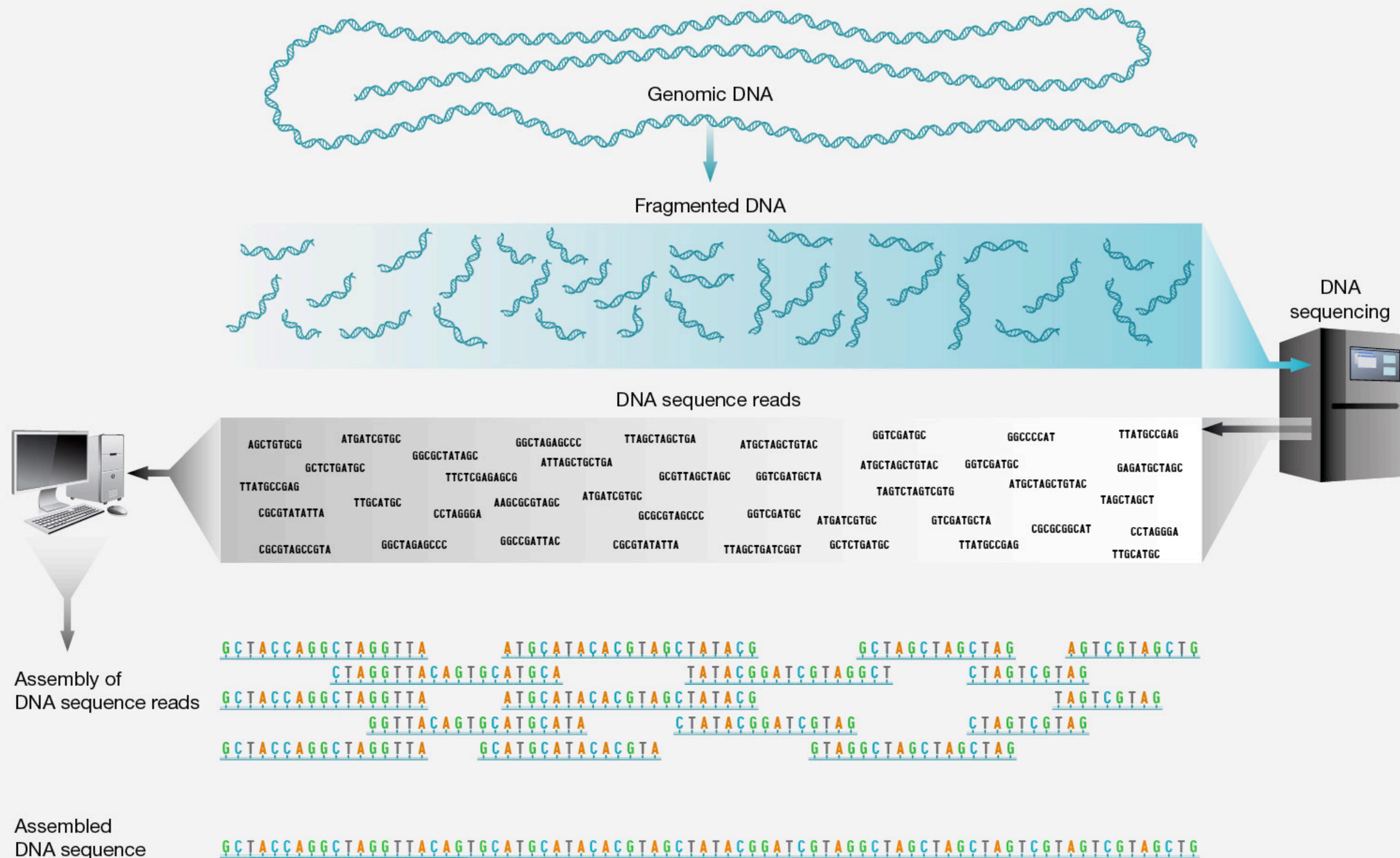
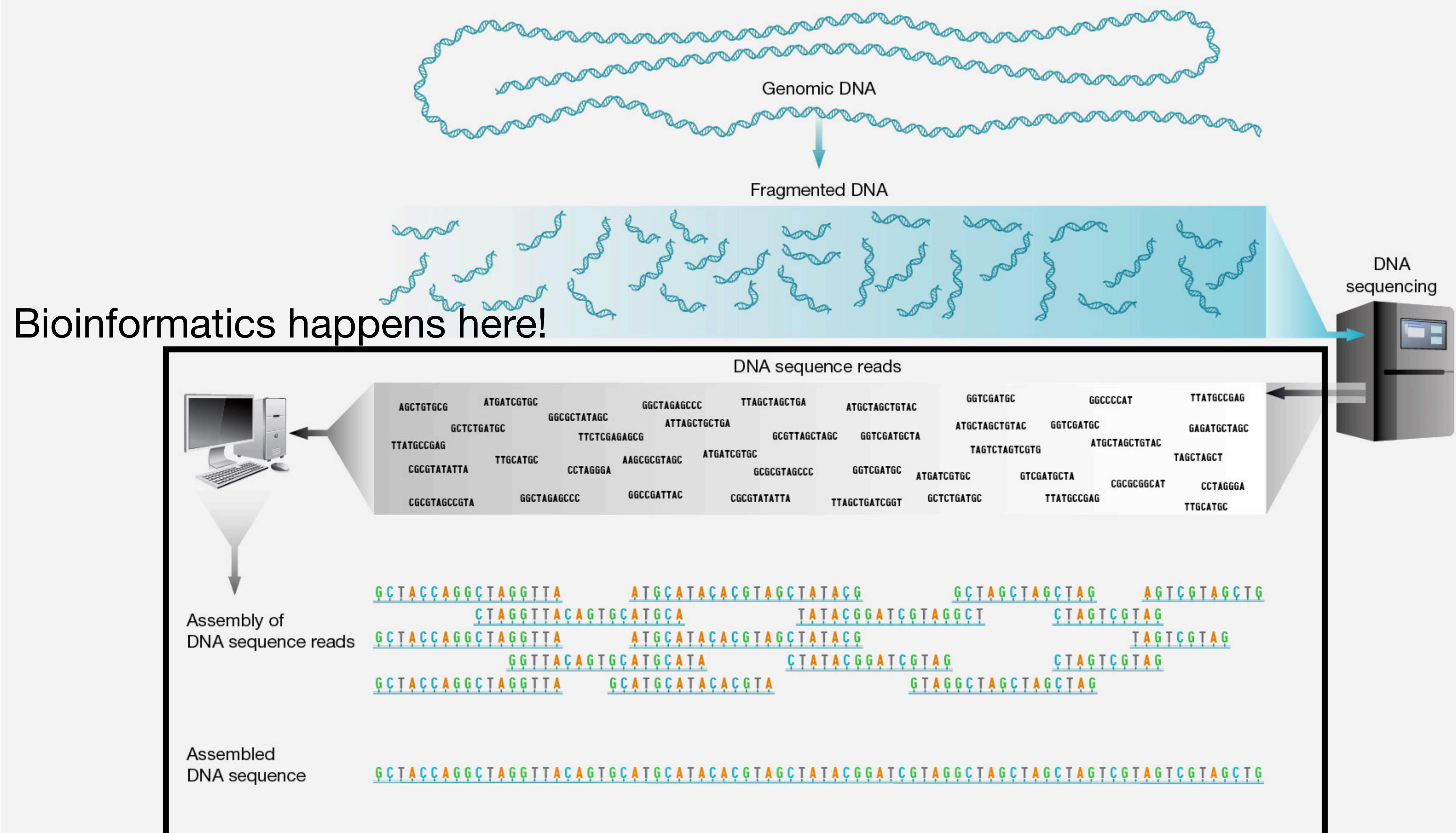


