Introduction to Whole Genome Sequencing

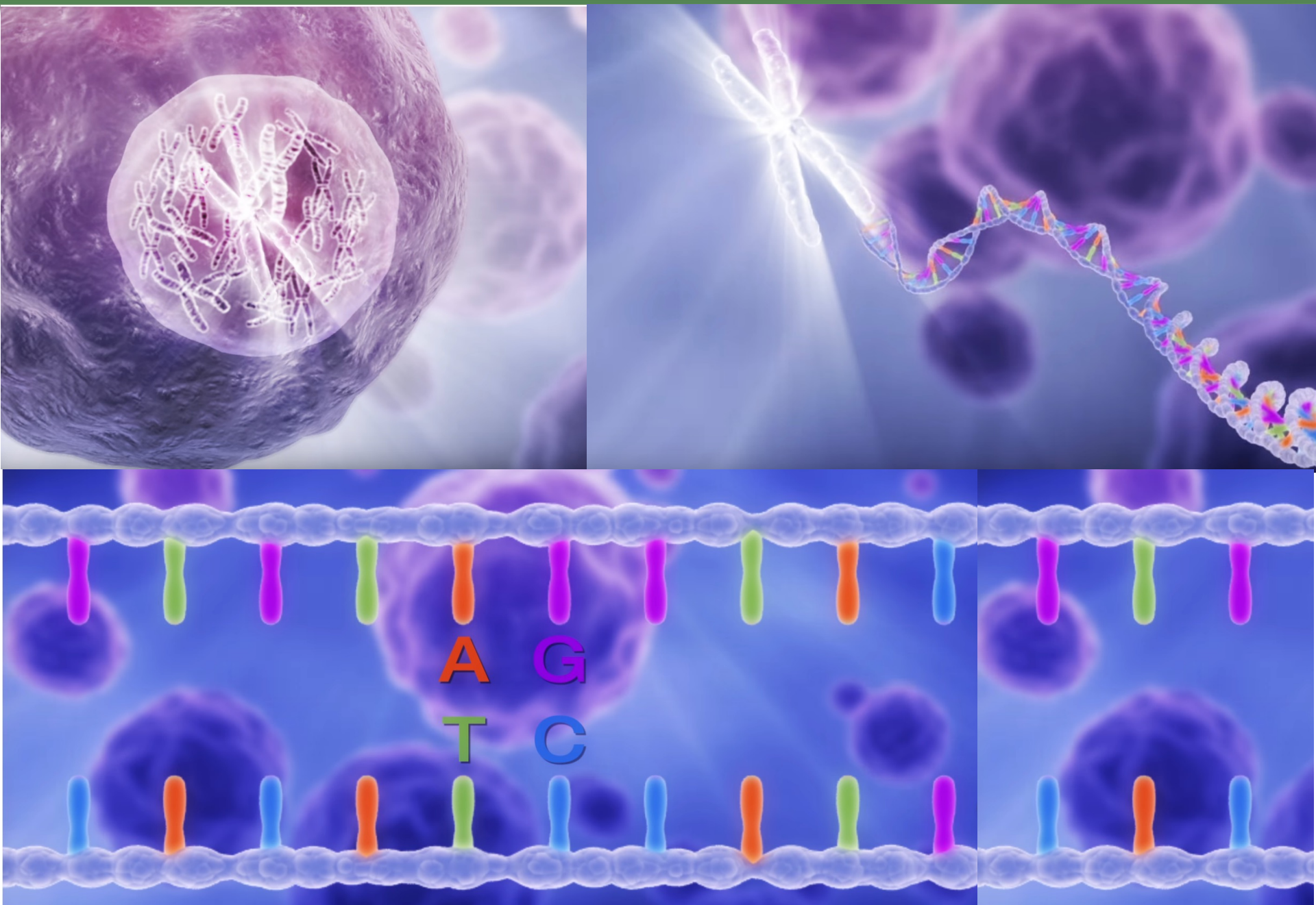
File formats

Rodrigo Pelicioni Savegnago

Department of Animal Science

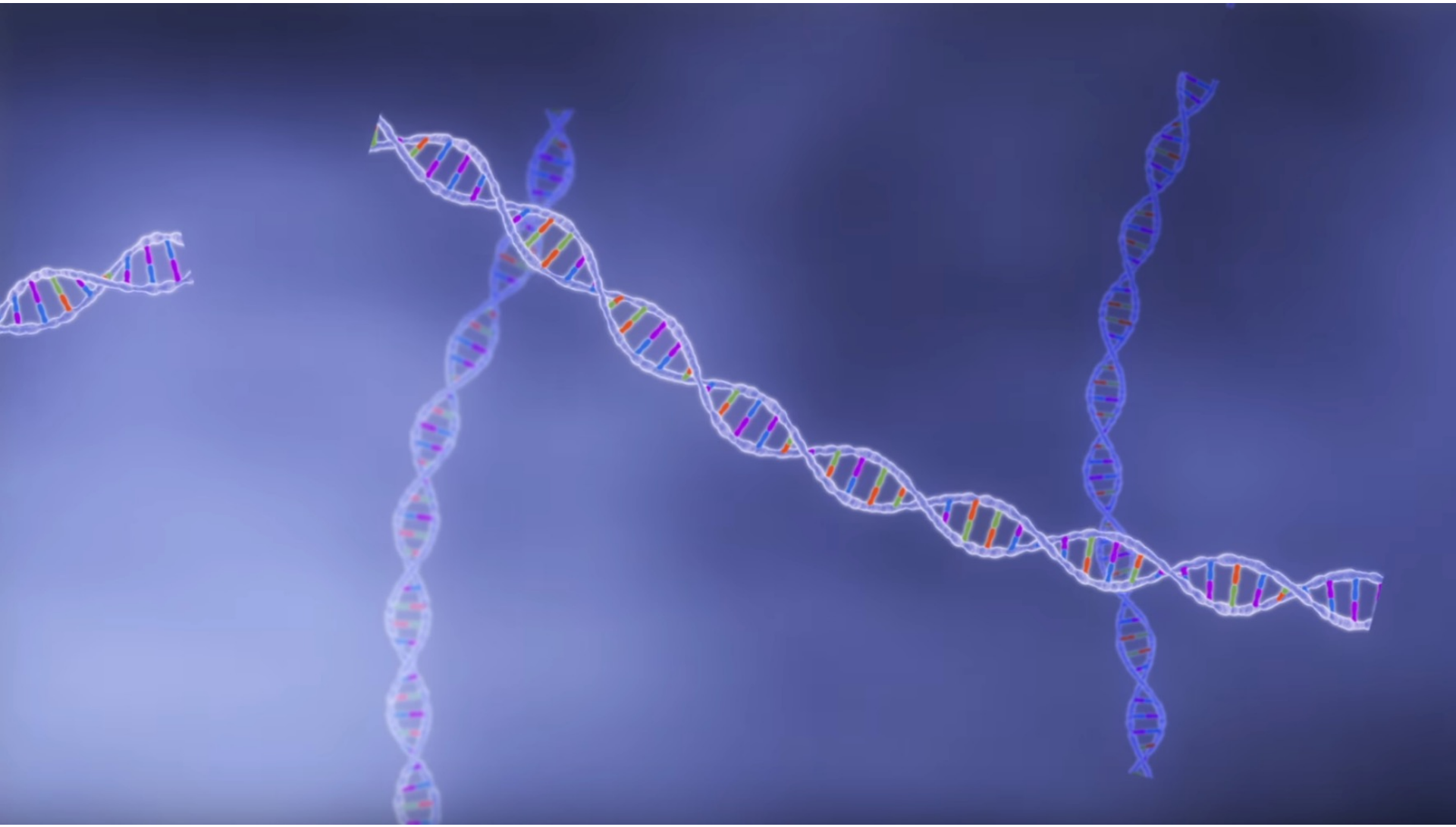
Intro Whole Genome Sequencing

