



How to Use the Human Pangenome Reference in AnVIL

Alice B. Popejoy, Julian Lucas, and Jean Monlong
Human Pangenome Reference Consortium (HPRC)



*Alice Popejoy, PhD
Assistant Professor, UC Davis Health*

#ACMGMtg23

*Alice Popejoy, PhD
Assistant Professor, UC Davis Health*



*Julian Lucas, BS
Graduate Student, UC Santa Cruz*

*Alice Popejoy, PhD
Assistant Professor, UC Davis Health*

*Julian Lucas, BS
Graduate Student, UC Santa Cruz*



***Jean Monlong, PhD
Postdoctoral Researcher, UC Santa Cruz***

Financial Disclosure

ALICE POPEJOY & JEAN MONLONG

Do not have any relevant disclosures.

Financial Disclosure

JULIAN LUCAS

Discloses the following relevant relationships with ineligible companies. Any potential interests have been mitigated:

Independent contractor (including contracted research) relationship with Prime Genomics (has ended).

Workshop Logistics

Slides
10 minutes

Discussion
15 minutes

Slides
10 minutes

Demo
40 minutes

Discussion
15 minutes

Introduction To The Human Pangenome Reference Consortium

Using The Human Reference Genome In Clinical Genetics

Overview of AnVIL and HPRC Resources

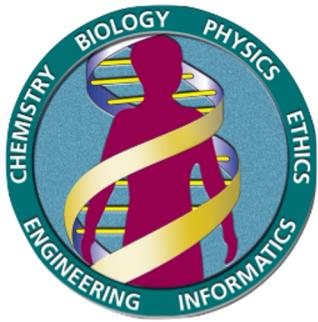
How To Use Pangenomes To Improve (Clinical) Variant Calling

The Pangenome Reference For Clinical Variant Interpretation

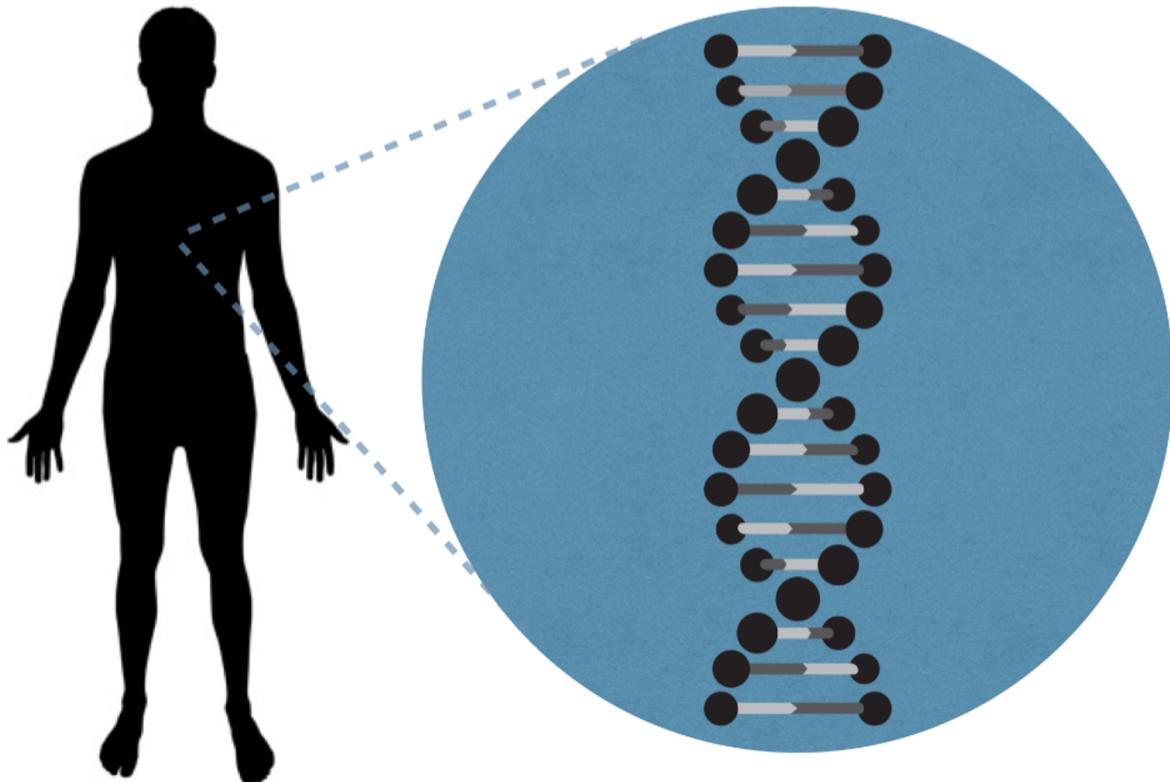
Introduction To The Human Pangenome Reference Consortium