Sequence analysis practical

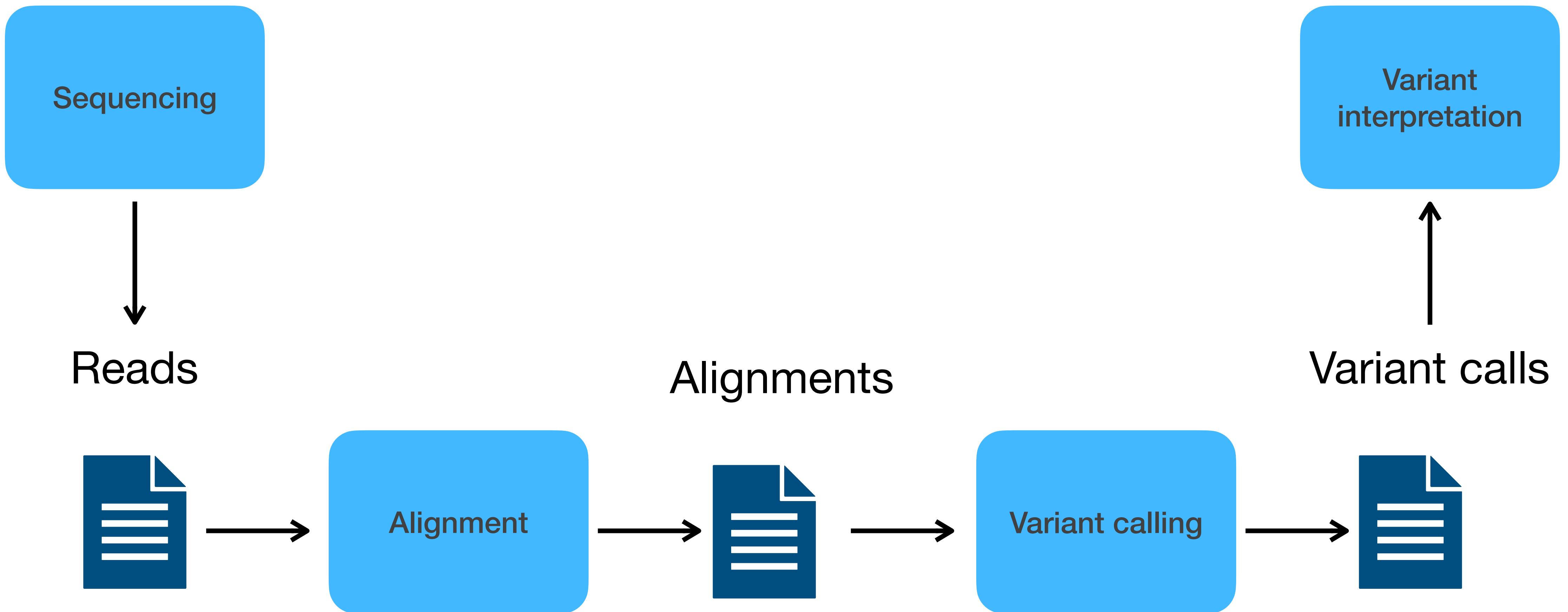
UZH-BIO392 day 08

Max Verbiest 2023-09-28

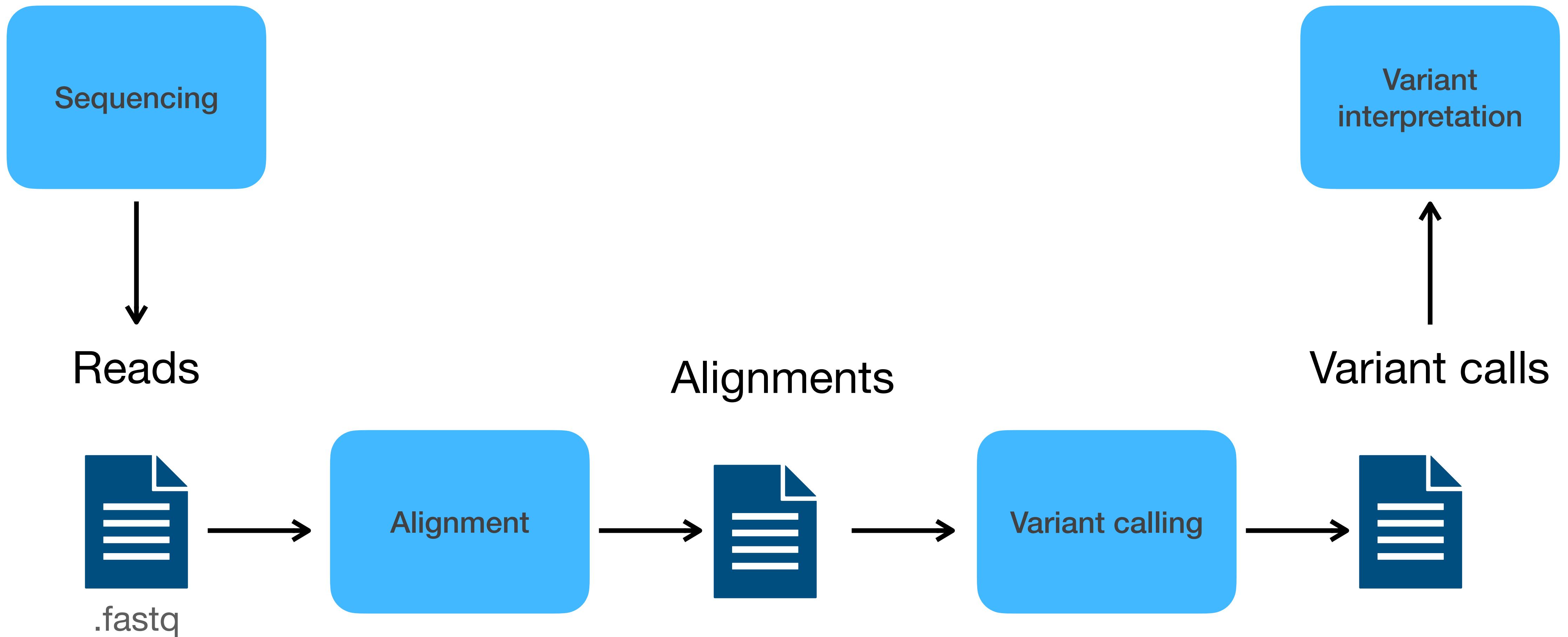




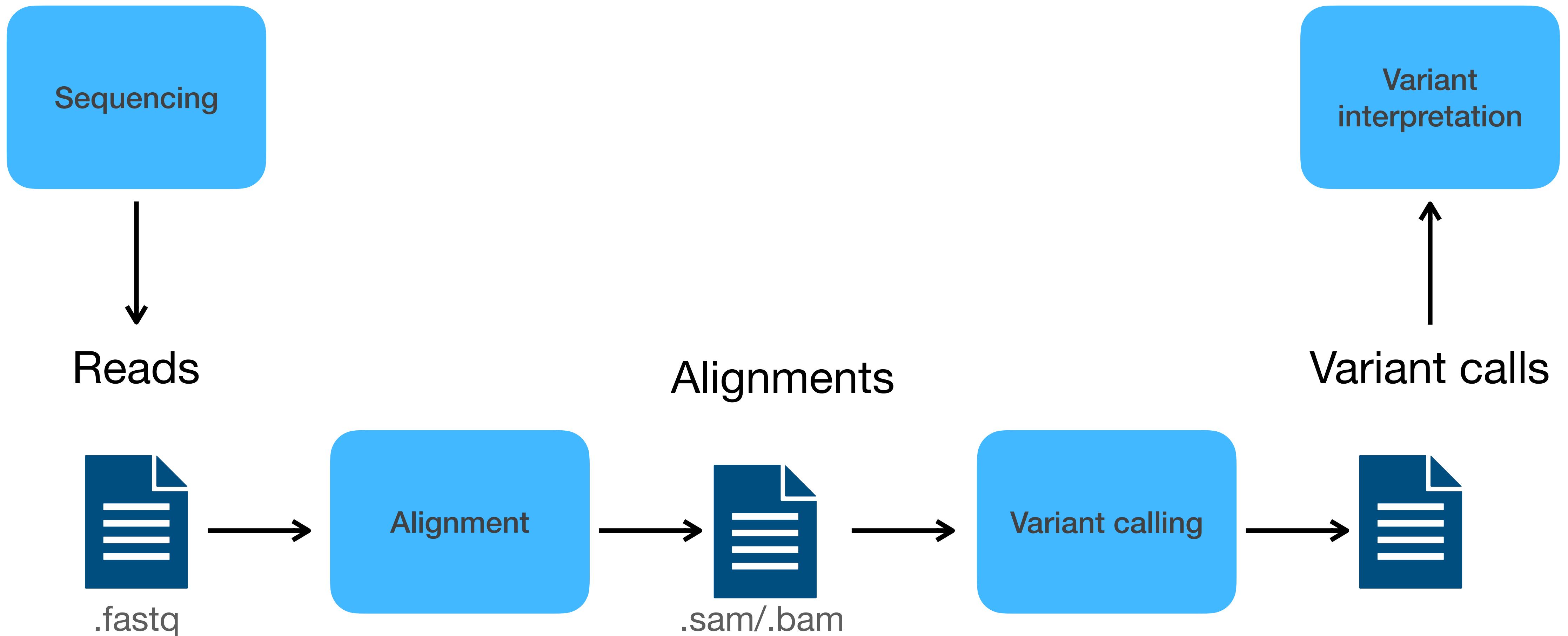
Generic bioinformatics pipeline



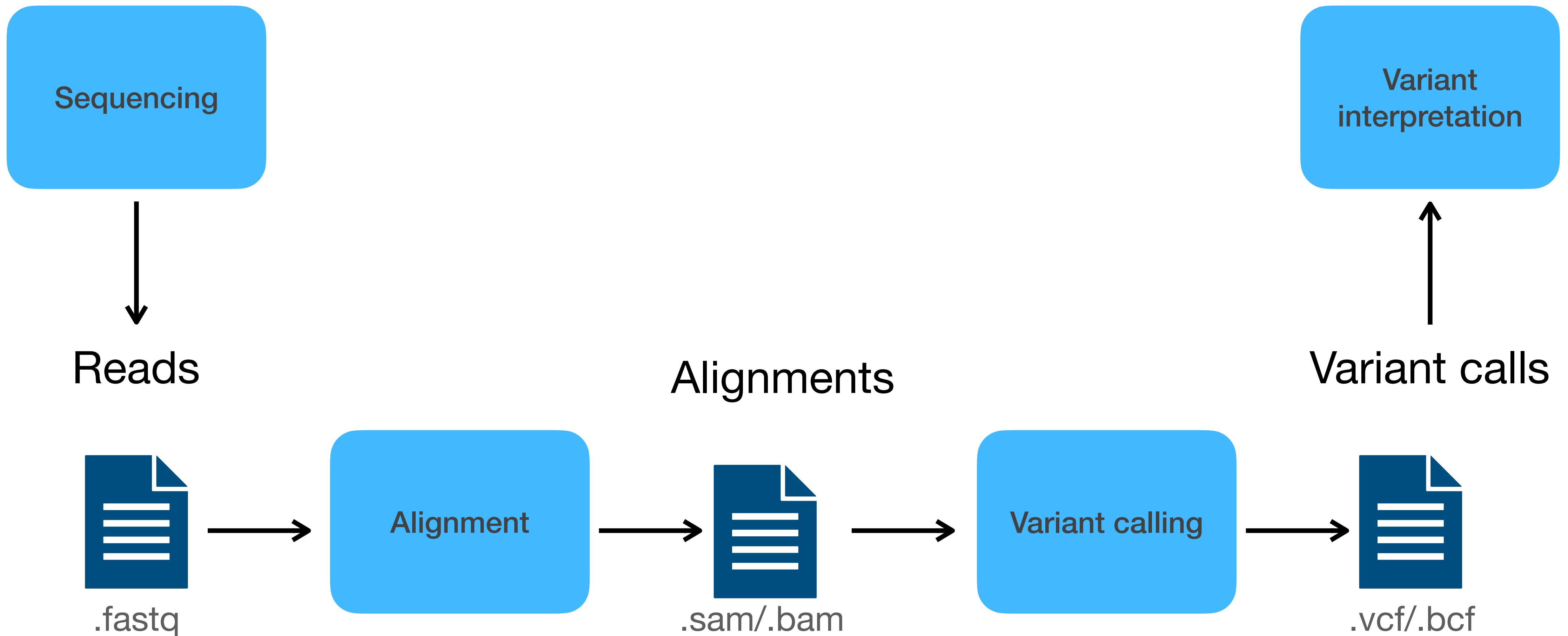
Generic bioinformatics pipeline



Generic bioinformatics pipeline



Generic bioinformatics pipeline



Program

Program

- Background: short tandem repeats

Program

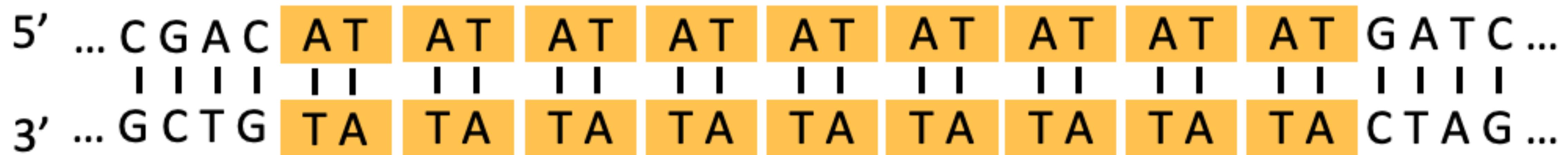
- Background: short tandem repeats
- Shotgun sequencing