MICHIGAN STATE UNIVERSITY



Intro Whole Genome Sequencing: reads

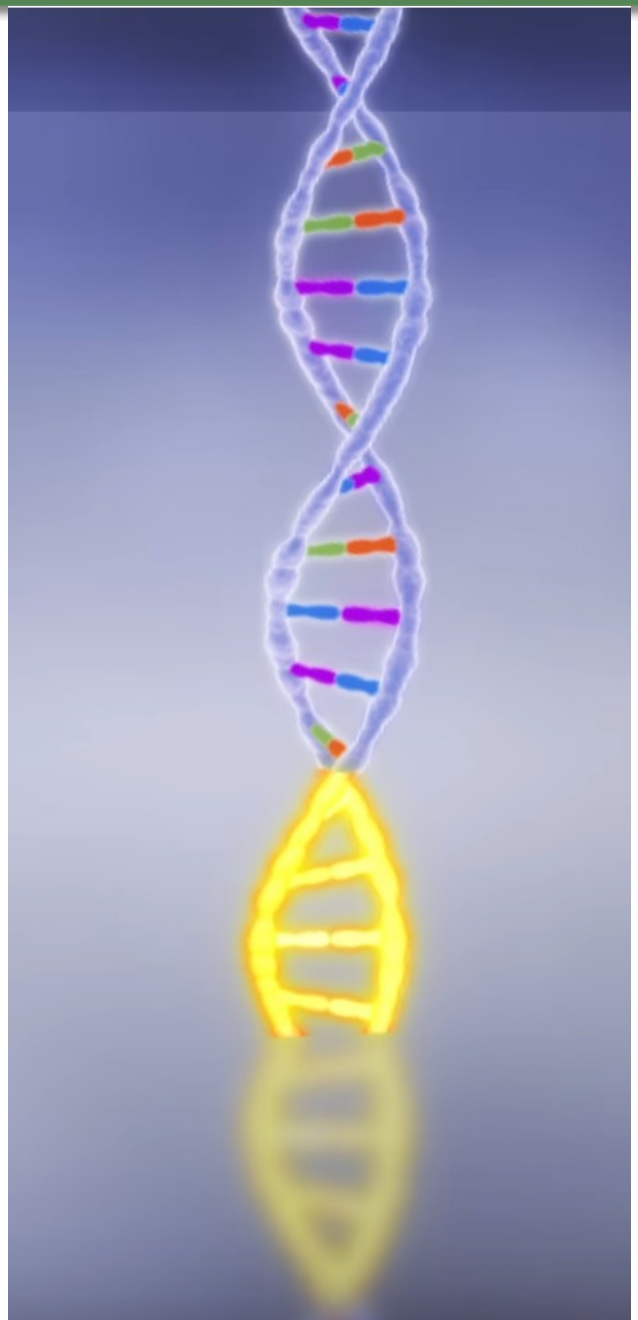
MICHIGAN STATE UNIVERSITY



Intro Whole Genome Sequencing: adapter

MICHIGAN STATE UNIVERSITY