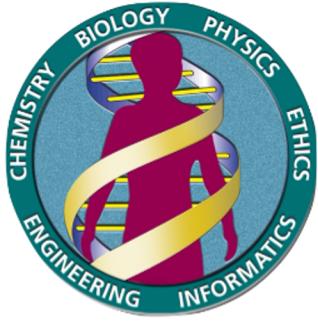
The Initial Human Genome Project



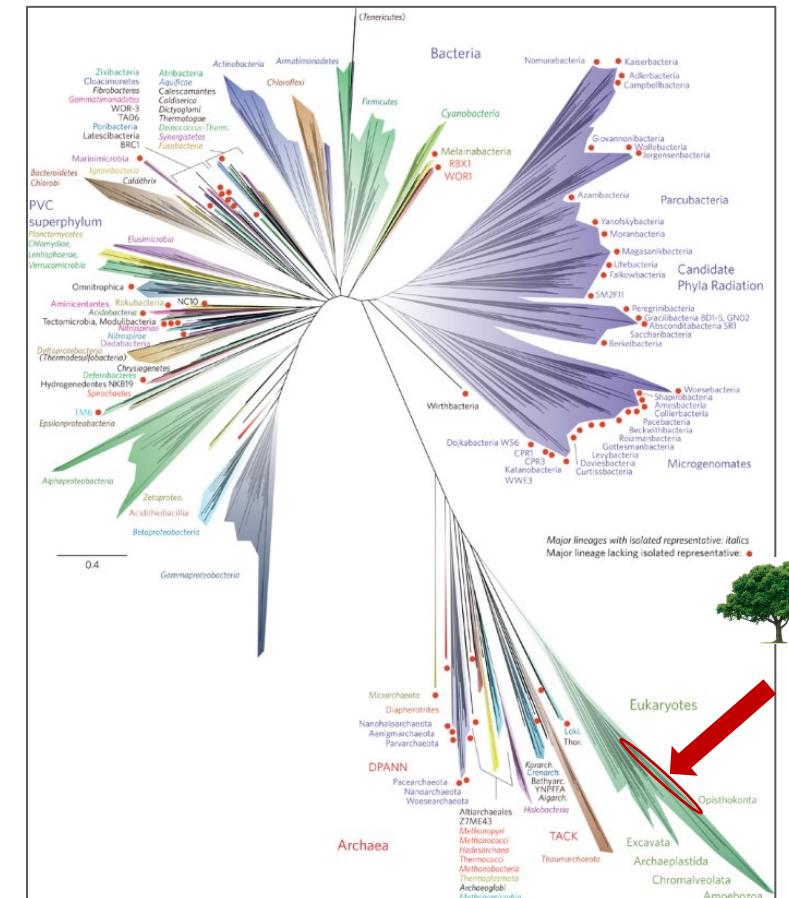
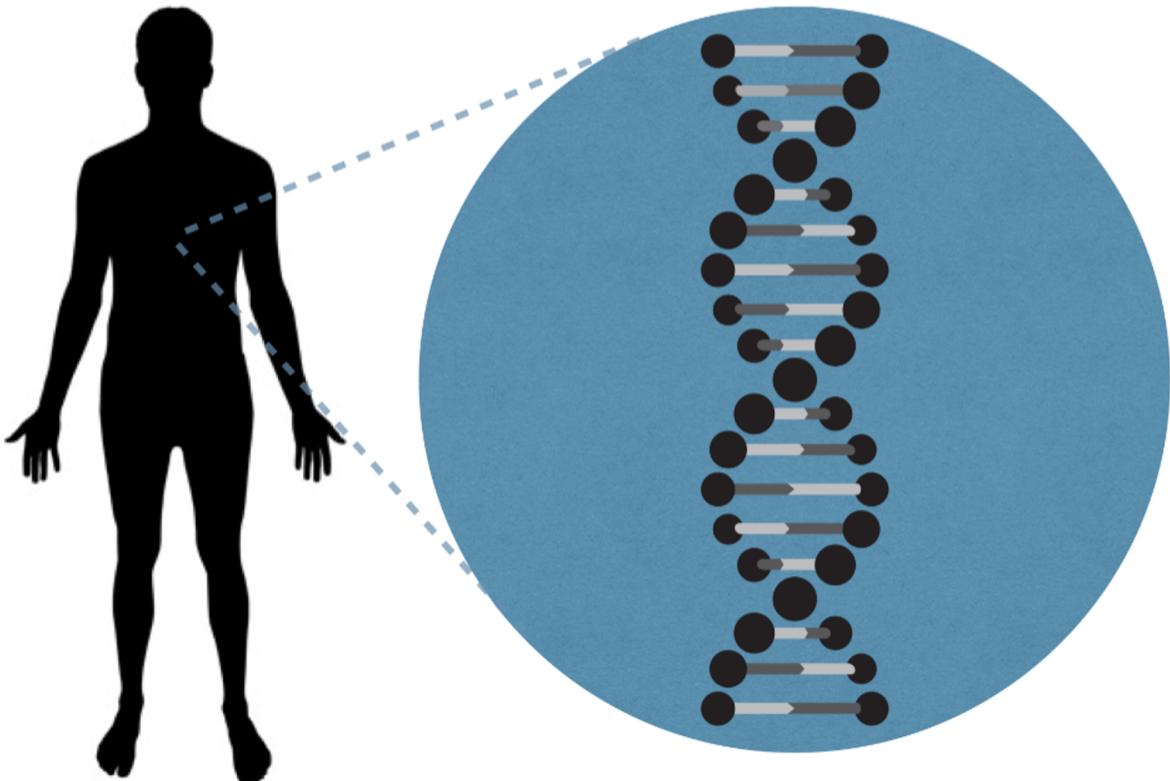
The Human Genome Project: **international, collaborative research program** whose goal was the complete mapping and understanding of all the genes of human beings.



The Initial Human Genome Project

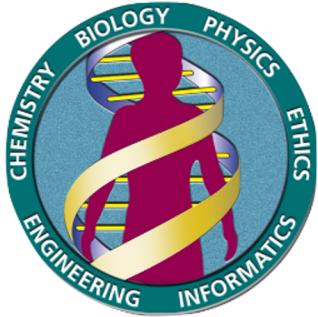


The Human Genome Project: **international, collaborative research program** whose goal was the complete mapping and understanding of all the genes of human beings.

