



Genomics in the Cloud

Book Club - Week 3

December 14, 2020

Agenda

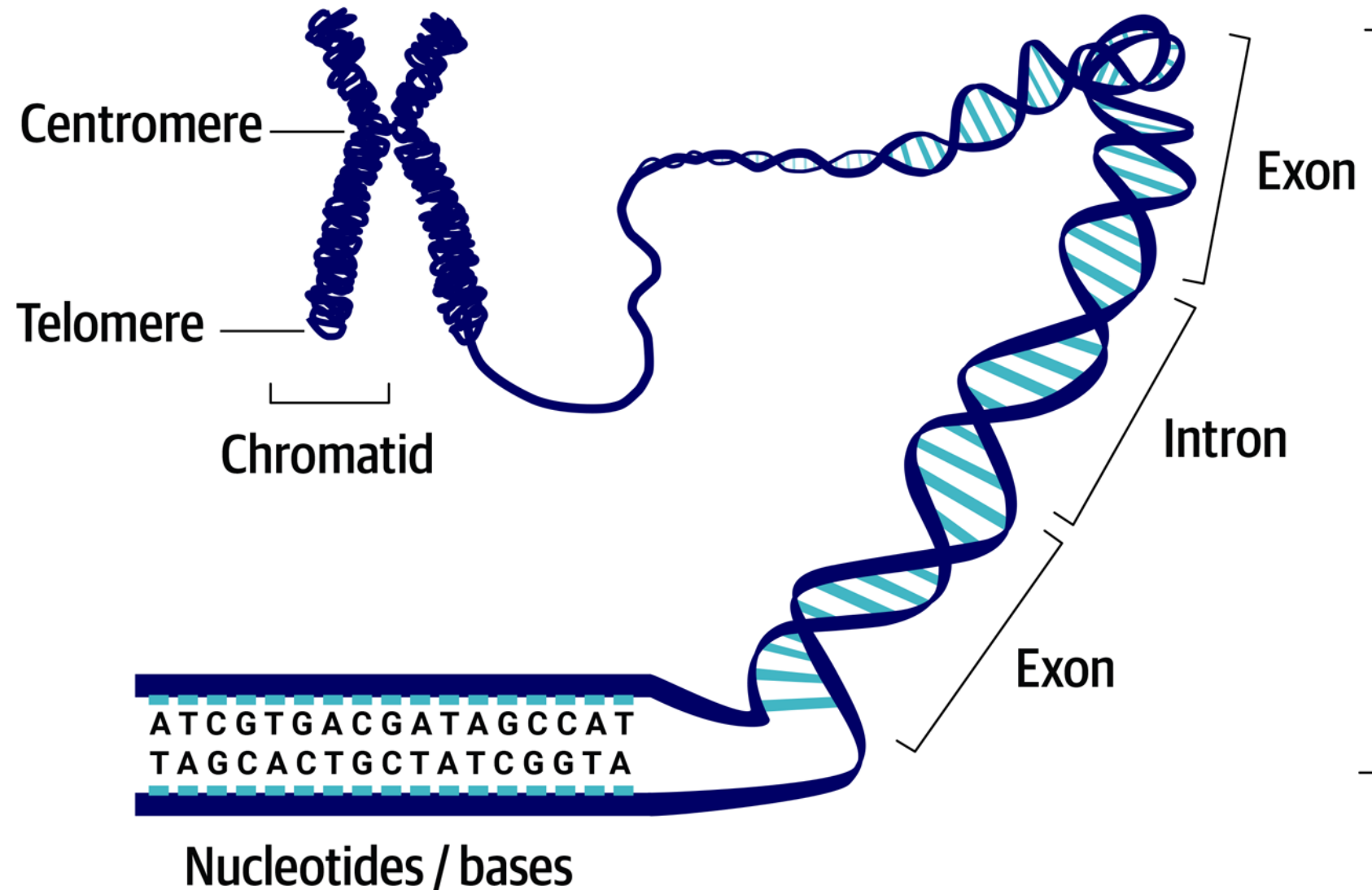
- Chapter 2: Genomics in a Nutshell
- Additional resources
- Open discussion



Chapter 2: Genomics in a Nutshell

Genomics in the Cloud by Geraldine A. Van der Auwera and Brian D. O'Connor (O'Reilly). Copyright 2020 The Broad Institute, Inc. and Brian O'Connor, 978-1-491-97519-0.

