Session 2– Lecture 1

Alignment to reference genomes

Qi Sun

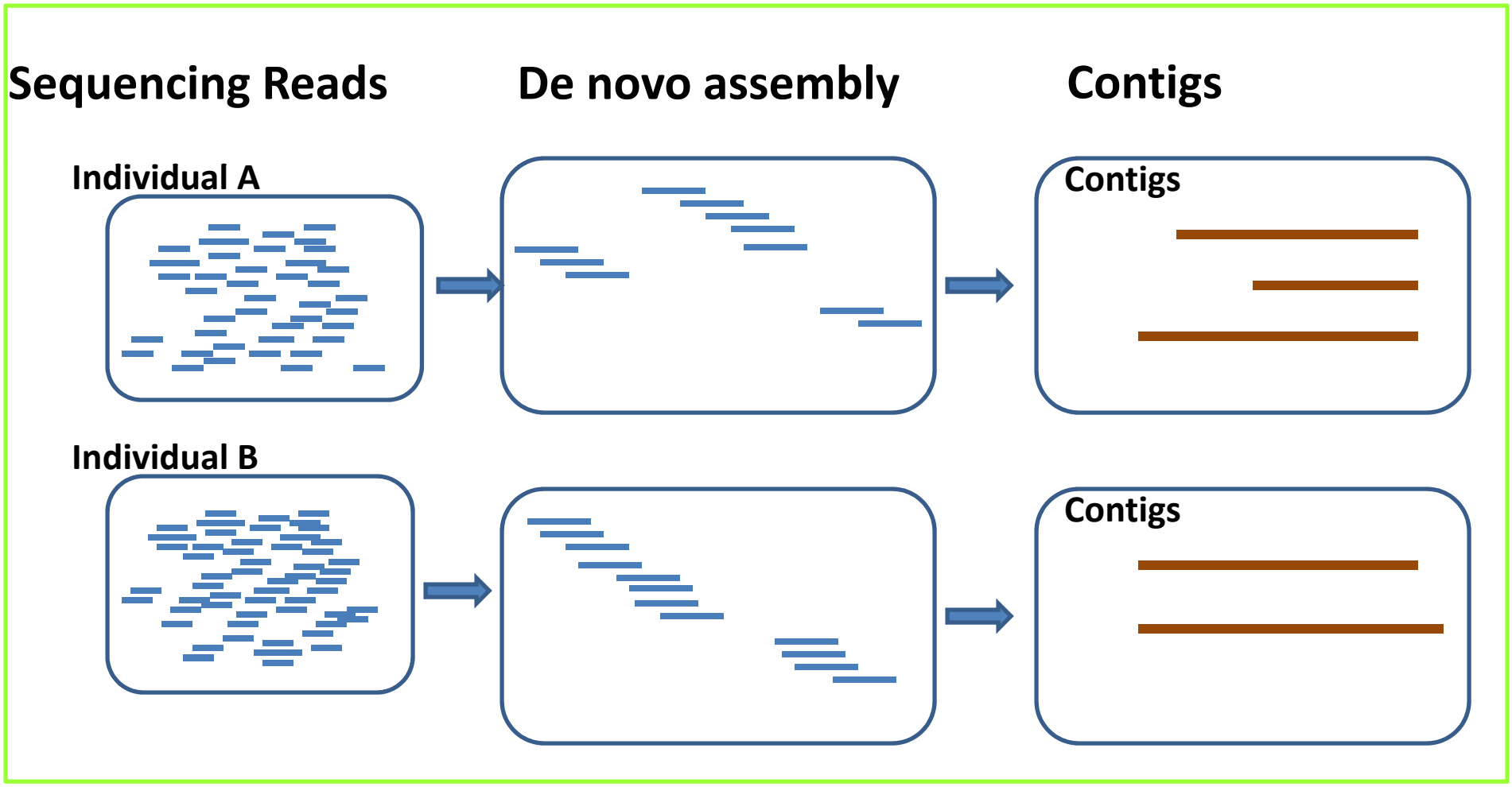
Computational Biology Service Unit
Cornell University

Outline

1. **Alignment and assembly**
2. **How alignment software work**
3. **Commonly used alignment software**
4. **Standard output file format**