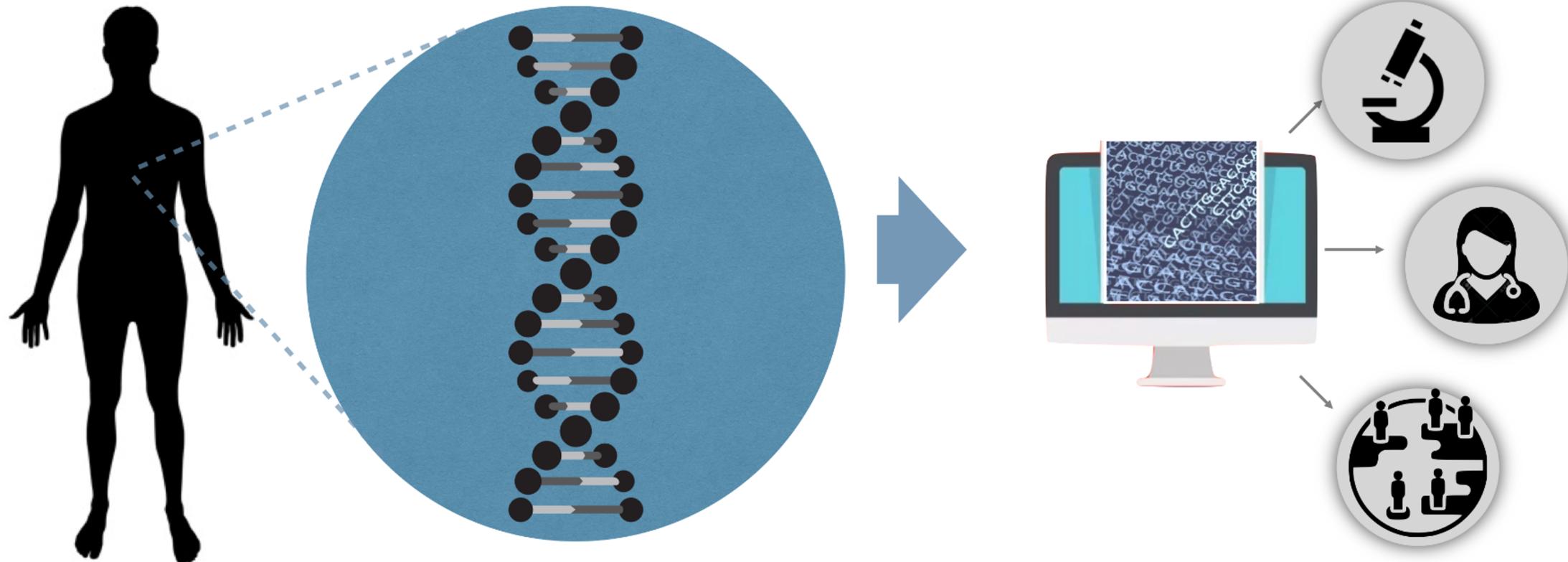


“Tree of Life” (Hug et al., 2016)

The Initial Human Genome Project



Centralized coordinate system: critical for **sharing genomic data and data analysis standards** among international researchers.



How do you use the human reference genome in your work (if at all)?

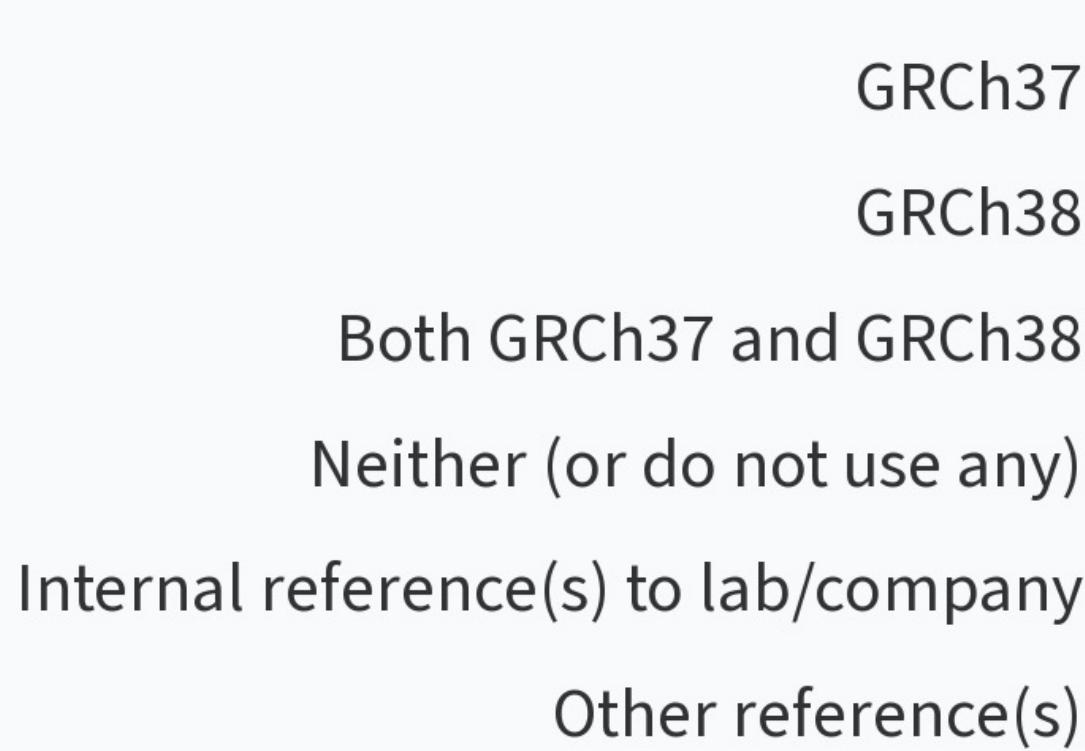
What do you find most useful or important about the current reference genome?