From chromosome to base



Useful mnemonics

All Those Genetic Codes (A-T G-C)

Exons are exciting

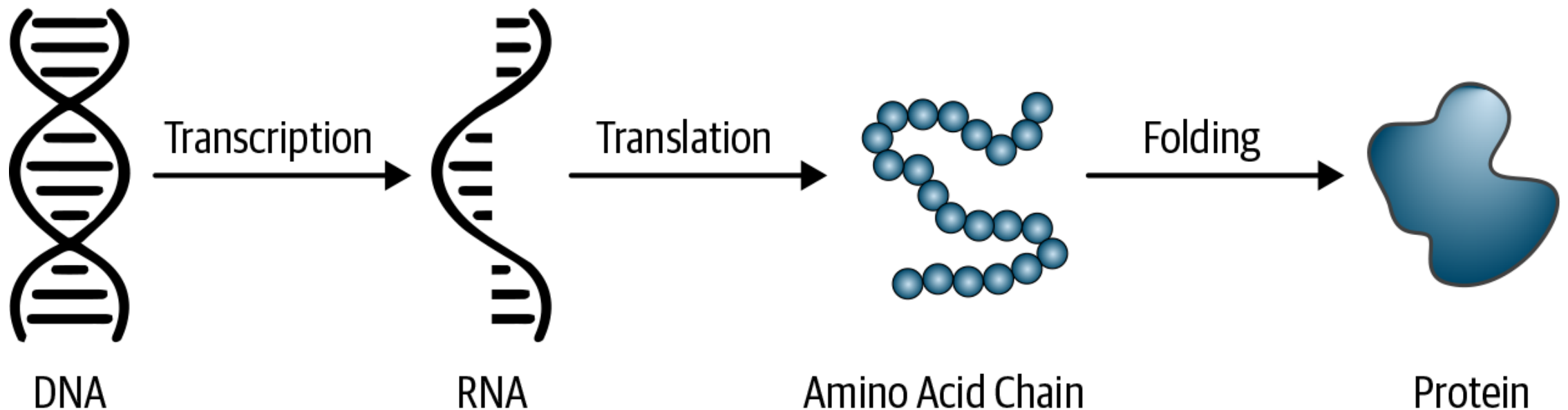
Introns are the bits in between



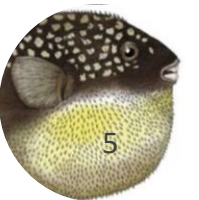
Central dogma of biology

Useful mnemonics

Translation
happens *later*



Central dogma: ...once "information" has passed into protein it cannot get out again. -- Crick (1958)



mRNA translation to amino acids

		Second letter				
		U	C	A	G	
First letter	U	UUU Phenylalanine	UCU Serine	UAU Tyrosine	UGU Cysteine	Third letter
		UUC	UCC	UAC	UGC	
		UUA Leucine	UCA	UAA Stop codon	UGA Stop codon	
		UUG	UCG	UAG Stop codon	UGG Tryptophan	
	C	CUU Leucine	CCU Proline	CAU Histidine	CGU Arginine	
		CUC	CCC	CAC	CGC	
		CUA	CCA	CAA Glutamine	CGA	
		CUG	CCG	CAG	CGG	
	A	AUU Isoleucine	ACU Threonine	AAU Asparagine	CGU Serine	
		AUC	ACC	AAC	CGC	
		AUA Methionine; start codon	ACA	AAA Lysine	CGA Arginine	
		AUG	ACG	AAG	CGG	
	G	GUU Valine	GCU Alanine	GAU Aspartic acid	GGU Glycine	
		GUC	GCC	GAC	GGC	
		GUA	GCA	GAA Glutamic acid	GGA	
		GUG	GCG	GAG	GGG	