Difference between alignment and assembly

Assembly process

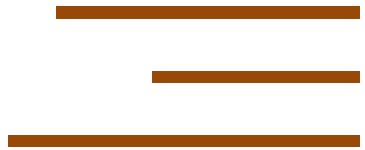


Difference between alignment and assembly

Assembly process

Individual A

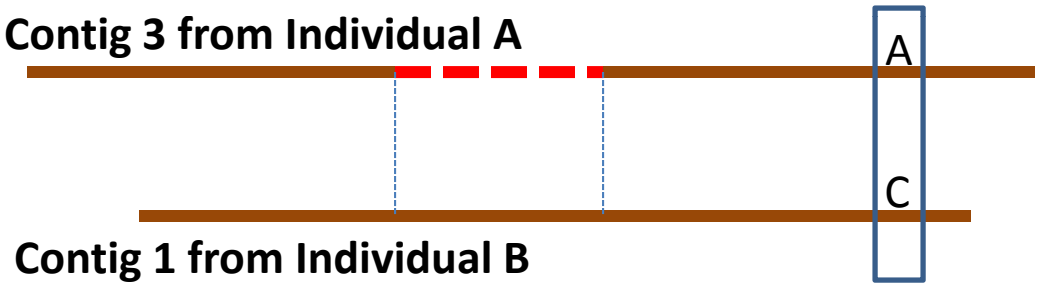
Contigs



Individual B

Contigs

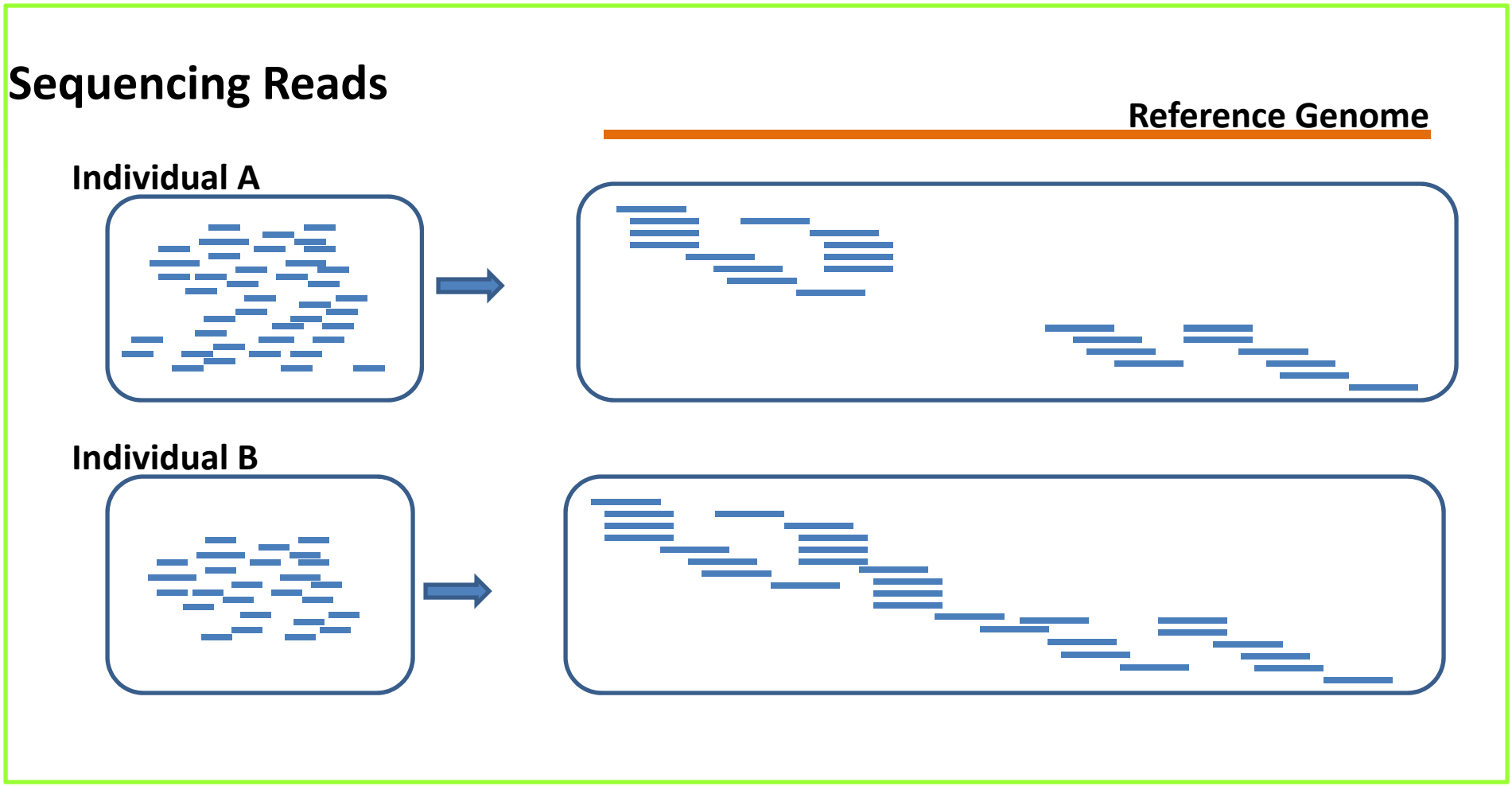
Align the contigs



*** Most often, assemblies are followed by gene annotation, and comparison in the annotated genes.**

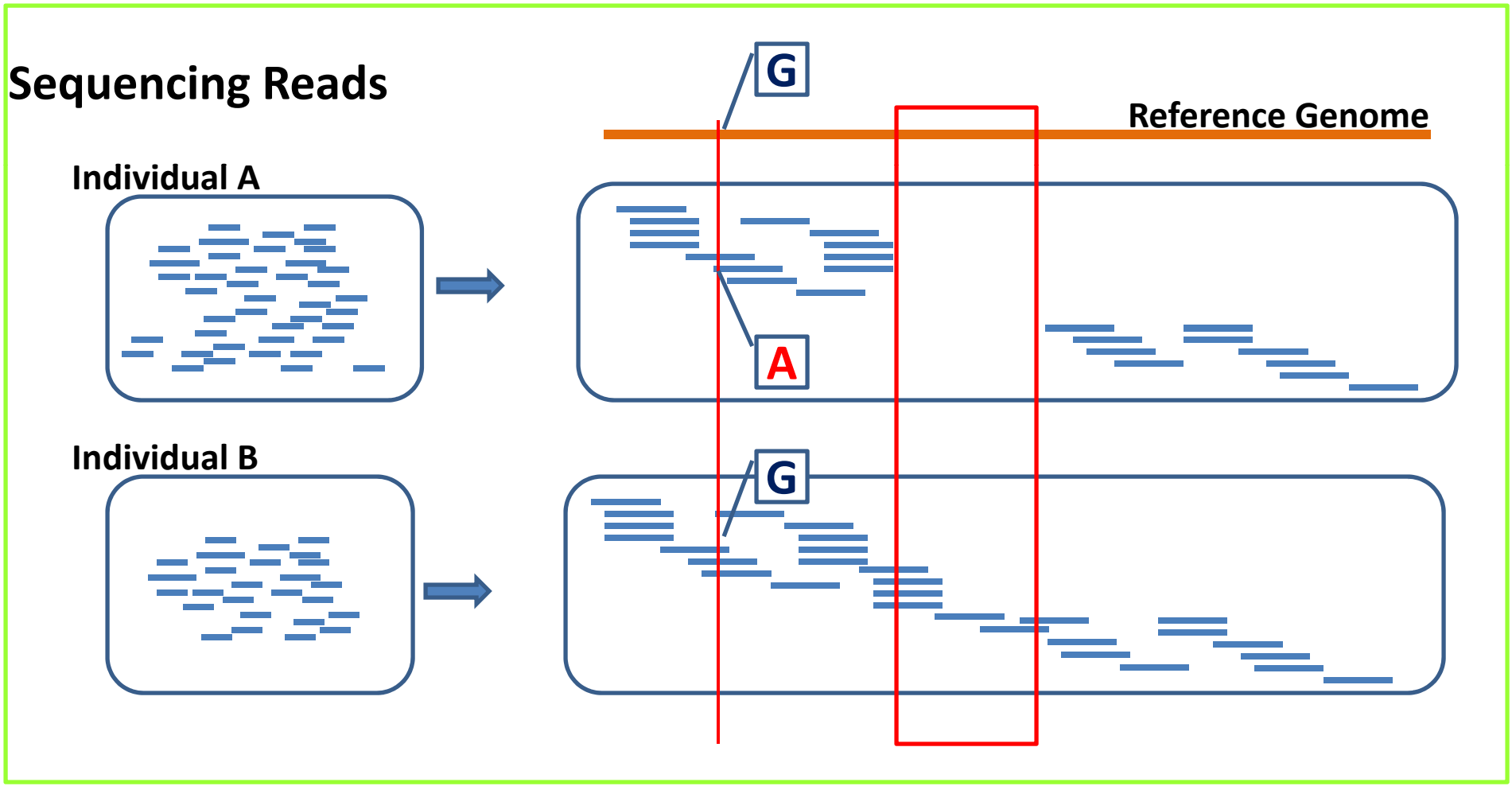