What concerns, frustrations, or limitations are you aware of regarding the human reference genome?

🌐 When poll is active, respond at **pollev.com/popejoy**

SMS Text **POPEJOY** to **+1 (747) 444-3548** once to join

If you do use the human reference genome, which build is most frequently used?



UPDATE! Data Structure: C- to an A+

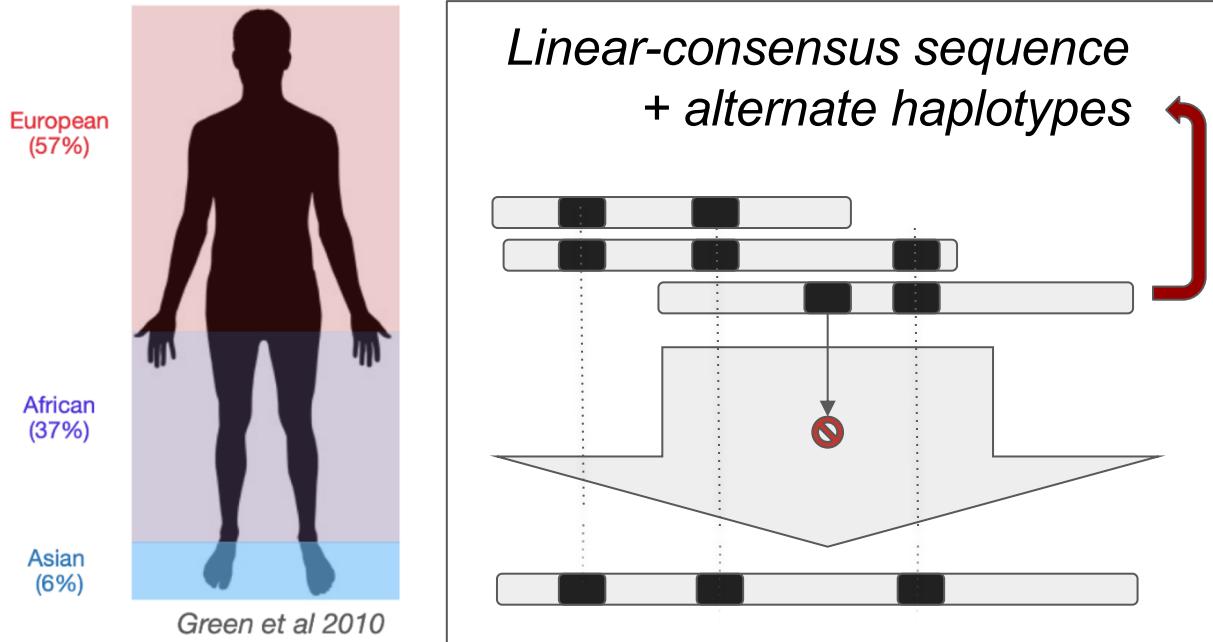
What is ‘*The Human Genome Reference*’ ?

- live genomes are dynamic and actively changing in our cells
- data representations of genomes collapse complexity to 2D
- current genome representations are linear; misleading
- global *diversity must be integrated* into the reference genome

Data Structure: C- to an A+

What is ‘*The Human Genome Reference*’ ?