Useful example

DNA: TACTTGATC

A sticks to T/U

RNA: AUGAACUAG

Amino acid: Asparagine



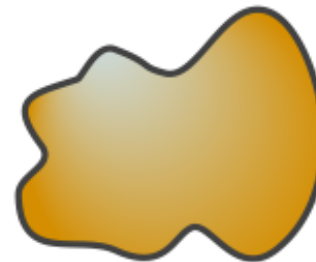
DNA mutations



Normal protein



or

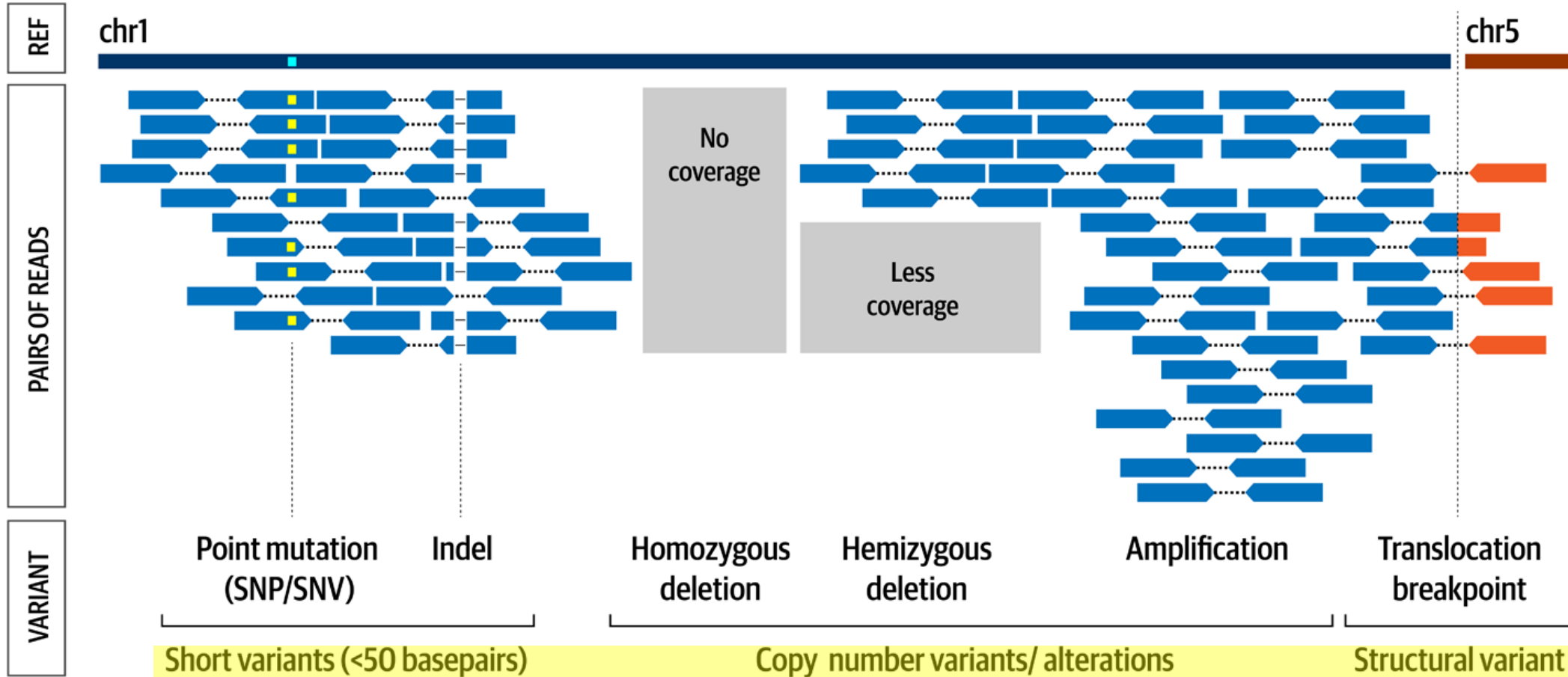


Abnormal protein

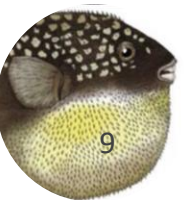
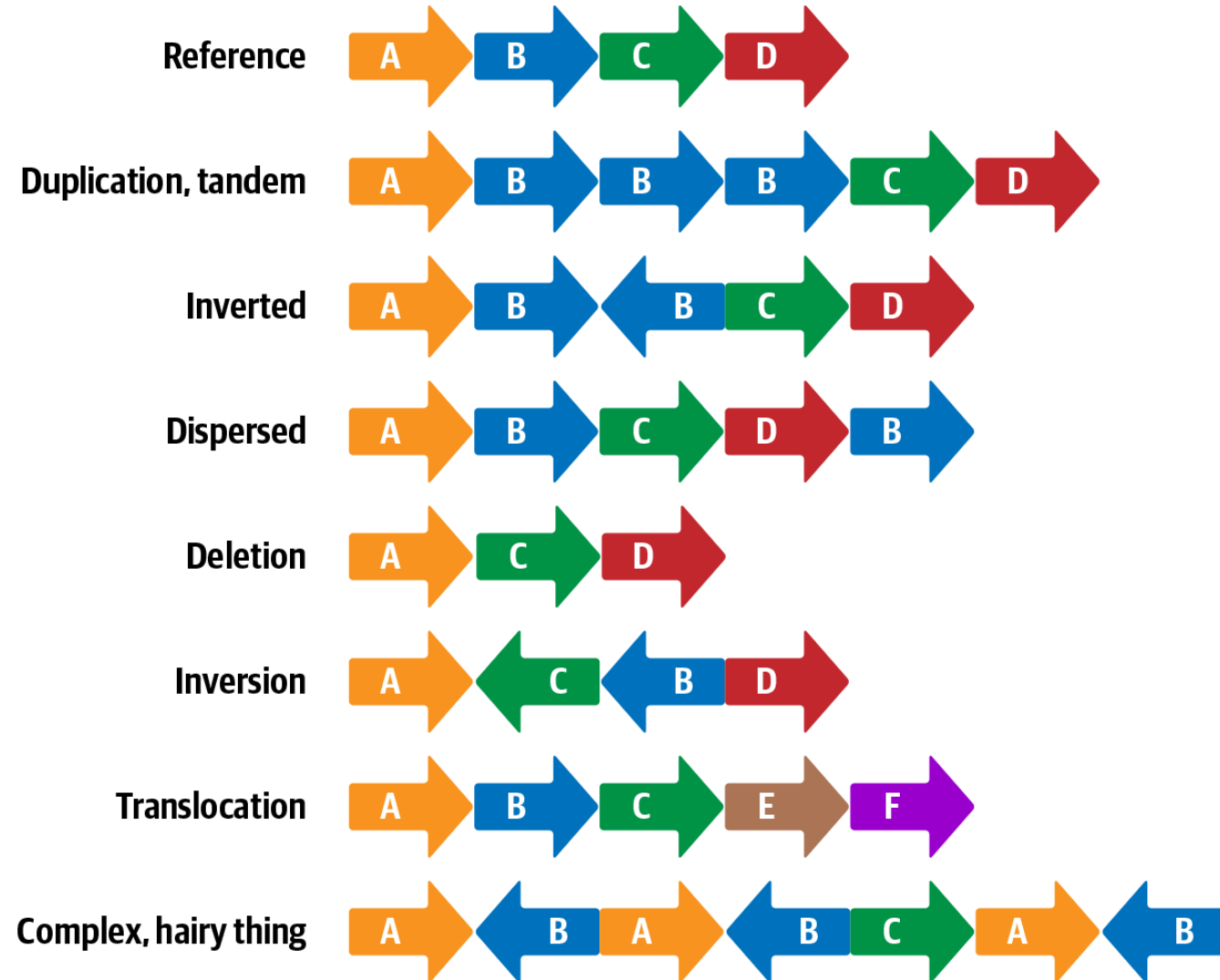
No protein