Difference between alignment and assembly

Alignment process



Difference between alignment and assembly

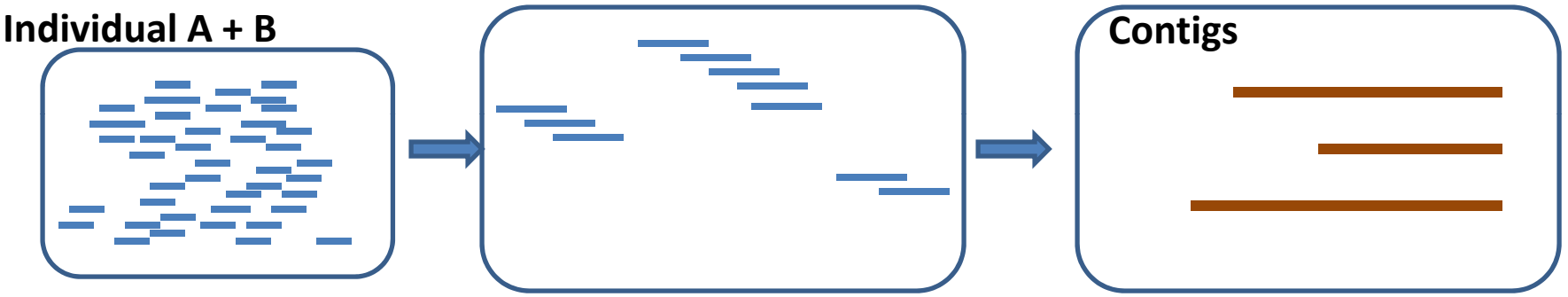
Alignment process



Combining the two approaches

Step1 : Assembly

Mix the reads from sample A and B, creating an assembly



Step2 : Alignment

Use the assembled contigs as reference, and align the reads from A and B to the reference.