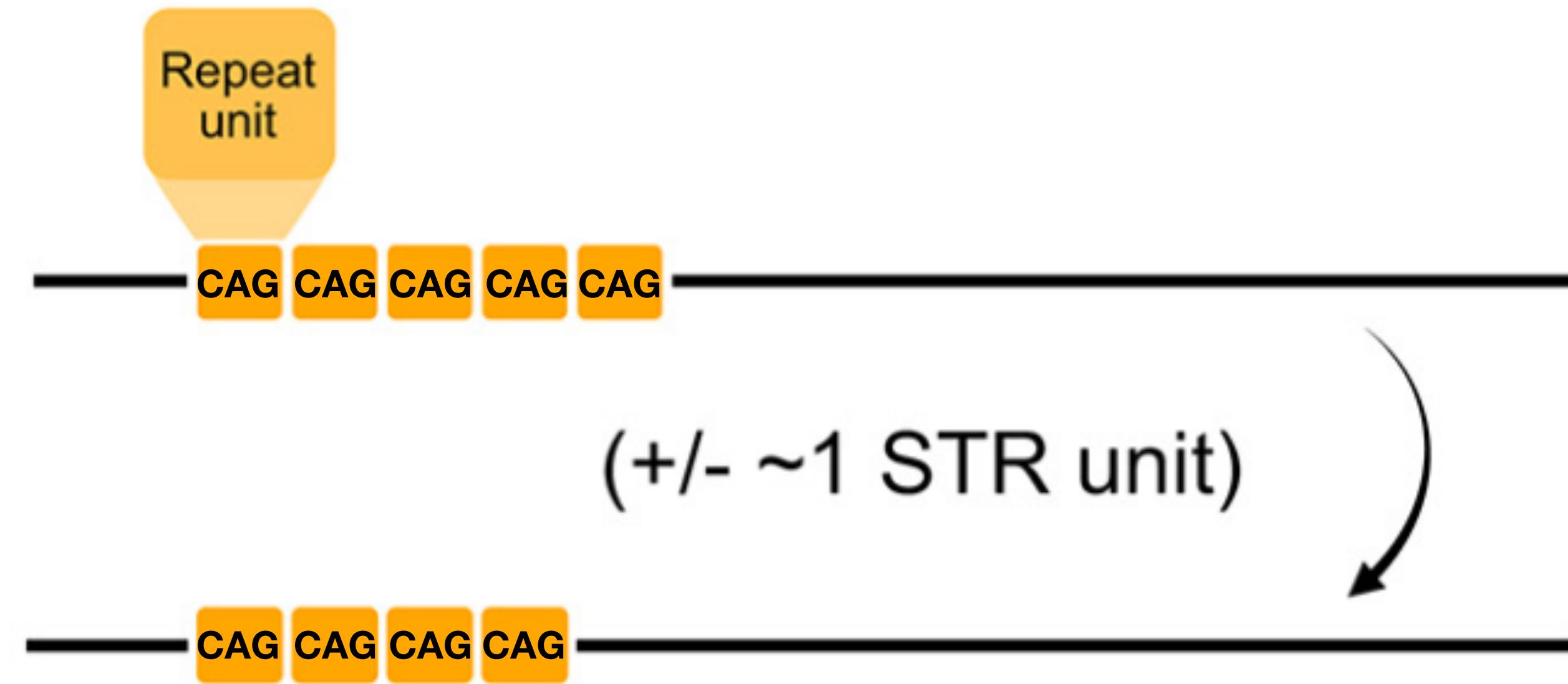
Program

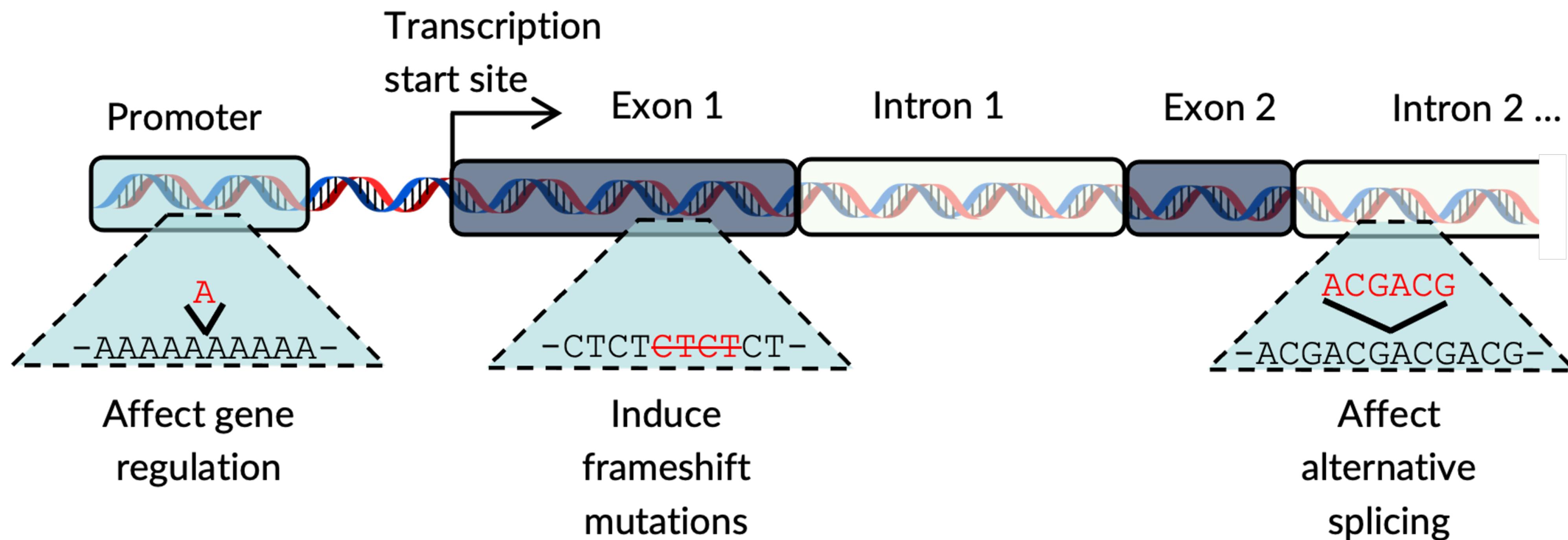
- Background: short tandem repeats
- Shotgun sequencing
- Bioinformatics pipeline
 - Processing reads
 - Alignment
 - Variant calling
 - Variant interpretation