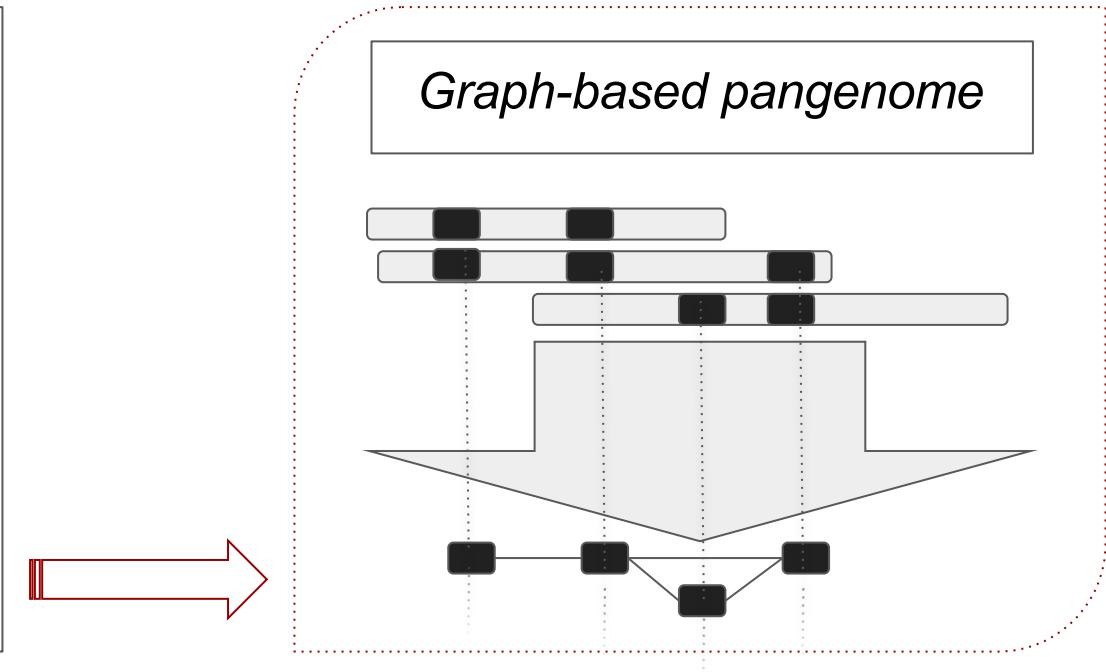
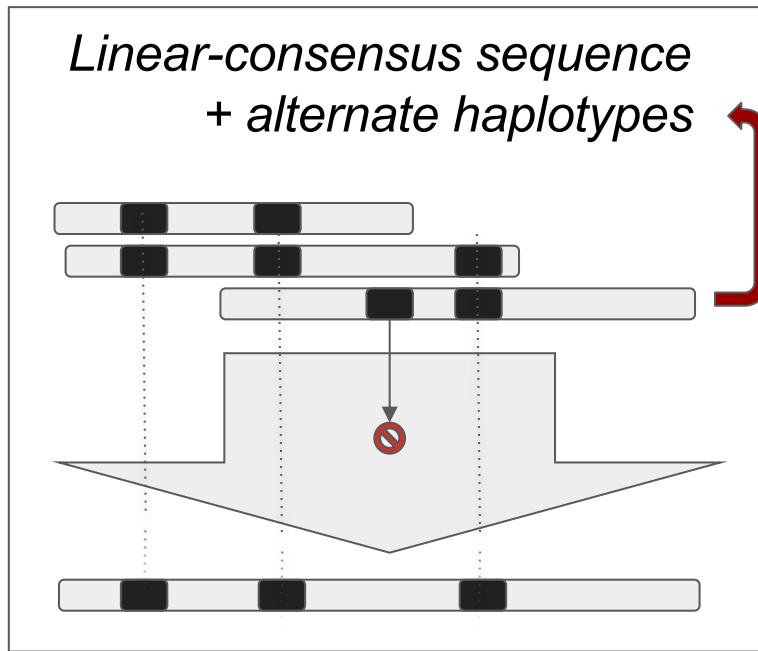
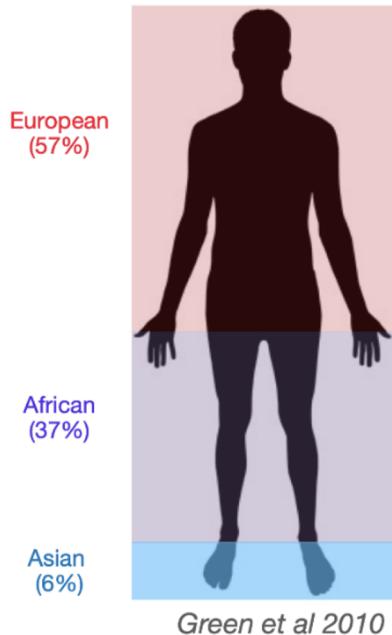
- live genomes are dynamic and actively changing in our cells
- data representations of genomes collapse complexity to 2D
- current genome representations are linear; misleading
- global *diversity must be integrated* into the reference genome



Data Structure: C- to an A+

What is ‘*The Human Genome Reference*’ ?

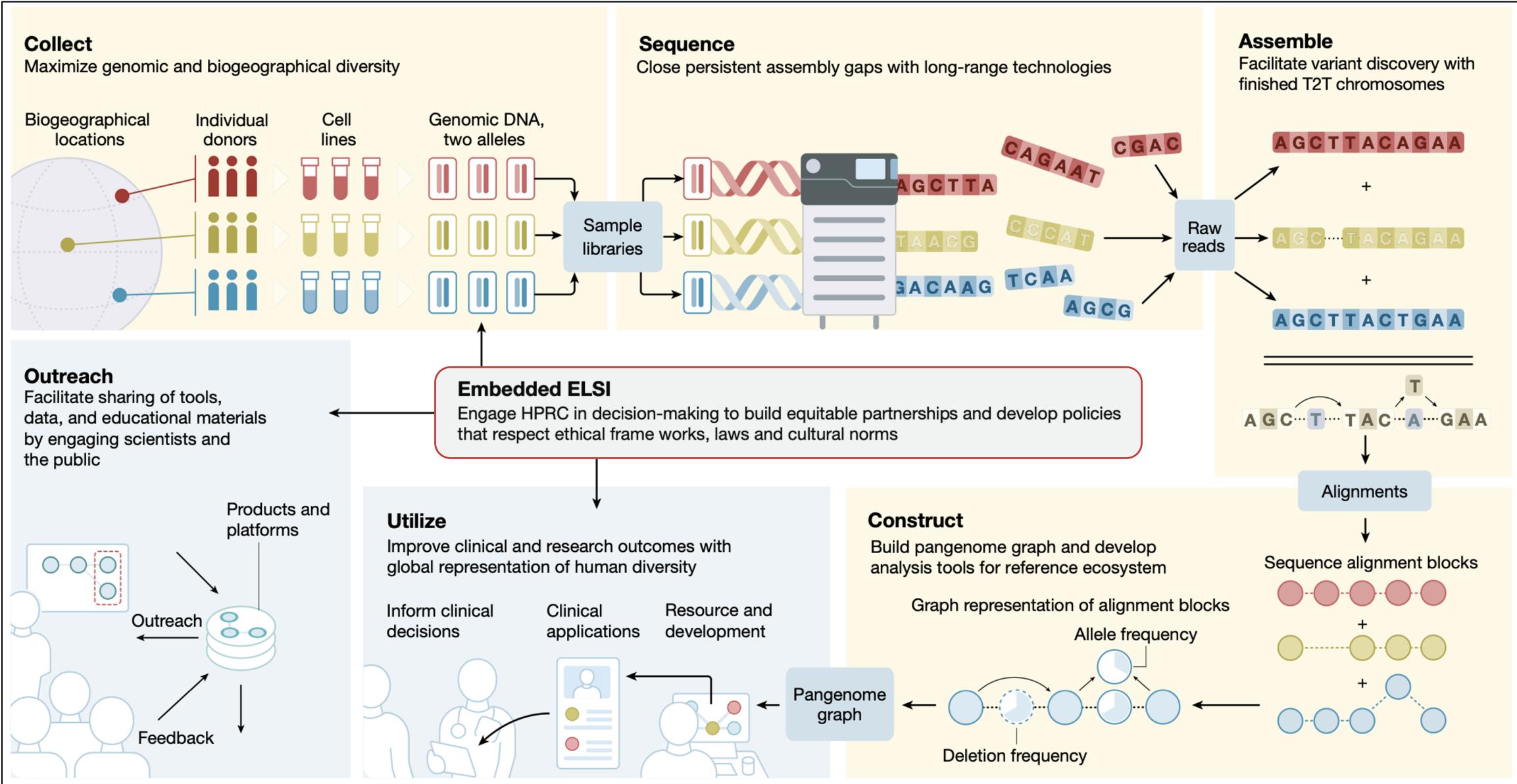
- live genomes are dynamic and actively changing in our cells
- data representations of genomes collapse complexity to 2D
- current genome representations are linear; misleading
- global *diversity must be integrated* into the reference genome