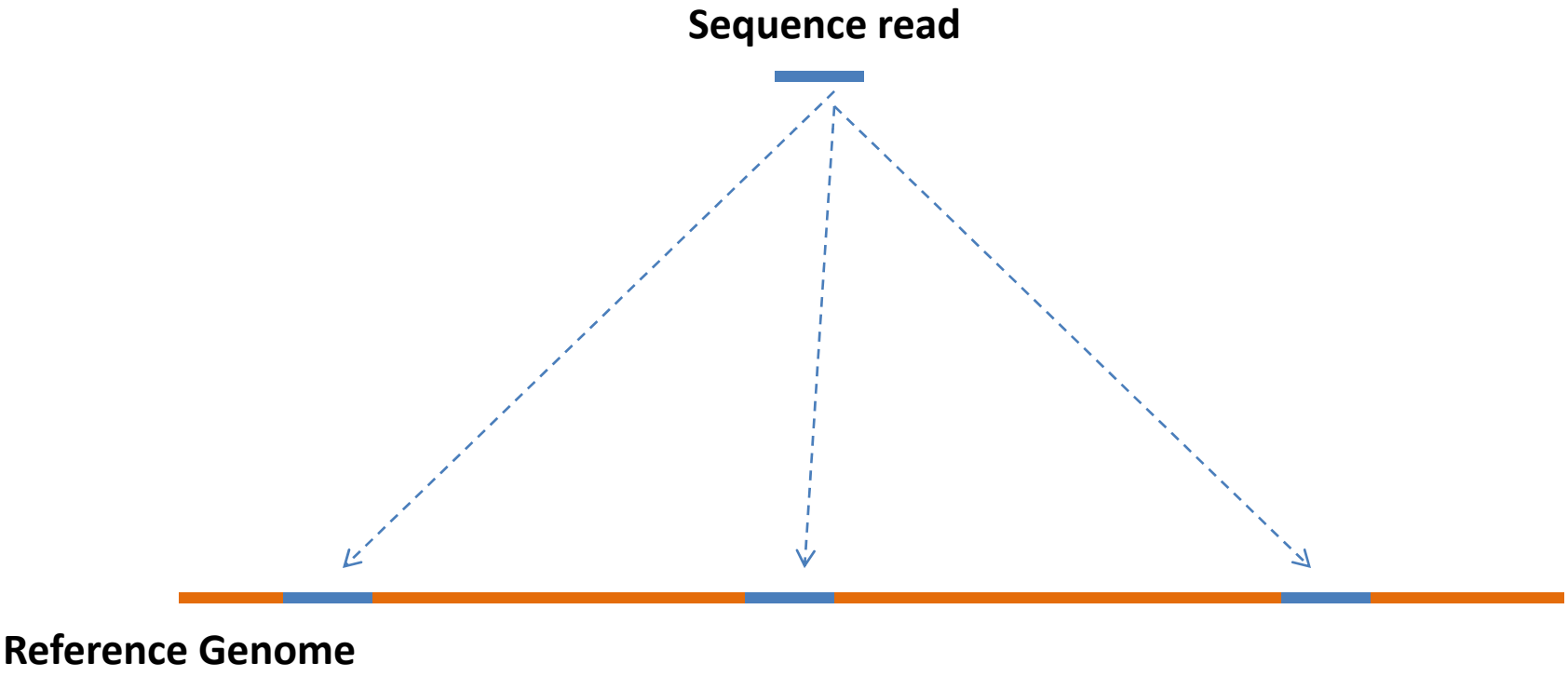
Limitation of alignment approach

Restricted by reference genome. Highly polymorphic regions or large insertions cannot be detected.

Alignment based approach is rarely used in comparative bacterial genome projects.

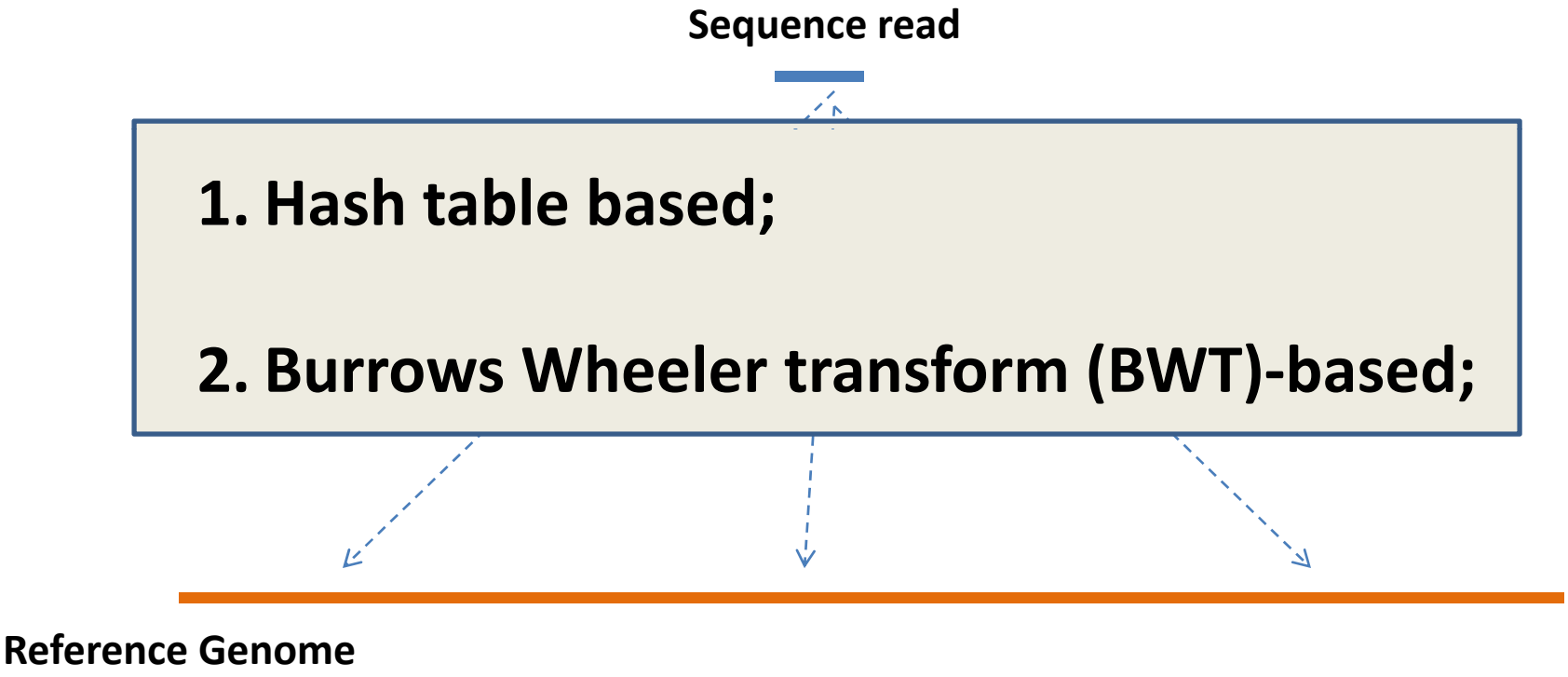
2-Step Alignment Strategy

Step 1: Mapping. Goal: Quickly identify candidates of hits



2-Step Alignment Strategy

Step 1: Mapping. Goal: Quickly identify candidates of hits



Issues in the mapping step

1. **The mapping step is heuristic. Not all reads that should have been aligned would be aligned, especially for reads in highly polymorphic regions, repetitive regions, or reads with indels.**