Intro Whole Genome Sequencing: adapter

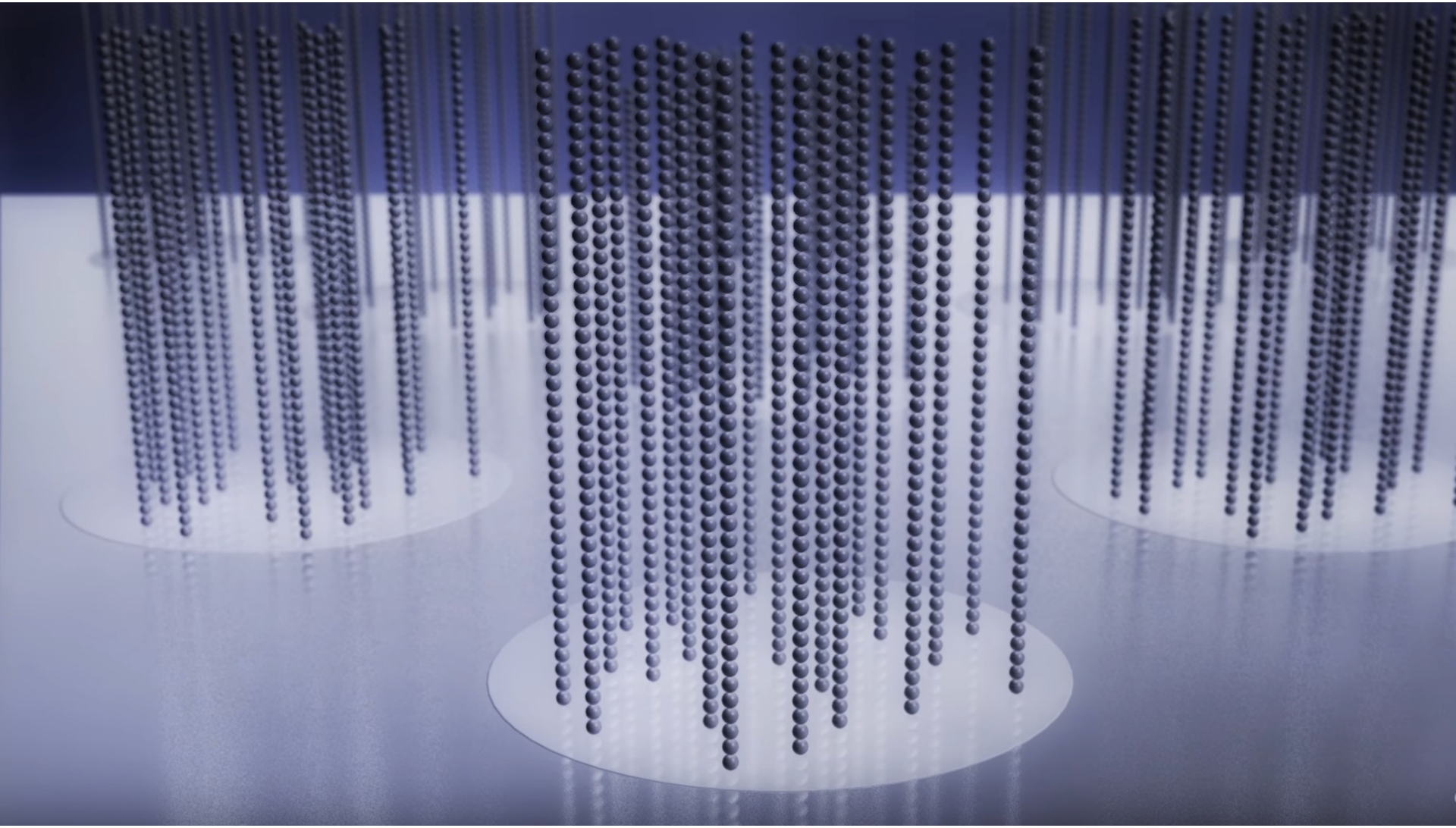


Chip, slide, flow cell…

HiSeq 2500

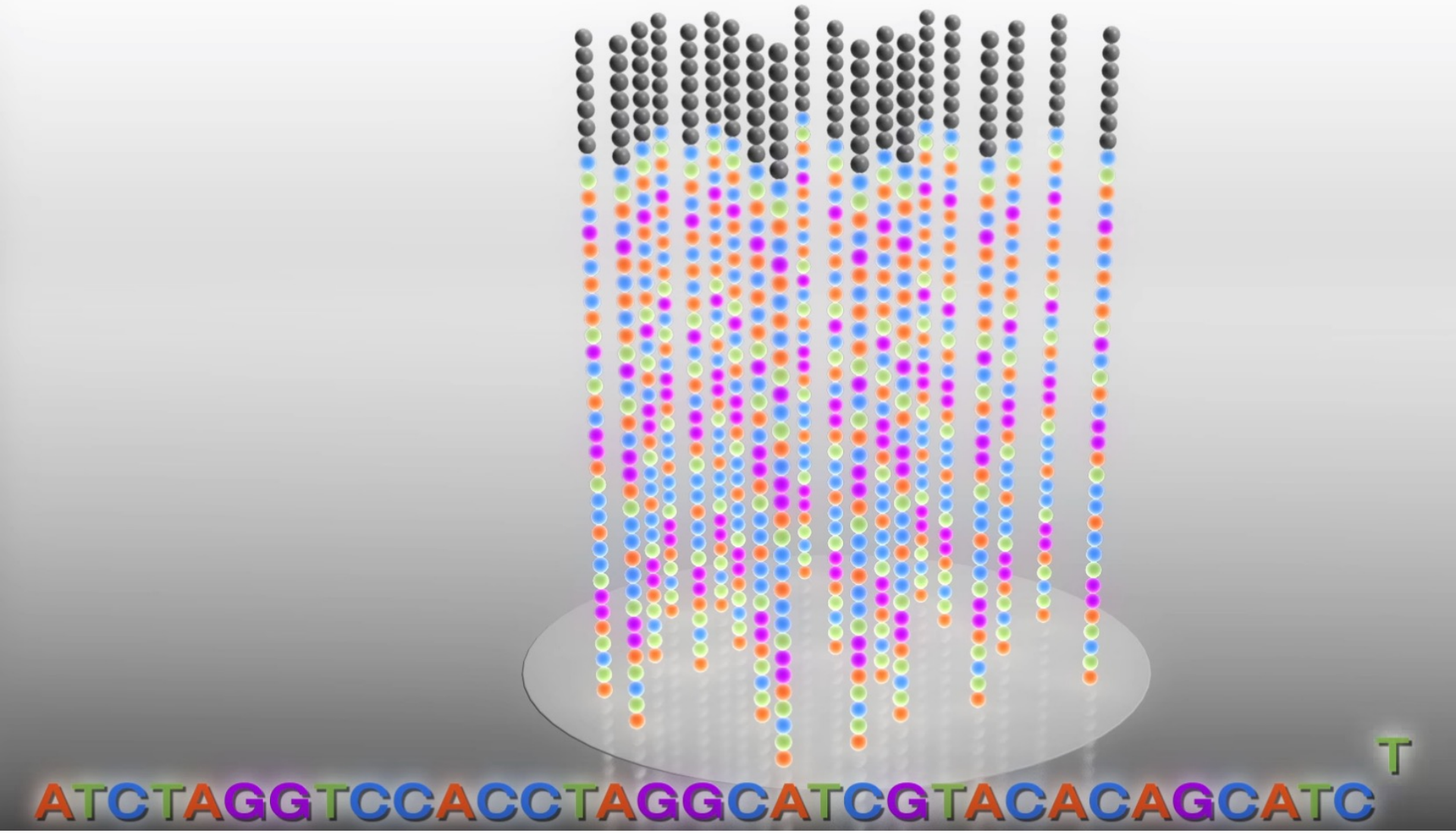
Intro Whole Genome Sequencing: amplification