Short tandem repeats

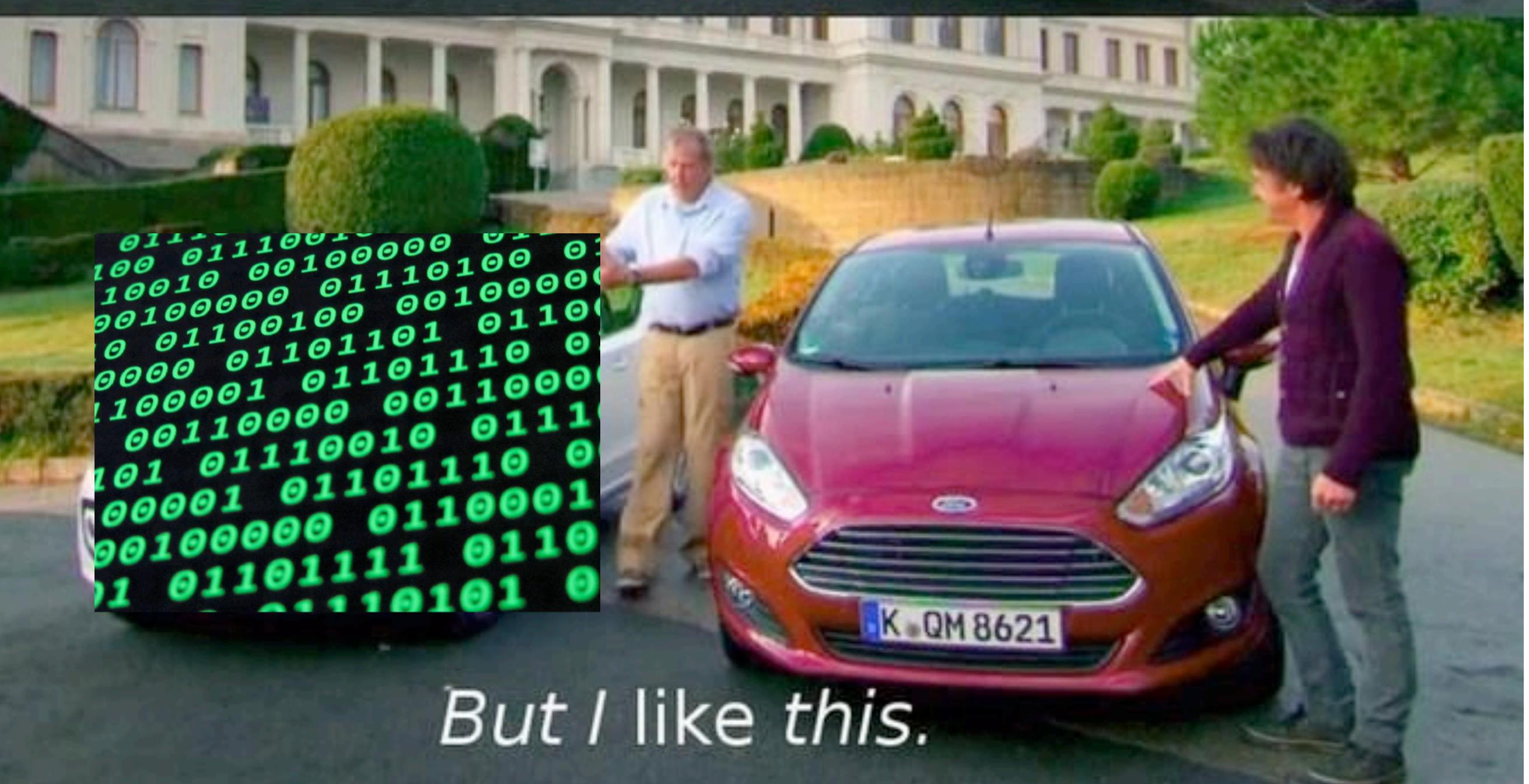
Repetitions of 1-6 bp DNA units



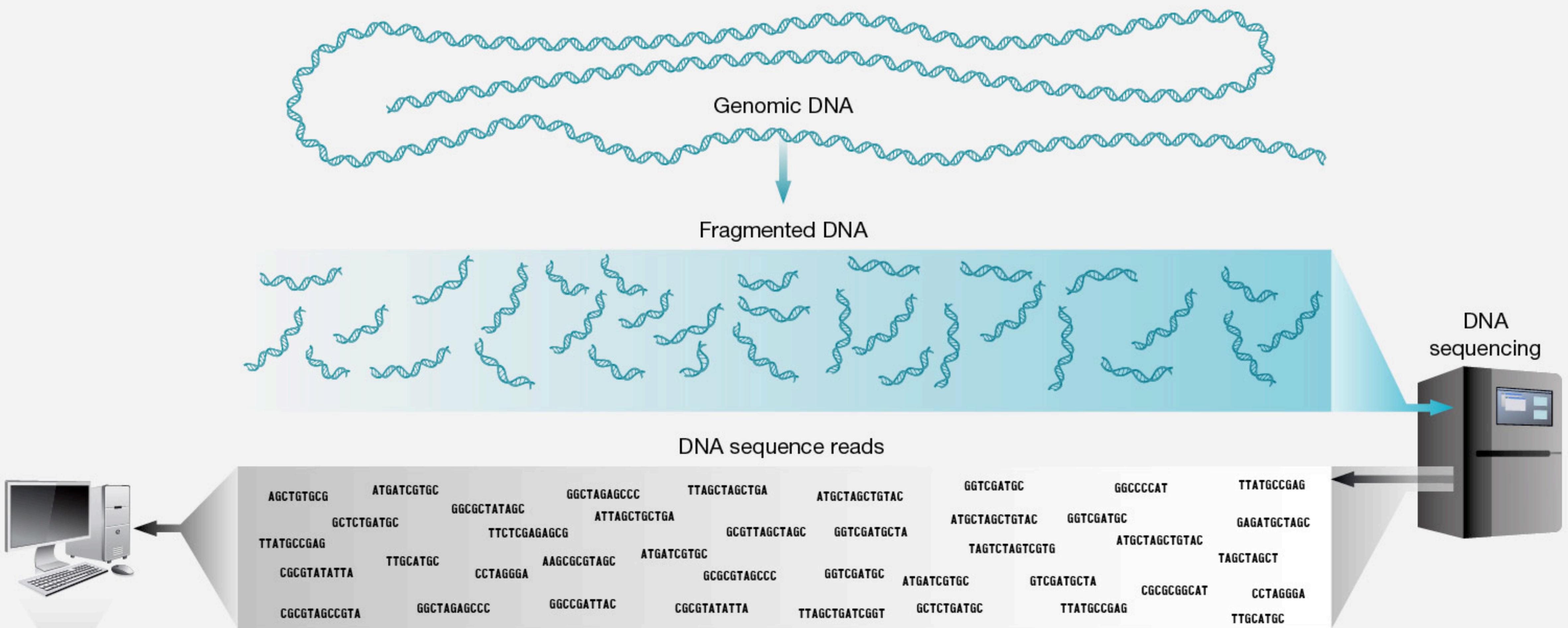




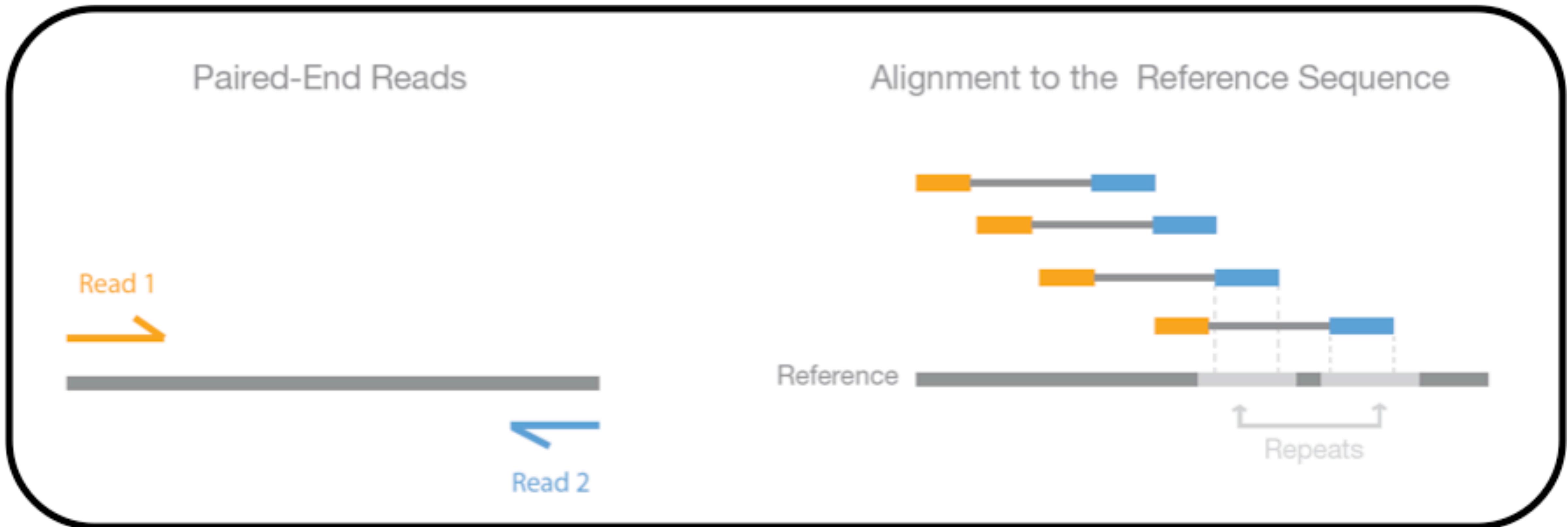
Effects of STR variation still poorly understood