What questions do you have about a pangenome graph? (Please vote for questions that you would like answered or add a unique question.)

Top

Human Pangenome Reference Consortium (HPRC)

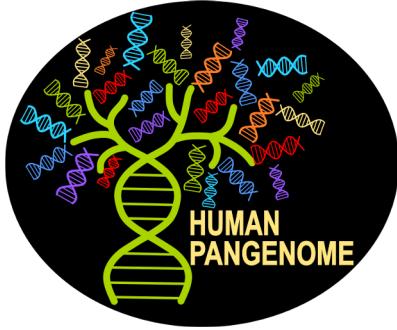


NIH-OSP Workshop (2021)

Integrating Bioethics and Biomedical Research

*Bioethics engagement over the lifecycle of biomedical research projects can enhance the conduct and **output of research**, ensuring **better science**, increased **inclusion** of participant, stakeholder, and societal values, and engendering greater **trust** in the biomedical research enterprise.*

*Bio ethicists, biomedical researchers, institutions, and funders all have an important role to play in building the **relationships and structures** that can enable bio ethicists and biomedical researchers to collaboratively advance science.*



Building a Trustworthy Resource for Genomics and Medicine

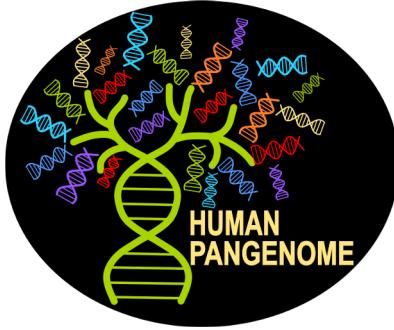
How might we represent global genomic diversity in a graph-based data structure while respecting ethical frameworks and data sovereignty?

Sampling and Representation

- selection criteria for diversity
- contributions to reference variation
- legacy sample reuse and immortal cell line creation

Engagement

- Communication with participants and the public
- clinical providers needs and constraints
- genomics researcher use interoperability
- cultural norms and expectations



Building a Trustworthy Resource for Genomics and Medicine

How might we represent global genomic diversity in a graph-based data structure while respecting ethical frameworks and data sovereignty?

Sampling and Representation

- selection criteria for diversity
- contributions to reference variation
- legacy sample reuse and immortal cell line creation

Engagement

- Communication with participants and the public
- clinical providers needs and constraints
- genomics researcher use interoperability
- cultural norms and expectations

Recruitment

- informed consent for broad further use
- criteria for inclusion and exclusion

Justice

- equitable benefit sharing
- Indigenous data sovereignty
- commercial entities and interests

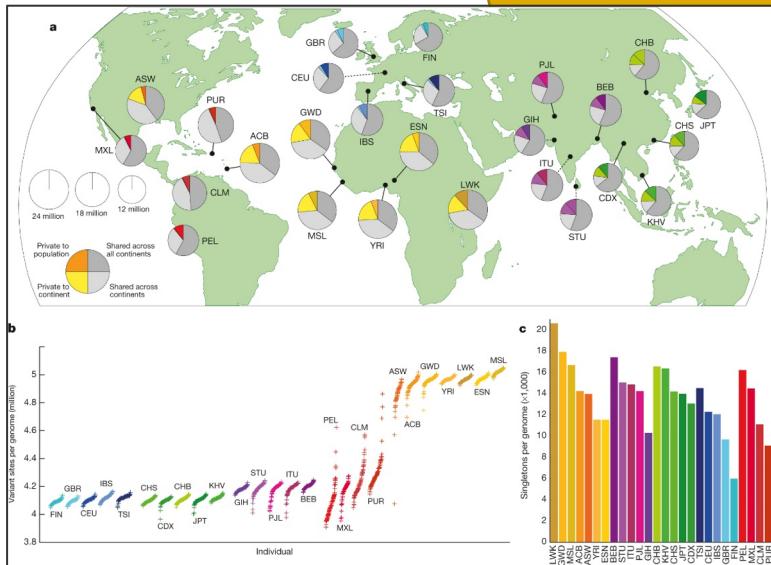
Collaborations

- international data sharing
- embargos and political implications

Technical – Logistical – Ethical – Social – Cultural – Legal – Regulatory – Commercial