MICHIGAN STATE UNIVERSITY



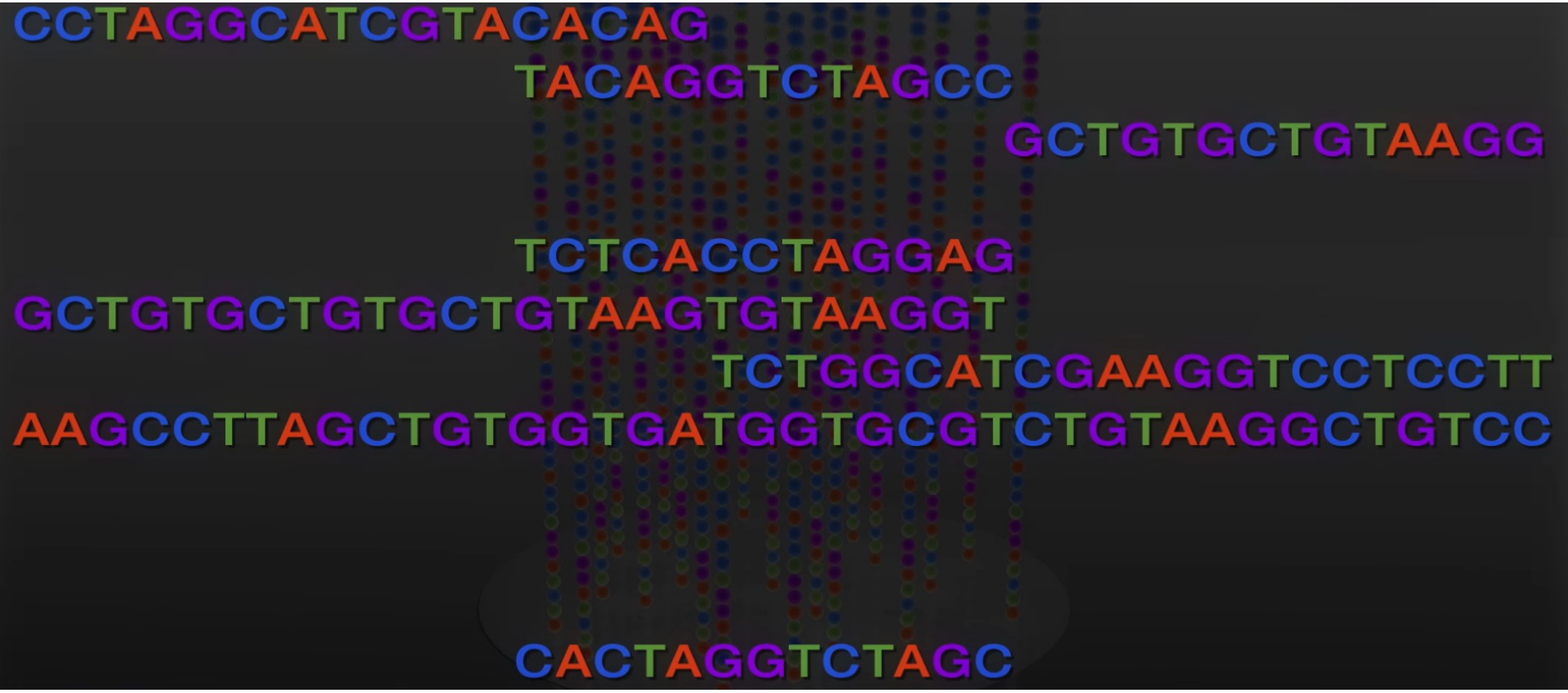
Intro Whole Genome Sequencing: signals

MICHIGAN STATE UNIVERSITY



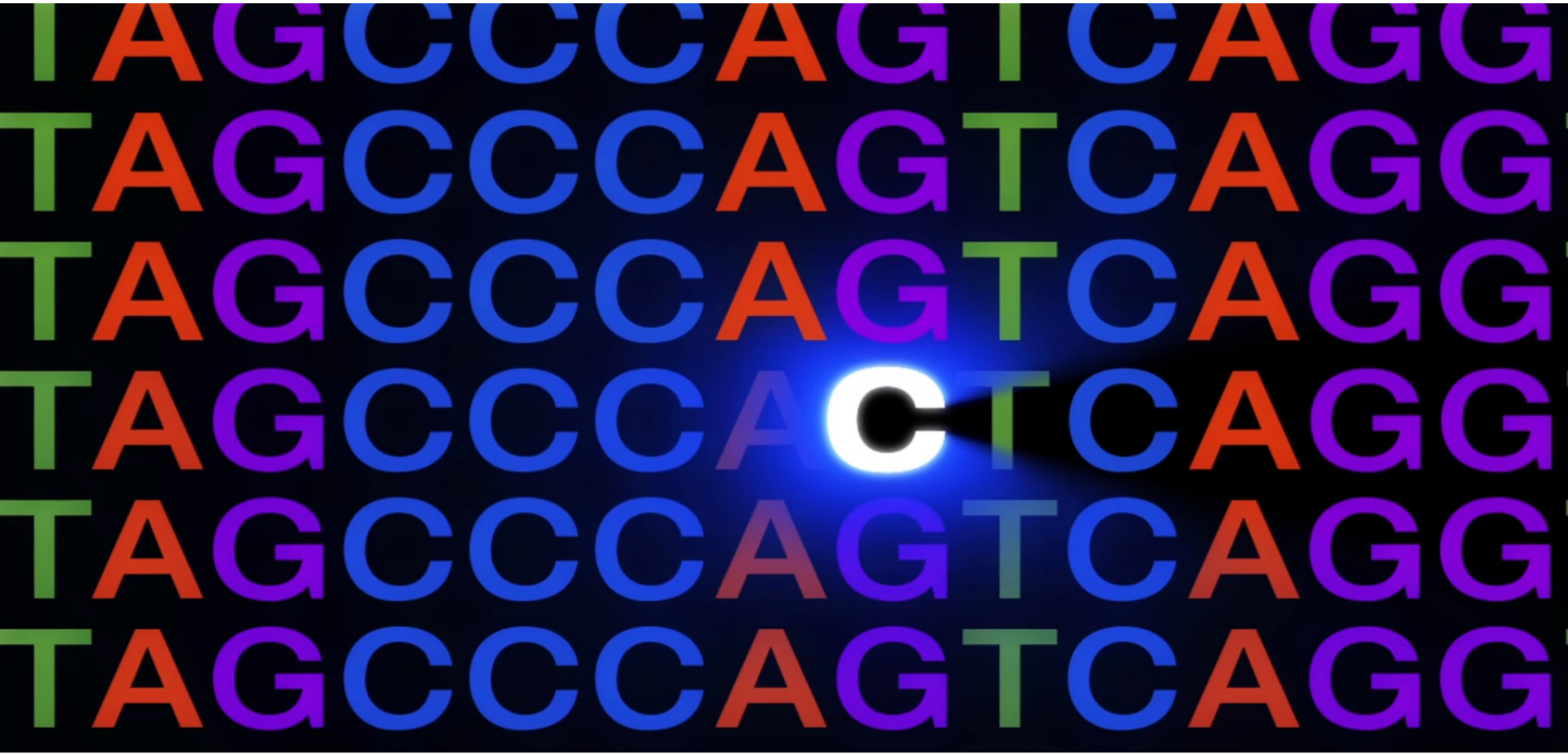
Intro Whole Genome Sequencing: bioinfo

MICHIGAN STATE UNIVERSITY



Intro Whole Genome Sequencing: bioinfo