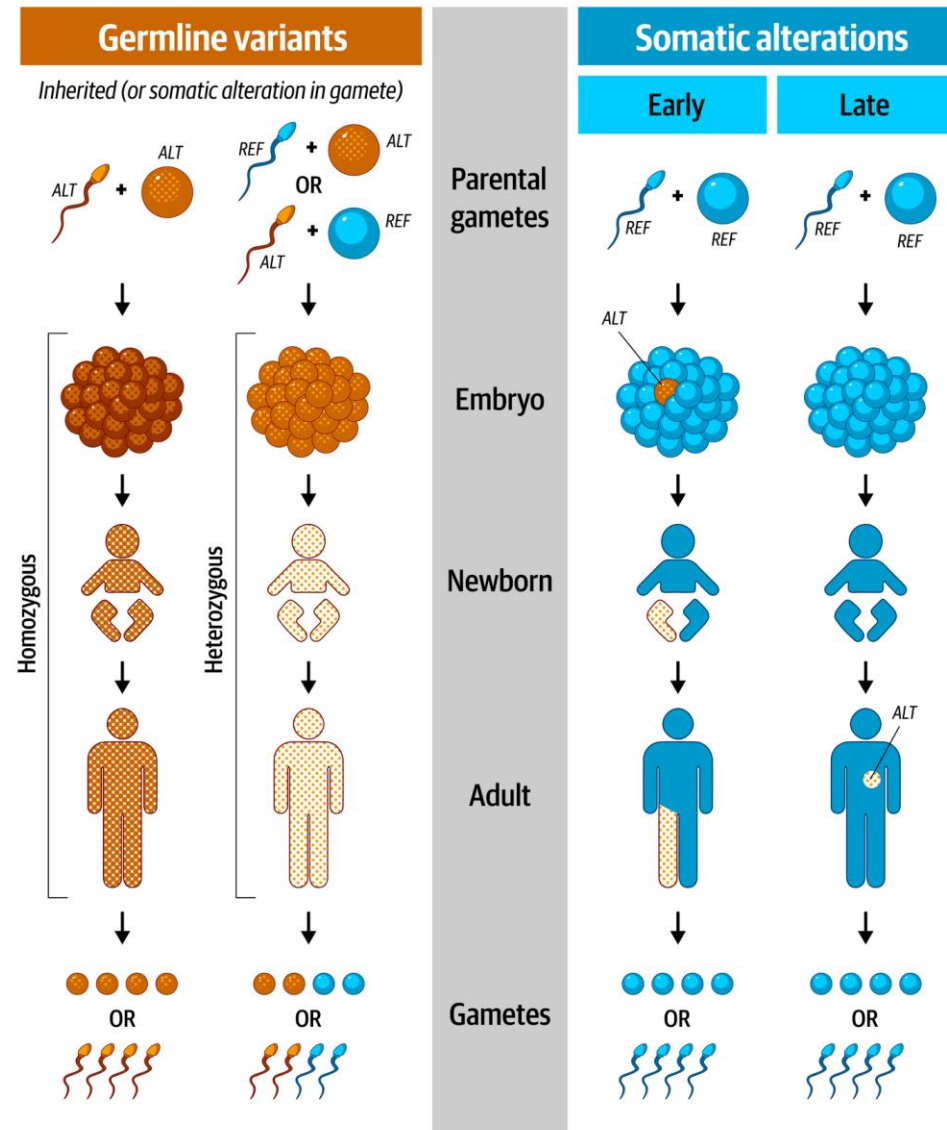
Variant / CNV / SV



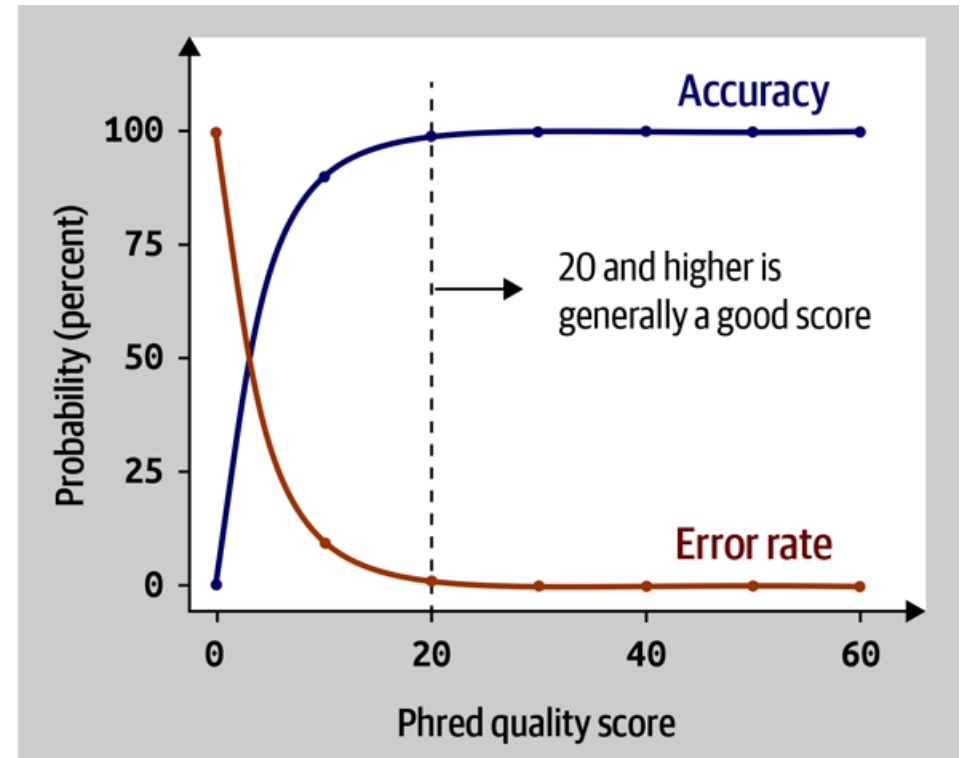