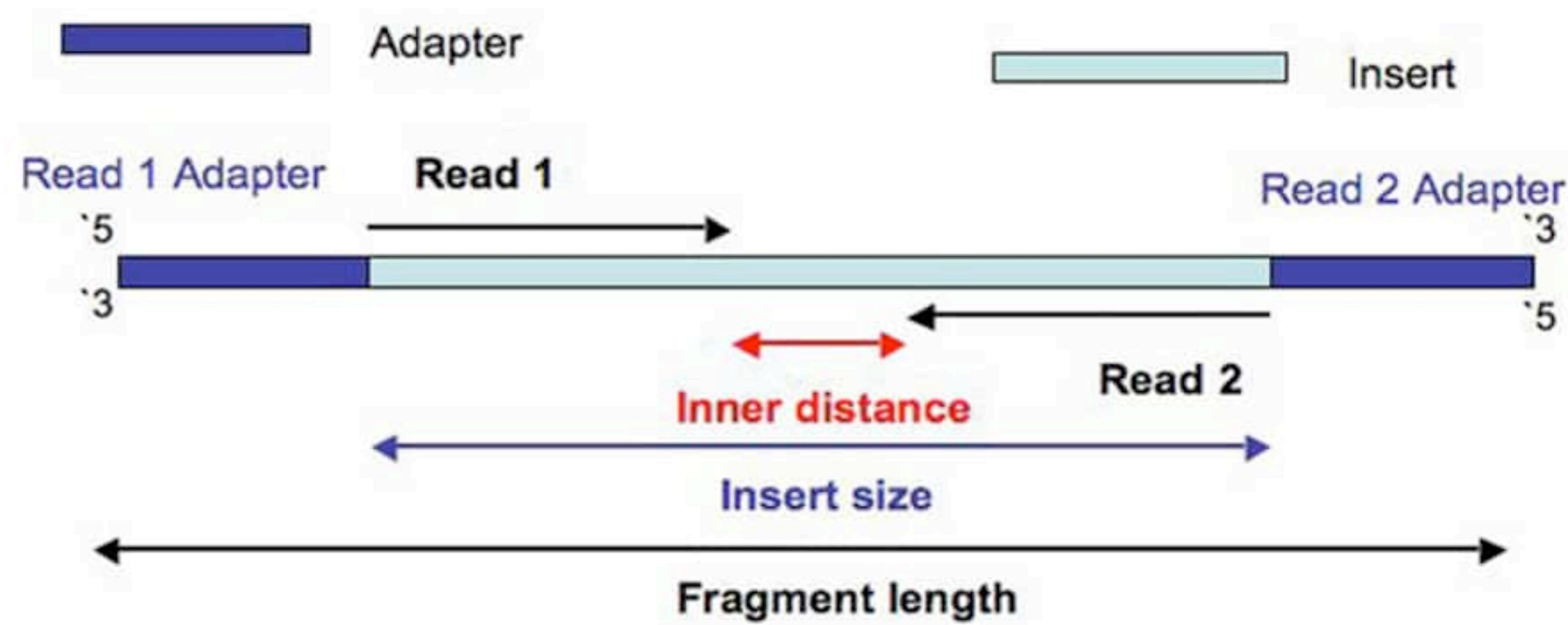
Shotgun sequencing



Paired-end reads



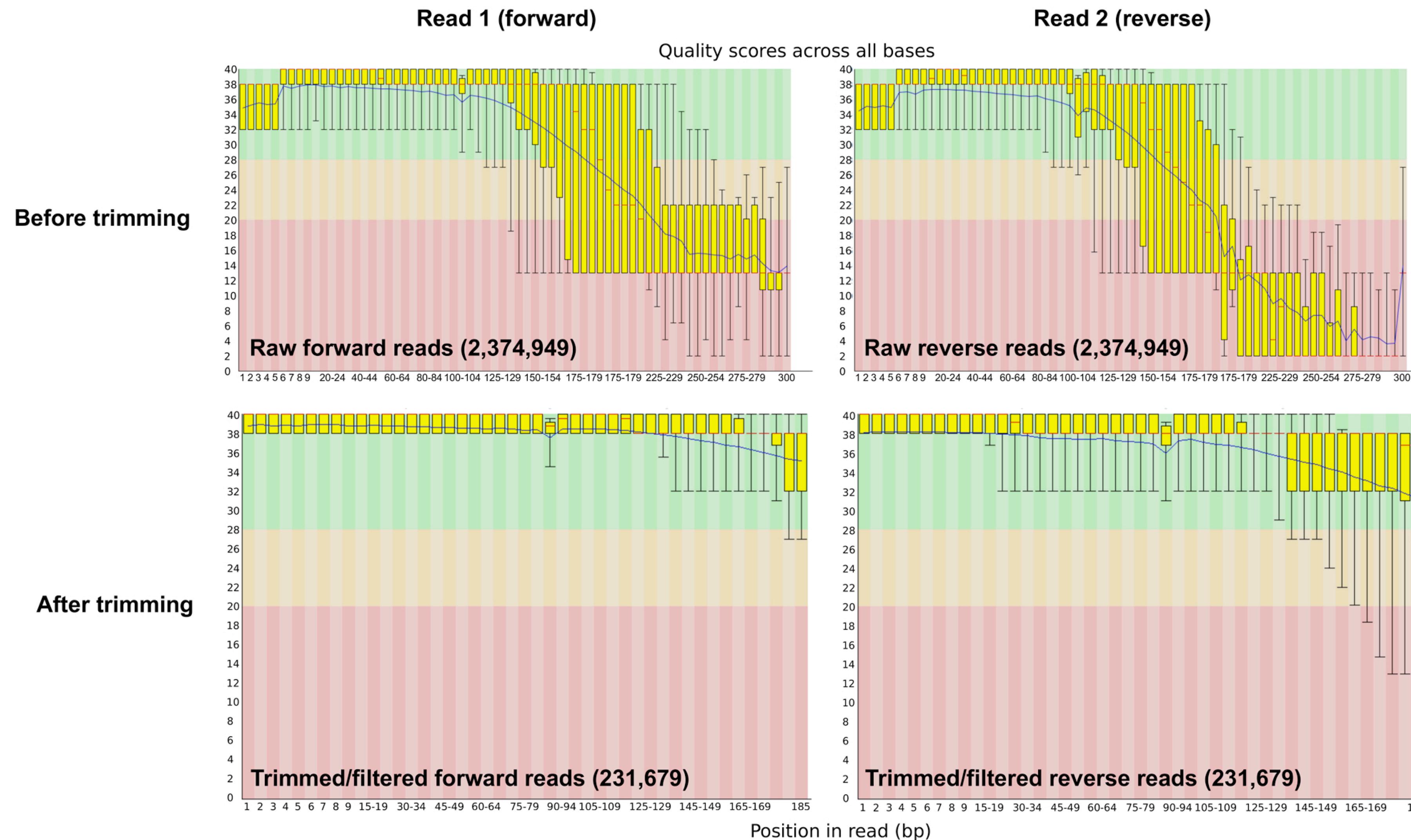
Paired-end reads: more detail



Reads are processed before analysis

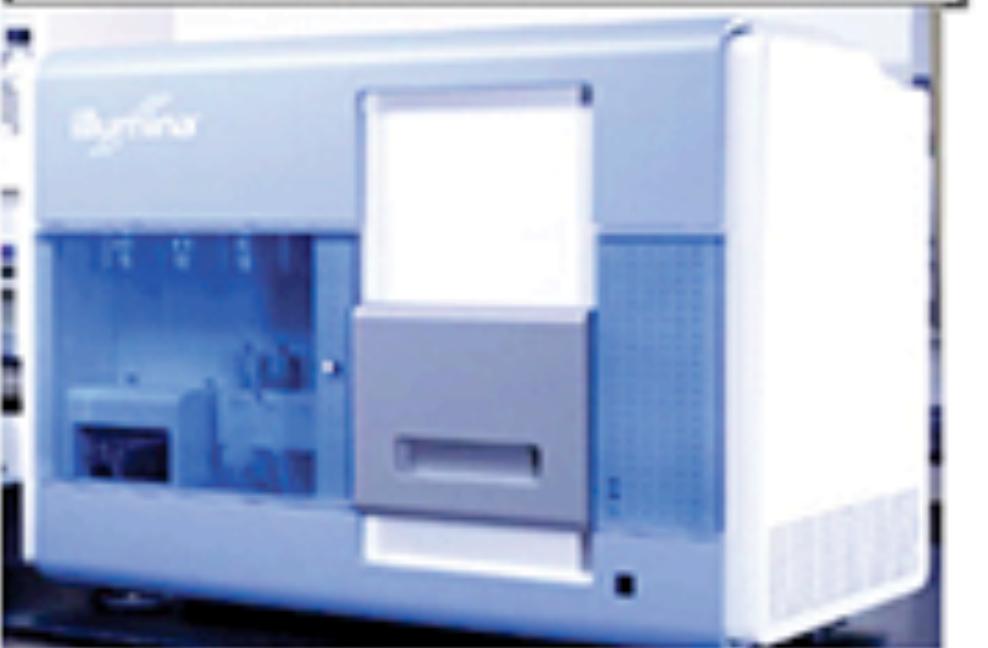
- Typical read processing steps:
 - Adapter trimming (remove leftovers from sequencing process)
 - Quality trimming (remove low Phred-score bases)
 - Length filter (remove too-short reads)

Read processing before/after



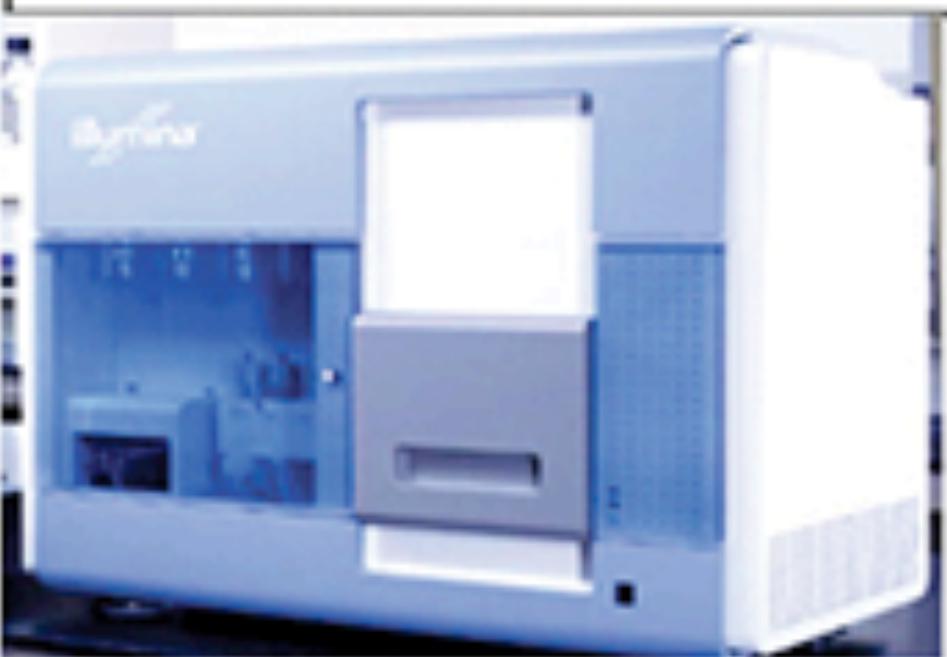
Short reads

.....
GGTCTGGATGC
CGGTCTGGATGC
GCGGTCTGGATG
GCGGTCTGGAT
GGCGGTCTGGAT
GGCGGTCTGGA
TCTATGCAGGGCCCT
TCTATGCAGGGCCCC
ATCTATGCAGGGCC
TATCTATGCAGGGC
TTATCTATGCAGGG
CTTATCTATGCAGGG