2. **Balance between alignment speed and alignment accuracy.**

2-Step Alignment Strategy

Step 2: Alignment and reporting. Goal: Score the alignment

Read:	AGGtCCGGGATACCGGGGAC	
Candidate 1 (chr1):	AGGGCCGGGAAACCGGGGAC	Score: -2
Candidate 2 (chr2):	AGGGCCGGGATACCGGGGAC	Score: -1

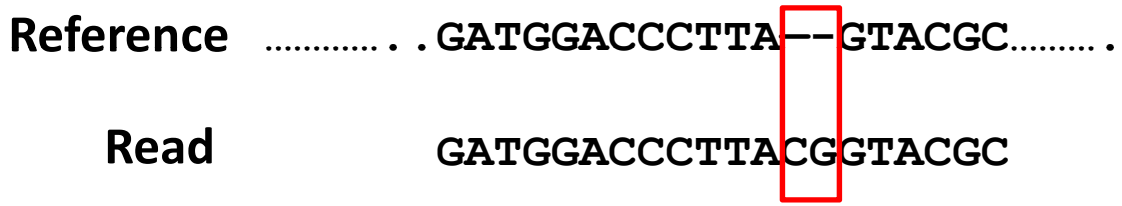
Issues in the alignment/reporting step

1. Some software would use the base quality score to evaluate the alignment. Others do not.

2. Software parameters relevant to this step: 1) Maximum mismatches that would be reported; 2) Reporting unique hits or multiple hits.

Features in alignment software

1. Gapped vs ungapped alignment:



Issues with gapped alignments

Issues with gapped alignments

- **Indels located at the edge of a read**

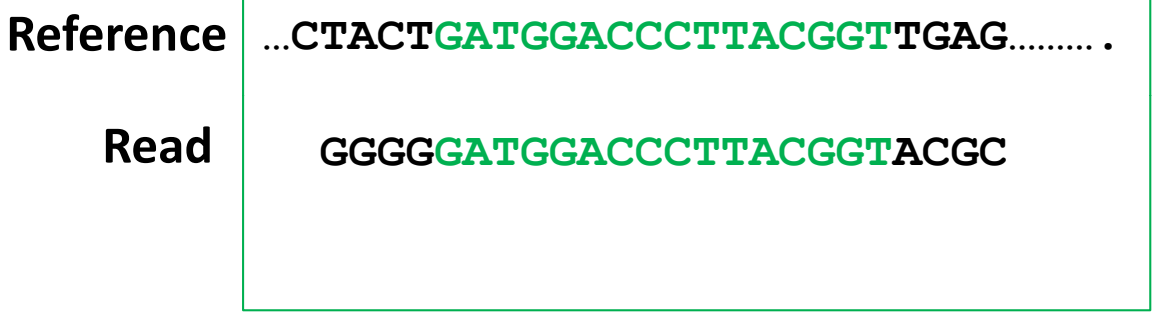
- **Reported positions of the INDELs**

……..GATGGACCCTTAAAAGTACGC…….

GATGGACCCTT--AACGGTACGC

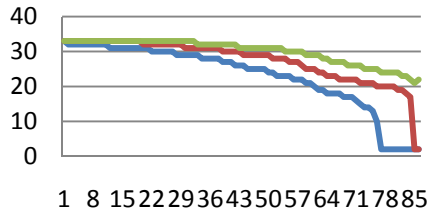
Features in alignment software

1. Global vs Local alignment



Situations where only part of the reads can be aligned

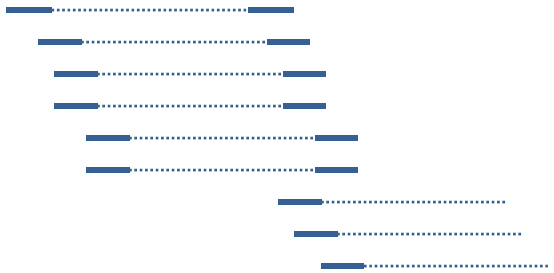
- Low quality part of the read



- Reads span breaking points in a chromosome re-arrangement event

- Reads span splicing junctions

Paired – end alignment



Software supporting paired-end alignment requires user input:

1. **Minimum and maximum size of the inserts**

2. **Options of reporting unpaired alignment**

Options of reporting the alignments

- **Unique alignment only**

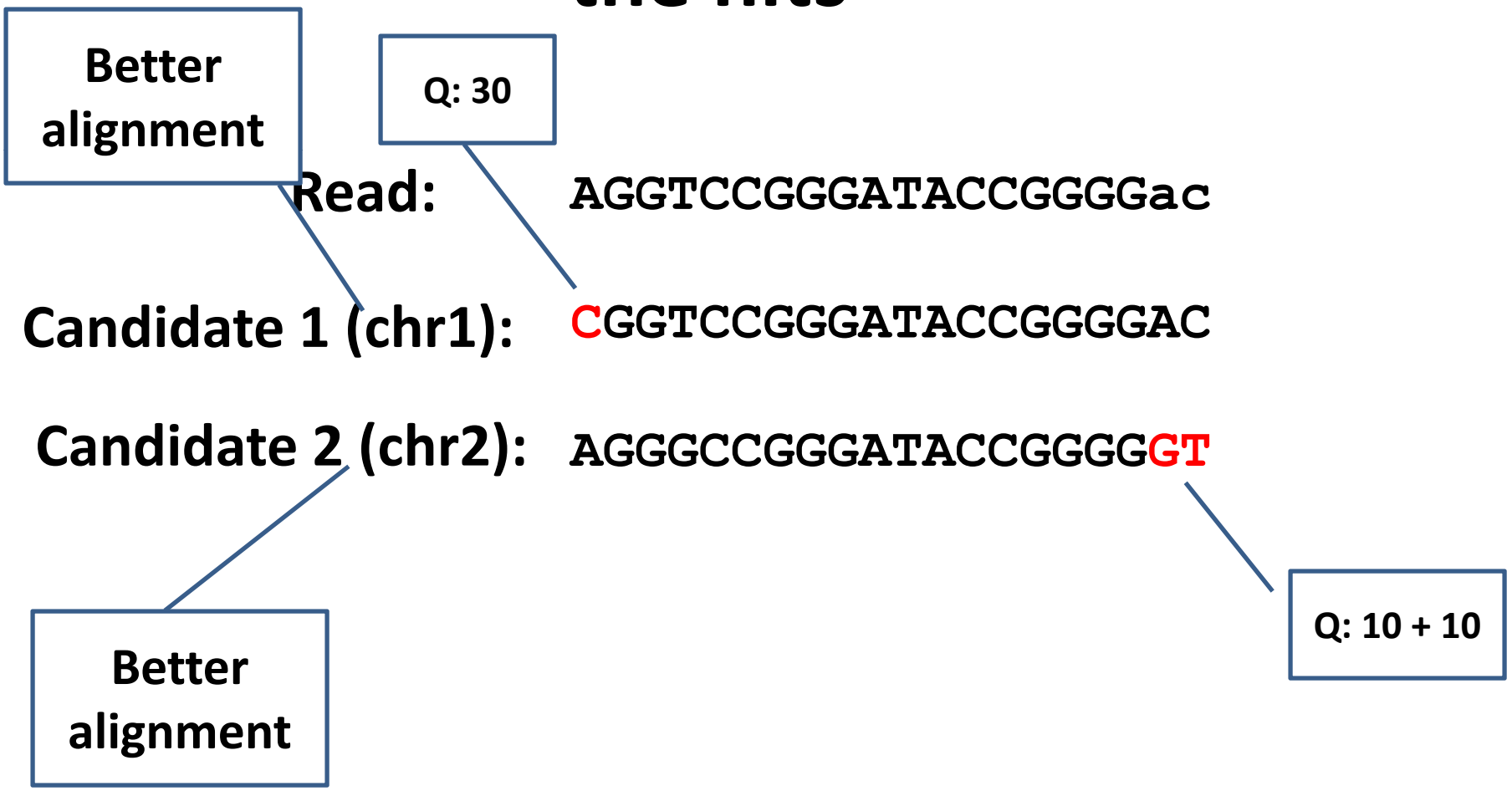
- **Reporting ambiguous hits**

a) **Randomly report one**

b) **Report all alignment above cutoff parameters**

c) **Report all alignments with best alignment scores**

Using base qualities to evaluate the hits



Hash size selection/ Guaranteed number of mismatches

- **Hash size too big: miss the reads with many mismatches.**

- **Hash size too small: slow; too many hits in the mapping step and fail to be aligned**

BWA

- **BWA aln:**

1) **Short reads up to 200 bp with errors <5%**
2) **global alignment;**
3) **Gapped alignment;**
4) **Base quality is not used in evaluating hits;**
5) **Can do paired-end;**
6) **Report ambiguous hits;**

- **BWASW**

1) **Longer reads with more mismatches;**
2) **Local alignment;**
3) **Slow and less accurate, does not work with paired -end**

Li H. and Durbin R. (2009) *Bioinformatics*, 25:1754

Bowtie

- **One of the fastest alignment software for short reads**
- **Not gapped-alignment;**
- **Base quality can be used evaluating alignment;**
- **Paired-end;**
- **Flexible reporting mode**

Langmead etal (2009) *Genome Biology*, **10:**R25

Other software:

- **ELAND** **part of the Illumina CASAVA software**
- **SOAP** **developed at BGI**
- **MAQ** **developed by Heng Li**

Specialized alignment software

- **Tophat: splicing junction alignment**

- **BSMAP: C->T tolerant alignment**

Commercial Alignment Software

Common Features:

Gapped alignment tools;
Reporting ambiguous hits;
Supporting paired-end alignment;
flexible reporting modes;
Not open source.

- **SlimSearch by RealTimeGenomics**

•Available in April, 2010. Cost: Unknown. CBSU is in its early access program.

- **Novoalign by Novocraft**

•Available now. Free for academic users.

SAM/BAM file format

Sample SAM File:

Header

```
@HD     VN:1.0
@SQ     SN:chr20 LN:62435964
@RG     ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891
@RG     ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891
```

Alignment

```
read_28833_29006_6945 99 chr20 28833 20 10M1D25M = 28993 195 \
AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG
<<<<<<<<<<<<<<<<<<<:<9/,&,22;;<<< \
NM:i:1 RG:Z:L1
read_28701_28881_323b 147 chr20 28834 30 35M = 28701 -168 \
ACCTATATCTTGGCCTTGGCCGATGCGGCCTTGCA
<<<<<;<<<<7;:<<<6;<<<<<<<<<<<7<<<< \
MF:i:18 RG:Z:L2
```

SAM/BA…at

Alignment section

1	QNAME	Query (pa…
2	FLAG	bitwise FL…
3	RNAME	Reference…
4	POS	1-based l… …ipped sequence
5	MAPQ	MAPping…
6	CIAGR	extended CIGAR string
7	MRNM	Mate Refe… …e as RNAME)
8	MPOS	1-based M…
9	ISIZE	Inferred i…
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QU… …se quality)
12	OPT	variable C… …VTYPE:VALUE

Strand; Paired-end; et al.