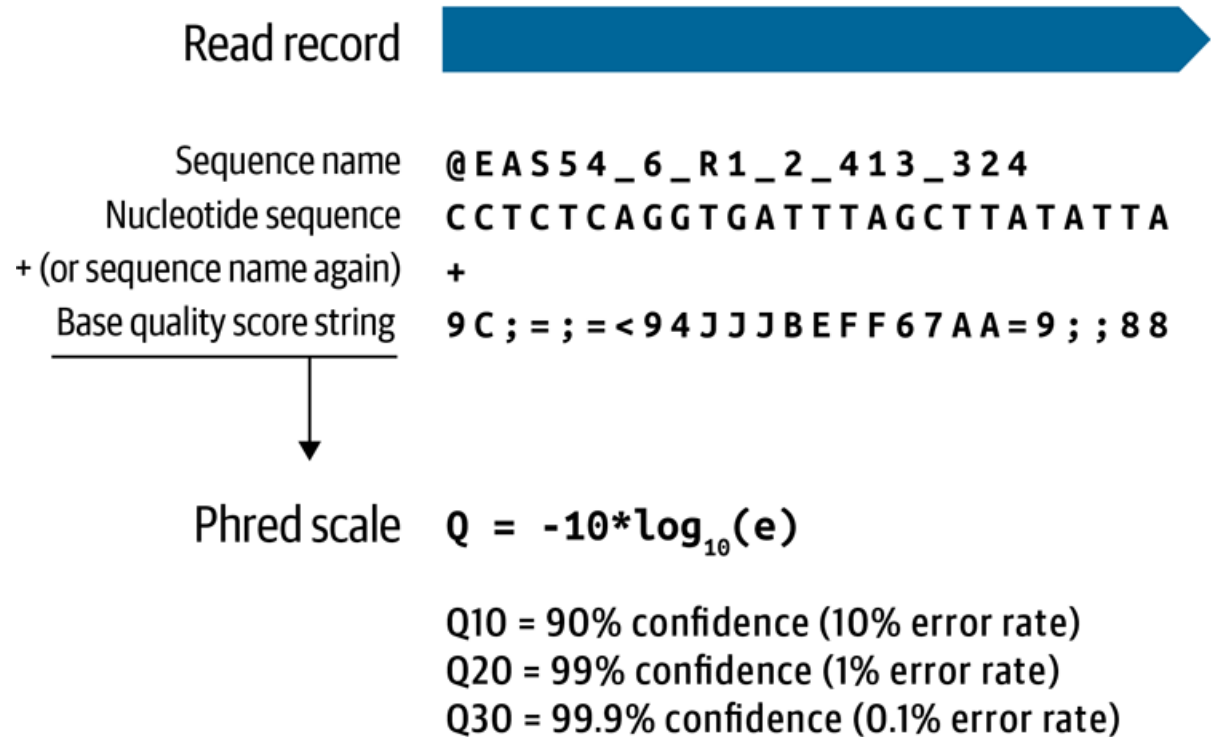
Structural variants