Short reads

.....
GGTCTGGATGC
CGGTCTGGATGC
GCGGTCTGGATG
GCGGTCTGGAT
GGCGGTCTGGAT
GGCGGTCTGGA
TCTATGCAGGGCCCT
TCTATGCAGGGCCC
ATCTATGCAGGGCC
TATCTATGCAGGGC
TTATCTATGCAGGG
CTTATCTATGCAGGG



GGCGGTCTAGATGCTTATCTATGCAGGGCCCT

Reference genome sequence

Short reads

```
.....  
GGTCTGGATGC  
CGGTCTGGATGC  
GCGGTCTGGATG  
GCGGTCTGGAT  
GGCGGTCTGGAT  
GGCGGTCTGGA  
TCTATGCAGGCCCT  
TCTATGCAGGCC  
ATCTATGCAGGC  
TATCTATGCAGGC  
TTATCTATGCAGG  
CTTATCTATGCAGG
```



Alignment of reads to the reference genome and SNP calling

SNP: A->G

GTCTGGATGCT	TCTATGCAGGCCCT
GGTCTGGATGC	TCTATGCAGGCC
CGGTCTGGATGC	ATCTATGCAGGC
GCGGTCTGGATG	TATCTATGCAGG
GCGGTCTGGAT	TTATCTATGCAGG
GCGGTCTGGAT	CTTATCTATGCAGG
GCGGTCTGGAT	CTTATCTATGCAGG
GCGGTCTGGA	CTTATCTATGCAGG

GGCGGTCTAGATGCTTATCTATGCAGGCCCT	

Reference genome sequence

Aligning reads to STRs

... CGAC **AT AT AT AT AT AT AT AT AT GATC...**

Reference genome

Aligning reads to STRs

Sequenced reads

The diagram illustrates the alignment of four sequenced reads against a reference genome. The reads are composed of alternating 'A' and 'T' nucleotides, represented by small boxes. Some boxes are highlighted in yellow, while others are black. The reference genome is shown at the bottom in a black box.

Read 1: CGAC **AT** AT AT AT AT AT

Read 2: AT AT AT AT AT AT GATC

Read 3: AC **AT** AT AT AT AT AT AT AT G

Read 4: GAC **AT** AT AT AT AT AT AT AT

Reference genome: ... CGAC **AT** AT AT AT AT AT AT AT GATC ...

Reference genome

Aligning short reads to STRs

Sequenced reads

AT AT AT AT AT AT AT

Aligned reads

CGAC	AT								
GAC	AT								
AC	AT	G							
	AT	GATC							
... CGAC	AT	GATC ...							

Reference genome

Aligning short reads to STRs

Sequenced reads

AT AT AT AT AT AT AT

Aligned reads

CGAC AT AT AT AT AT AT AT

GAC AT AT AT AT AT AT AT AT

AC AT AT AT AT AT AT AT AT G

AT AT AT AT AT AT AT GATC

... CGAC AT AT AT AT AT AT AT AT GATC ...

Only 1 informative read!

Reference genome

Aligning short reads to STRs

Sequenced reads

??? 

Aligned reads

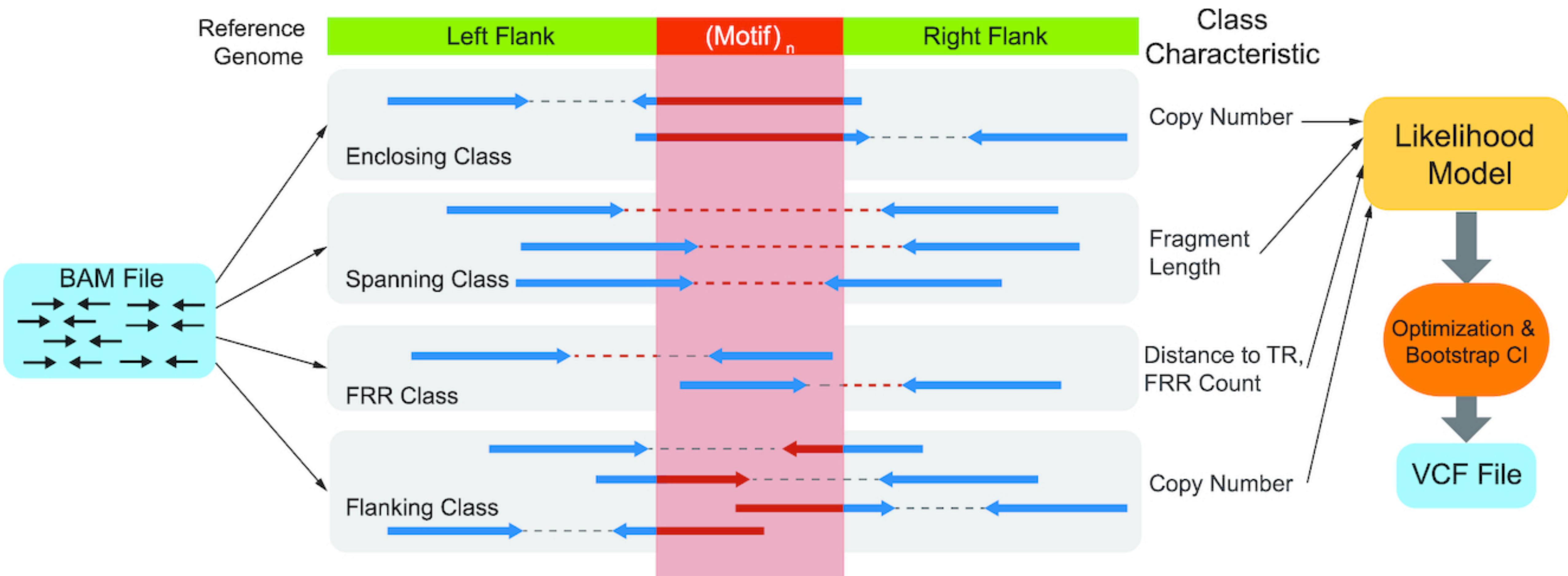
CGAC	AT							
GAC	AT							
AC	AT							
	AT	GATC						
								

Only 1 informative read!

Reference genome

Specialised tools are needed!