Map position

Indels; Junctions; et al

Read sequence & base qualities

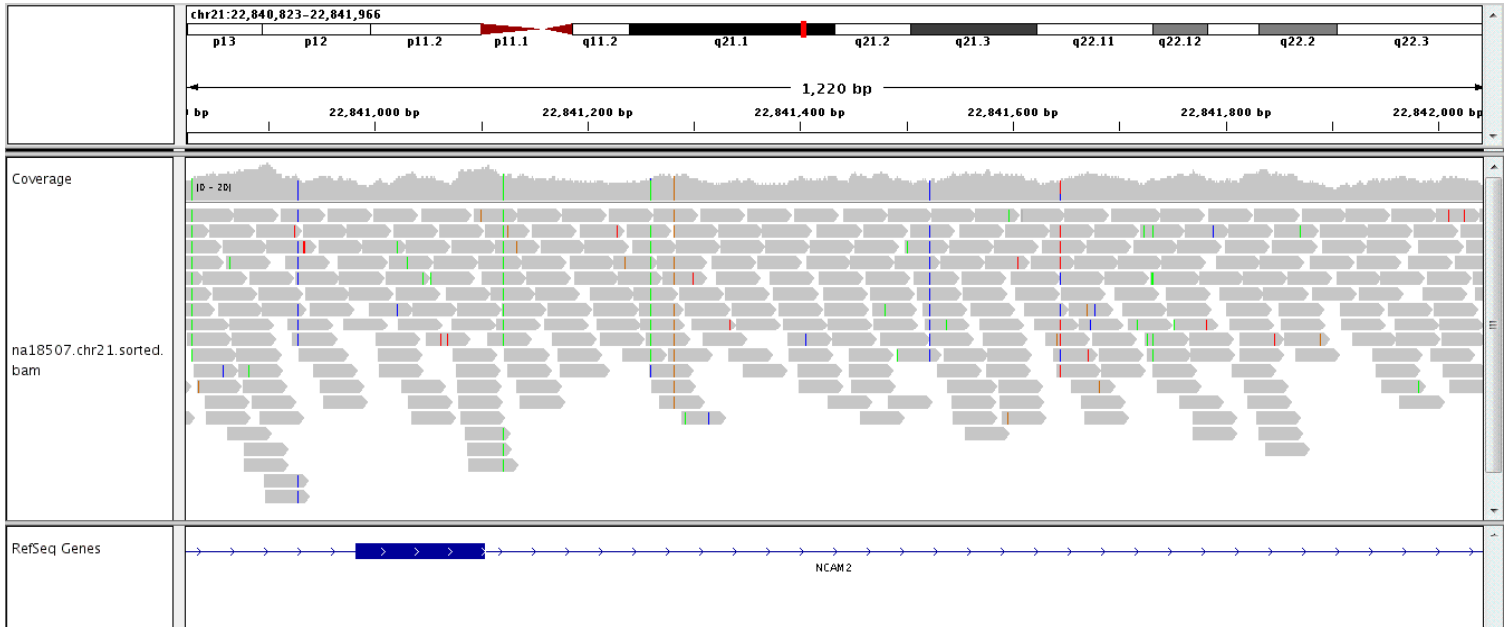
Departures from the standard

How to represent reads with multiple hits in SAM file?

BOWTIE/TOPHAT represents multiple hits with multiple lines, one hit per line.

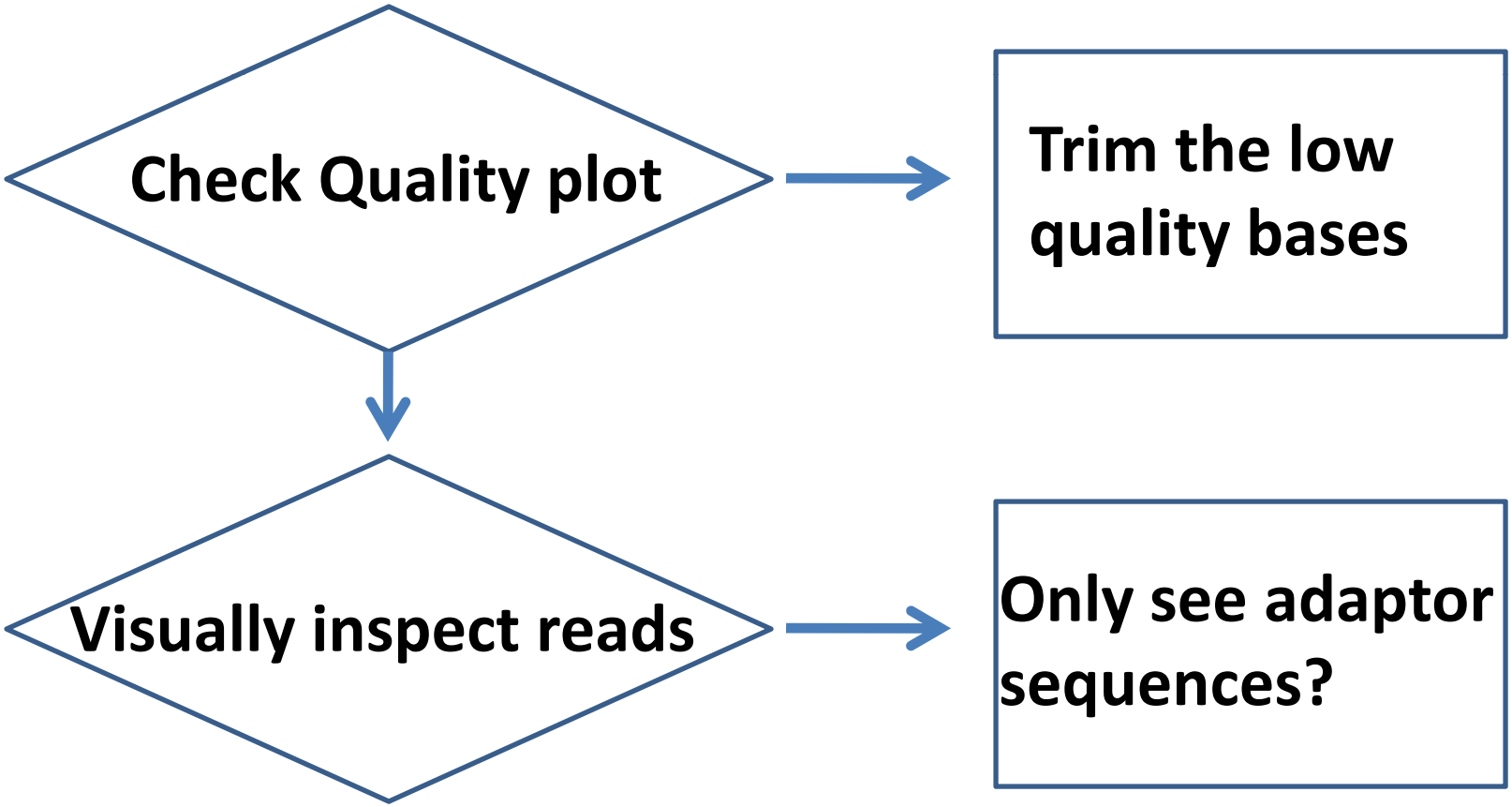
BWA represents multiple hits with single. The alternative alignments are reported with optional XA tag.