MICHIGAN STATE UNIVERSITY



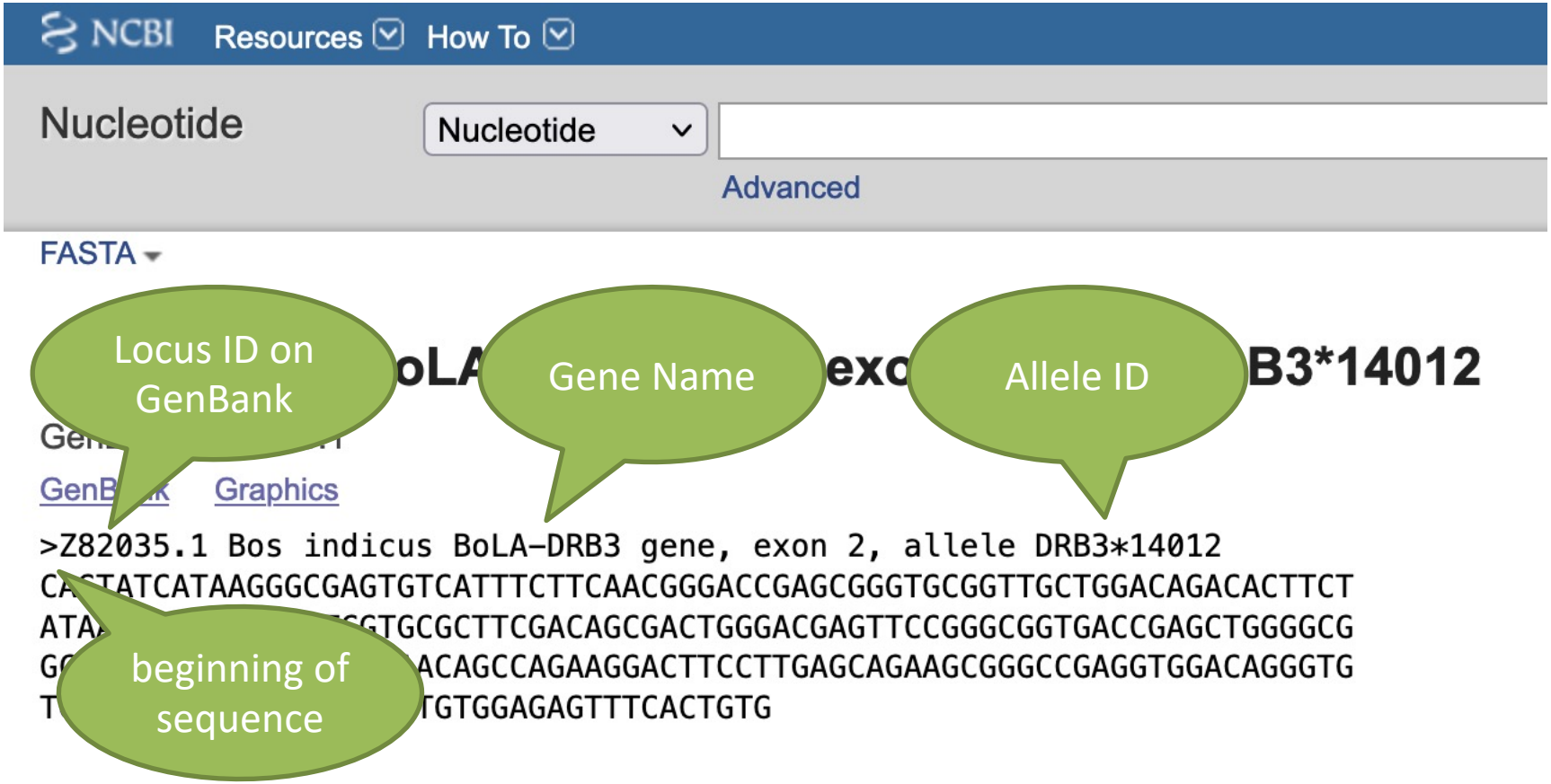
File formats

- Fasta
- Fastq
- SAM / BAM
- VCF / GVCF
- GTF / GFF
- BED

Fasta format (.fasta, .fa, .fna, frn)

MICHIGAN STATE UNIVERSITY

Text-based for storing biological sequence



Fasta format (.fasta, .fa, .fna, frn)