Germline variants vs somatic alterations



FASTQ file format



FASTQ files are about 200 GBytes (60 GBytes compressed)



Mapping FASTQ to a reference → SAM/BAM

Header lines starting with @ symbol describing various metadata for *all* reads

@HD VN:1.6 SO:coordinate	– BAM header line
@SQ SN:chr1 LN:248956422	– Reference sequence dictionary entries
@SQ SN:chr2 LN:242193529	
@RG ID:RG1 SM:SAMPLE_A	– Read group(s)

Records containing structured read information (1 line per read/record)

Read name	Position	CIGAR	Read sequence	Metadata
SLX1:1:127:63:4	99	chr1 10052169	60	23M3D10M = 14 10
GAAGATACTGGTT	768832'48::::	RG:Z:A ...		

Flags MAPQ Mate information Phred quality scores

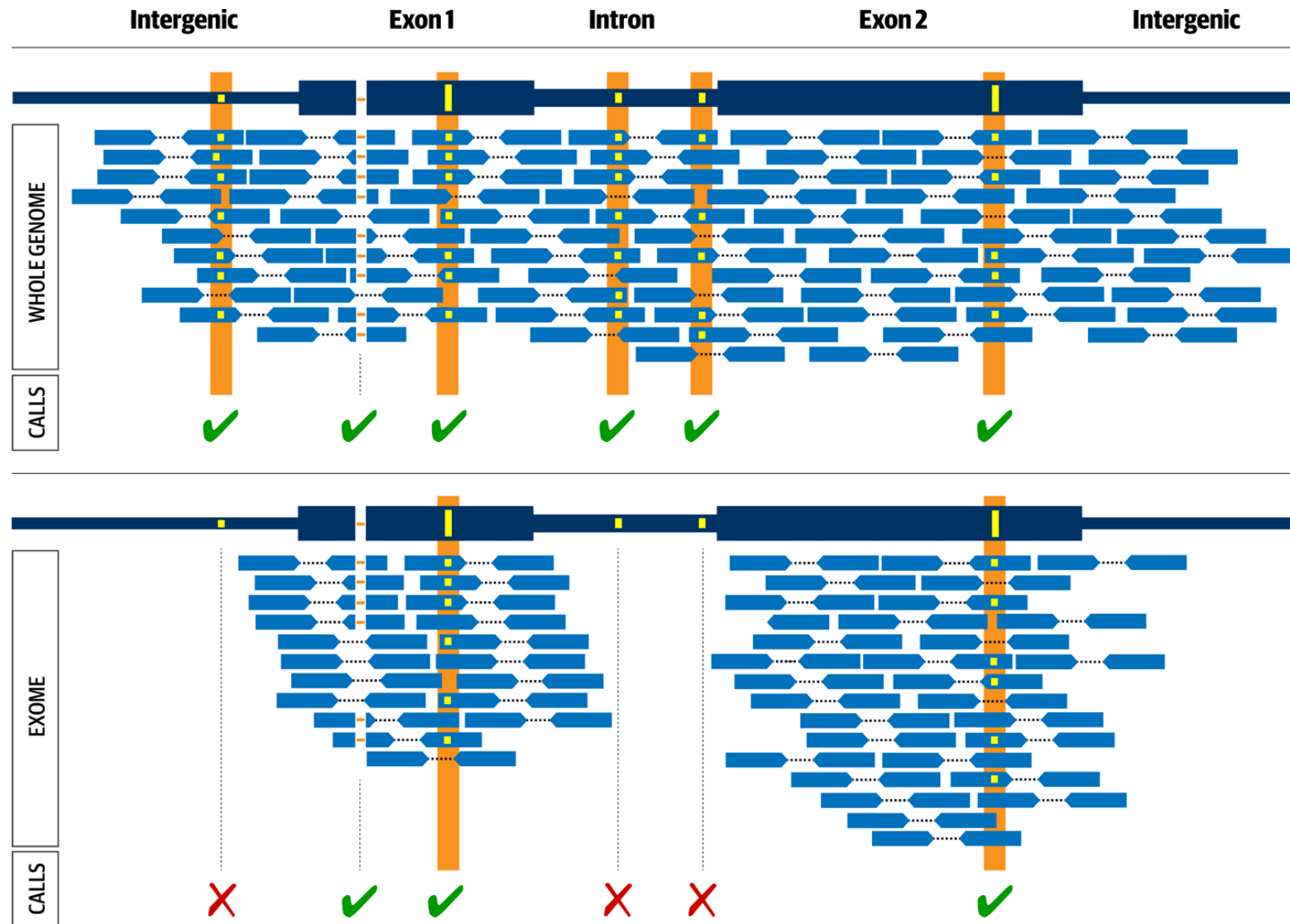
SAM = Sequence Alignment Map

BAM is the binary version (about 80 GBytes uncompressed)

CRAM is a compressed BAM file (30–60% smaller than BAM)



Whole genome vs exome sequencing



VCF file format

```
##fileformat=VCFv4.1
##reference=10000GenomesPilot-NCBI36
##INFO=<ID=DP,Number1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number2,Type=Float,Description="Allele Frequency">
##INFO=<ID=DB,Number0,Type=Flag,Description="dbSNP membership">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