Challenge: Diversity in Genome Reference Data

1000 Genomes
Auton et al. (2015)



Known

Genomic variation in existing reference resources

Known Unknown

Geographic regions
Global populations
Effective population size
Continental ancestry proportions
Self-identified group representation

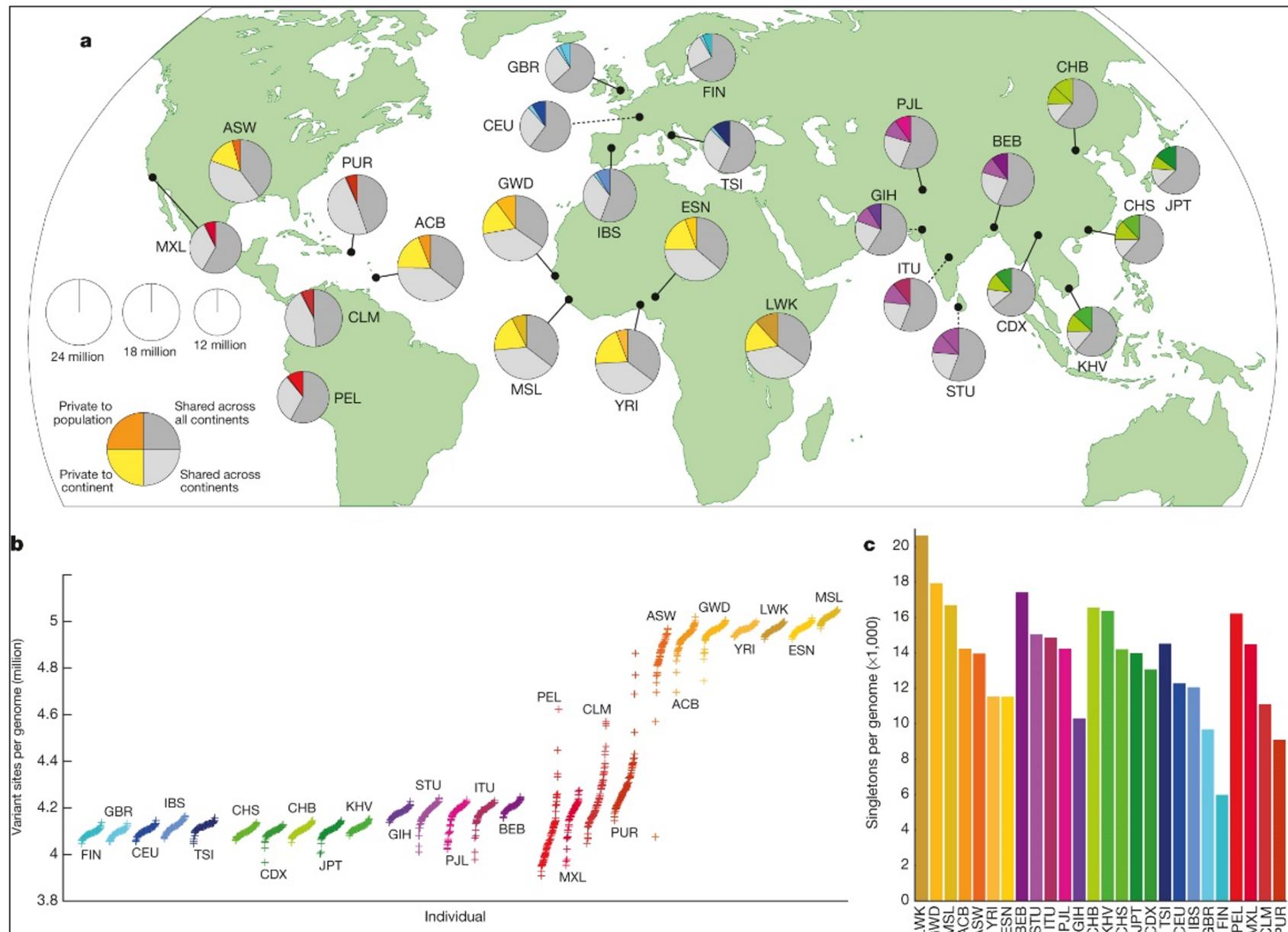
Unknown

Global variation that has yet to be sampled or identified

Human Genomic Diversity is Shared (Gray Slices)

