DNA-seq alignments and genome analysis

Quantitative Methods Workshop 2026

Erik Owen

January 5, 2026

Page Lab | Computational and Systems Biology

Massachusetts Institute of Technology

Whitehead Institute

Welcome to today's genomics roadmap:

- Sequencing technology: how reads (and errors!) are made
  - Covered in prework section
- File formats: FASTA, FASTQ, SAM, BAM, VCF
- Align reads to reference genome
- Call variants + interpret evidence
- Understand depth & coverage

- Case studies:
  - Pathogenic SNP in *HBB* locus
  - Repeat-heavy complexity in *MTERF3* locus

Sequencing Technologies 2026 Overview

Type	Length (bp)	Method	Error (per bp)	Cost
Sanger	~500	SBS (Seq by Synth)	10^{-5}	$4/read
Illumina	~100	SBS - bridge PCR	10^{-3}	$2.3 - $50/gigabase
Aviti	~100	SBS - rolling PCR	10^{-4}	~$15/gigabase
PacBio HiFi	10^{4}	SBS	$> 10^{-3}$	$300/gb
Oxford Nanopore	$> 10^{4-6}$	Sensing technology	10^{-2}	$30/gb

Sequencing Technologies 2026 Overview

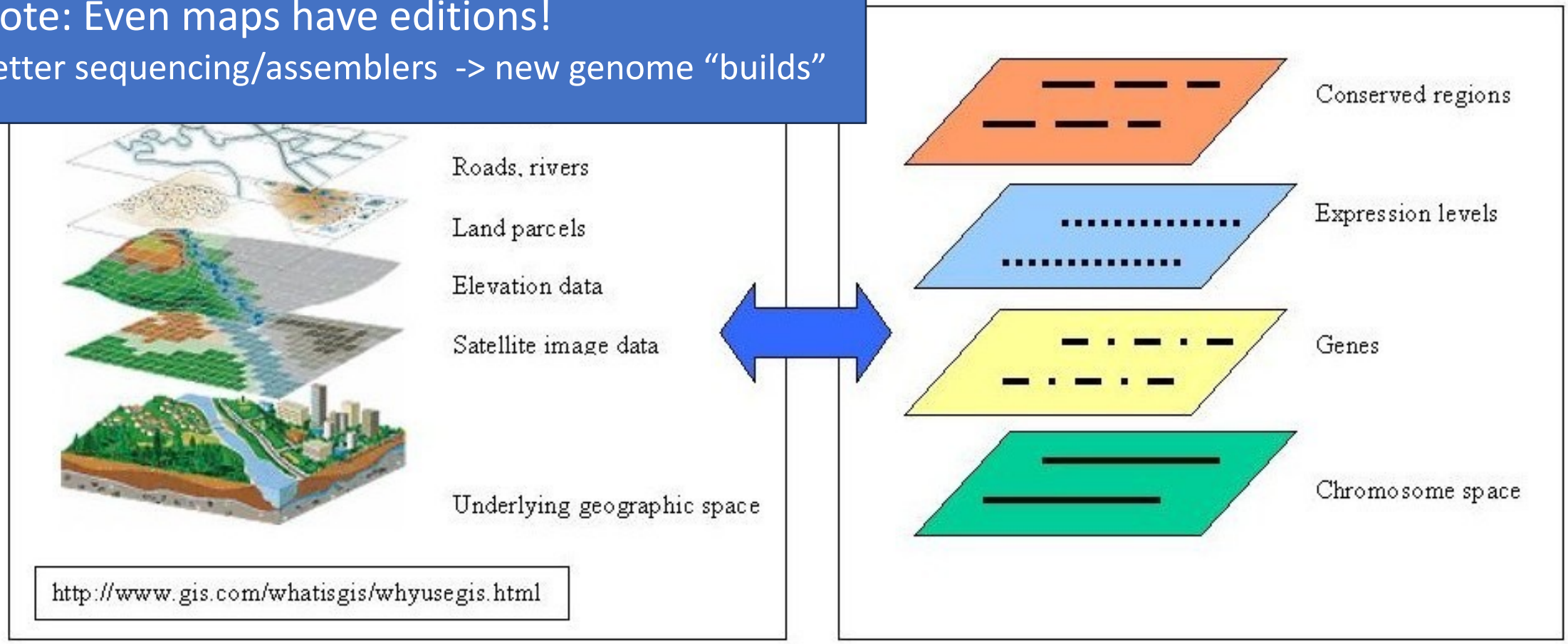
Type	Length (bp)	Method	Error (per bp)	Cost
Sanger	~500	SBS (Seq by Synth)	10^{-5}	$4/read
Illumina	~100	SBS - bridge PCR	10^{-3}	$2.3 - $50/gigabase
Aviti	~100	SBS - rolling PCR	10^{-4}	~$15/gigabase
PacBio HiFi	10^{4}	SBS	$> 10^{-3}$	$300/gb
Oxford Nanopore	$> 10^{4-6}$	Sensing technology	10^{-2}	$30/gb

Today's focus!

Reference genomes are like maps:
Features on either map are coordinate based

Note: Even maps have editions!

Better sequencing/assemblers -> new genome "builds"



Dolan, Mary E et al. "Genomes as geography: using GIS technology to build interactive genome feature maps." *BMC bioinformatics* vol. 7 416. 19 Sep. 2006, doi:10.1186/1471-2105-7-416

Genomic coordinates require a build

- A coordinate is like a street address:

  MIT

  77 Massachusetts Ave.

  Cambridge, MA 02139

- Genome coordinates need:
  - A build: {hg19, hg38, hs1}
  - A chromosome:
  - Position range:
  - For variants, what change?

- Here's a range describing the locus for *XIST*:

  hg38 chrX:73,820,656-73,852,714

- Here's a variant example:

  hg38 chr1:55,555,555 G>A

  - *Note*: for position range, 0-based vs. 1-based matters based on file type (more later)

No two people are 100% identical; we carry germline & somatic mutations

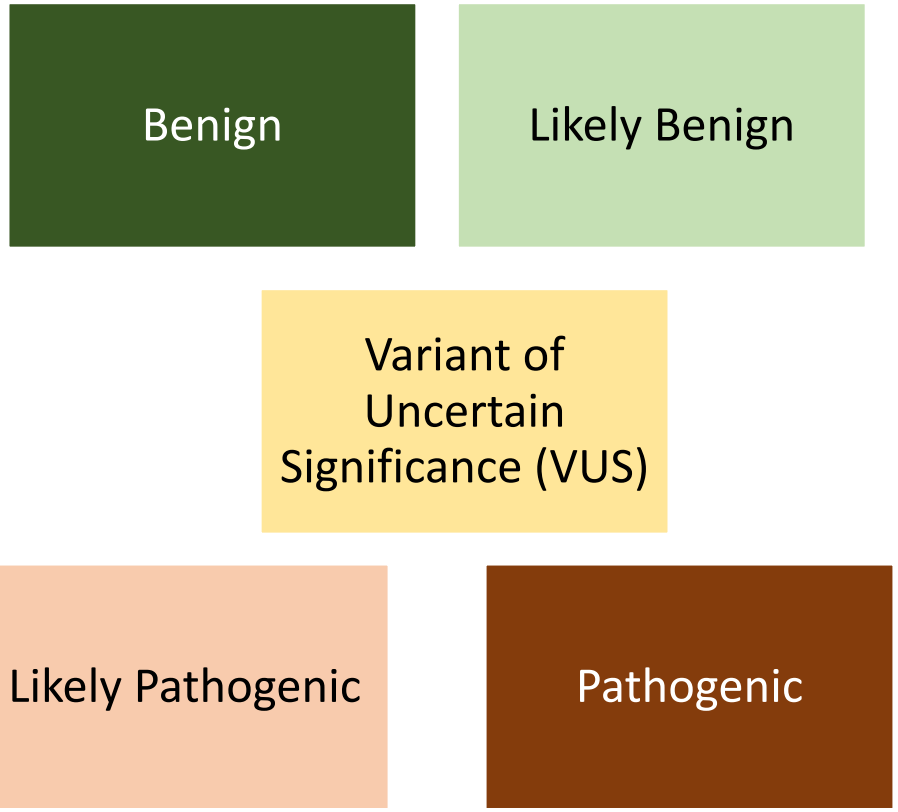
- Human genome is $\approx 3\times10^9$ bp long!
- However, at base-pair level, humans are ~99.9% identical!
- Two state locus $\{N_1, N_2\}$: which state you are in == your allele!
- There are $\approx 3\times10^6$ SNPs per person > ~1 SNP/1000 bp
- Variants include
  - Single nucleotide variants (SNVs) – single base pair (bp)
  - indels (insertions + deletions) -- <50 bp
  - Structural variants (SVs) – >50 bp – all the way to chr level (aneuploidy)

- **Driving question of genetics: how do genetic variants drive a measurable trait (aka a phenotype)?**

A variant call is an evidenced-based claim

- Variant := difference from reference
- Variant call := interpretation of evidence

- Interpretation labels:



Today's goal: learn how to find a variant!

Checkpoint!
Answer questions in LMS

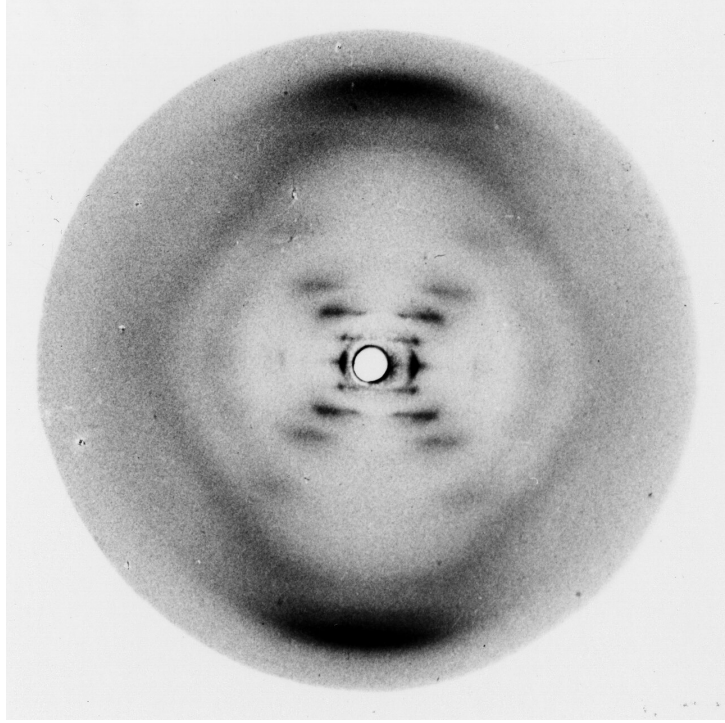


Photo 51, Raymond Gosling and Rosalind Franklin 1952

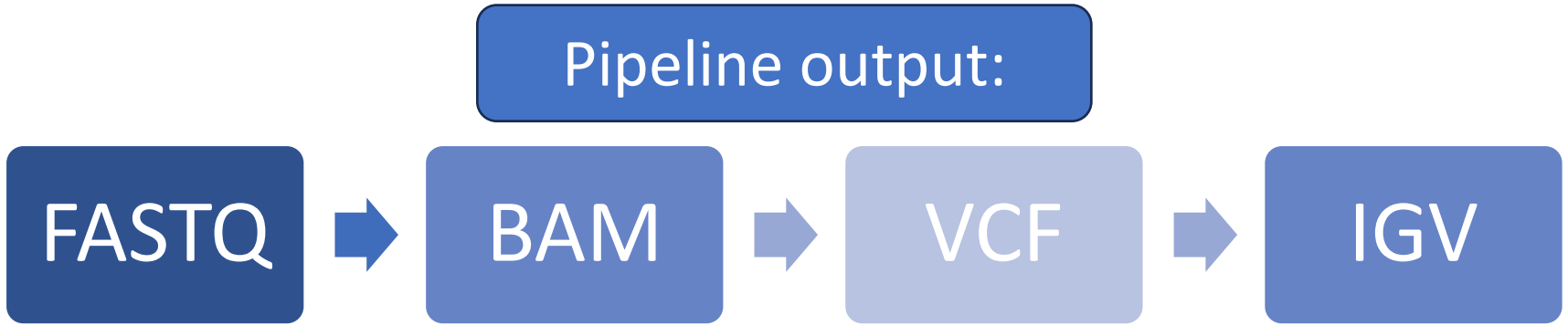
Case studies: same pipeline, different trust

Locus 1: *HBB* (on chr11)

- *HBB* codes for Hemoglobin subunit beta
- Has functional evidence for how a SNV causes sickle cell disease

Locus 2: *MTERF3* (on chr8)

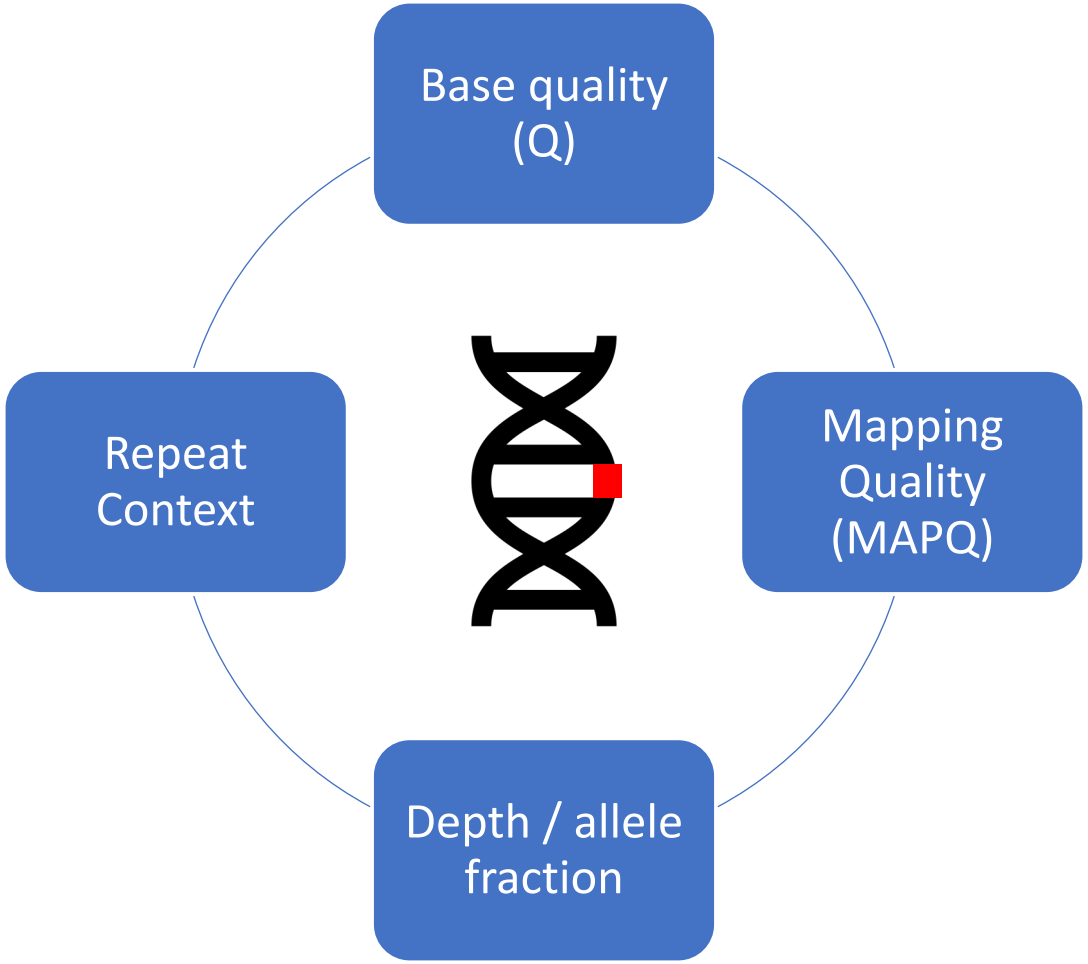
- *MTERF3* codes for Mitochondrial Transcription Termination Factor 3
- Repeat-heavy region which will let us study ambiguous alignments **if we have time**

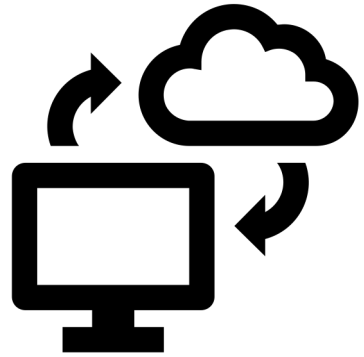


Using notebook with colab, you will generate:

1. QC summary from FASTQ
2. Alignment (BAM) & index
3. Coverage & MAPQ summaries
4. Variant calls (VCF)
5. Visualization of variants

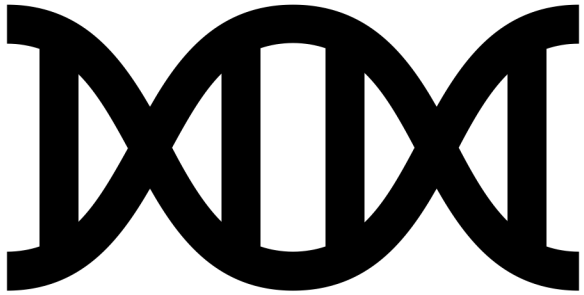
Before you trust a call, check:





Now follow instructions in the README of the linked git repo for *Environment Setup*

https://github.com/erik-owen/2026_QMW_Genomics/tree/main



Complete the *Sequencing & File formats* portion in colab

https://github.com/erik-owen/2026_QMW_Genomics/tree/main

Alignment Module

- Aligners
  - algorithms that place each read on a reference by optimizing a score
  - matches == good
  - mismatches, indels, clipping == bad
- Repeats create multiple equally good placements
  - Leads to “multi-mapping” where aligner can’t be confident
- MAPQ metric
  - Confidence in placement of (trimmed) read
  - Low MAPQ == “this read could be put in multiple places in the reference”

Aligner walkthrough demo: Bowtie2

- Bowtie2 is a short-read alignment tool
- Good at aligning reads of 50-100s/1000s of characters

- https://bowtie-bio.sourceforge.net/bowtie2/index.shtml

- If you would like to learn how the genome indexing works "under the hood", it's based on the Burrows-Wheeler transform

End-to-end vs local alignment

End-to-End involves all characters in read

```
Alignment:


Read:
  GACTGGGCGATCTCGACTTCG
  |||||  |||||||||| |||
  GACTG--CGATCTCGACATCG
Reference:
```

Local alignment allows trimming/clipping

```
Alignment:


Read:
  ACGGTTGCGTTAA-TCCGCCACG
       |||||||||| ||||||
  TAACTTGCGTTAAATCCGCCTGG
Reference:
```

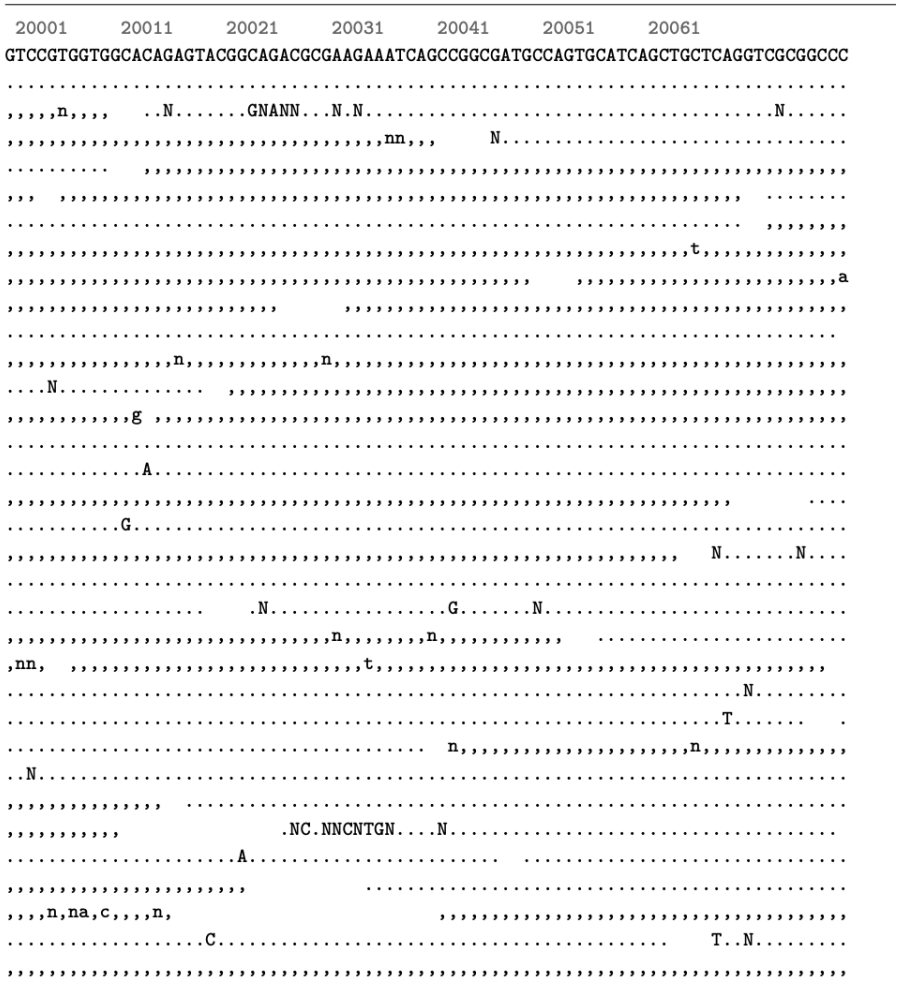
https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#end-to-end-alignment-versus-local-alignment

Now let's work through the *Alignment* portion of the notebook

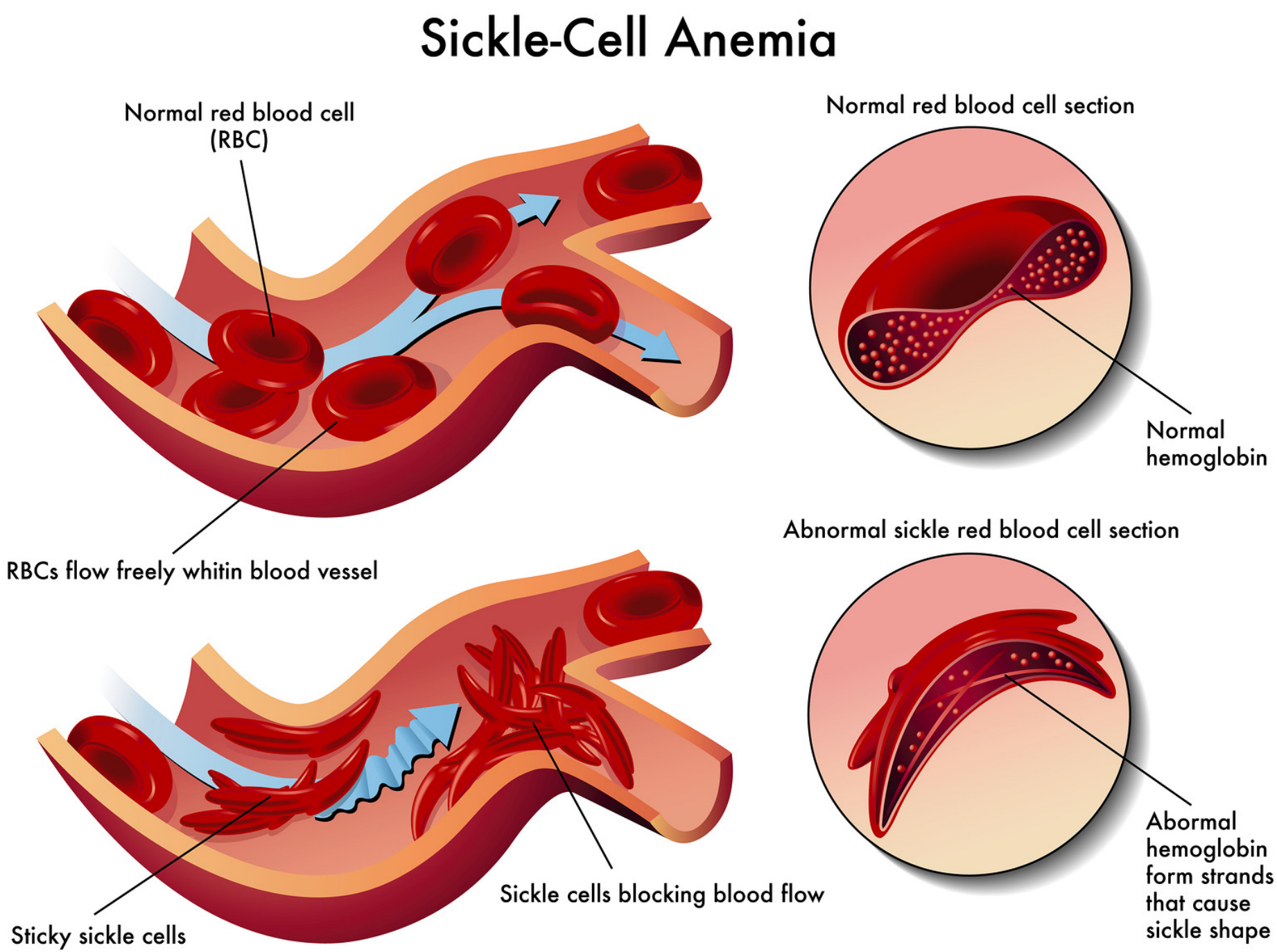
https://github.com/erik-owen/2026_QMW_Genomics/tree/main

Pileups, Variants, & Interpretation

- Pileup our reads!
- Our coverage graph showed we have alignment to our reference FASTA
- Now we want to look to find if we have variants in the individual we sequenced from relative to our reference

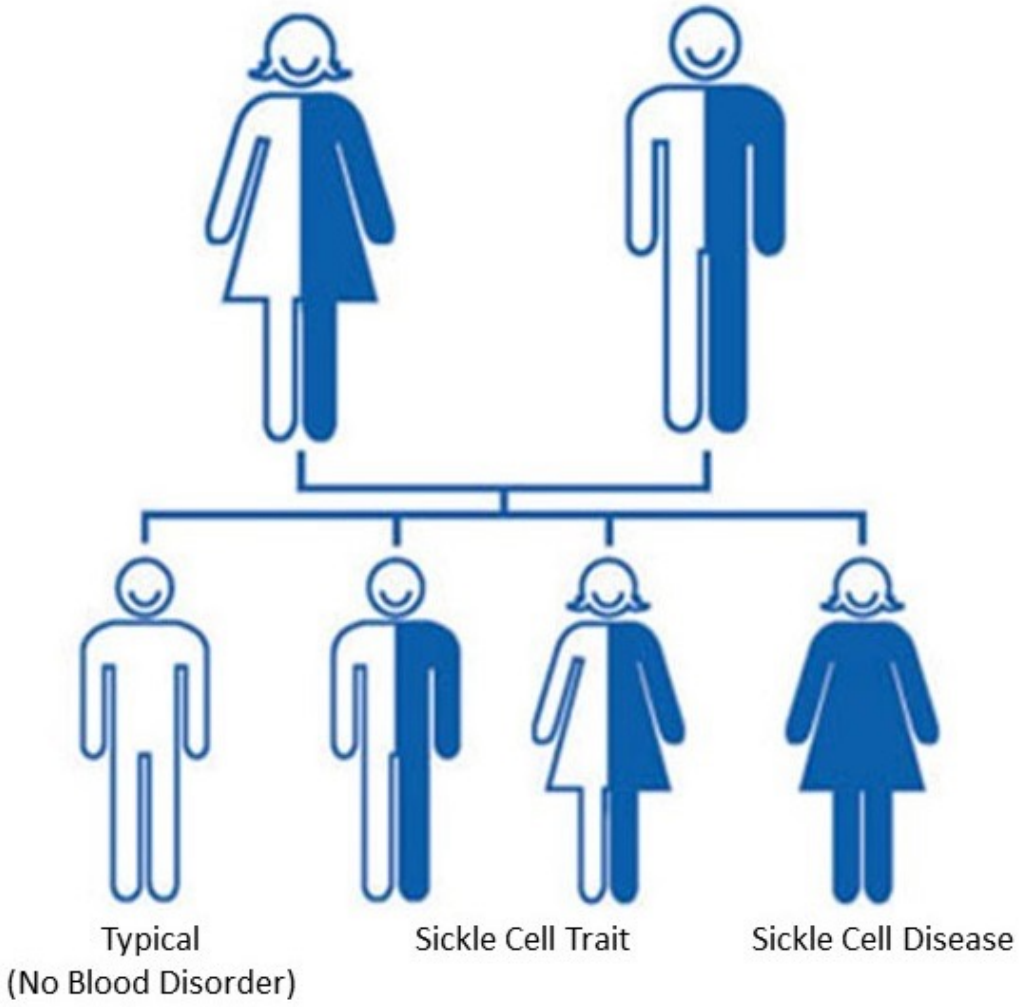


Case Study


Sickle-Cell Anemia
Normal red blood cell (RBC)
RBCs flow freely whitin blood vessel
Normal red blood cell section
Normal hemoglobin
Abnormal sickle red blood cell section
Sticky sickle cells
Sickle cells blocking blood flow
Abormal hemoglobin form strands that cause sickle shape

Sickle Cell Disease appears as an autosomal recessive pattern

- Sickle Cell Trait can sometimes also lead to pain crises
- Sickle cell disease affects ~7.7 million people worldwide
- Causes >34k annual deaths; contributes to >375k deaths
- However, sickling is protective against malaria, so there's a "heterozygote advantage"



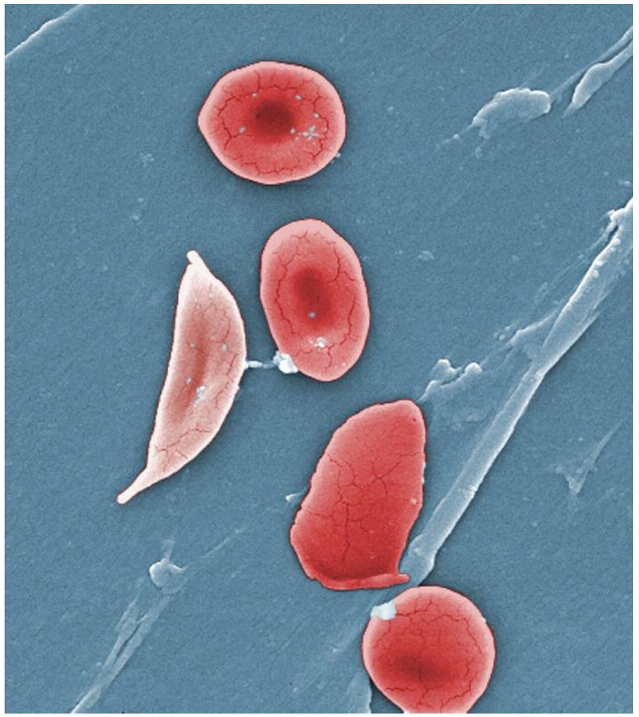
https://doi.org/10.1016%2FS2352-3026%2823%2900118-7

https://www.cdc.gov/sickle-cell/sickle-cell-trait/index.html

Case study: Sickle Cell Disease

Patient 1: healthy control

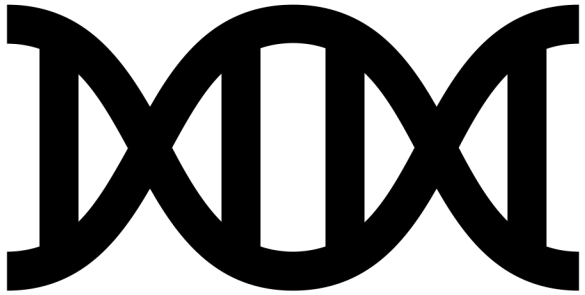
- No sickle cell



Patient 2: from disease cohort trio

- Patient is healthy
- However, daughter carries sickle cell disease

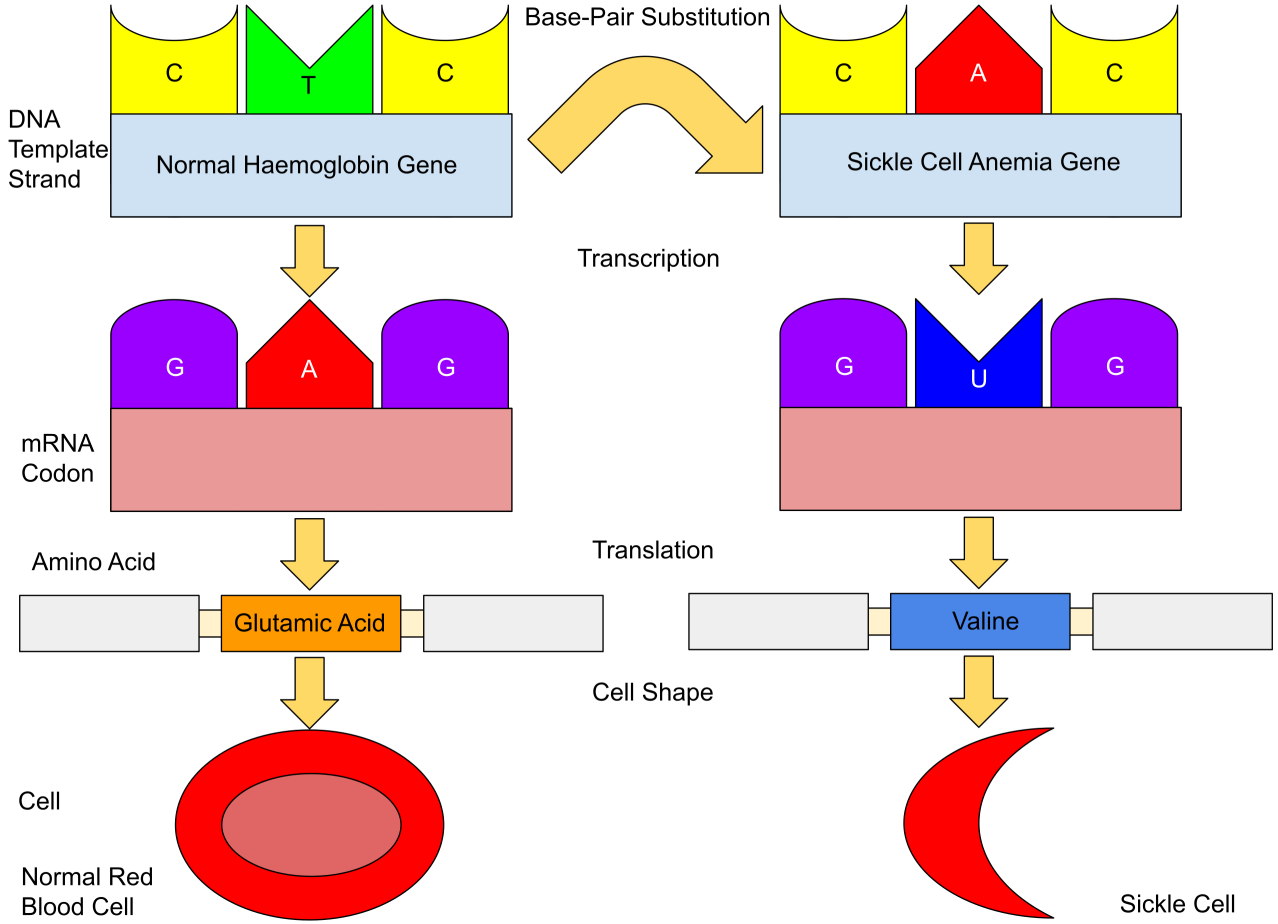
- We want to know if patient parent is a carrier of a variant that would cause the sickle cell phenotype?



Complete the *Pileups & Variant Calling* portion in colab

https://github.com/erik-owen/2026_QMW_Genomics/tree/main

Functional impact of HbS mutation



https://en.wikipedia.org/wiki/Sickle_cell_disease#/media/File:Sickle_Cell_Anemia.svg

Molecular change in HbS was described in 1956 – before sequencing!

A SPECIFIC CHEMICAL DIFFERENCE BETWEEN THE GLOBINS OF NORMAL HUMAN AND SICKLE-CELL ANÆMIA HÆMOGLOBIN

By Dr. V. M. INGRAM

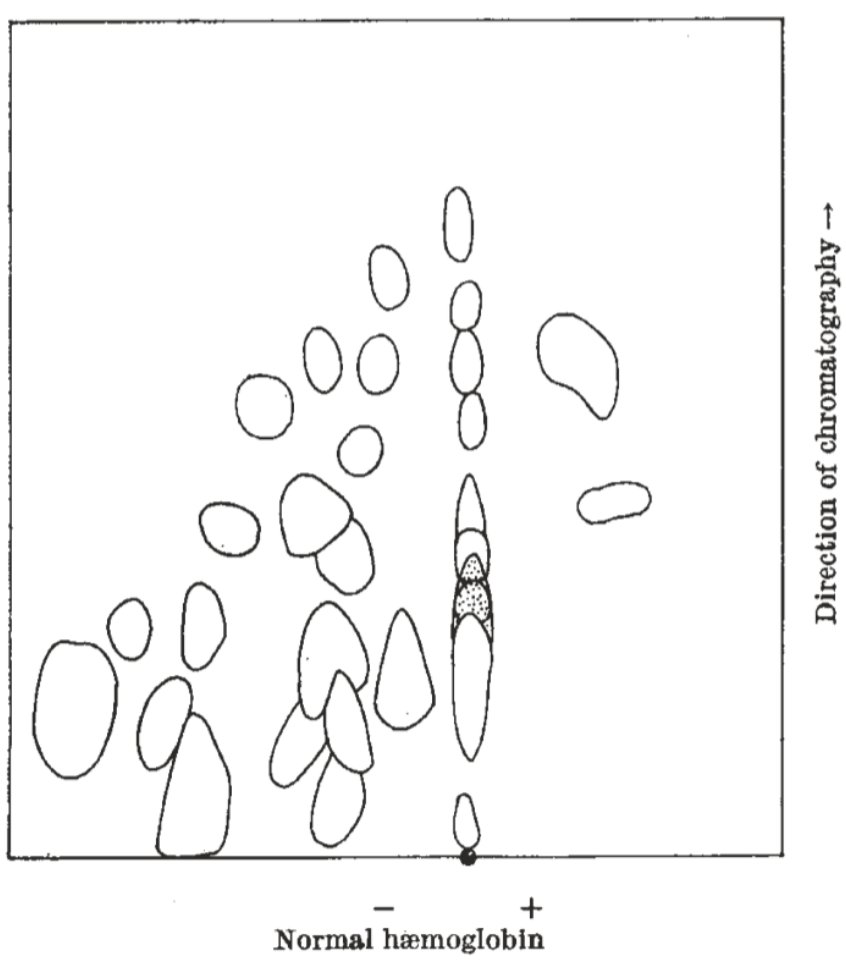
Medical Research Council Unit for the Study of the Molecular Structure of Biological Systems, Cavendish Laboratory, University of Cambridge

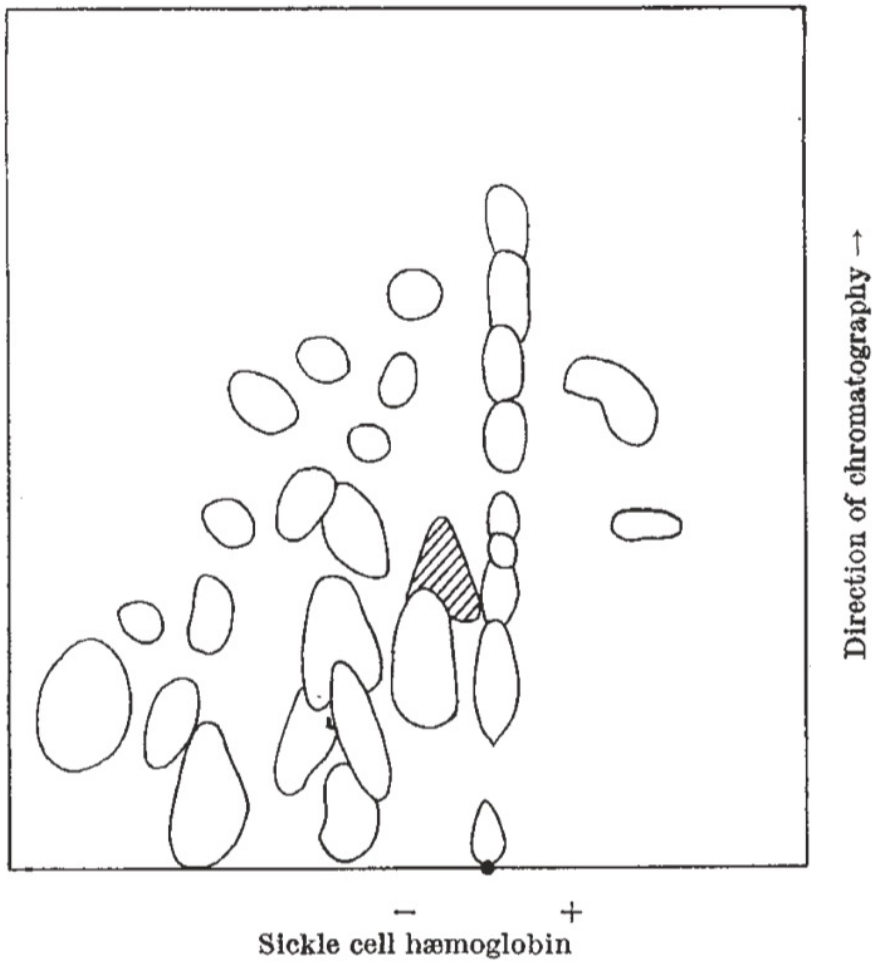
A NEW and rapid technique of characterizing the chemical properties of a protein in considerable detail has been devised ; by its application a specific difference is found in the sequence of amino-acid residues of normal and sickle-cell hæmoglobin. This difference appears to be confined to one small section of one of the polypeptide chains.

Of all the abnormal human hæmoglobins, the one that has been most intensively studied is hæmoglobin *S* from patients with sickle-cell anæmia. In 1949 Pauling and his collaborators[1] demonstrated by electrophoretic experiments that at neutral *p*H the hæmoglobin *S* molecule has a net charge which is more positive by three units compared with the normal molecule, hæmoglobin *A*. It has since been suggested[2] that this difference is really due to hæmoglobin *S* having fewer free carboxyl groups than does hæmoglobin *A*. It is also known that in the reduced state the abnormal protein has a much lower solubility[3]. However, careful determinations of the amino-acid composition of the two proteins[4,5] did not show any significant differences between them within the accuracy of the methods employed. Comparison of the N-terminal[6] and C-terminal[7] amino-acids and of the sulphydryl groups[8] was equally disappointing. On this evidence alone, it is not possible to decide whether the difference between the proteins, which is in any event small, lies in the amino-acid

INGRAM, V. A Specific Chemical Difference Between the Globins of Normal Human and Sickle-Cell Anæmia Hæmoglobin. *Nature* **178**, 792–794 (1956). https://doi.org/10.1038/178792a0

Molecular change in HbS was described in 1956 – before sequencing!



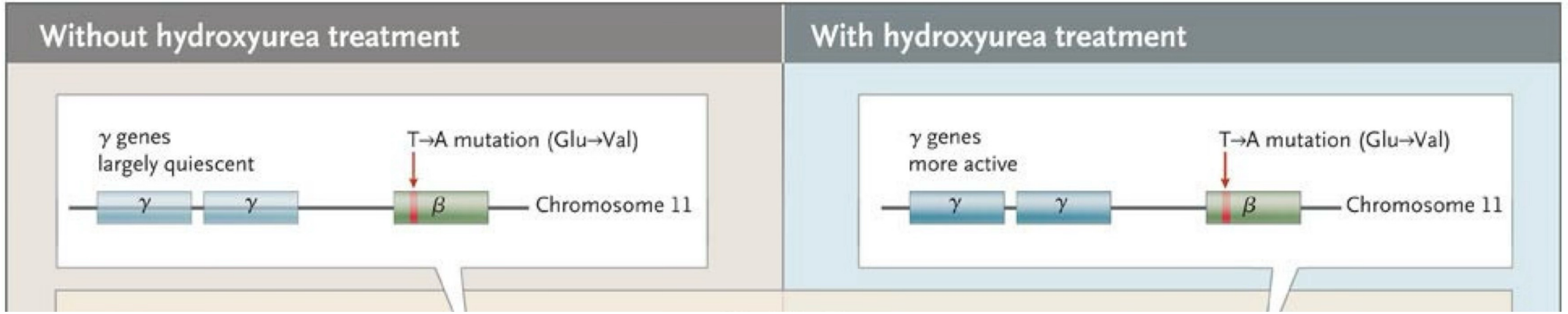


INGRAM, V. A Specific Chemical Difference Between the Globins of Normal Human and Sickle-Cell Anæmia Hæmoglobin. *Nature* **178**, 792–794 (1956). https://doi.org/10.1038/178792a0

What is the future of sickle cell disease management?

Slides adapted from HMS case conference by Maya Talukdar, PhD

What is the future of sickle cell disease management?



› Science. 2008 Dec 19;322(5909):1839-42. doi: 10.1126/science.1165409. Epub 2008 Dec 4.

Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A

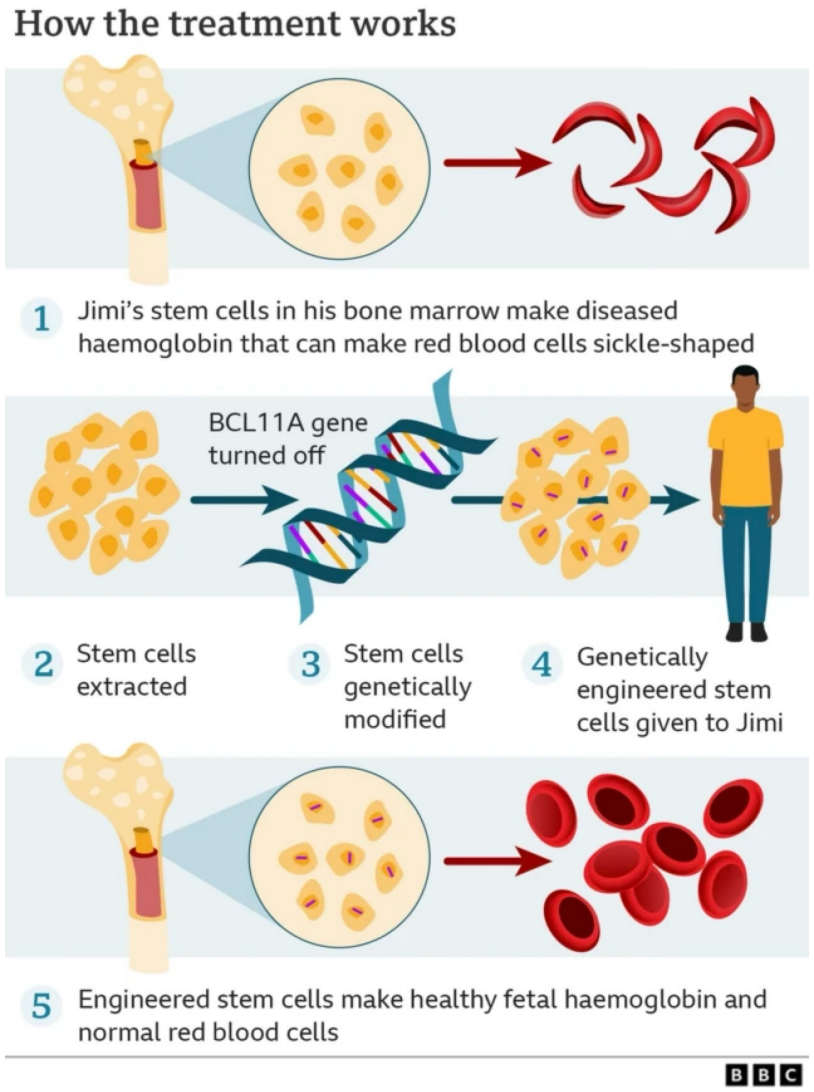
Vijay G Sankaran [1], Tobias F Menne, Jian Xu, Thomas E Akie, Guillaume Lettre, Ben Van Handel, Hanna K A Mikkola, Joel N Hirschhorn, Alan B Cantor, Stuart H Orkin

Affiliations + expand

PMID: 19056937 DOI: 10.1126/science.1165409

What genetic variants do people who naturally have very high levels of HbF carry?

What is the future of sickle cell disease management?



CONCLUSIONS

Treatment with exa-cel eliminated vaso-occlusive crises in 97% of patients with sickle cell disease for a period of 12 months or more. (CLIMB SCD-121; ClinicalTrials.gov number, NCT03745287.)

FDA NEWS RELEASE

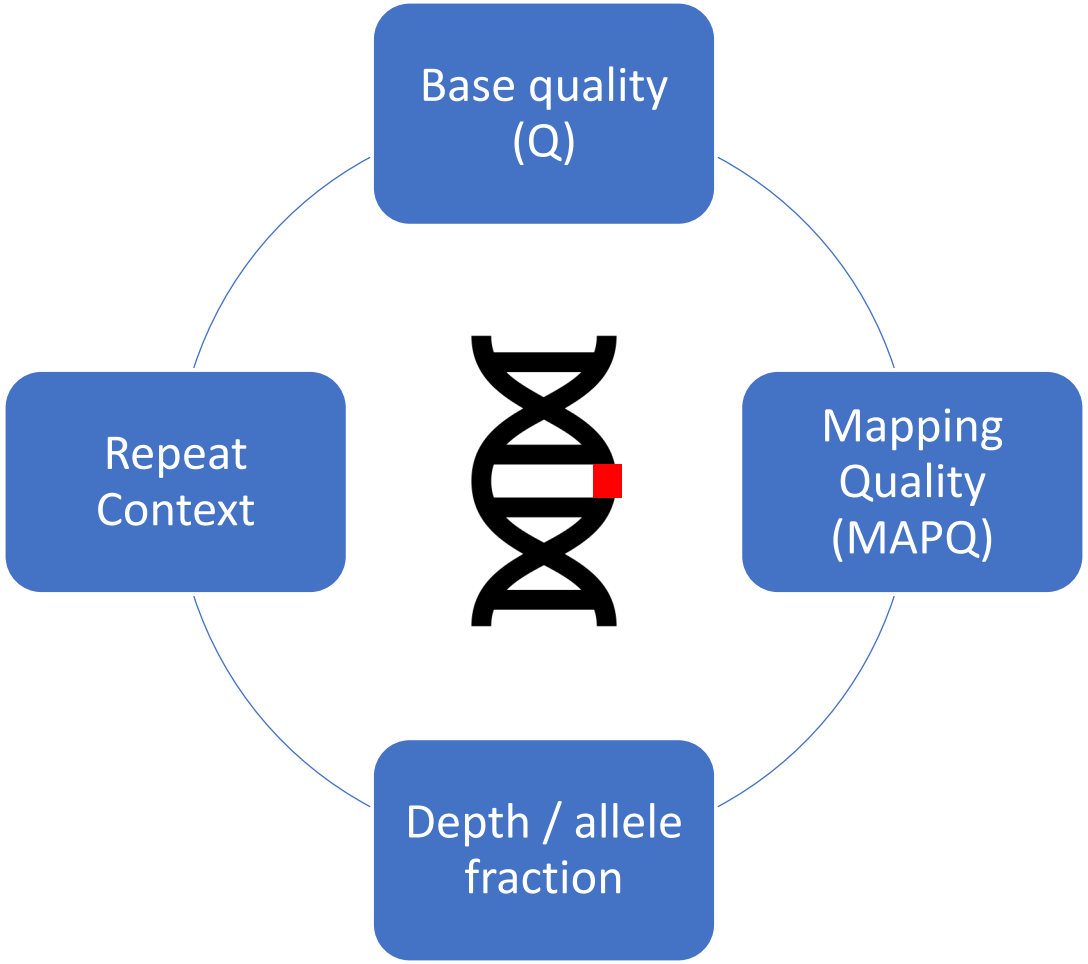
FDA Approves First Gene Therapies to Treat Patients with Sickle Cell Disease

For Immediate Release: December 08, 2023

First Patient Begins Newly Approved Sickle Cell Gene Therapy

A 12-year-old boy in the Washington, D.C., area faces months of procedures to remedy his disease. "I want to be cured," he said.

Wrapup: Before you trust a call, check:



What's next?

- Long read sequencing and aligners like minimap2
- STAR for RNA-seq
- Structural variants

- Assemblers

- Variant interpretation: ClinVar
  https://www.clinicalgenome.org/data-sharing/clinvar/

Thank you!!
Hope you're excited about genomics!!!

- Special Acknowledgements:
  - Mentorship in the Page Lab
    - Especially D.W. Bellott and H. Skaletsky
  - HST.508 Population and Quantitative Genomics
    - Professors Tami Liberman and Leonid Mirny

- Feel free to chat after class or email about:
  - genomics & bioethics
  - industry vs. academia
  - grad school
- email: eowen (at) mit (dot) edu