```

Header

Variant records

One sample

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA000001	NA000002	NA000003
20	14370	rs6054257	G	A	29	PASS	DP=14;AF=0.5	GT:GQ:DP	0/0:48:1	1/0:48:8	1/1:43:5
20	1230237	.	T	.	47	PASS	DP=13	GT:GQ:DP	0/0:54:7	0/0:48:4	0/0:61:2
20	1234567	.	GT	G	50	PASS	DP=9	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

Site/population-level annotations

Sample-level annotations

VCF = Variant Call Format (about 2.5 GBytes compressed)



Additional resources

- FASTQ file format
 - https://en.wikipedia.org/wiki/FASTQ_format
- SAM file format
 - [https://en.wikipedia.org/wiki/SAM_\(file_format\)](https://en.wikipedia.org/wiki/SAM_(file_format))
- CRAM file format
 - https://www.ebi.ac.uk/sites/ebi.ac.uk/files/groups/ena/documents/cram_format_1.0.1.pdf
- VCF file format
 - https://en.wikipedia.org/wiki/Variant_Call_Format
- Functionally equivalent pipelines
 - <https://www.nature.com/articles/s41467-018-06159-4>
- About human genome reference builds
 - <https://gatk.broadinstitute.org/hc/en-us/articles/360035890951-Human-genome-reference-builds-GRCh38-or-hg38-b37-hg19>





Thank you for joining us today!

Next week: Chapter 3

Next meeting: December 21, 2020