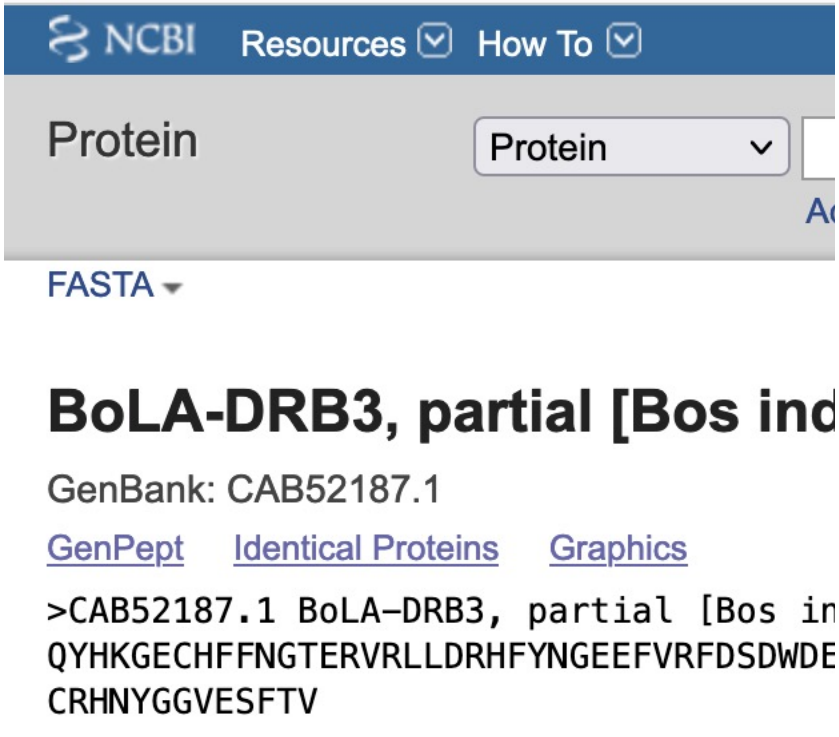
MICHIGAN STATE UNIVERSITY

```
>chr1
…
TGGACTTGTGGCAGGAATGAAATCCTTAGACCTGTGCTGTC
CAATATGGTAGCCACCAGGCACATGCAGCCACTGAGCACTT
GAAATGTGGATAGTCTGAATTGAGATGTGCCATAAGTGTAA
AATATGCACCAAATTTCAAAGGCTAGAAAAAAAGAATGTAA
AATATCTTATTATTTTATATTGATTACGTGCTAAAATAACC
ATATTTGGGATATACTGGATTTTAAAAATATATCACTAATT
TCAT
…
>chr2
…
>chr3
…
```

Fasta format (.fasta, .fa, .fna, frn)

MICHIGAN STATE UNIVERSITY

Text-based for storing I



Amino Acid	3-Letters	1-Letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

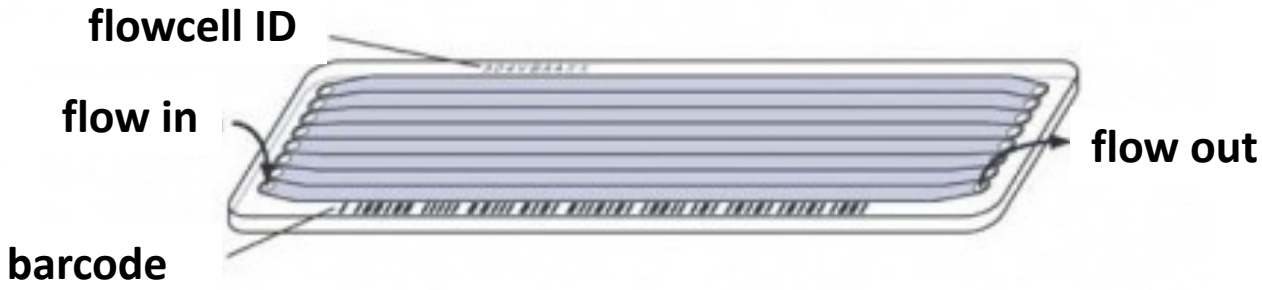
Fastq format

NOTE: for paired-end runs, there is a second file with one-to-one corresponding headers and reads

Fastq format: flowcell overview

Chip, slide, flow cell…

HiSeq 2500



Bronner et al. Improved Protocols for Illumina Sequencing**. Current Protocols in Human Genetics,** 2014.

Fastq format: Base Call Quality (Phred Scores)

Phred quality score (*Q)* with base-calling error probability P

$$Q = -10 \log_{10} P$$

Q score	Probability of base error	Base confidence
10	0.1	90%
20	0.01	99%
30	0.001	99.9%
40	0.0001	99.99%

SAM/BAM Format: Header

MICHIGAN STATE UNIVERSITY

```
[benpass align_genotype]$ samtools view -H allY.recalibrated.merge.bam
@HD   VN:1.0     GO:none    SO:coordinate
@SQ   SN:chrM    LN:16571
@SQ   SN:chr1    LN:249250621
@SQ   SN:chr2    LN:243199373
@SQ   SN:chr3    LN:198022430
…
@SQ   SN:chr19   LN:59128983
@SQ   SN:chr20   LN:63025520
@SQ   SN:chr21   LN:48129895
@SQ   SN:chr22   LN:51304566
@SQ   SN:chrX    LN:155270560
@SQ   SN:chrY    LN:59373566
…
@RG   ID:86-191  PL:ILLUMINA     LB:IL500   SM:86-191-1
@RG   ID:BsK010  PL:ILLUMINA     LB:IL501   SM:BsK010-1
@RG   ID:Bsk136  PL:ILLUMINA     LB:IL502   SM:Bsk136-1
@RG   ID:MAK001  PL:ILLUMINA     LB:IL503   SM:MAK001-1
@RG   ID:NG87    PL:ILLUMINA     LB:IL504   SM:NG87-1
…
@RG   ID:SDH023  PL:ILLUMINA     LB:IL508   SM:SDH023
@PG   ID:GATK IndelRealigner     VN:2.0-39-gd091f72    CL:knownAlleles=[] targetIntervals=tmp.intervals.l
@PG   ID:bwa     PN:bwa     VN:0.6.2-r126
```