E.g., GangSTR



End product: VCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	mut
chr5	298	.	aaaaaaaaaaaa	.	.	.	END=309;RU=a;PERIOD=1;REF=12;GRID=9,15;STUTTERUP=0.05;STUTTERDOWN=0.05;STUT		
chr5	7241	.	aaaaaaaaaa	.	.	.	END=7249;RU=a;PERIOD=1;REF=9;GRID=6,12;STUTTERUP=0.05;STUTTERDOWN=0.05;STUT		
chr5	9390	.	aaaaaaaaaa	.	.	.	END=9399;RU=a;PERIOD=1;REF=10;GRID=7,13;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	10062	.	ttttttttttttt	END=10077;RU=t;PERIOD=1;REF=16;GRID=13,19;STUTTERUP=0.05;STUTTERDOWN=0.05;STU	
chr5	10673	.	aaaaaaaaaaaaaaa	END=10688;RU=a;PERIOD=1;REF=16;GRID=13,19;STUTTERUP=0.05;STUTTERDOWN=0.05;STU	
chr5	15411	.	ttttttttttttttttt	END=15439;RU=t;PERIOD=1;REF=29;GRID=26,32;STUTTERUP=0.05;STUTTERDOWN=0.05;STU	
chr5	15503	.	ttttttttt	.	.	.	END=15512;RU=t;PERIOD=1;REF=10;GRID=7,13;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	16887	.	tttttttttt	.	.	.	END=16897;RU=t;PERIOD=1;REF=11;GRID=8,14;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	17044	.	ttttttttttttt	.	.	.	END=17058;RU=t;PERIOD=1;REF=15;GRID=12,18;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	19394	.	aaaaaaaaaa	.	.	.	END=19402;RU=a;PERIOD=1;REF=9;GRID=6,12;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	20647	.	aaaaaaaaaa	.	.	.	END=20655;RU=a;PERIOD=1;REF=9;GRID=6,12;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	20965	.	aaaaaaaaaa	.	.	.	END=20974;RU=a;PERIOD=1;REF=10;GRID=7,13;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	21005	.	atatatatat	.	.	.	END=21014;RU=at;PERIOD=2;REF=5;GRID=2,8;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	22490	.	tttttttt	.	.	.	END=22498;RU=t;PERIOD=1;REF=9;GRID=6,12;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	24686	.	tttttttttt	.	.	.	END=24697;RU=t;PERIOD=1;REF=12;GRID=9,15;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	25452	.	aaaaaaaaaaa	.	.	.	END=25462;RU=a;PERIOD=1;REF=11;GRID=8,14;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	32035	.	aaaaaaaaaa	.	.	.	END=32043;RU=a;PERIOD=1;REF=9;GRID=6,12;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	32296	.	ttttttttttttt	.	.	.	END=32310;RU=t;PERIOD=1;REF=15;GRID=12,18;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	32477	.	tttttttt	.	.	.	END=32485;RU=t;PERIOD=1;REF=9;GRID=6,12;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	32685	.	atatatatat	.	.	.	END=32694;RU=at;PERIOD=2;REF=5;GRID=2,8;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	37544	.	tacatacataca	.	.	.	END=37555;RU=taca;PERIOD=4;REF=3;GRID=1,6;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	41515	.	tttttttt	.	.	.	END=41523;RU=t;PERIOD=1;REF=9;GRID=6,12;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	46982	.	ttttttttttttttttt	END=47004;RU=t;PERIOD=1;REF=23;GRID=20,26;STUTTERUP=0.05;STUTTERDOWN=0.05;STU	
chr5	47414	.	ttttttttttt	.	.	.	END=47426;RU=t;PERIOD=1;REF=13;GRID=10,16;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	47459	.	tttttttttt	.	.	.	END=47469;RU=t;PERIOD=1;REF=11;GRID=8,14;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	47471	.	ttttttttttt	.	.	.	END=47482;RU=t;PERIOD=1;REF=12;GRID=9,15;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	47958	.	tttttttttt	.	.	.	END=47968;RU=t;PERIOD=1;REF=11;GRID=8,14;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	53447	.	aaaaaaaaaaaaaaa	.	.	.	END=53461;RU=a;PERIOD=1;REF=15;GRID=12,18;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	53475	.	aaaaaaaaaaa	.	.	.	END=53485;RU=a;PERIOD=1;REF=11;GRID=8,14;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	57340	.	tttttttt	.	.	.	END=57348;RU=t;PERIOD=1;REF=9;GRID=6,12;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	58315	.	ttttttttttttt	.	.	.	END=58328;RU=t;PERIOD=1;REF=14;GRID=11,17;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	60745	.	tttttttt	.	.	.	END=60753;RU=t;PERIOD=1;REF=9;GRID=6,12;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	60865	.	tttttttt	.	.	.	END=60874;RU=t;PERIOD=1;REF=10;GRID=7,13;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	61581	.	aaaaaaaaaaaaaaa	END=61598;RU=a;PERIOD=1;REF=18;GRID=15,21;STUTTERUP=0.05;STUTTERDOWN=0.05;STU	
chr5	61748	.	aaaaaaaaaaaaaaa	END=61765;RU=a;PERIOD=1;REF=18;GRID=15,21;STUTTERUP=0.05;STUTTERDOWN=0.05;STU	
chr5	65546	.	ttttttttttttt	END=65562;RU=t;PERIOD=1;REF=17;GRID=14,20;STUTTERUP=0.05;STUTTERDOWN=0.05;STU	
chr5	65866	.	aaaaaaaaaaaaaa	.	.	.	END=65878;RU=a;PERIOD=1;REF=13;GRID=10,16;STUTTERUP=0.05;STUTTERDOWN=0.05;STU		
chr5	66016	.	ttttttttttttt	END=66032;RU=t;PERIOD=1;REF=17;GRID=14,20;STUTTERUP=0.05;STUTTERDOWN=0.05;STU	
chr5	66553	.	ttttttttttttttt	END=66572;RU=t;PERIOD=1;REF=20;GRID=17,23;STUTTERUP=0.05;STUTTERDOWN=0.05;STU	

For all things genome: Ensembl

e!Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Human (GRCh38.p14) ▾

Login/Register

Search Human (Homo sapiens)

Search all categories ▾ Search... Go

e.g. PPP2R2A or 8:26291508-26372680 or rs699 or osteoarthritis

Genome assembly: GRCh38.p14 (GCA_000001405.29)

More information and statistics
Download DNA sequence (FASTA)
Convert your data to GRCh38 coordinates
Display your data in Ensembl

Other assemblies
GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart Go

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

More about this genebuild
Download FASTA files for genes, cDNAs, ncRNA, proteins
Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins
Update your old Ensembl IDs

Pax6 INS FOXP2 BRCA2 DMD ssh Example gene

Example transcript

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

More about comparative analysis
Download alignments (EMF)

Regulation

What can I find? DNA methylation, transcription factor binding sites, histone modifications, and regulatory features such as enhancers and repressors, and microarray annotations.

More about the Ensembl regulatory build and microarray annotation
Experimental data sources
Download all regulatory features (GFF)

Variation

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes

More about variation in Ensembl
Download all variants (GVF)
Variant Effect Predictor VeIP

ATCGAGCT ATCCAGCT ATCGAGAT Example variant

Example phenotype

Example structural variant

For all things genome: Ensembl

Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Human (GRCh38.p14) ▾

Login/Register

Search Human (Homo sapiens)

Search all categories ▾ Search... Go

e.g. PPP2R2A or 8:26291508-26372680 or rs699 or osteoarthritis

Genome assembly: GRCh38.p14 (GCA_000001405.29)

More information and statistics
Download DNA sequence (FASTA)
Convert your data to GRCh38 coordinates
Display your data in Ensembl

Other assemblies
GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart Go

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

More about comparative analysis
Download alignments (EMF)

Regulation

What can I find? DNA methylation, transcription factor binding sites, histone modifications, and regulatory features such as enhancers and repressors, and microarray annotations.

More about the Ensembl regulatory build and microarray annotation
Experimental data sources
Download all regulatory features (GFF)

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

More about this genebuild
Download FASTA files for genes, cDNAs, ncRNA, proteins
Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins
Update your old Ensembl IDs

Pax6 INS FOXP2 BRCA2 DMD ssh Example gene

Example transcript

Variation

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes

More about variation in Ensembl
Download all variants (GVF)
Variant Effect Predictor VeIP

ATCGAGCT ATCCAGCT ATCGAGAT Example variant

Example phenotype

Example structural variant



Variant Effect Predictor (VEP)

Web Tools

- Web Tools
 - BLAST/BLAT
 - Variant Effect Predictor
 - Linkage Disequilibrium Calculator
 - Variant Recoder
 - File Chameleon
 - Assembly Converter
 - ID History Converter
 - VCF to PED Converter
 - Data Slicer

[!\[\]\(24ce36ad8a1745263e2734b9313a9dc2_img.jpg\) Configure this page](#)[!\[\]\(0d417cfc0d70f74bf5febdb4ffef61a7_img.jpg\) Custom tracks](#)[!\[\]\(0f56135fbafc50fedfcfff94393fe0f3_img.jpg\) Export data](#)[!\[\]\(240c12821e227464ff6b7614924c0018_img.jpg\) Share this page](#)[!\[\]\(5c15ba9377947e38b2e2544b77c612d5_img.jpg\) Bookmark this page](#)

Variant Effect Predictor ?

[New job](#)

Species:



Homo_sapiens



Assembly: GRCh38.p14

[Change species](#)

If you are looking for VEP for Human GRCh37, please go to [GRCh37 website](#).

Name for this job (optional):

Input data:

Either paste data:
1 65568 . A C . . .
2 265023 . C T . . .
3 319780 . GA G . . .

[Run instant VEP for current line >](#)

Examples: [Ensembl default](#), [VCF](#), [Variant identifiers](#), [HGVS notations](#), [SPDI](#)

Or upload file:

[Browse...](#) No file selected.

Or provide file URL:

 Ensembl/Gencode transcripts Ensembl/Gencode basic transcripts RefSeq transcripts

Transcript database to use: