WHOLE EXOME SEQUENCING ANALYSIS

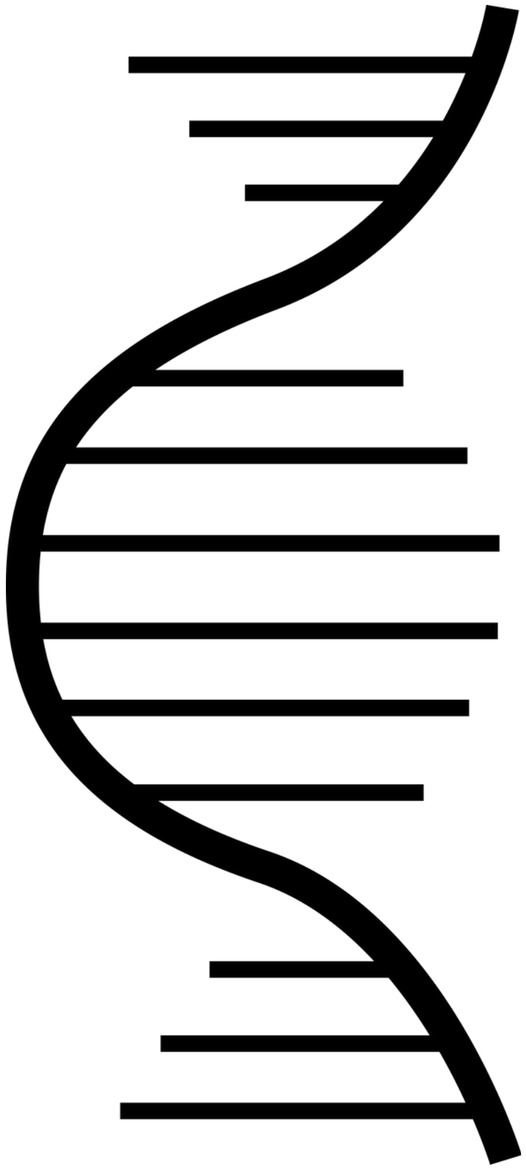
ACIBADEM ÜNİVERSİTESİ

TRANSPAN

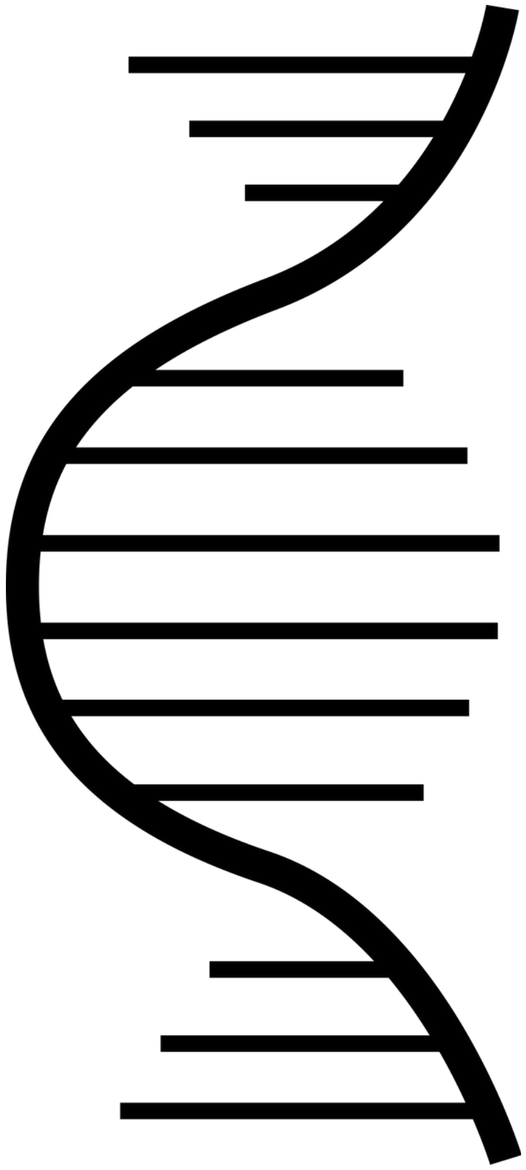
COST ACTION CA21116

Elif ÖZ

LIST OF CONTENTS

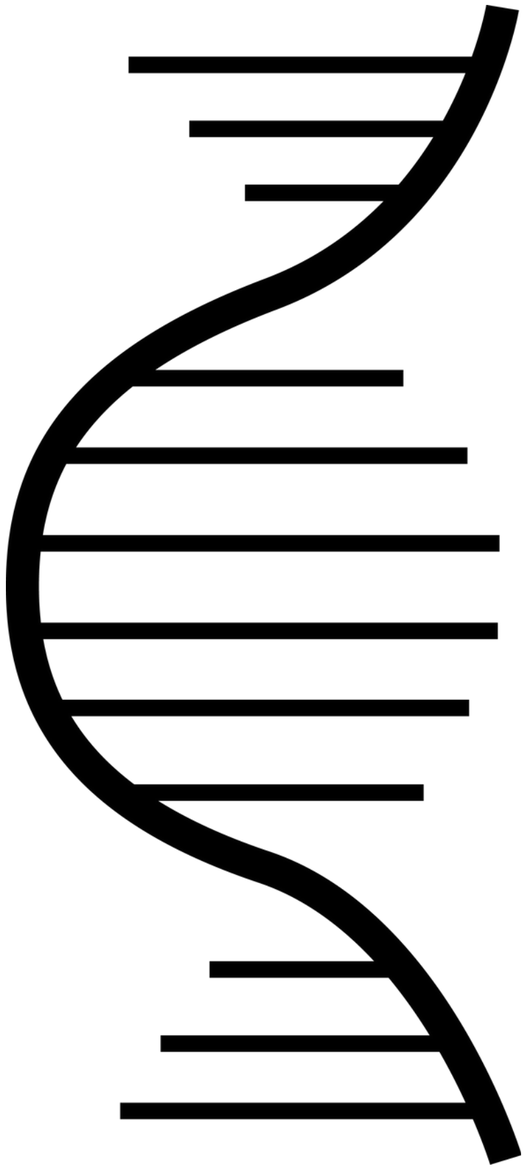


HOW MANY NUCLEOTIDES DOES A HUMAN GENOME HAVE?

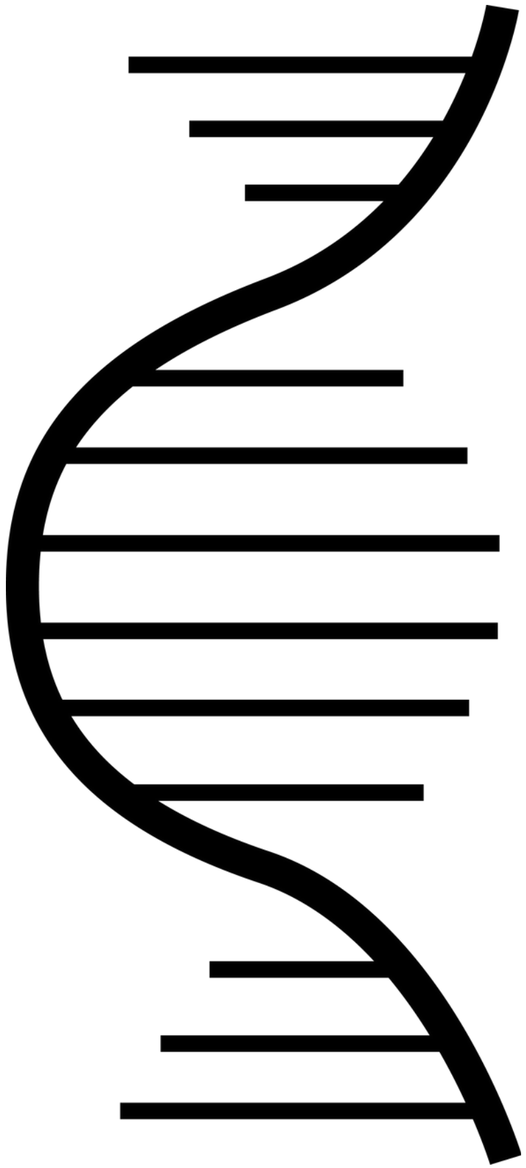


HOW MANY NUCLEOTIDES DOES A HUMAN GENOME HAVE?

3.2 billion
base pairs



HOW MANY GENES DOES A HUMAN GENOME HAVE?



HOW MANY GENES DOES A HUMAN GENOME HAVE?

23,500

~1% of the genome

30 million base pairs

99.6 % identical genome

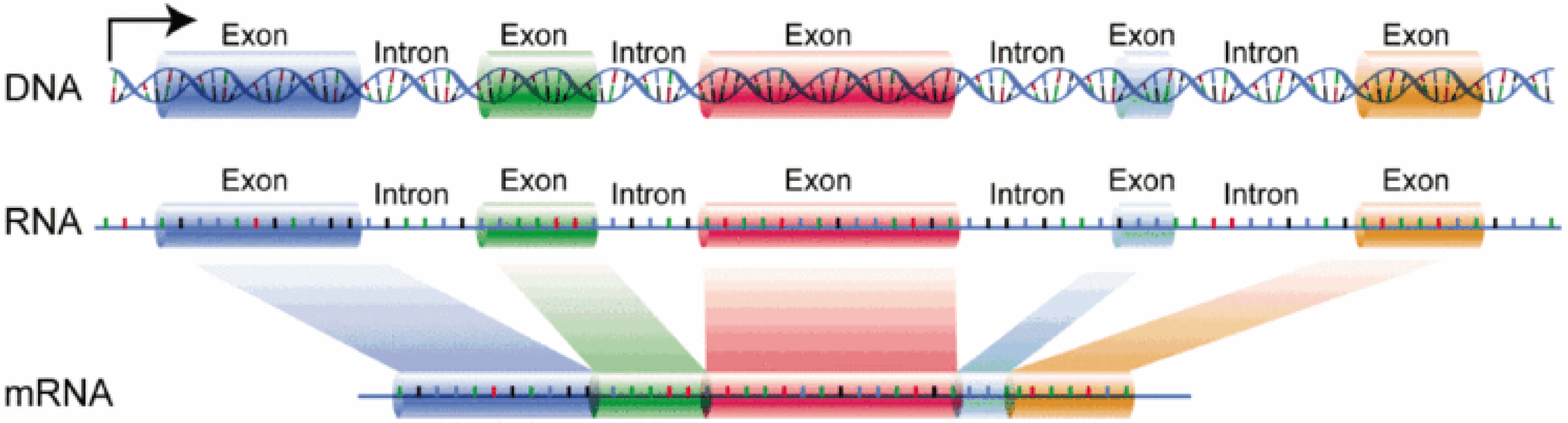
0.4% variation

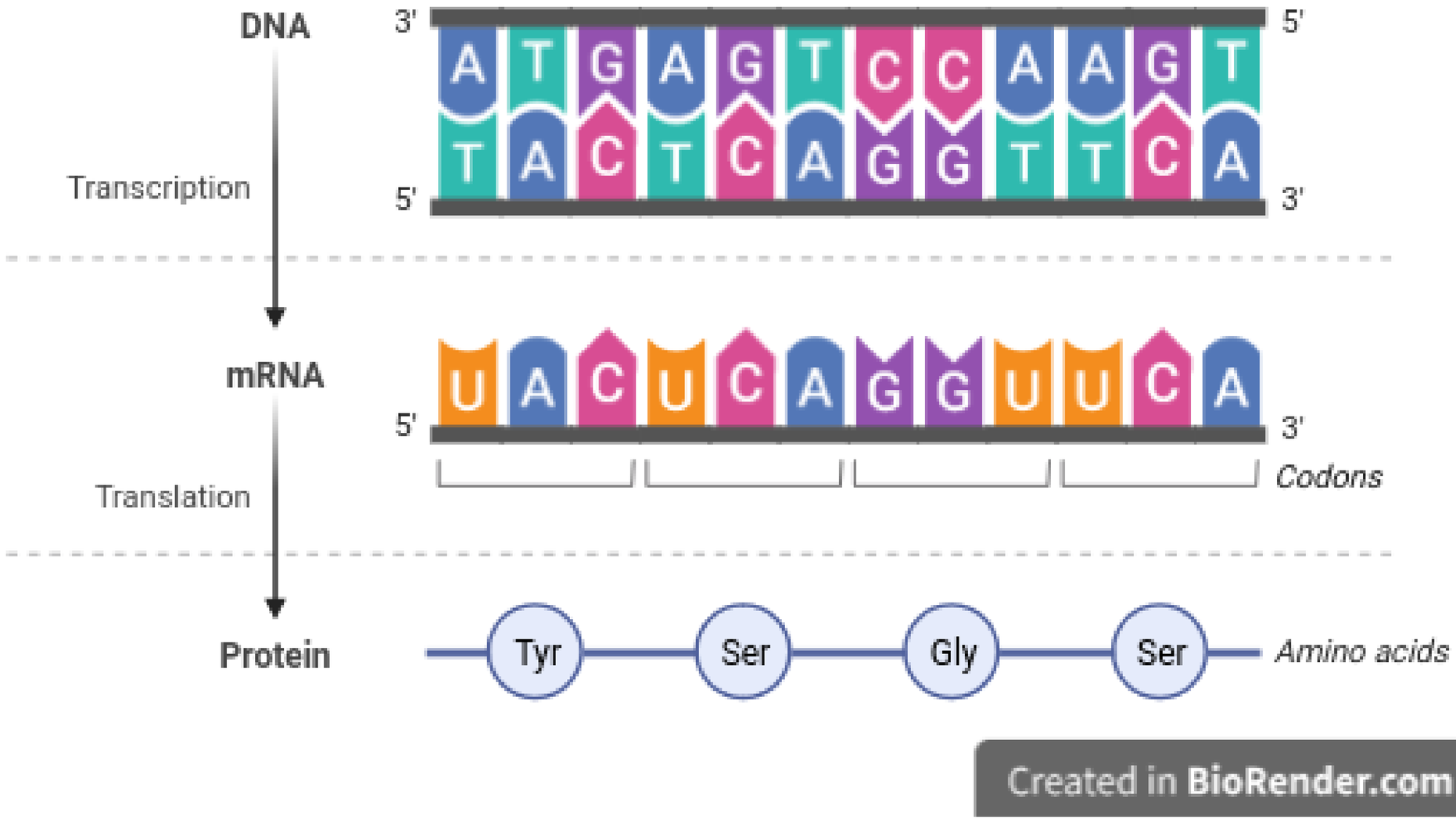
12 million base pairs

~4 to 5 million SNPs in a person's genome

01

Central Dogma: From DNA to Proteins



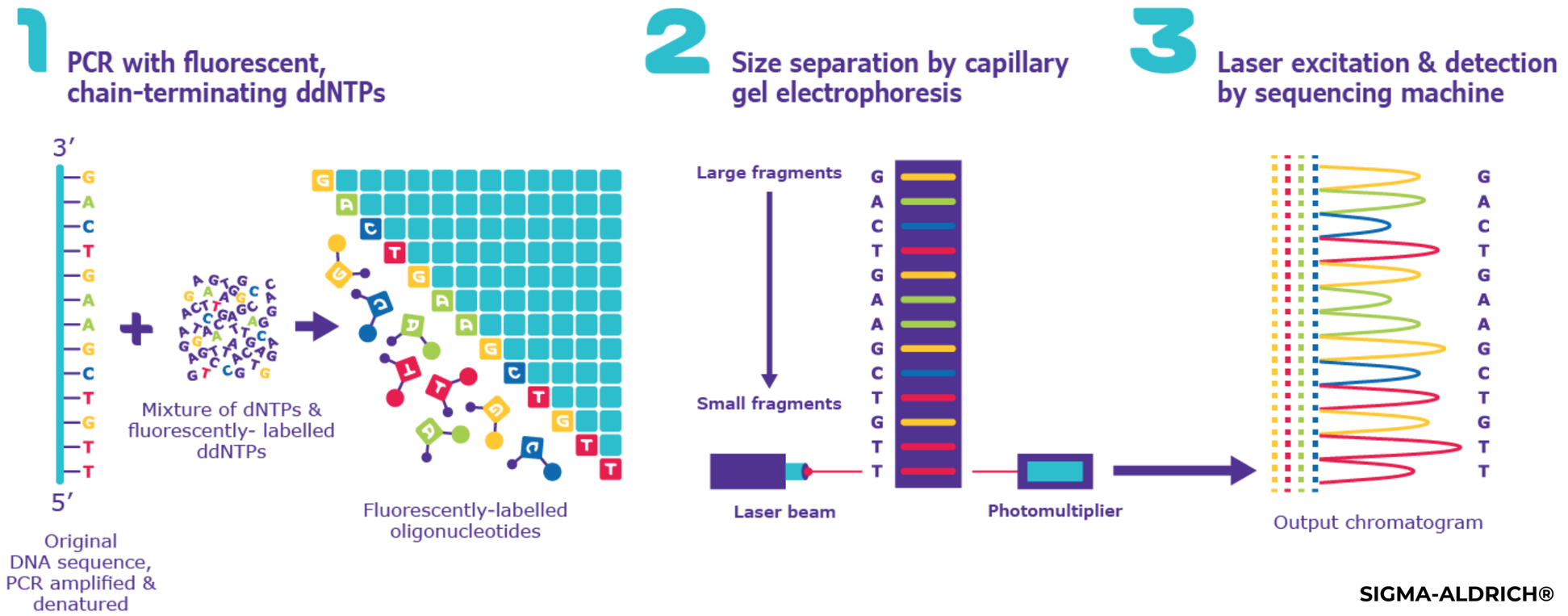

DNA
3'
5'
A T G A G T C C A A G T
T A C T C A G G T T C A
5'
3'
Transcription
mRNA
5'
U A C U C A G G U U C A
3'
Codons
Translation
Protein
Tyr
Ser
Gly
Ser
Amino acids
Created in BioRender.com
bio

Codon Chart

First Base	Second Base: U	Second Base: C	Second Base: A	Second Base: G	Third Base
U	UUU, UUC – Phenylalanine (Phe/F) UUA, UUG – Leucine (Leu/L)	CUU, CCU, CAU, CGU – Serine (Ser/S)	AUU, ACU – Tyrosine (Tyr/Y) **AAU** – STOP **AGU** – STOP	GUU, GCU – Cysteine (Cys/C) **GAU** – STOP GGU – Tryptophan (Trp/W)	U C A G
C	CUU, CUC, CUA, CUG – Leucine (Leu/L)	CUC, CCC, CAC, CGC – Proline (Pro/P)	AUC, ACC – Histidine (His/H) AAC, AGC – Glutamine (Gln/Q)	GUC, GCC, GAC, GGC – Arginine (Arg/R)	U C A G
A	AUU, AUC, AUA – Isoleucine (Ile/I) **AUG** – Methionine (Met/M)	CUA, CCA, CAA, CGA – Threonine (Thr/T)	AUA, ACA – Asparagine (Asn/N) AAA, AGA – Lysine (Lys/K)	GUA, GCA – Serine (Ser/S) GAA, GGA – Arginine (Arg/R)	U C A G
G	GUU, GUC, GUA, GUG – Valine (Val/V)	CUG, CCG, CAG, CGG – Alanine (Ala/A)	AUG, ACG – Aspartic acid (Asp/D) AAG, AGG – Glutamic acid (Glu/E)	GUG, GCG, GAG, GGG – Glycine (Gly/G)	U C A G

FIRST GENERATION SEQUENCING

SANGER SEQUENCING

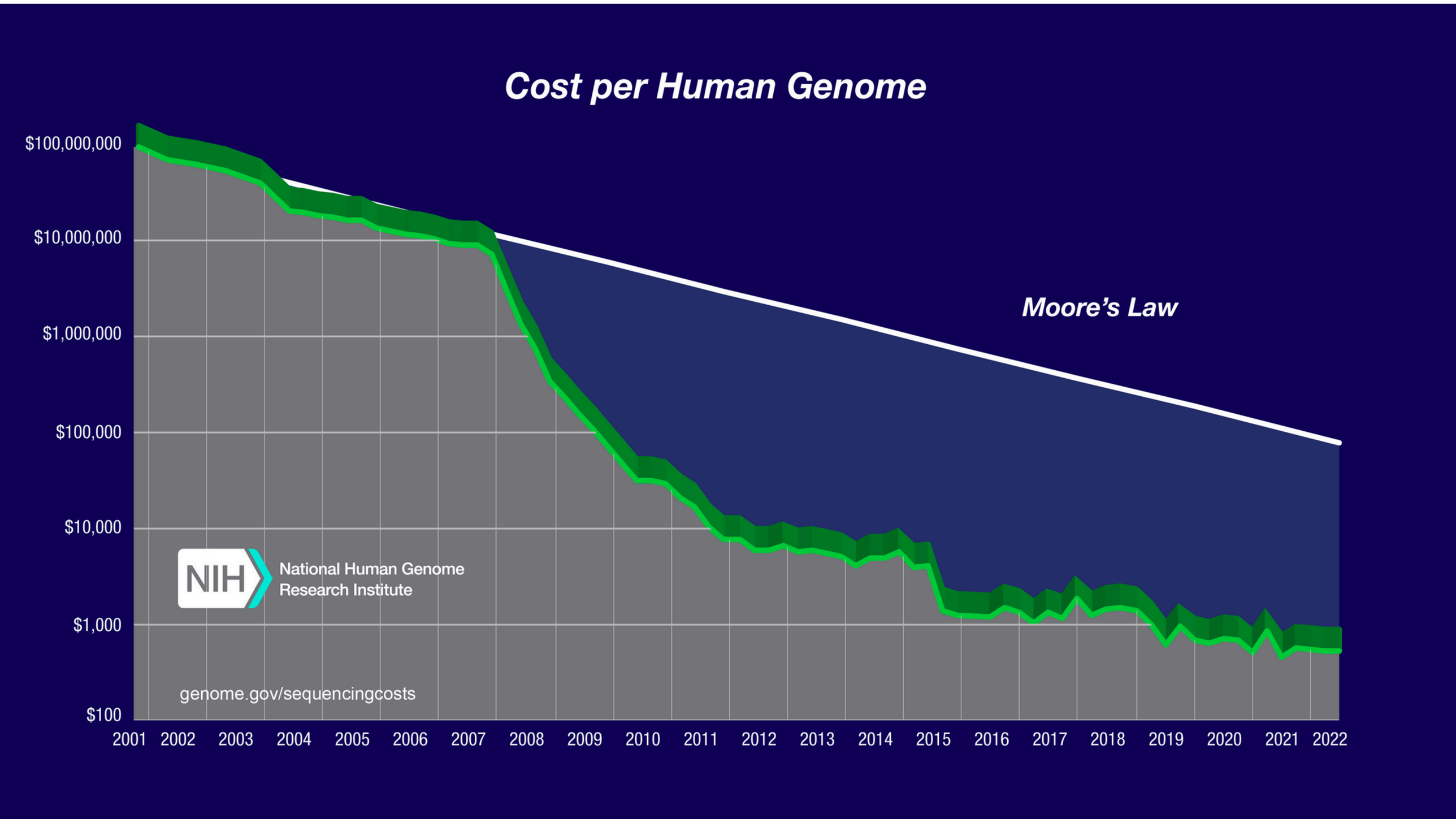


SIGMA-ALDRICH®

02

High Throughput Sequencing

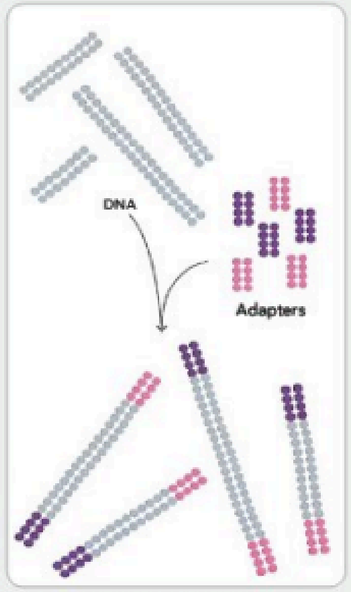
- Human Genome Project (1990-2003)
    - sequence and map the entire human genome
- 1000 Genomes Project (2008-2015)
    - create a comprehensive catalog of human genetic variation


Cost per Human Genome
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
$1,000
$100
Moore's Law
NIH
National Human Genome
Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022

NEXT GENERATION SEQUENCING

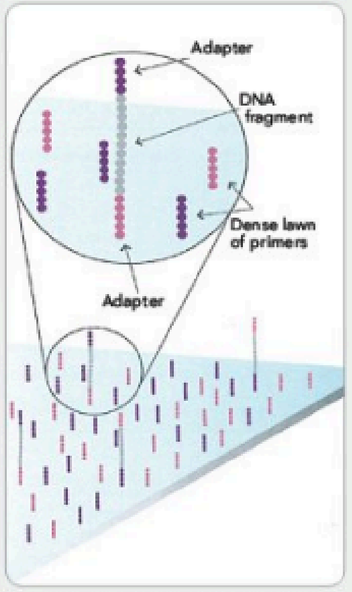
- sequencing millions of DNA fragments simultaneously
- short read (50-300 bps)
- Massively parallel sequencing
- Sequencing by synthesis
- Roche 454 pyrosequencing, Illumina, SOLID

1. PREPARE GENOMIC DNA SAMPLE



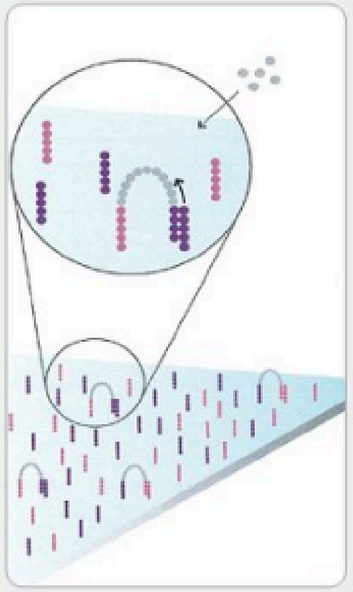
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE



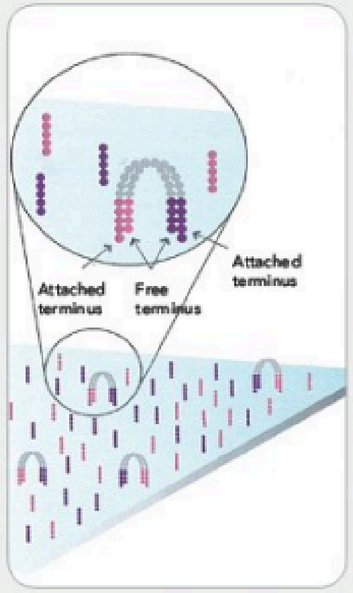
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGE AMPLIFICATION



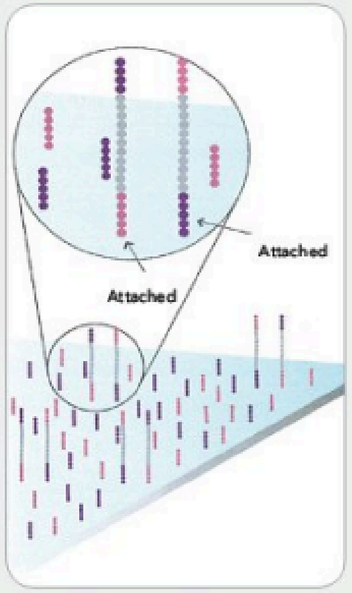
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

4. FRAGMENTS BECOME DOUBLE STRANDED



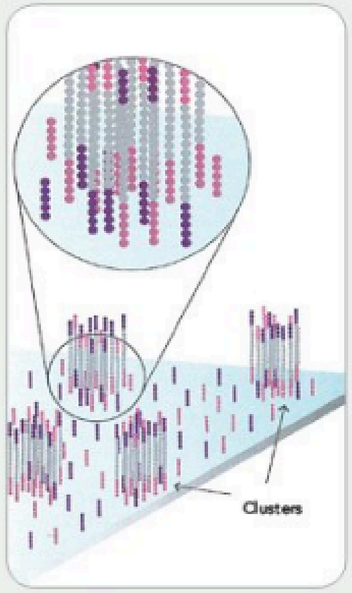
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES



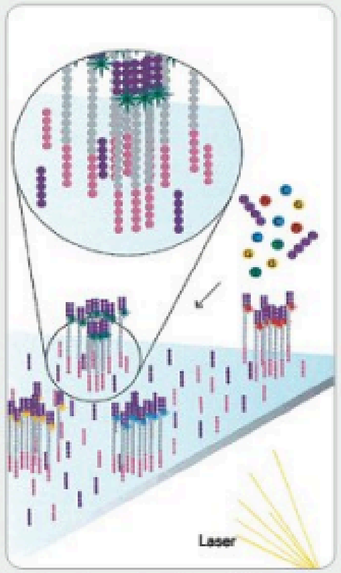
Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION



Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

7. DETERMINE FIRST BASE

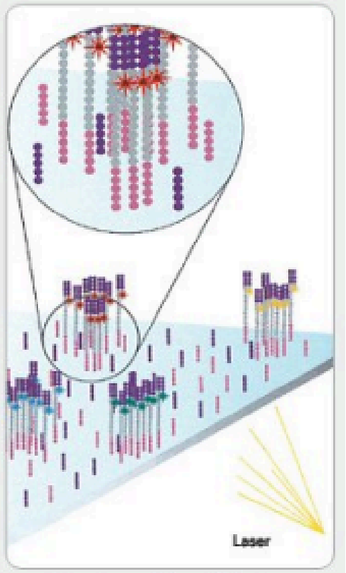


First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

8. IMAGE FIRST BASE

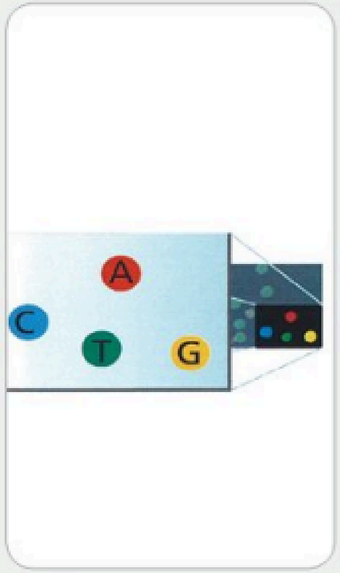
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

9. DETERMINE SECOND BASE



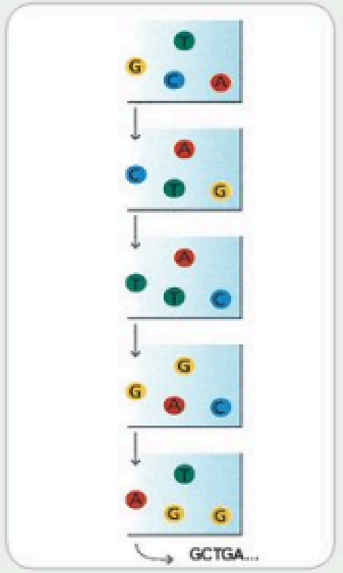
Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

10. IMAGE SECOND CHEMISTRY CYCLE



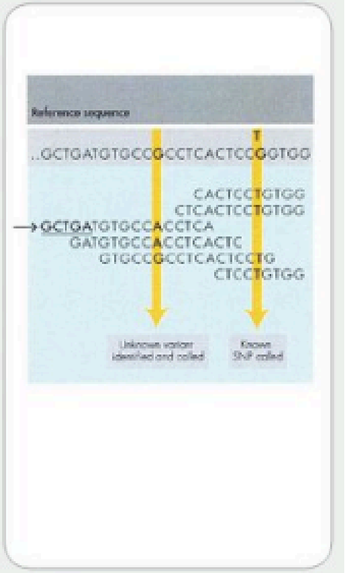
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES



Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.

12. ALIGN DATA



Align data, compare to a reference, and identify sequence differences.

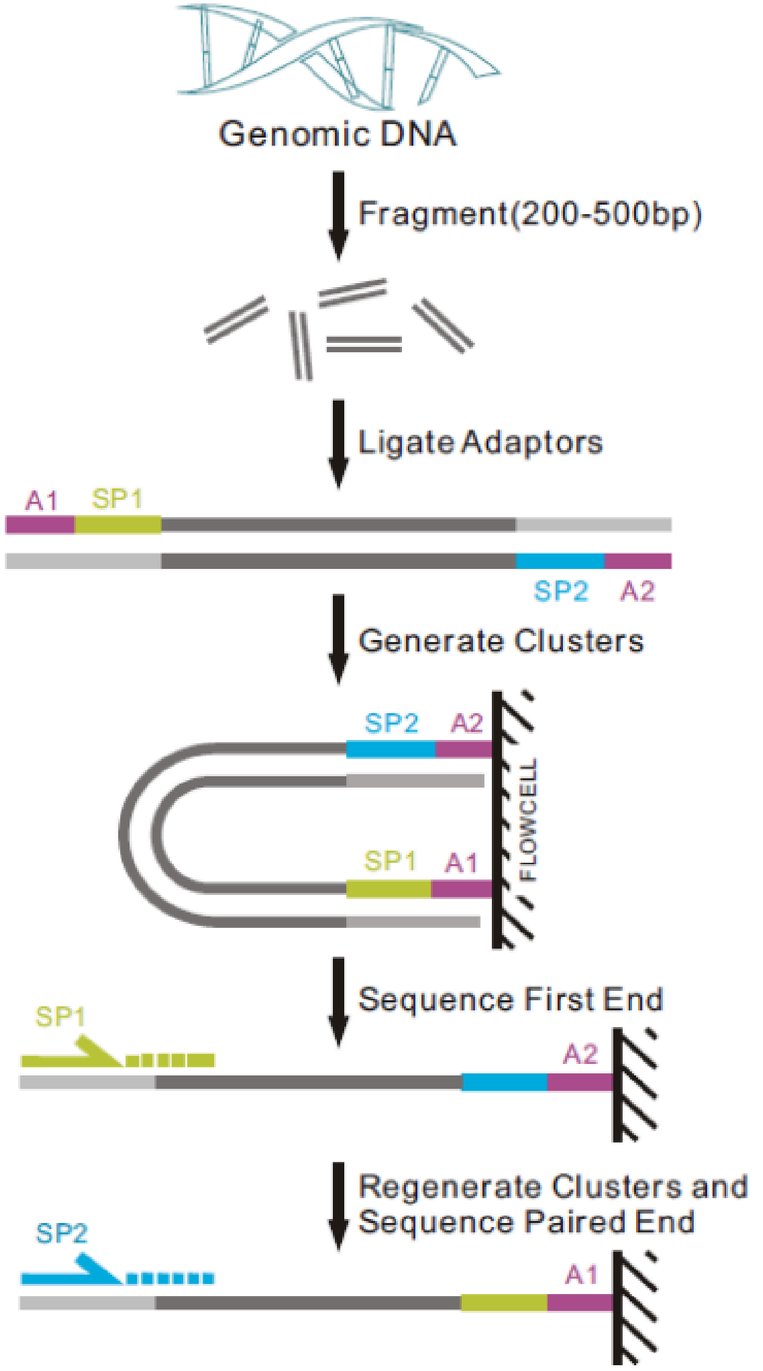


Figure 1-2-1 Pipeline of paired-end sequencing (www.illumina.com)

ADVANTAGES OF PAIRED END SEQUENCING

- Improved alignment accuracy
- Better Detection of Structural Variations
- Higher coverage
- Better Resolution of Complex Regions
- Enhanced Detection of Fusion Genes
- Reduced Error Rate

ADVANTAGES OF PAIRED END SEQUENCING

Scenario 1: Single-end Read

Sequence: ATATATATGGGTTTGG
Read: ATAT

ADVANTAGES OF PAIRED END SEQUENCING

Scenario 1: Single-end Read

Sequence: ATATATATGGGTTTGG□

Read: ATAT

Alignment: ATATATATGGGTTTGG□
ATAT
ATAT
ATAT

ADVANTAGES OF PAIRED END SEQUENCING

Scenario 2: Paired-end Read

Sequence: ATATATATGGGTTTGG□

Read: ATAT□

Paired Read: TTGG
□
Distance between reads: 4bp

ADVANTAGES OF PAIRED END SEQUENCING

Scenario 2: Paired-end Read

Sequence: ATATATATGGGTTTGG□

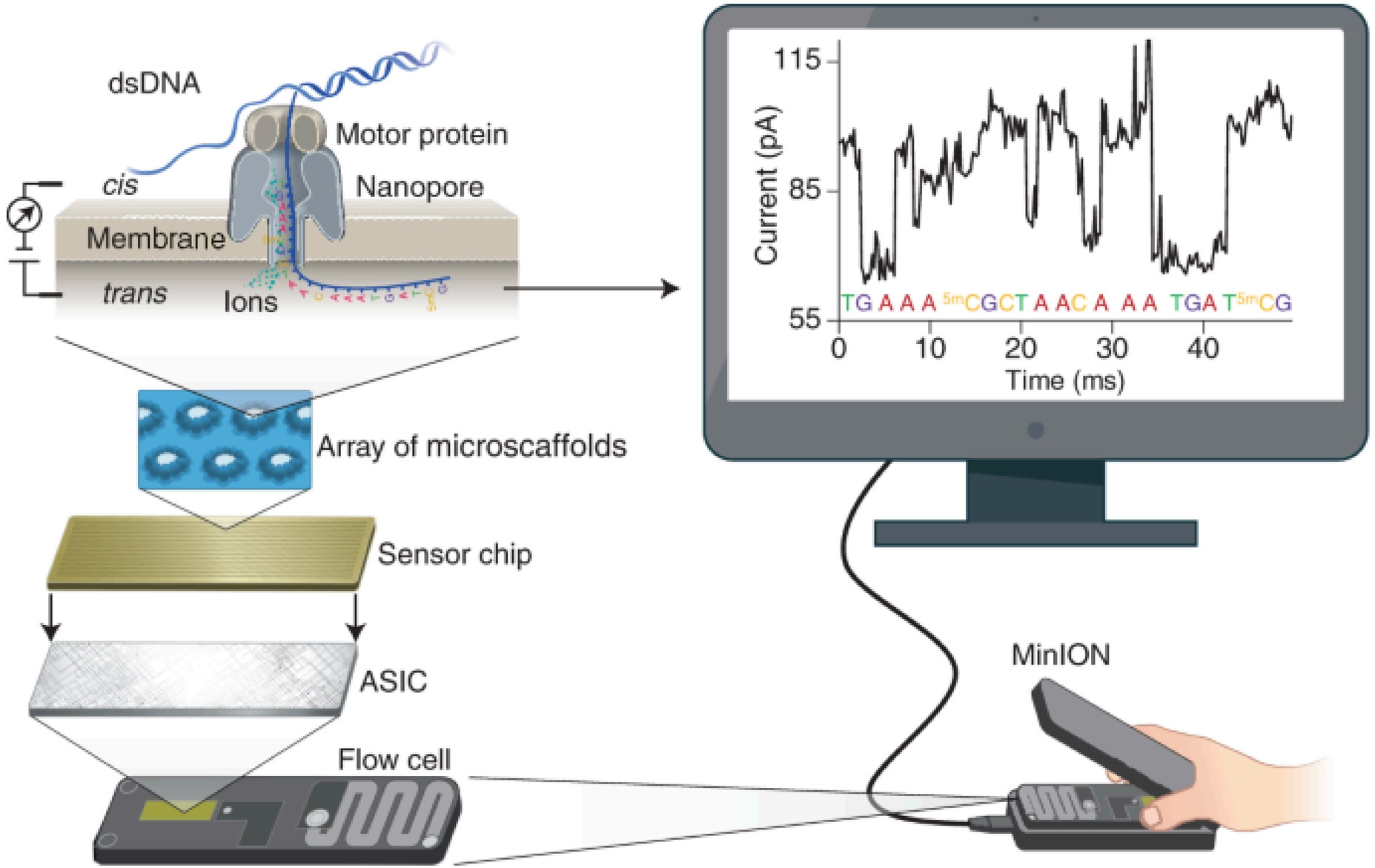
Read: ATAT□

Paired Read: TTGG
□
Distance between reads: 4bp

Alignment: ATATATATGGGTTTGG□
ATAT - - - - TTGG

THIRD GENERATION SEQUENCING

- sequencing single molecules of DNA or RNA directly, without the need of amplification
- longer reads (1000-20,000 bps)
- PacBio, Oxford Nanopore



Wang, Y., Zhao, Y., Bollas, A. et al. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol 39, 1348–1365 (2021). https://doi.org/10.1038/s41587-021-01108-x

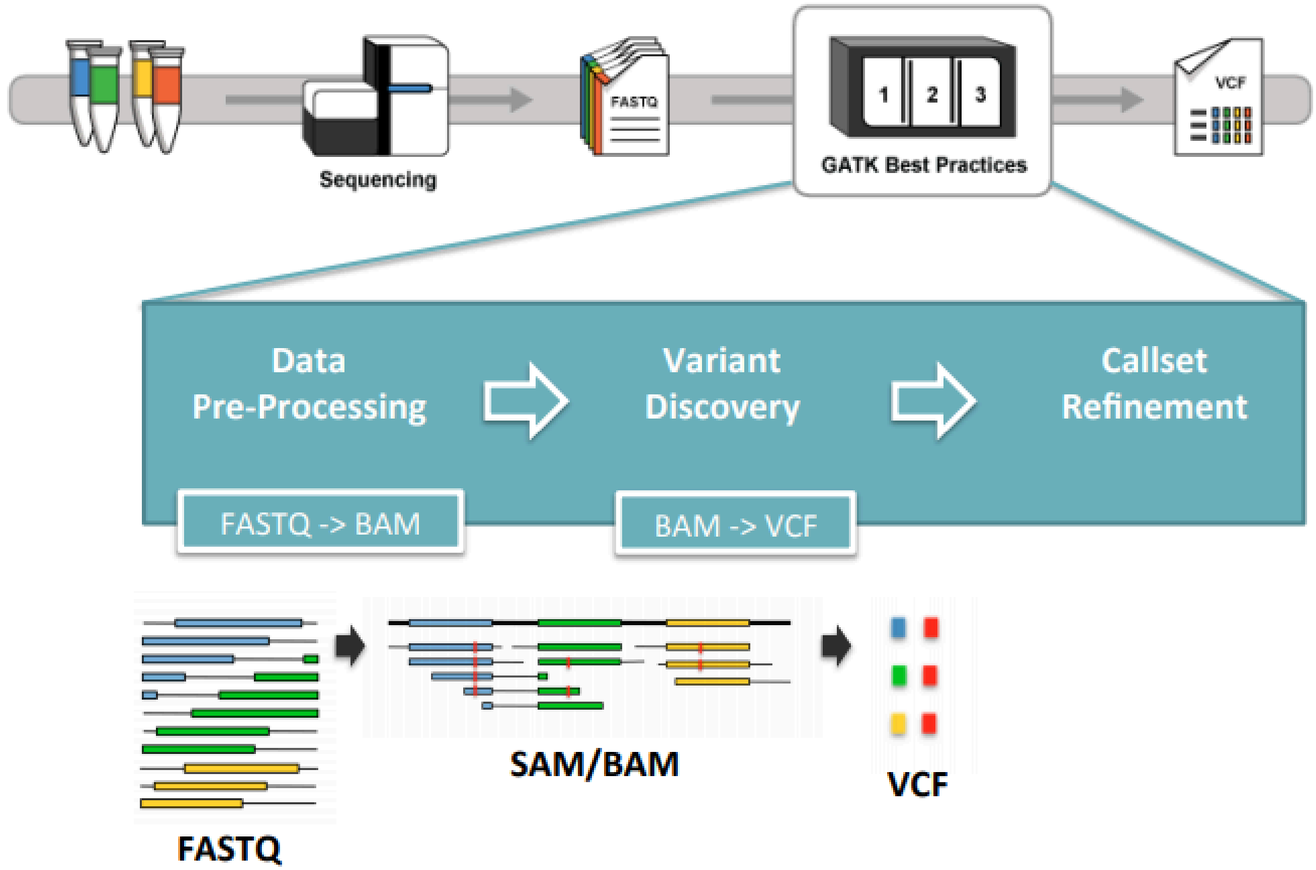
FROM BIOLOGICAL SAMPLE TO INFORMATICS DATA

- Library preparation
  - DNA isolation
  - Shear into fragments
  - Attach to adapters

- Sequencing
  - Short read sequencing (50-300 bp)

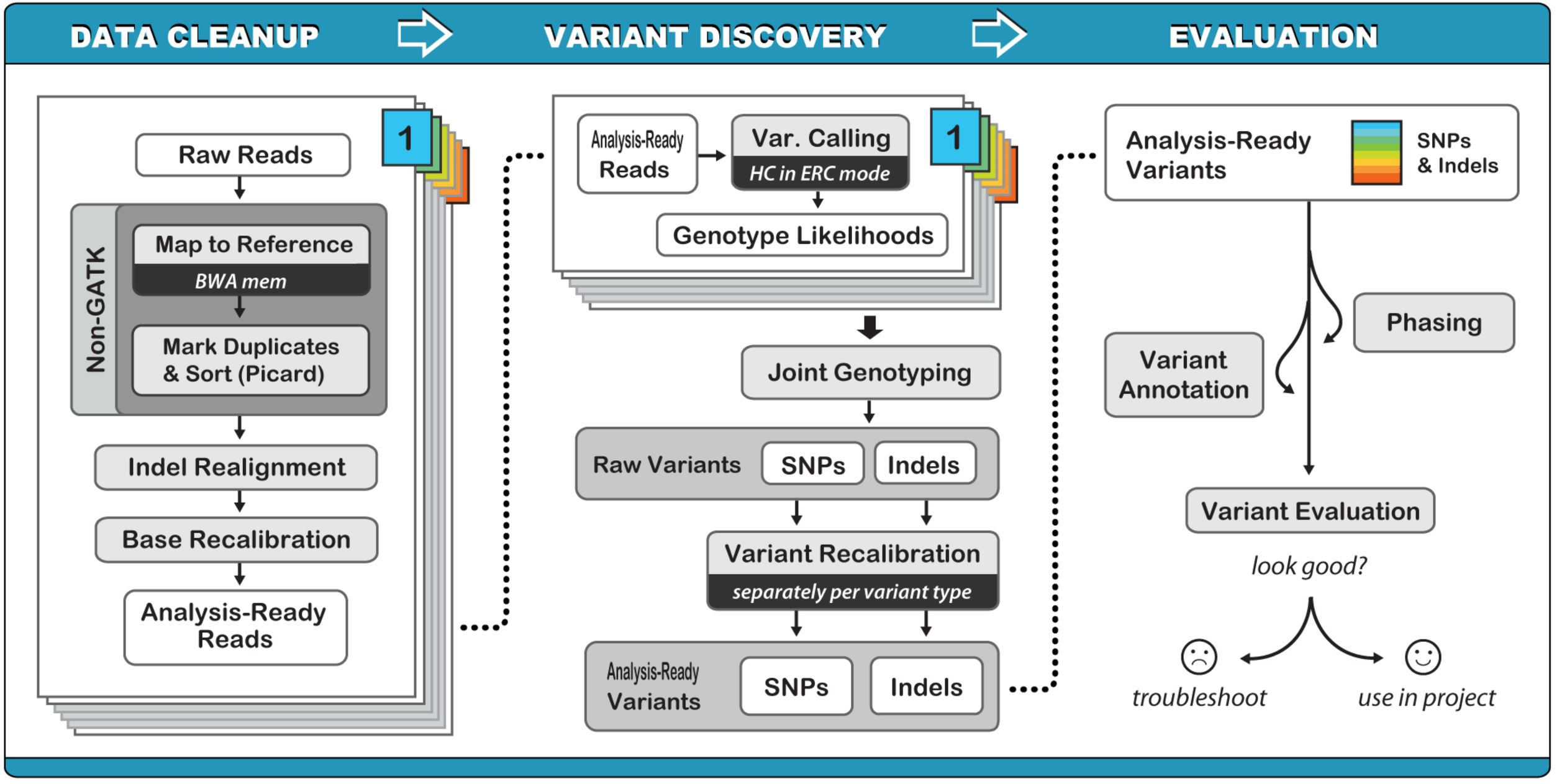
03

GATK Best Practices

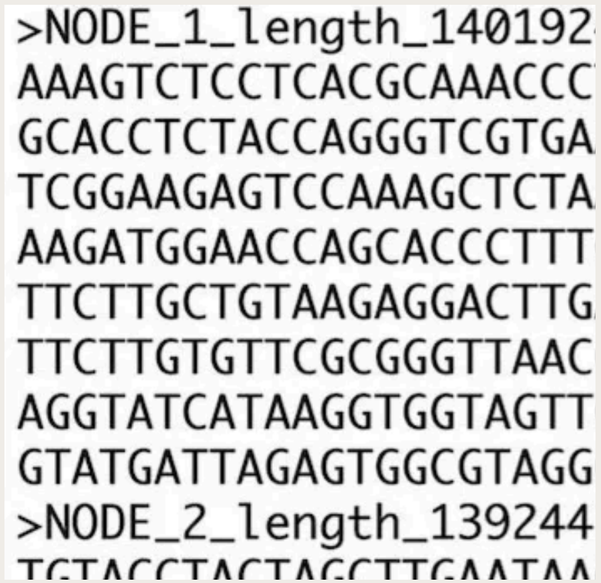


GENOME ANALYSIS TOOLKIT

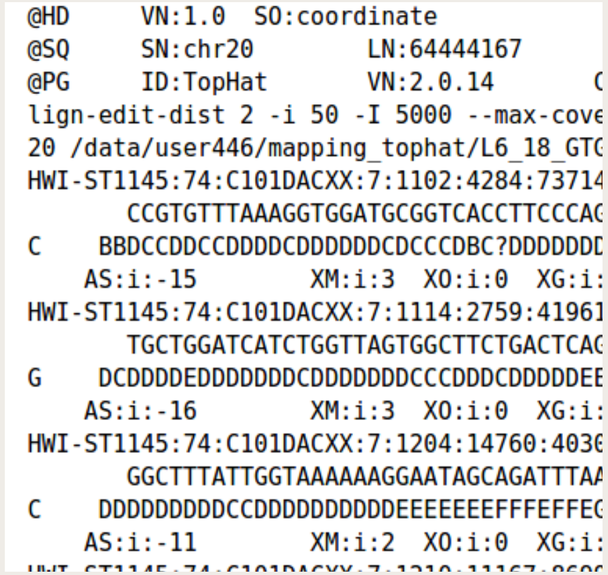
- Offers “Best Practices”:
  - pipelines optimized for accuracy and performance
- Offers specialized pipelines for both germline and somatic variant calling
- Enables efficient multi-sample analysis with joint genotyping


DATA CLEANUP
VARIANT DISCOVERY
EVALUATION
1
Raw Reads
Non-GATK
Map to Reference
BWA mem
Mark Duplicates & Sort (Picard)
Indel Realignment
Base Recalibration
Analysis-Ready Reads
1
Analysis-Ready Reads
Var. Calling
HC in ERC mode
Genotype Likelihoods
Joint Genotyping
Raw Variants
SNPs
Indels
Variant Recalibration
separately per variant type
Analysis-Ready Variants
SNPs
Indels
Analysis-Ready Variants
SNPs & Indels
Phasing
Variant Annotation
Variant Evaluation
look good?
troubleshoot
use in project

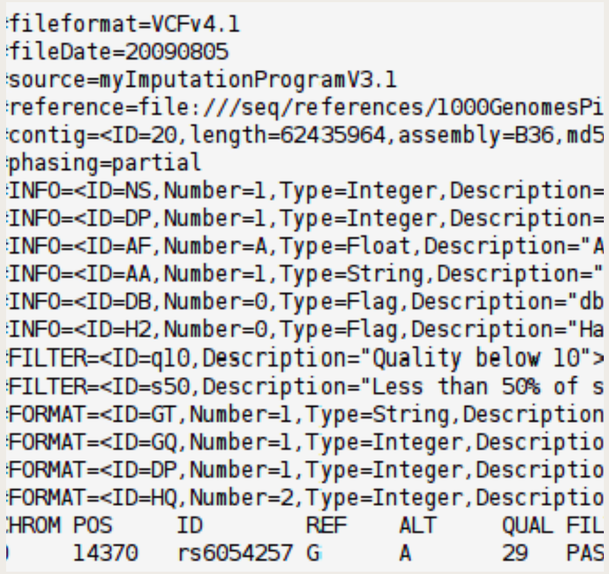
FILE FORMATS



FASTQ

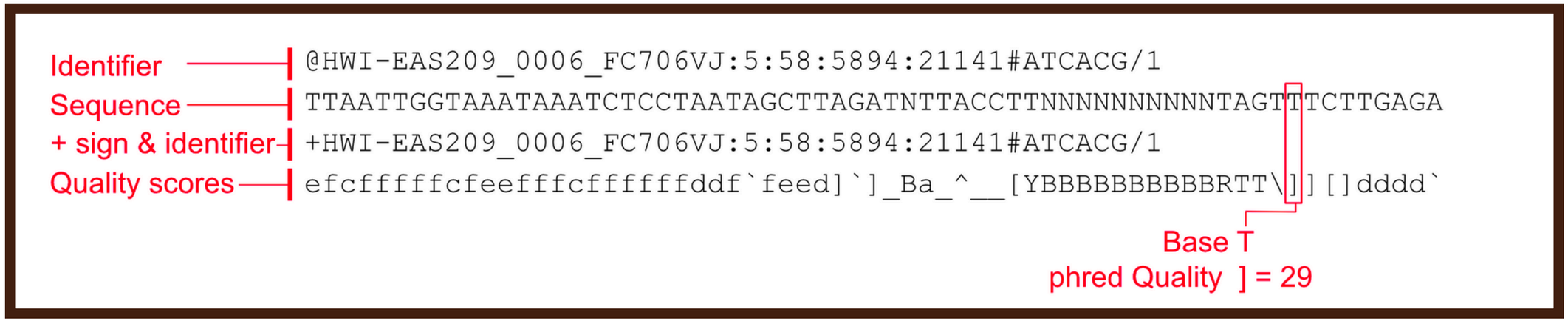


SAM/BAM



VCF

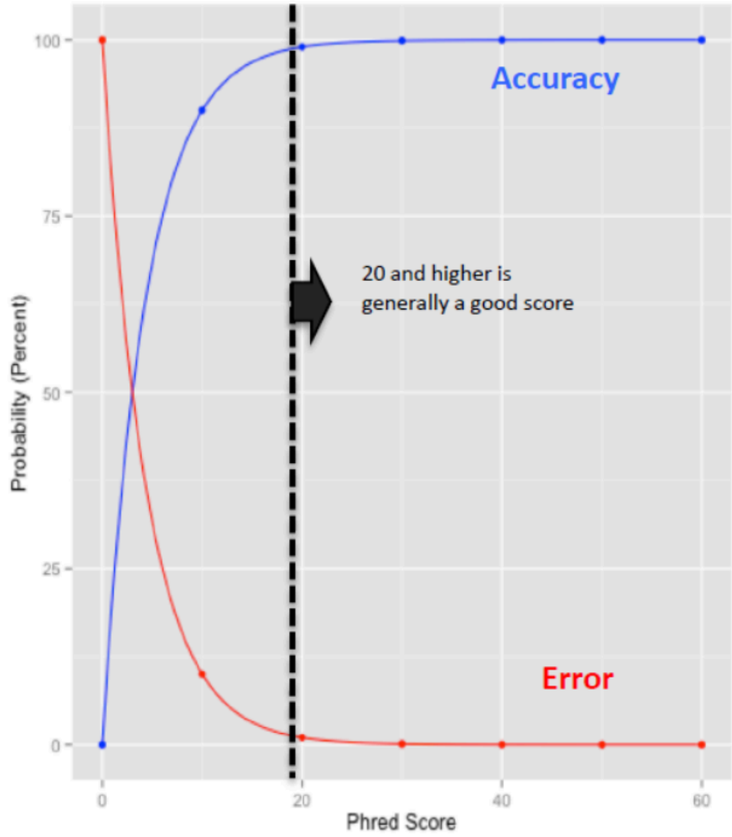
FASTQ FORMAT



Phred Scale

Q = -10 log10(P)

Q10 = 90% Confidence (10% Error rate)
Q20 = 99% Confidence (1% Error rate)
Q30 = 99.9% Confidence (0.1% Error rate)

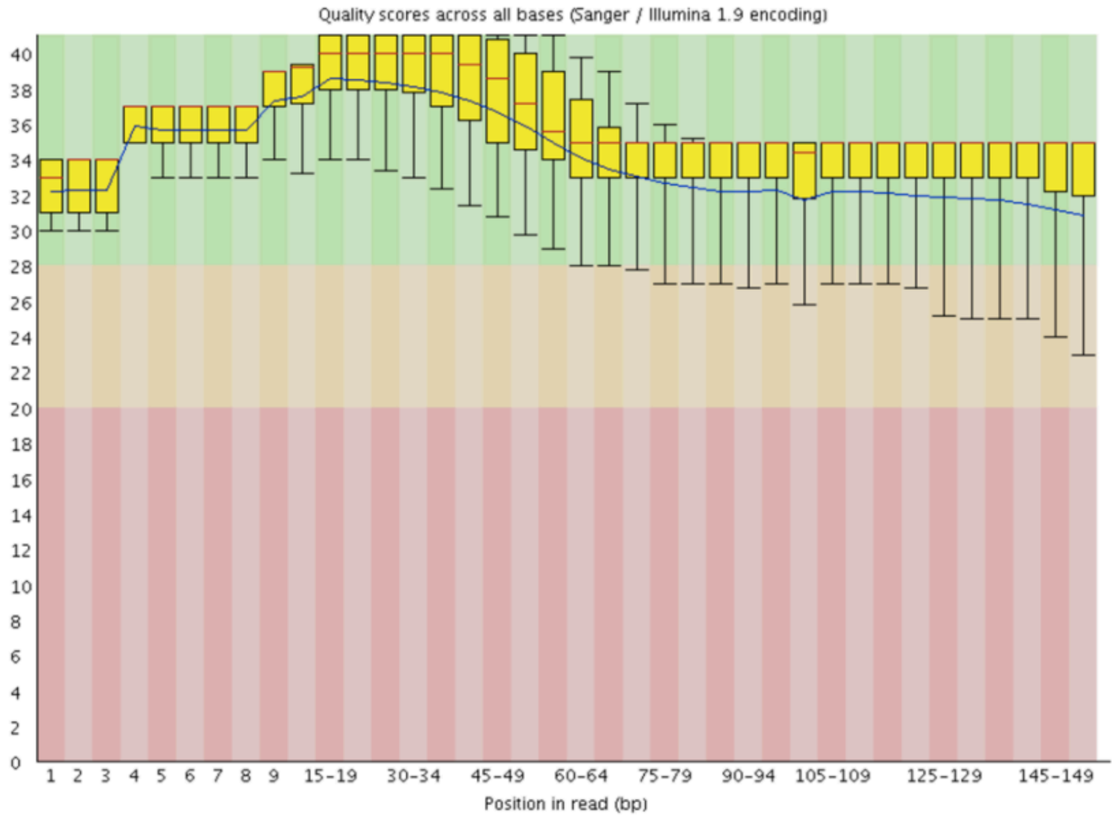


QUALITY CONTROL

- The process of evaluating and improving data by removing identifiable errors from it
- QC cannot turn bad data into good data, and we can never salvage what appears to be a total loss
- Developed by Babraham Institute, FastQC is the most commonly used quality control visualization tool

QUALITY CONTROL

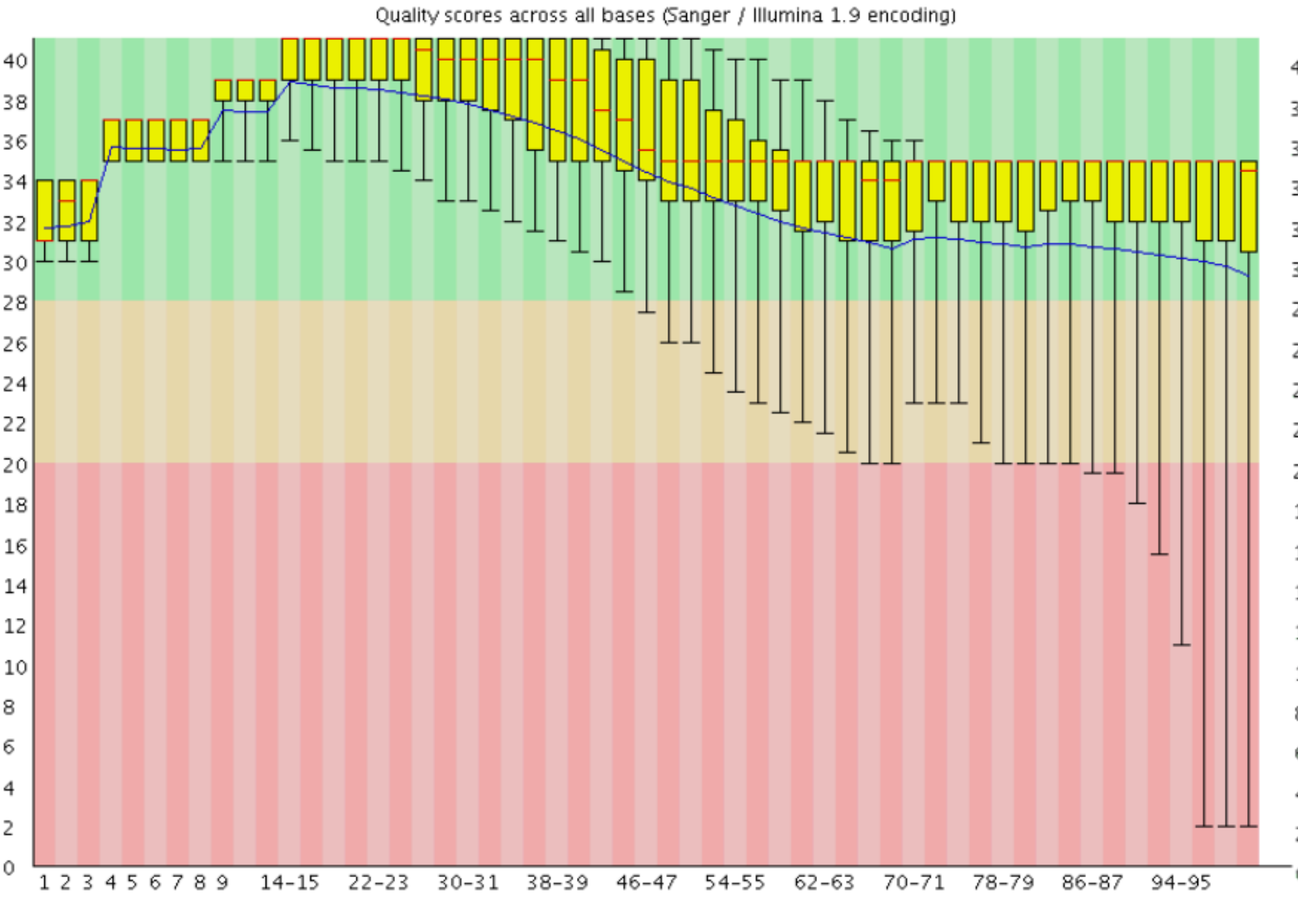
- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

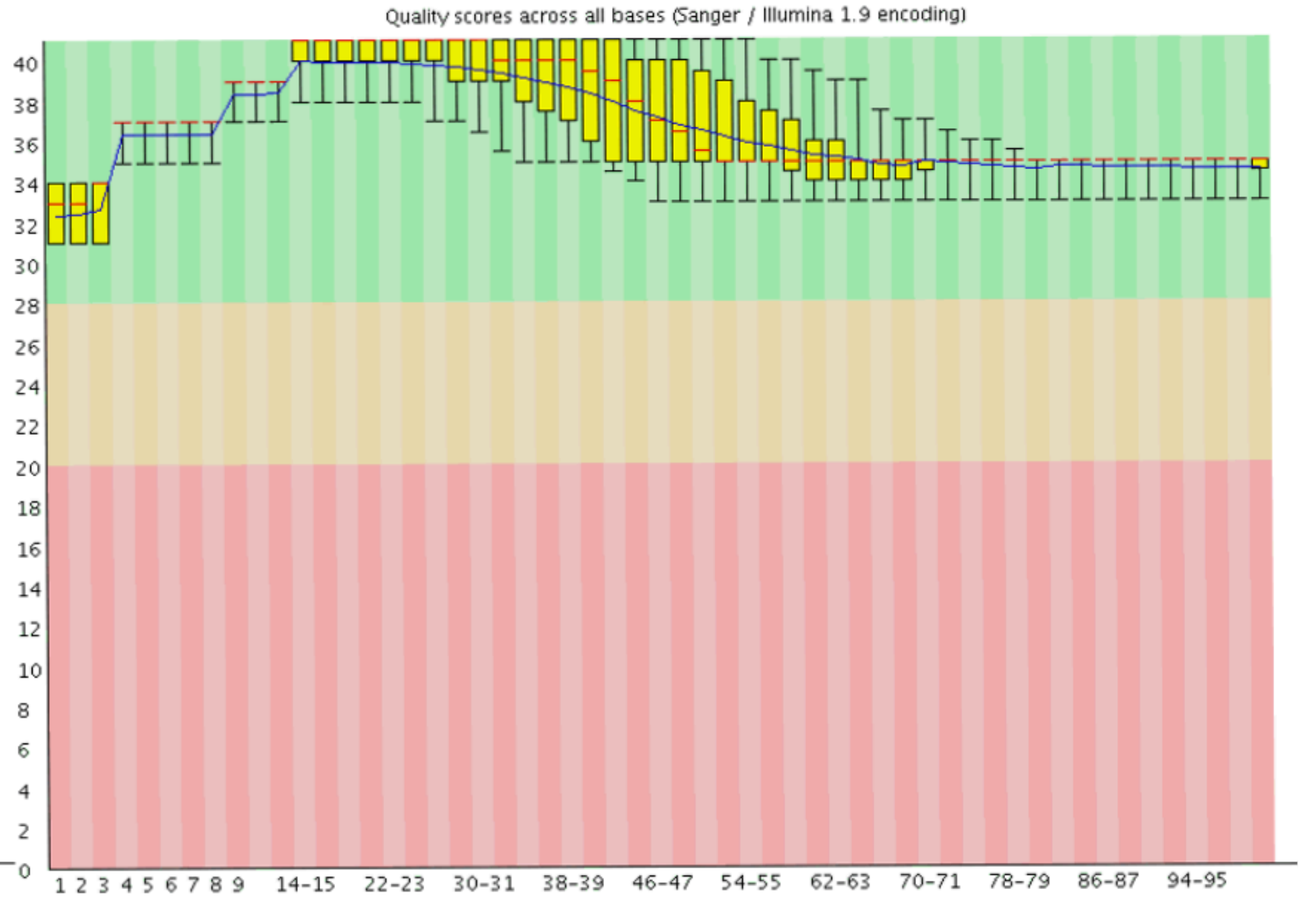


TRIMMING

- The removal of adapter sequences
- Can be done with **Trim Galore**, trimmomatic, BBDuk, Fastx Toolkit, etc.

TRIMMING

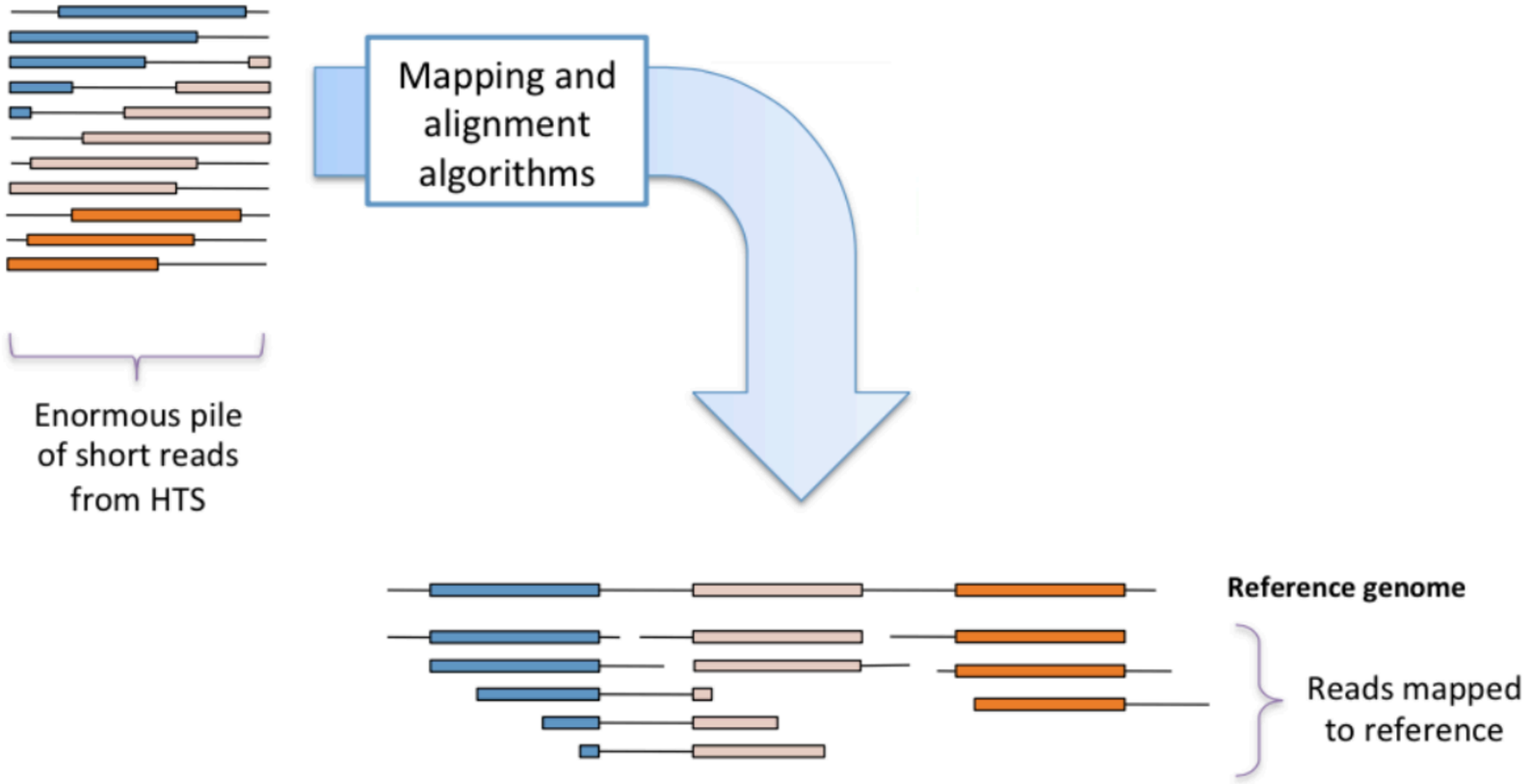

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9
14-15
22-23
30-31
38-39
46-47
54-55
62-63
70-71
78-79
86-87
94-95


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9
14-15
22-23
30-31
38-39
46-47
54-55
62-63
70-71
78-79
86-87
94-95

ALIGNMENT

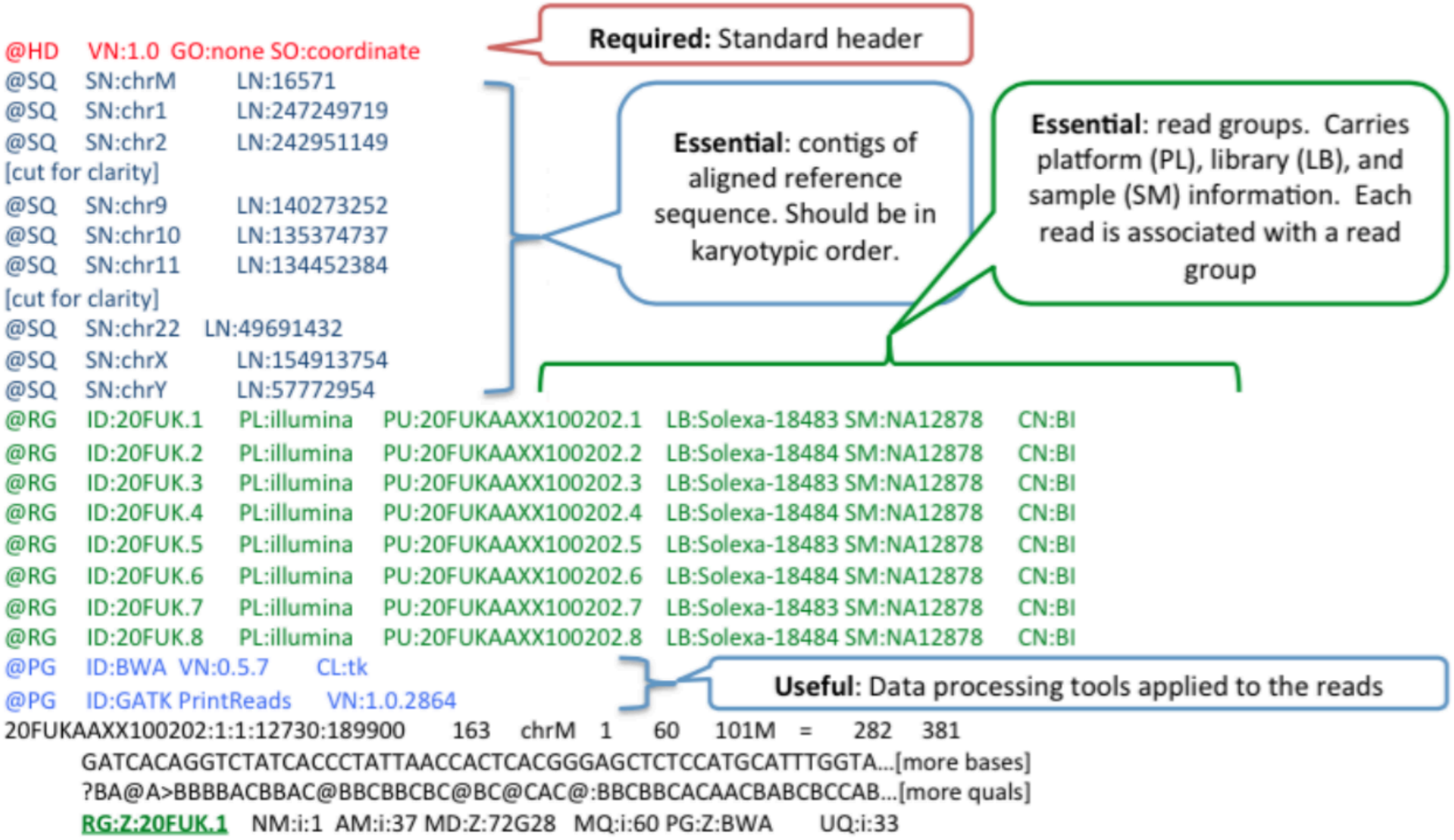
- aligning the raw sequencing reads to a reference genome to determine where each read originated in the genome
- tools like **BWA (Burrows Wheeler Aligner)** or Bowtie are commonly used to map sequencing reads to a standard reference genome

ALIGNMENT


Mapping and alignment algorithms
Enormous pile of short reads from HTS
Reference genome
Reads mapped to reference

SAM/BAM FORMAT

SEQUENCE ALIGNMENT MAP / BINARY ALIGNMENT MAP



1	2	3	4	5	6	7	8	9	10	11	12
SRR067577.2766	99	chr14	73240003	60	101M	=	73240004	102	GCTA…	FHG@…	NM:I:0

1. Query Name
2. FLAG
3. Reference Name
4. Position
5. Mapping Quality
6. CIGAR
7. Mate Name
8. Position of Mate
9. Template Length (Reference)
10. Sequence
11. Quality String
12. Predefined Tags

1	2	3	4	5	6	7	8	9	10	11	12
SRR067577.2766	99	chr14	73240003	60	101M	=	73240004	102	GCTA…	FHG@…	NM:I:0

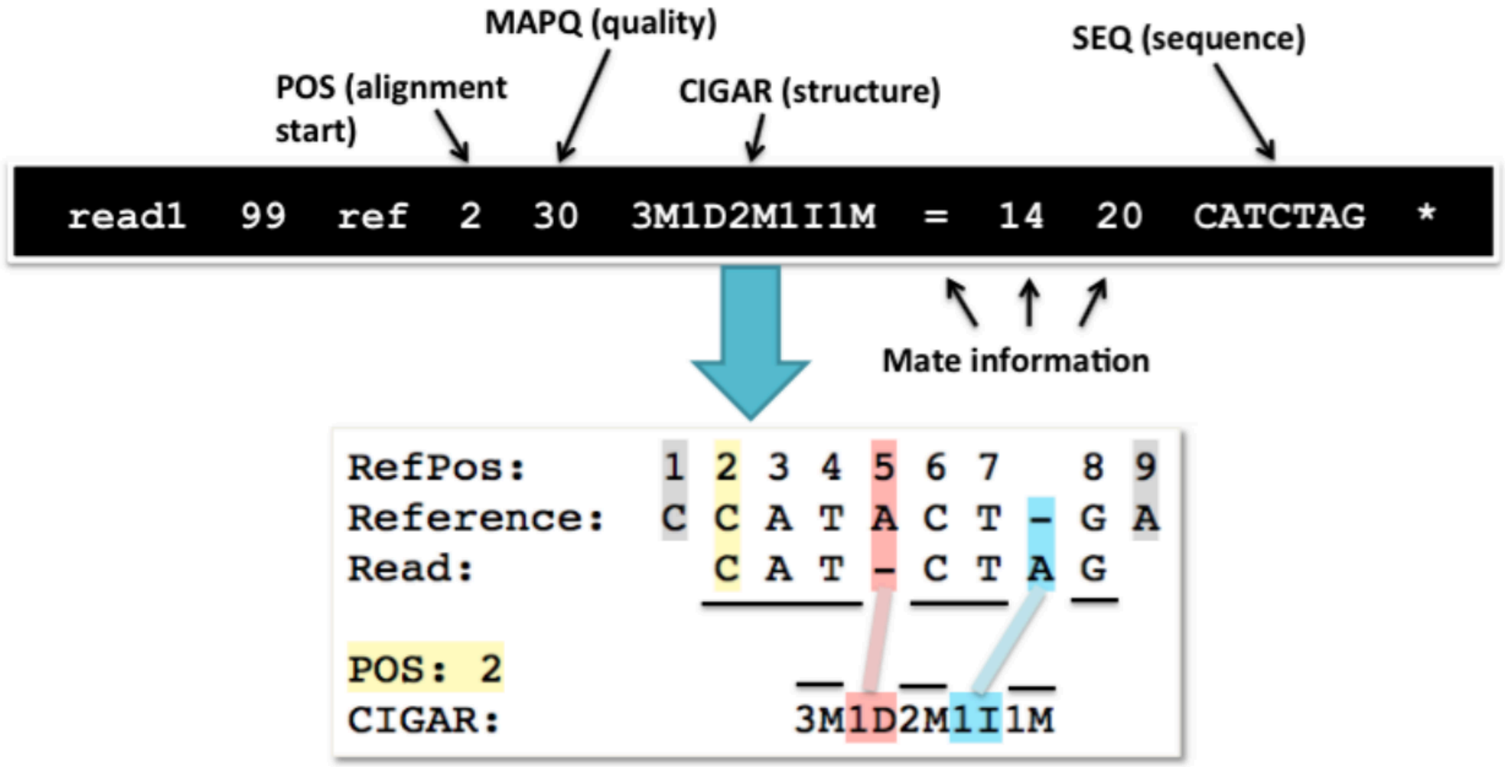
Decimal	Binary	Exp.	Meaning
1	1	2^0	This is a paired read
2	10	2^1	This read is part of a pair that aligned properly*
4	100	2^2	This read was not aligned
8	1000	2^3	This read is part of a pair and its mate was not aligned
16	10000	2^4	This read aligned in the reverse direction**
32	100000	2^5	This read is part of a pair and its mate aligned in the reverse direction**
64	1000000	2^6	This read is the first in the pair (read 1)
128	10000000	2^7	This read is the second in pair (read 2)
256	100000000	2^8	The given alignment is a secondary alignment***
512	1000000000	2^9	Read failed quality check (such as Illumina quality filtering)
1024	10000000000	2^{10}	Read was flagged as a duplicate (such as a PCR duplicate)
2048	100000000000	2^{11}	Supplementary alignment (Exact meaning varies by aligner)

FLAG value is 99 (64 + 32 + 2 + 1) indicating that:

- The read is the first in pair (read 1)
- The paired-end mate of this read mapped in the reverse direction
- The read was part of a properly aligning pair
- The read was paired

https://zymoresearch.eu/blogs/blog/what-are-sam-and-bam-files

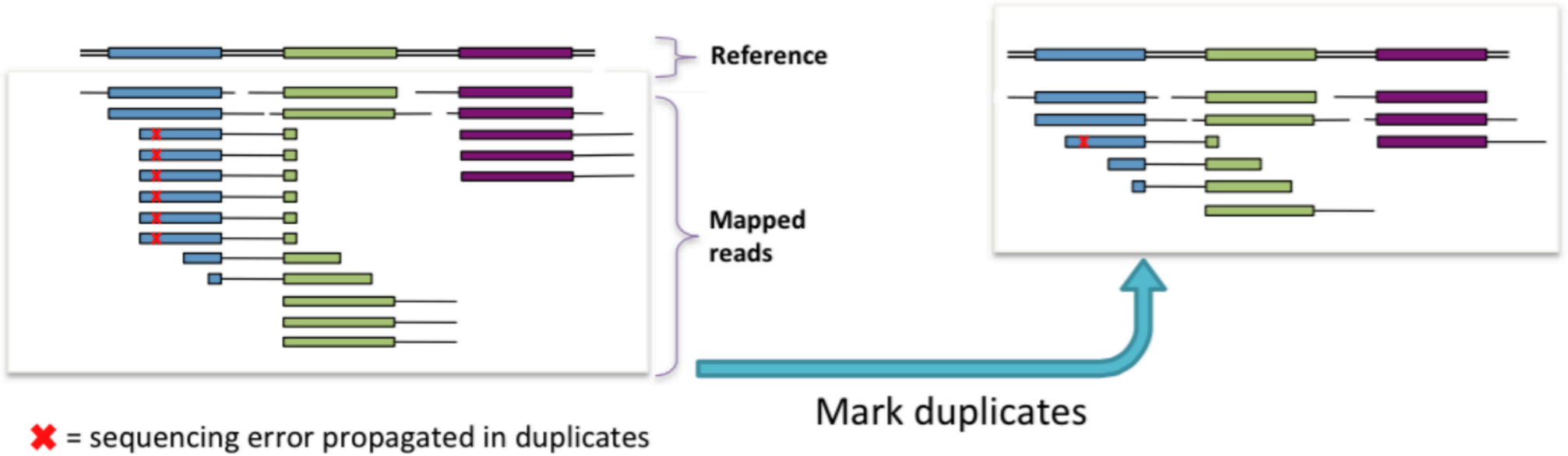
CIGAR (Compact Idiosyncratic Gapped Alignment Report)



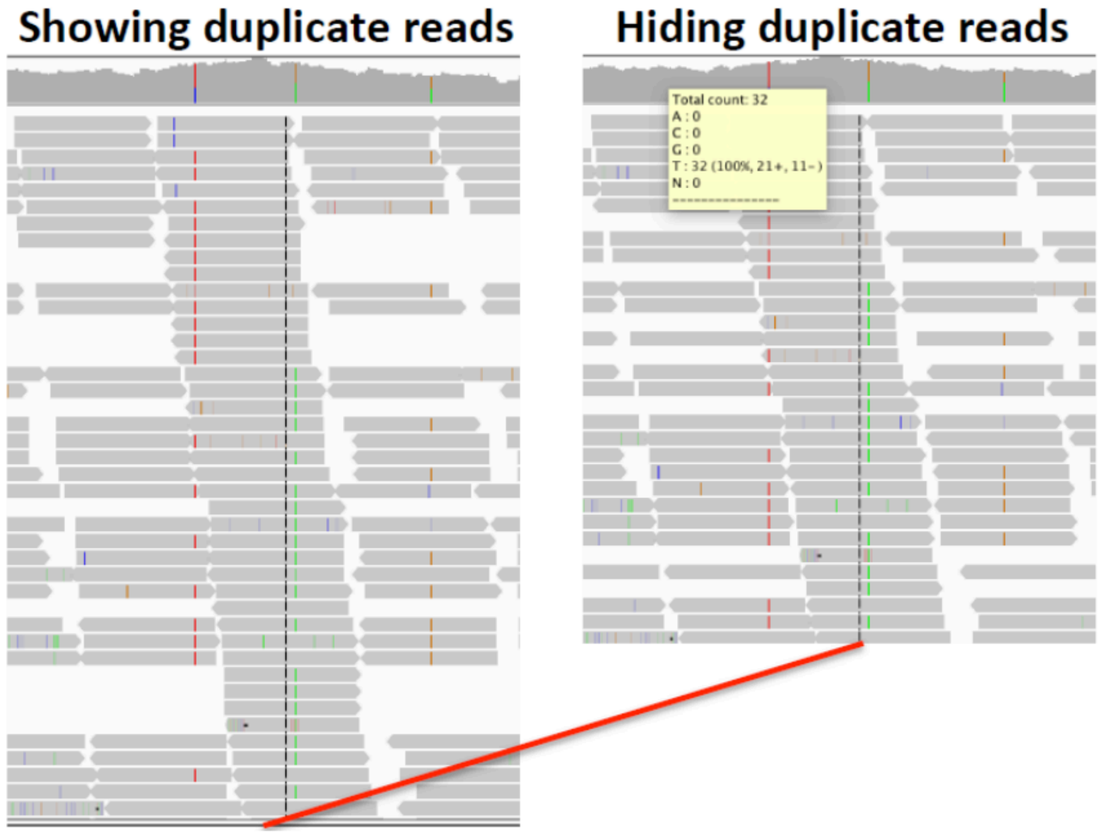
MARKING&REMOVING DUPLICATES

- Duplicates do not represent independent DNA fragments but are rather technical artifacts
- artificially inflate coverage and bias variant calling
- sets of reads pairs that have the same alignment start and end
- can lead to overrepresentation of certain alleles, resulting in false-positive variant calls

MARKING&REMOVING DUPLICATES



MARKING&REMOVING DUPLICATES



- Duplicate status is indicated in SAM flag
- Duplicates are not removed, just tagged (unless you request removal)
- Downstream tools can read the tag and choose to ignore those reads
- Most GATK tools ignore duplicates by default

BASE QUALITY SCORE RECALIBRATION

- adjusts base quality scores to account for known sources of errors, improving the reliability of variant calls
- provide better confidence estimates for true variants

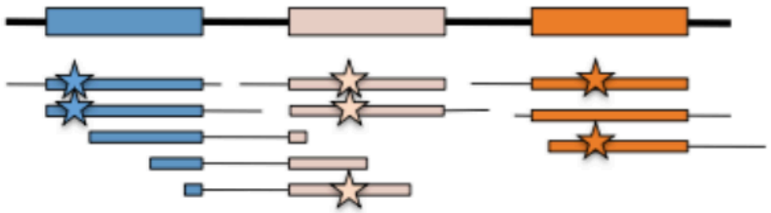
BASE QUALITY SCORE RECALIBRATION

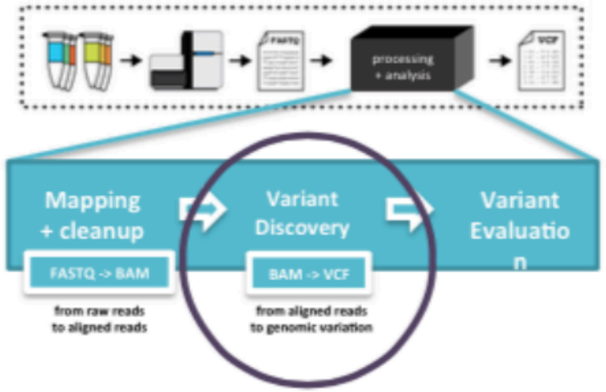
- First Step: Analyze Covariates (Error Detection)
  - examines how base quality scores are influenced by various factors, such as the sequencing cycle or the nucleotide context
  - generates a recalibration model, which maps out the biases and errors that need correction.
- Second Step: Recalibrate Base Scores (Error Correction)
  - the recalibration model is applied to the actual data to adjust the base quality scores.

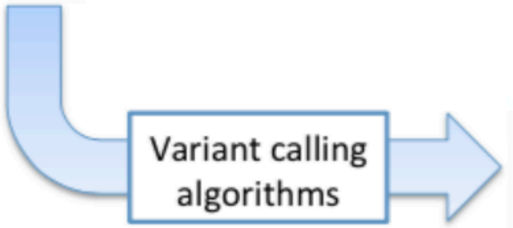
VARIANT CALLING

- Identification of genetic variants within the sequenced regions
- by comparing aligned reads (from the BAM file) to the reference genome

VARIANT CALLING







site 1 description
+ sample genotypes

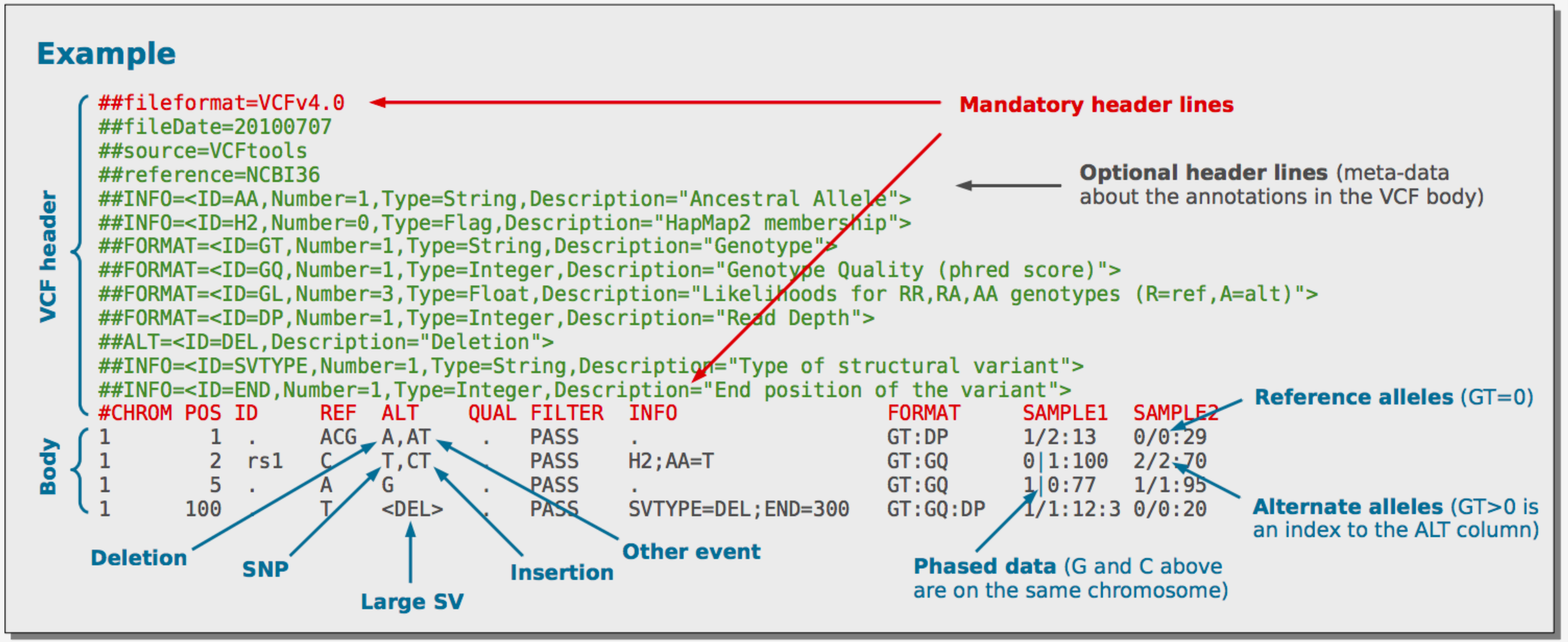
site 2 description
+ sample genotypes

site 3 description
+ sample genotypes

GENOTYPING

- Assigning of genotype information
- Analyze the proportion of reads supporting each allele to determine the most likely genotype
- Removal of reference homozygous regions (0/0)

VCF FORMAT


Example
VCF header
Body
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT . PASS . GT:DP 1/2:13 0/0:29
1 2 rs1 C T,CT . PASS H2;AA=T GT:GQ 0|1:100 2/2:70
1 5 . A G . PASS . GT:GQ 1|0:77 1/1:95
1 100 . T <DEL> . PASS SVTYPE=DEL;END=300 GT:GQ:DP 1/1:12:3 0/0:20
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Deletion
SNP
Large SV
Insertion
Other event
Phased data (G and C above are on the same chromosome)

GT (GENOTYPE)

REF: A
ALT: C, T

0/0 : A/A
1/0: C/A
1/1: C/C
2/0: T/A
2/1: T/C
2/2: T/T

DP (DEPTH)

- The number of times a specific base in the target region is sequenced
- Typically expressed as an integer (e.g., 30x depth means each base is sequenced 30 times on average)
- For WES, a depth of 30x–50x is usually considered sufficient for accurate variant calling
- For WGS, a depth of 30x is considered high-quality
- For clinical-grade sequencing, a depth of 100x or more is often preferred for critical or diagnostic regions to ensure very high confidence in detecting variants

COVERAGE

- The percentage of the target region that has been successfully sequenced
- Typically expressed as a percentage (e.g., 90% coverage means 90% of the target region has been sequenced)
- ≥95% coverage of the target region is typically considered good

GQ (GENOTYPE QUALITY)

- The confidence in the assigned genotype for a specific variant
- Often calculated using the likelihood of the observed data given the genotype compared to the likelihood of the data given an alternative genotype
- Typically expressed as Phred-scaled scores
- ≥20 is considered acceptable for variant calling
- ≥30 is ideal for high-confidence calls

MQ (MAPPING QUALITY)

- The reliability of the read's alignment to the reference genome
- Reflects how well a read aligns to a specific position, considering factors like the number of mismatches and the presence of multiple alignments
- also use a Phred scale
- ≥30 is commonly accepted as high-quality
- ≥50 is often preferred for very high-confidence alignments

ANNOTATION

- Assigning biological meaning to variants by linking them to known genes, their functions, and potential clinical significance
- Helps classify variants as benign, pathogenic, or of uncertain significance
- Provide insights into their roles in diseases

ANNOTATION

- **vcfanno**
  - Annotates ClinVar
  - Required for vcf2db
- **Ensembl Variant Effect Predictor (VEP)**
  - Adds genomic HGVS nomenclature
  - Adds frequency information from different databases
- **vcf2db**
  - Create Gemini compatible database for hg38
- **gemini**
  - Adds impact severity of the mutation
- **InterVar**
  - Clinical interpretation of genetic variants by the ACMG-AMP 2015 guidelines.
  - Uses **Annovar** for annotation

04

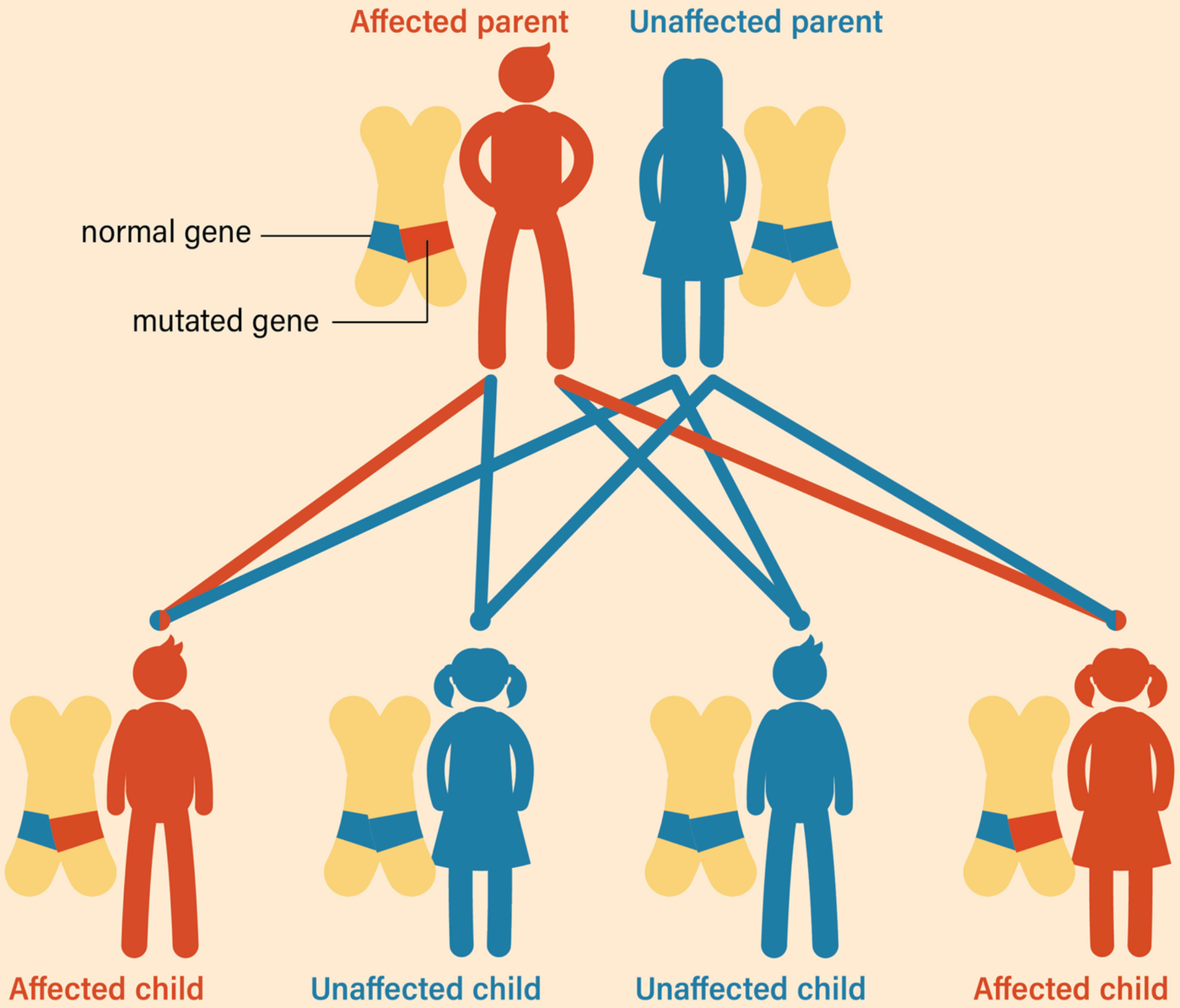
ACMG Guidelines

- The American College of Medical Genetics and Genomics
- Internationally accepted guidelines for the interpretation of variants

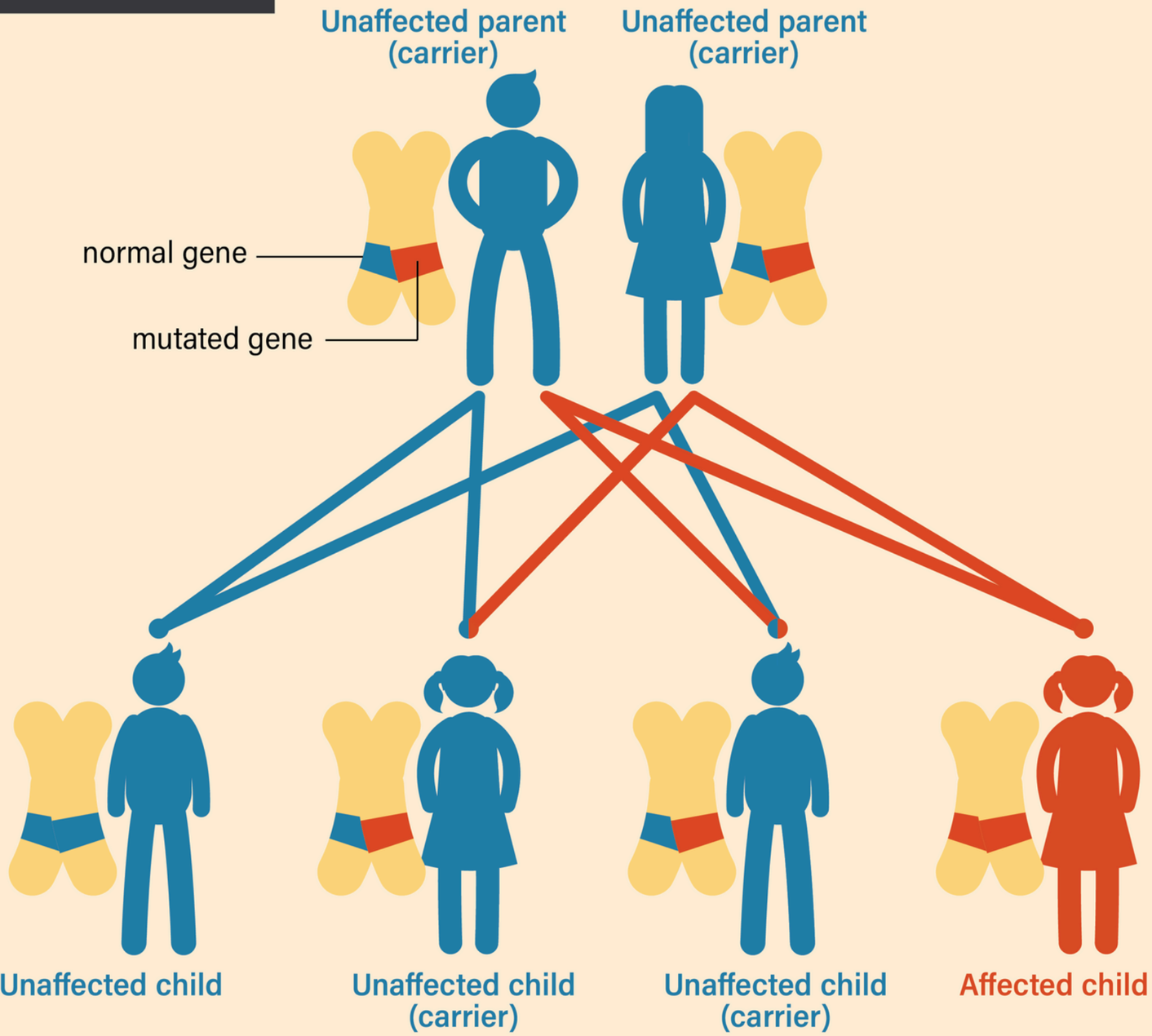
PATHOGENICITY ASSUMPTIONS

- Seen as rare or not seen in the healthy population
- Obey the inheritance pattern of the disease
- At the exonic or splicing region
- Affect the function of the protein
  - Found at;
    - a mutational hotspot
    - an important functional domain
  - The affected gene to be intolerant to the type of mutation of interest
  - Not shown as benign previously

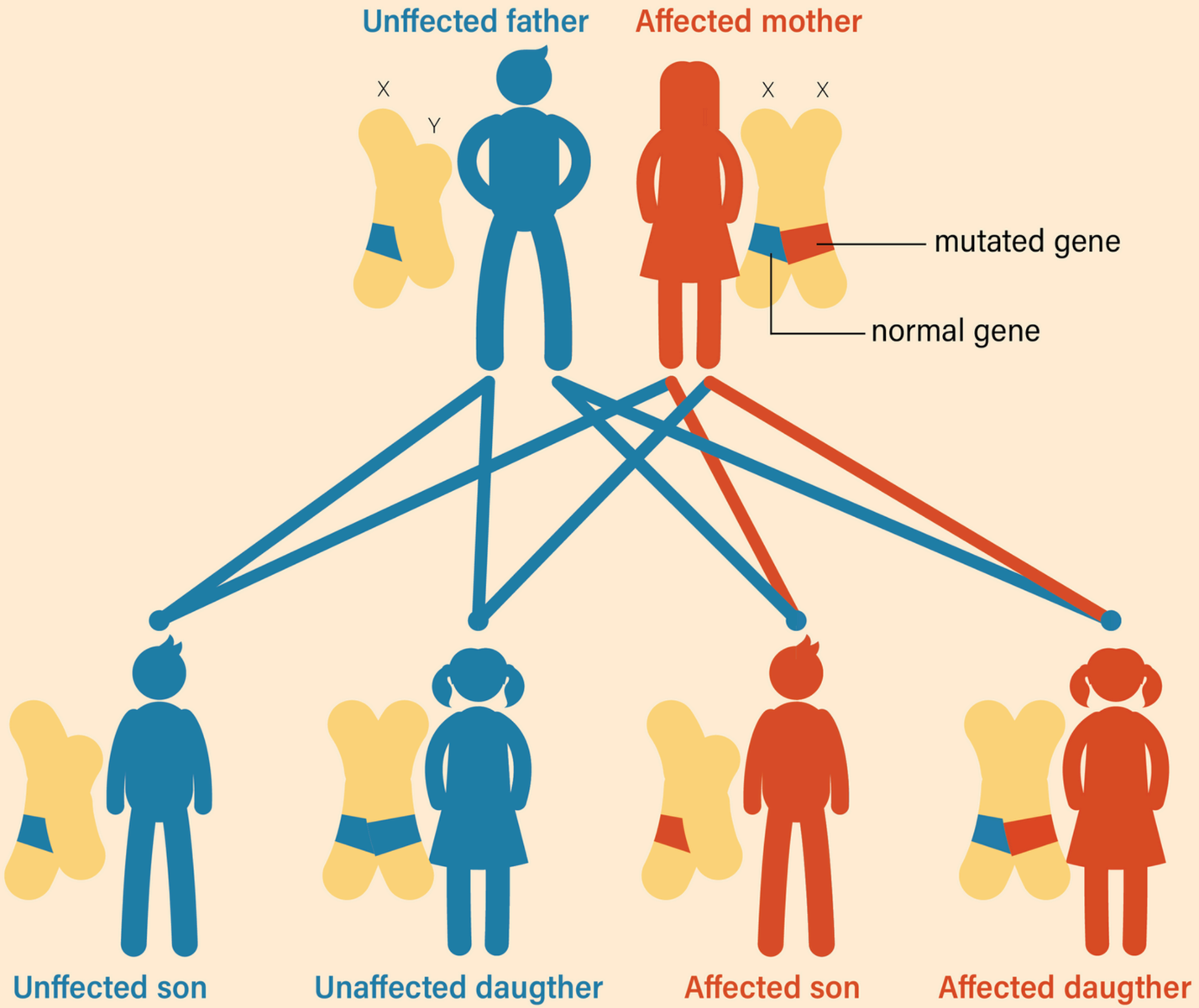
Autosomal dominant inheritance



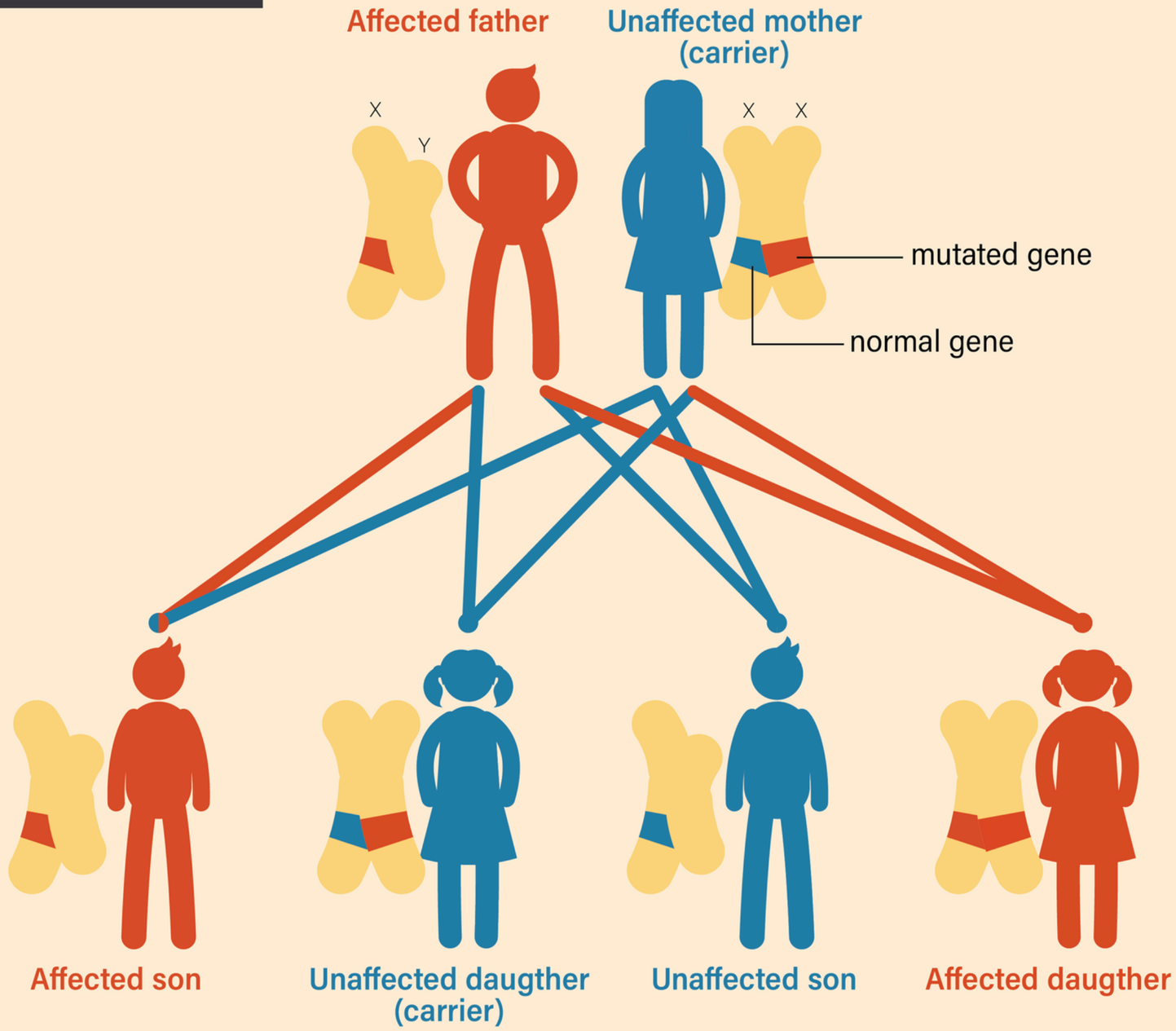
Autosomal recessive inheritance



X-linked dominant inheritance



X-linked recessive inheritance



MUTATION TYPES

Synonymous

Missense

In-frame indel

Loss of Function

- Nonsense
- Frameshift indels
- Splicing

BENIGN CRITERIA

Stand-alone

- Frequency of 0.005 for dominant genes
- Frequency of 0.01 for recessive genes

Strong

- Healthy example for fully penetrant diseases
- Experimental evidences
- Lack of segregation

Supporting

- Missense when LoF is known mechanism
- Inframe indels in repetitive regions
- In silico tool predictions
- Synonym mutations
- Mutations in; UTR, intron, intergenic

PATHOGENICITY CRITERIA

Very Strong

- Loss of function
- If LoF is known mechanism!

Strong

- Known pathogenic missense
- De novo
- Experimentally proven damage at variant level

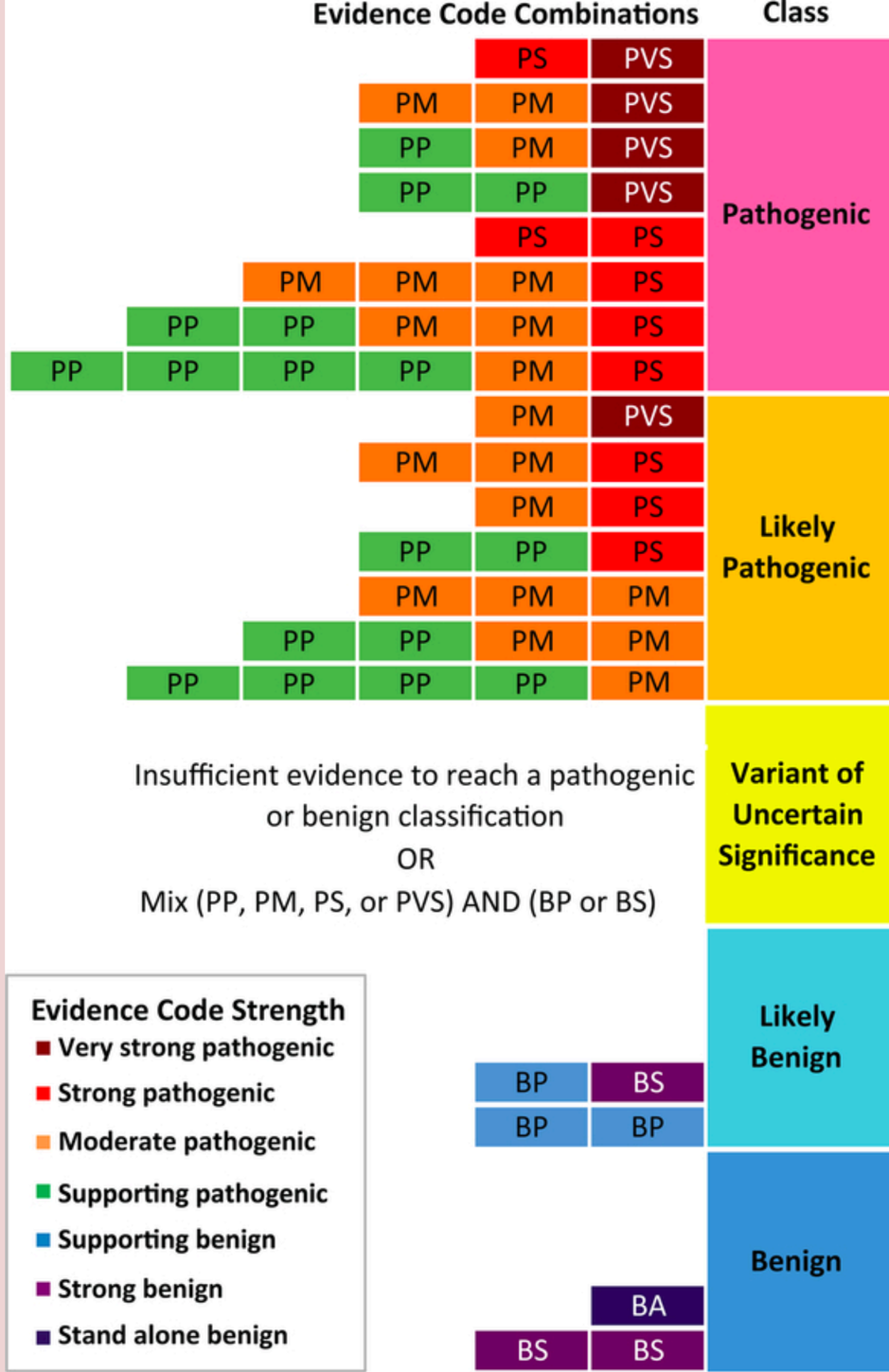
Moderate

- Mutational hotspots or well-known functional domain
- Inframe indels

Supporting

- Missense in a gene that has a low missense rate
- In silico tool predictions
- Experimentally proven damage at gene level

PATHOGENICITY INTERPRETATION



We want to hear about how you use gnomAD and your wish list! Please take 5 minutes to fill out our user survey.

genome aggregation database

gnomAD v2.1.1 Search by gene, region, or variant

Please note that gnomAD v2.1.1 and v3.1.1 have substantially different but overlapping sample compositions and are on different genome builds. For more information, see "Should I switch to the latest version of gnomAD?"

Examples

- Gene: PCSK9
- Transcript: ENST00000302118
- gnomAD v2.1.1 variant: 1-55516888-G-GA
- gnomAD v3.1.1 variant: 1-55051215-G-GA

DLG2 discs large MAGUK scaffold protein 2

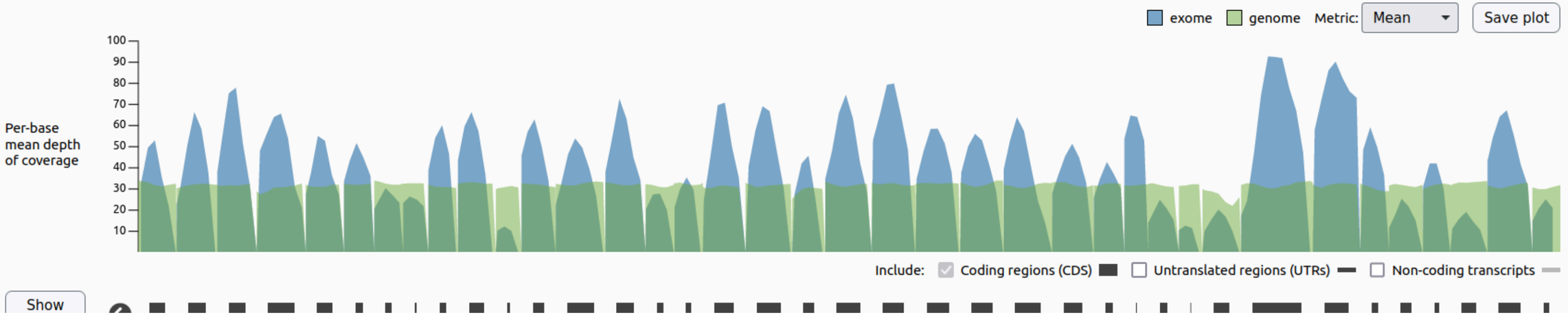
Dataset: gnomAD v2.1.1 | gnomAD SVs v2.1

Genome build GRCh37 / hg19
Ensembl gene ID ENSG00000150672.12
Ensembl canonical transcript ENST00000376104.2
Other transcripts ENST00000376106.3, ENST00000532653.1, and 30 more
Region 11:83166055-85338966
External resources Ensembl, UCSC Browser, and more

Constraint

Category	Expected SNVs	Observed SNVs	Constraint metrics
Synonymous	195.5	182	Z = 0.76 o/e = 0.93 (0.82 - 1.05)
Missense	535.2	400	Z = 2.08 o/e = 0.75 (0.69 - 0.81)
pLoF	61.2	13	pLI = 0.71 o/e = 0.21 (0.14 - 0.34)

Constraint metrics based on Ensembl canonical transcript (ENST00000376104.2).



DECIPHER GRCh38

About | Browse | DDD (UK) | Search DECIPHER | Help | Join | Log in

Mapping the clinical genome

Explore DECIPHER

It's free and you don't need to log in

DECIPHER is used by the clinical community to share and compare phenotypic and genotypic data. The DECIPHER database contains data from 46,207 patients who have given consent for broad data-sharing; DECIPHER also supports more limited sharing via consortia. Have a look at the numbers.

Anyone can browse publicly-available patient data on DECIPHER and request to be put in contact with the responsible clinician. Why? Because sharing benefits everyone.

Explore DECIPHER's genome browser

Delve into the Human Phenotype Ontology

Search all open-access DECIPHER data

Join DECIPHER

Be part of the sharing community

Projects affiliated to DECIPHER can deposit and share patients, variants, and phenotypes to invite collaboration and facilitate diagnosis. Once deposited, you can use DECIPHER to identify and prioritise potential matches, and you can request notifications as soon as new matches arrive.

As well as influencing individual patient outcomes, use of DECIPHER has contributed to over 2600 published articles since 2004. It's still free, and you are in control of what data to make public.

Join now

Find out more

Already a member?

Log in to access your patient data

Email address

Email

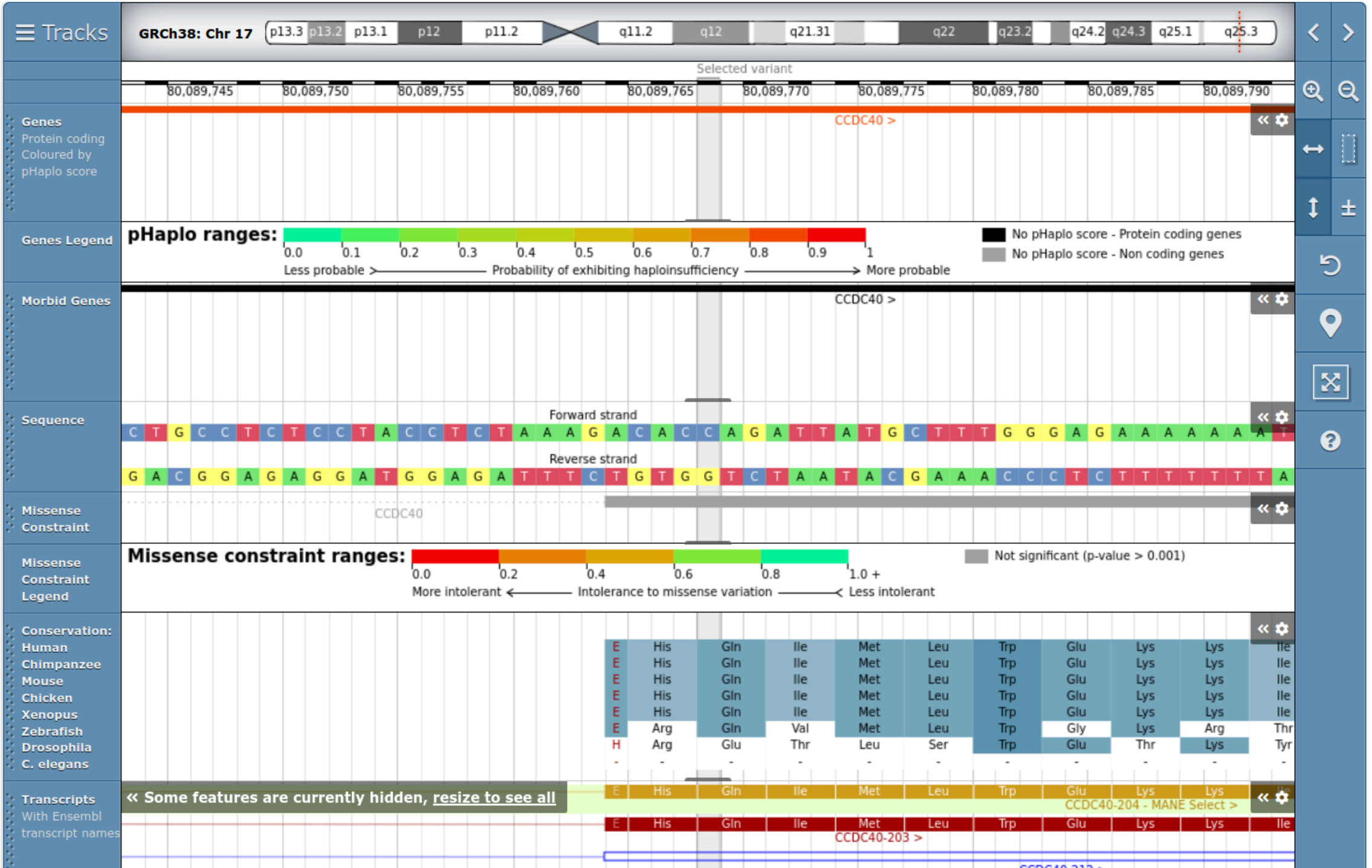
Password

Password

Log in

Reset your password

Feedback


Tracks
GRCh38: Chr 17
p13.3 p13.2 p13.1 p12 p11.2 q11.2 q12 q21.31 q22 q23.2 q24.2 q24.3 q25.1 q25.3
Selected variant
80,089,745 80,089,750 80,089,755 80,089,760 80,089,765 80,089,770 80,089,775 80,089,780 80,089,785 80,089,790
Genes
Protein coding
Coloured by
pHaplo score
CCDC40 >
Genes Legend
pHaplo ranges:
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Less probable
Probability of exhibiting haploinsufficiency
More probable
No pHaplo score - Protein coding genes
No pHaplo score - Non coding genes
Morbid Genes
CCDC40 >
Sequence
Forward strand
C T G C C T C T C C T A C C T C T A A A G A C A C C A G A T T A T G C T T T G G G A G A A A A A A A A T
Reverse strand
G A C G G A G A G G A T G G A G A T T T C T G T G G T C T A A T A C G A A A C C C T C T T T T T T T A
CCDC40
Missense Constraint
Missense constraint ranges:
0.0 0.2 0.4 0.6 0.8 1.0 +
More intolerant
Intolerance to missense variation
Less intolerant
Not significant (p-value > 0.001)
Missense Constraint Legend
Conservation:
Human
Chimpanzee
Mouse
Chicken
Xenopus
Zebrafish
Drosophila
C. elegans
E His Gln Ile Met Leu Trp Glu Lys Lys Ile
E His Gln Ile Met Leu Trp Glu Lys Lys Ile
E His Gln Ile Met Leu Trp Glu Lys Lys Ile
E His Gln Ile Met Leu Trp Glu Lys Lys Ile
E His Gln Ile Met Leu Trp Glu Lys Lys Ile
E Arg Gln Val Met Leu Trp Gly Lys Arg Thr
H Arg Glu Thr Leu Ser Trp Glu Thr Lys Tyr
Transcripts
With Ensembl
transcript names
« Some features are currently hidden, resize to see all
CCDC40-204 - MANE Select >
CCDC40-203 >

50 YEARS OMIM
Human Genetics Knowledge for the World

OMIM®

An Online Catalog of Human Genes and Genetic Disorders

Updated May 8, 2023

Search OMIM for clinical features, phenotypes, genes, and more...

Advanced Search : OMIM, Clinical Synopses, Gene Map

Need help? : Example Searches, OMIM Search Help, OMIM Video Tutorials

Mirror site : https://mirror.omim.org

OMIM is supported by a grant from NHGRI, licensing fees, and generous contributions from people like you.

Make a donation!

McKUSICK-NATHANS Department of Genetic Medicine

JOHNS HOPKINS MEDICINE

Case study

- 6-year-old boy
- Global developmental delay
- Generalized hypotonia
- Speech delay
- Strabismus: bilateral esotropia
- Decreased pain response
- Hyperactive DTR
- Mild facial dysmorphisms: frontal bossing, low-set ears

FILTERING CRITERIA

- Max allele frequency **< 0.001**
- Exclude variants with ClinVar clinical significance of **benign** or **likely benign**
- Exclude variants with ACMG classification of **benign** or **likely benign**
- Primary candidates -> knock-down (homozygous LoF)
- Check ClinVar Disease Name and check OMIM
- Include other missense and heterozygous mutations by relaxing the criteria and check ClinVar and OMIM again