samtools to view bam header

sort order

reference sequence names with lengths

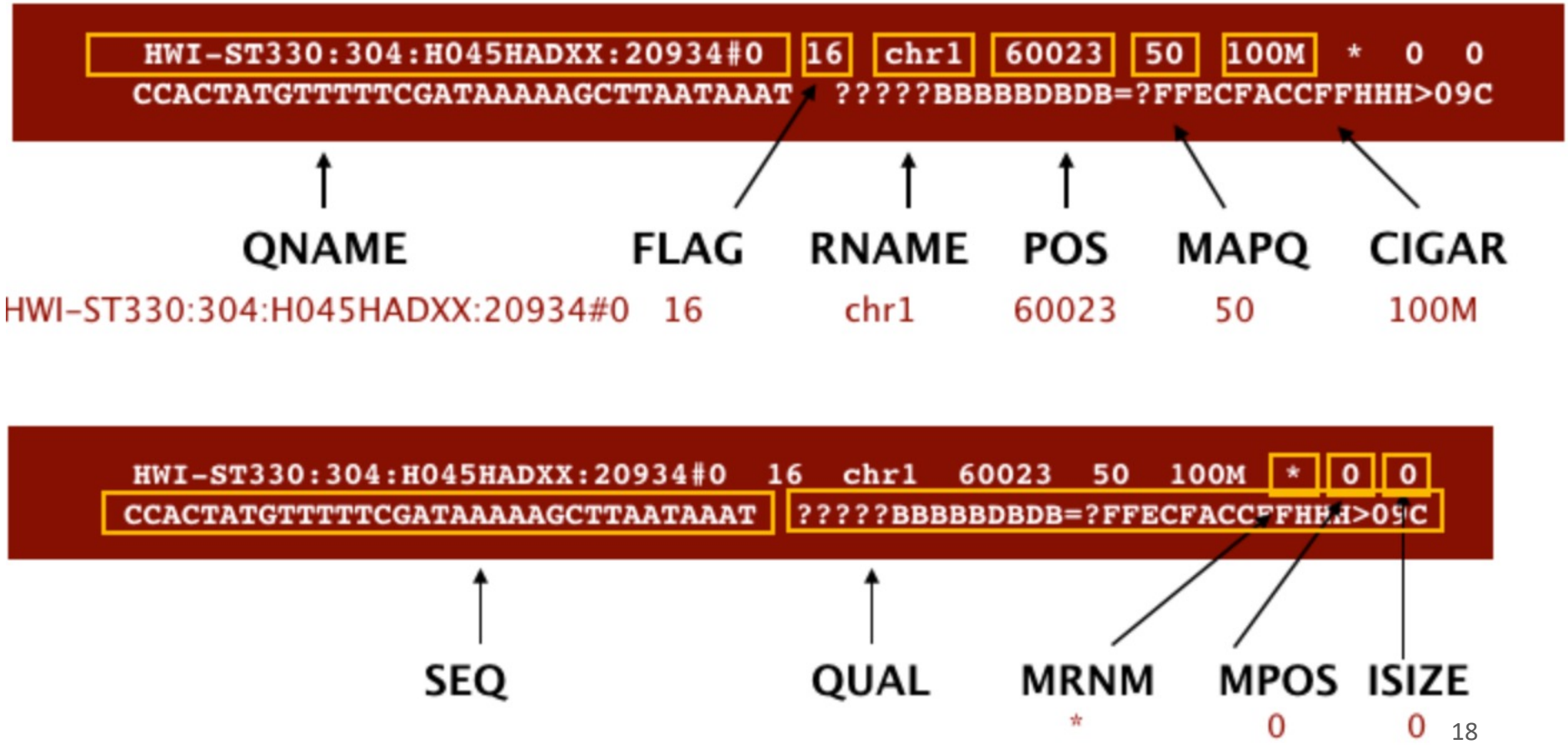
read groups with platform, library and sample information

program (analysis) history

SAM/BAM Format: Alignment Records

MICHIGAN STATE UNIVERSITY

Sequence Alignment Map (SAM) is a text-based format originally for storing biological sequences aligned to a reference sequence



SAM/BAM Format: Alignment Records

MICHIGAN STATE UNIVERSITY

Before alignment

```
RefPos:     1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Reference:  C C A T A C T G A A  C  T  G  A  C  T  A  A  C
Read: ACTAGAATGGCT
```

After alignment

```
RefPos:     1 2 3 4 5 6 7   8 9 10 11 12 13 14 15 16 17 18 19
Reference:  C C A T A C T   G A A  C  T  G  A  C  T  A  A  C
Read:             A C T A G A A  -  T  G  G  C  T
```