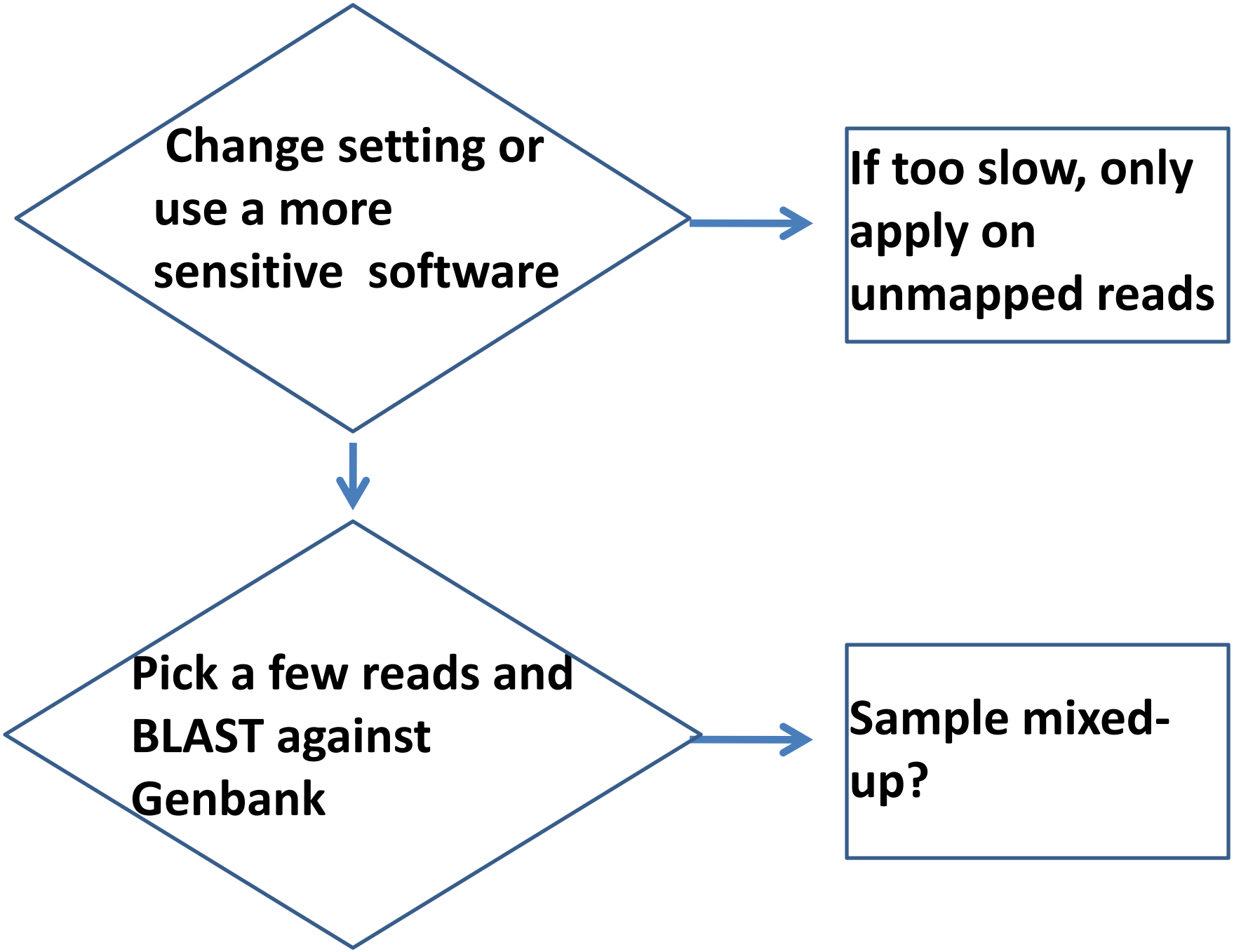
Visualization of BAM file through IGV



Summary

- Use a fast alignment tool in the first try. (BWA, Bowtie, Tophat, et al.). Normally ¾ of the reads should align to the genome.




Change setting or use a more sensitive software
If too slow, only apply on unmapped reads
Pick a few reads and BLAST against Genbank
Sample mixed-up?

Commercial solutions:

1) Alignment software like SlimSearch and Novoalign;

2) Specialized cloud computing service with combined hardware and software solutions, GenomeQuest et al.