step2: use the index to map (i.e., find alignment locations) reads

→ Read TGGTCT

{3,10} ?

T is an erroneous bases ?

Alignments will be tolerated to mismatches

Genome Index

Kmer/hash	Genomic position
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

EDIT DISTANCE (LEVENSHTEIN DISTANCE)

How many edits (changes) must be made to a word or kmer to make it match (align) to another word or kmer?

What is the difference between Edit distance & Hamming distance?

CURLED
HURLED → Edit distance = 1. Substitute C for H

SHORT
SHO-T → Edit distance = 1. Delete R

TGTTACGG
GGTTGACTA
?

TG-TT-ACGG
-GGTTGACTA

Edit distance = 5

TGTT-ACGG
GGTTGACTA

Edit distance = 4

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

HASH-BASED MAPPING:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
C	A	T	G	G	T	C	A	T	T	G	G	T	T	C	C

Step2: use the index to map (i.e., find alignment locations) reads

→ Read TGGTCT

{3,10} ?

Or :T is a True variation on the read and should be reported and studied.

Genome Index

Kmer/hash	Genomic position
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

Note?

Thought experiment: what is a good choice of hash size (k for k-mers) for building a hash table to facilitate sequence mapping to the human genome?

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

Note?

Thought experiment: what is a good choice of hash size (k for k-mers) for building a hash table to facilitate sequence mapping to the human genome?

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

Note?

k=1?

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

Note?

k=3? (4^3 possibilities)

AAA, AAC, AAG, … , TTT

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

Note?

k=10?

4^{10} (1,048,576)

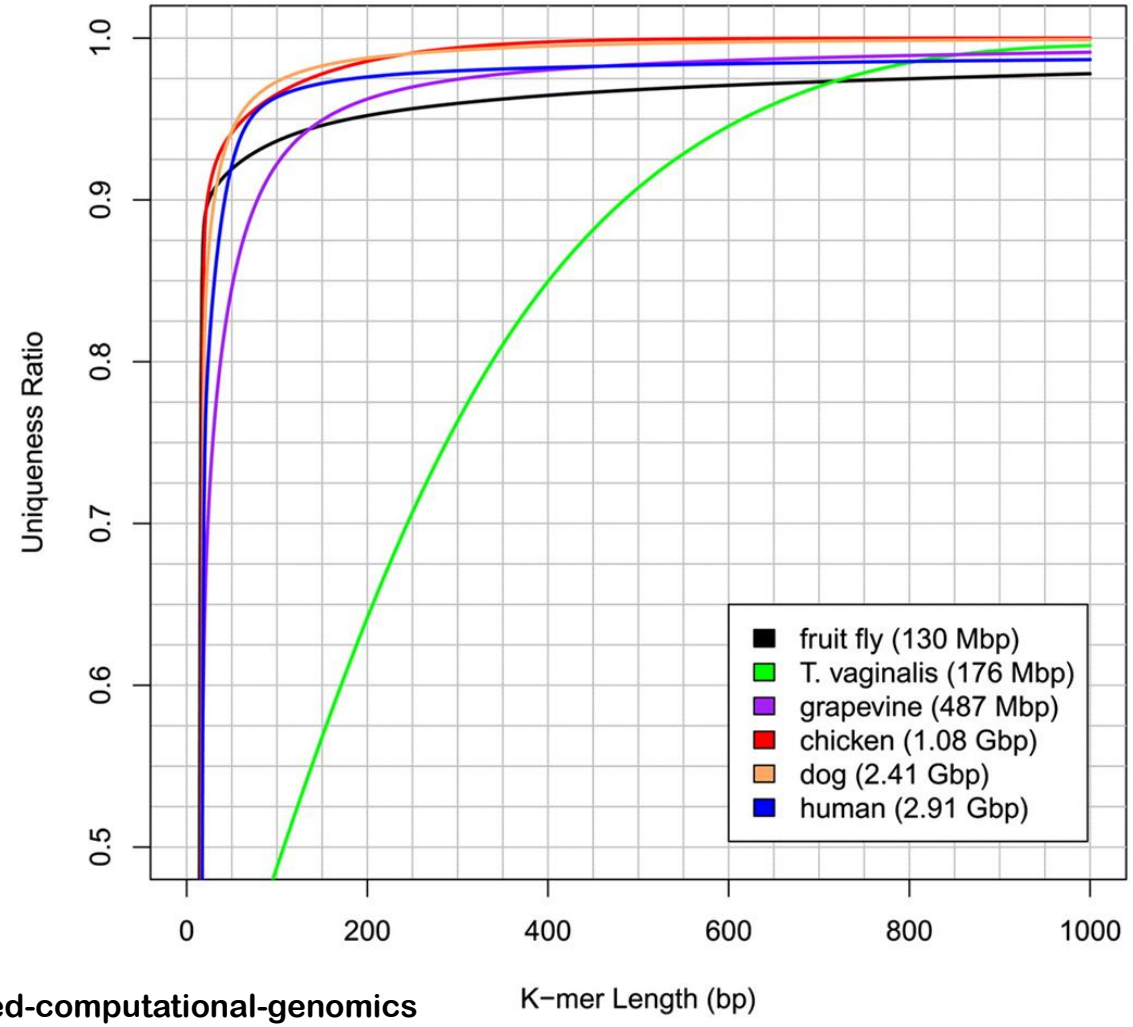
Every one of these is present in the human genome at least once

http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-167

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

Note?

It takes a very long k-mer to be unique in most genomes!



http://genome.cshlp.org/content/20/9/1165

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

Note?

- We have a sequencing run with over 100 million reads.
- After processing, the reads are between 20 and 25 nucleotide long.
- We would like to know if these sequences are in the human genome, and if so where.
- For a 20-mer such as ACGTGTGACGTGATCTGAGC takes about 10 seconds.
- Querying 100 million sequences would take more than 30 years.
- Without any indexing techniques, the whole genome will be scanned for every query.
- Hash-based indexing, access time is fast, does not depend on the text size.

Source: http://blog.thegrandlocus.com/2016/07/a-tutorial-on-burrows-wheeler-indexing-methods

Note?

- Kmers length ranges from 20 to 25 chars, each character will be represented in two bits (A:00, C:01, G:10, T:11). 20 to 25 kmers will take 40 to 50 bits of storage.

- The human genome contains over 3.2 billion nucleotides, so we need at least 108 GB (in reality many 20 to 25-mers are repeated so this number would be lower).

- If the storage required for the locations and the overhead for the dictionary will be added, the total size will be over 200 GB.

Source: http://blog.thegrandlocus.com/2016/07/a-tutorial-on-burrows-wheeler-indexing-methods

Thank you!