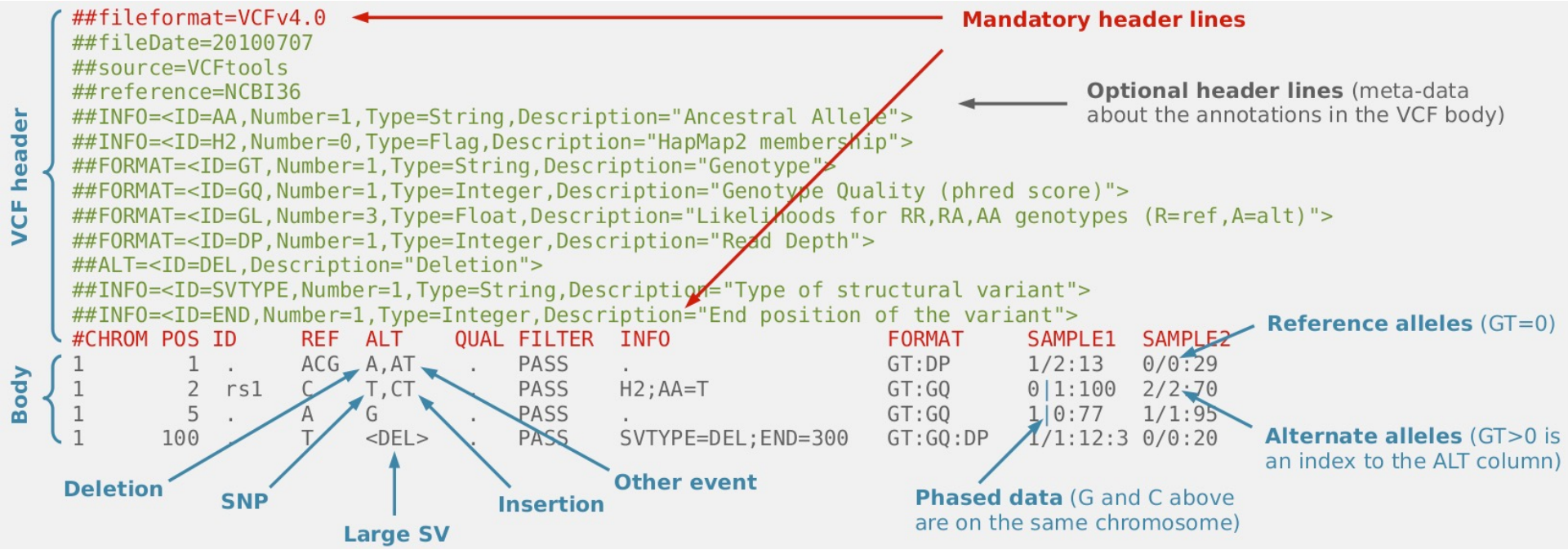
POS: 5

CIGAR: 3M 1I 3M 1D 5M

VCF File

MICHIGAN STATE UNIVERSITY

Variant Calling Format is a tab-delimited text file that is used to describe single nucleotide variants (SNVs) as well as insertions, deletions, and other sequence variations.



https://bioinf.comav.upv.es/courses/sequence_analysis/snp_calling.html

GVCF File

MICHIGAN STATE UNIVERSITY

GVCF

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12878
20 10000204 . A <NON_REF> . . END=10000210 GT:DP:GQ:MIN_DP:PL 0/0:33:84:31:0,84,1260
20 10000211 . C T,<NON_REF> 326.77 . BaseQRankSum=2.340;ClippingRankSum=-1.162;DP=35;
MLEAC=1,0;MLEAF=0.500,0.00;MQ=60.00;MQRankSum=0.623;ReadPosRankSum=0.152
GT:AD:DP:GQ:PL:SB 0/1:21,14,0:35:99:355,0,526,418,568,986:12,9,7,7
20 10000212 . A <NON_REF> . . END=10000216 GT:DP:GQ:MIN_DP:PL 0/0:35:90:33:0,90,1350
```

VCF

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12877 NA12878 NA12882
20 10000117 . C T 1606.16 . AC=4;AF=0.667;AN=6;BaseQRankSum=1.66;ClippingRankSum=0.340;DP=85;
FS=5.718;MLEAC=4;MLEAF=0.667;MQ=60.36;MQRankSum=1.45;QD=18.90;ReadPosRankSum=1.62;SOR=1.503
GT:AD:DP:GQ:PL 0/1:17,15:32:99:399,0,439 0/1:11,12:23:99:291,0,292 1/1:0,30:30:90:948,90,0
20 10000211 . C T 1765.16 . AC=4;AF=0.667;AN=6;BaseQRankSum=2.34;ClippingRankSum=-1.147e+00;
DP=97;FS=0.809;MLEAC=4;MLEAF=0.667;MQ=60.00;MQRankSum=1.21;QD=18.58;ReadPosRankSum=0.152;SOR=0.831
GT:AD:DP:GQ:PL 0/1:13,10:23:99:243,0,341 0/1:21,14:35:99:355,0,526 1/1:0,37:37:99:1199,111,0
20 10000439 . T G 1982.13 . AC=5;AF=0.833;AN=6;BaseQRankSum=1.31;ClippingRankSum=0.549;DP=103;
FS=0.000;MLEAC=5;MLEAF=0.833;MQ=60.00;MQRankSum=0.972;QD=19.82;ReadPosRankSum=1.56;SOR=0.839
GT:AD:DP:GQ:PL 0/1:18,12:30:99:208,0,455 1/1:0,29:29:86:795,86,0 1/1:1,40:41:99:1010,110,0
```

BED Format

BED files are used to define capture regions in the assembly. The BED file column specifications are as follows:

Custom BED File Setup			
chrom (required)	**chromStart (required)**	**chromEnd (required)**	**Column 4 and beyond (ignored)**
The name of the chromosome or scaffold. Numbers are preferred, but chr or ch prefixes are allowed.	Starting position for the feature.	Ending position for the feature.	Data in these columns are ignored.

```
1   1     10
1   50    100
1   105   110
```

GTF / GFF3 Format

MICHIGAN STATE UNIVERSITY

GTF = gene transfer format
GFF = general feature format

File format used for describing genes and other features of DNA, RNA and protein sequences.

1 transcribed_unprocessed_pseudogene gene 11869 14409 . + . gene_id "ENSG00000223972"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype "transcribed_unprocessed_pseudogene";

1 processed_transcript transcript 11869 14409 . + . gene_id "ENSG00000223972"; transcript_id "ENST00000456328"; gene_name "DDX11L1"; gene_sourc e "havana"; gene_biotype "transcribed_unprocessed_pseudogene"; transcript_name "DDX11L1-002"; transcript_source "havana";

General GFF structure

Position index	Position name	Description
1	sequence	The name of the sequence where the feature is located.
2	source	Keyword identifying the source of the feature, like a program (e.g. Augustus or RepeatMasker) or an organization (like TAIR).
3	feature	The feature type name, like "gene" or "exon". In a well structured GFF file, all the children features always follow their parents in a single block (so all exons of a transcript are put after their parent "transcript" feature line and before any other parent transcript line). In GFF3, all features and their relationships should be compatible with the standards released by the Sequence Ontology Project.
4	start	Genomic start of the feature, with a **1-base offset**. This is in contrast with other 0-offset half-open sequence formats, like BED.
5	end	Genomic end of the feature, with a **1-base offset**. This is the same end coordinate as it is in 0-offset half-open sequence formats, like BED.[*citation needed*]

GTF / GFF3 Format

MICHIGAN STATE UNIVERSITY

6	score	Numeric value that generally indicates the confidence of the source in the annotated feature. A value of "." (a dot) is used to define a null value.
7	strand	Single character that indicates the strand of the feature; it can assume the values of "+" (positive, or 5'->3'), "-", (negative, or 3'->5'), "." (undetermined).
8	phase	phase of CDS features; it can be either one of 0, 1, 2 (for CDS features) or "." (for everything else). See the section below for a detailed explanation.
9	attributes	All the other information pertaining to this feature. The format, structure and content of this field is the one which varies the most between the three competing file formats.

Hands on

MICHIGAN STATE UNIVERSITY

- Working with fasta, fastq, sam/bam, vcf, and gtf
- Specific bioinformatic programs
- Linux language