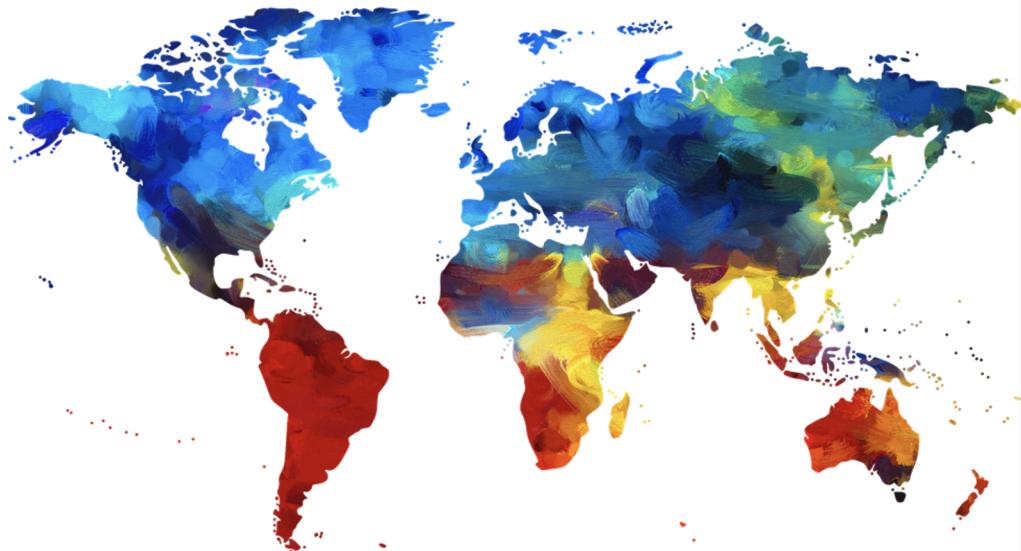
Population

African Caribbean in Barbados [ACB]
African Ancestry in SW USA [ASW]
Bengali in Bangladesh [BEB]
British From England and Scotland [GBR]
Chinese Dai in Xishuangbanna, China [CDX]
Colombian in Medellín, Colombia [CLM]
Esan in Nigeria [ESN]
Finnish in Finland [FIN]
Gambian in Western Division – Mandinka [GWD]
Gujarati Indians in Houston, Texas, USA [GIH]
Han Chinese in Beijing, China [CHB]
Han Chinese South [CHS]
Iberian Populations in Spain [IBS]
Indian Telugu in the U.K. [ITU]
Japanese in Tokyo, Japan [JPT]
Kinh in Ho Chi Minh City, Vietnam [KHV]
Luhya in Webuye, Kenya [LWK]
Mende in Sierra Leone [MSL]
Mexican Ancestry in Los Angeles CA USA [MXL]
Peruvian in Lima Peru [PEL]
Puerto Rican in Puerto Rico [PUR]
Punjabi in Lahore, Pakistan [PJL]
Sri Lankan Tamil in the UK [STU]
Toscani in Italia [TSI]
Yoruba in Ibadan, Nigeria [YRI]

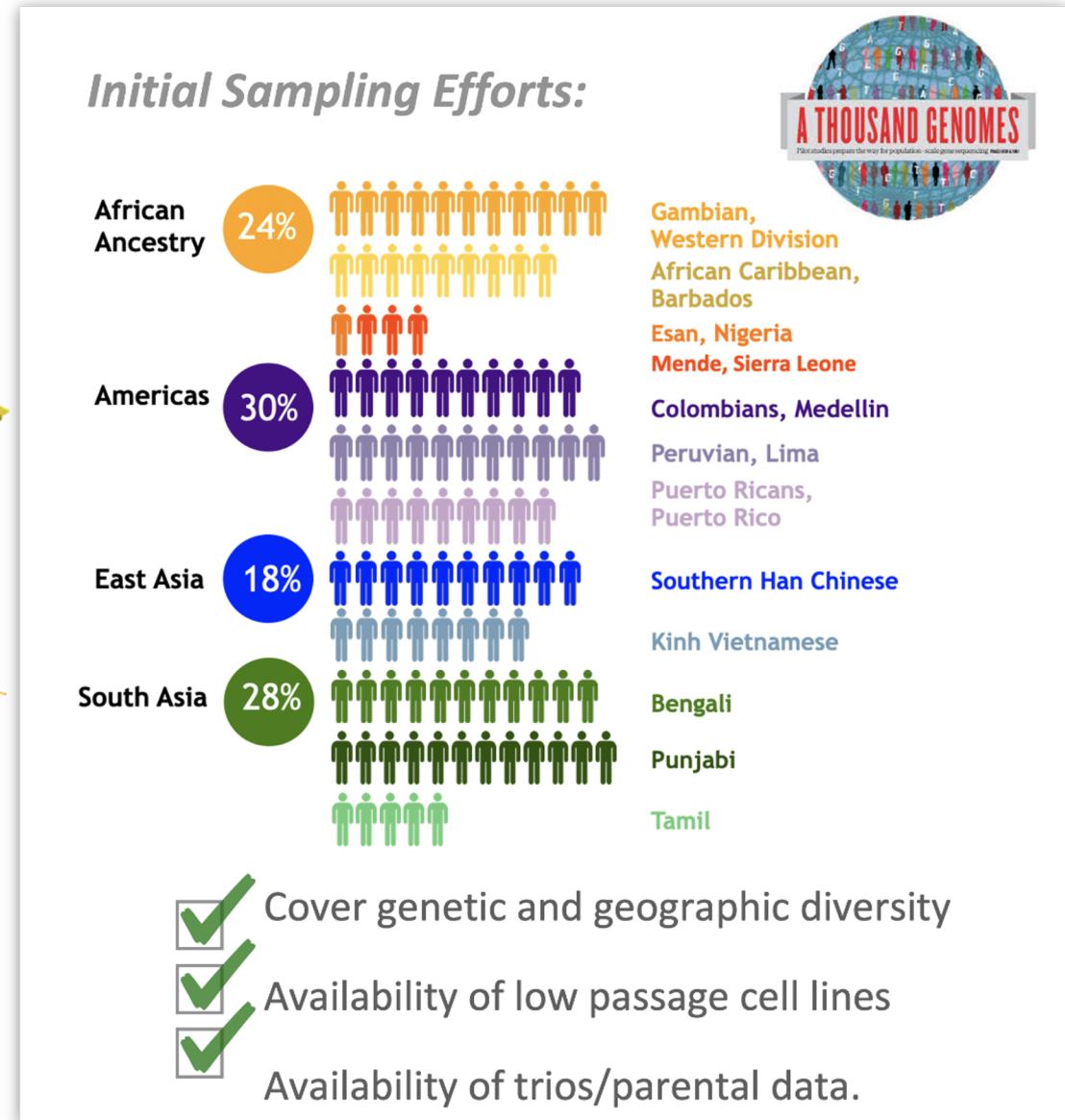


HPRC's Initial Sampling Efforts

Population Representation and Sampling

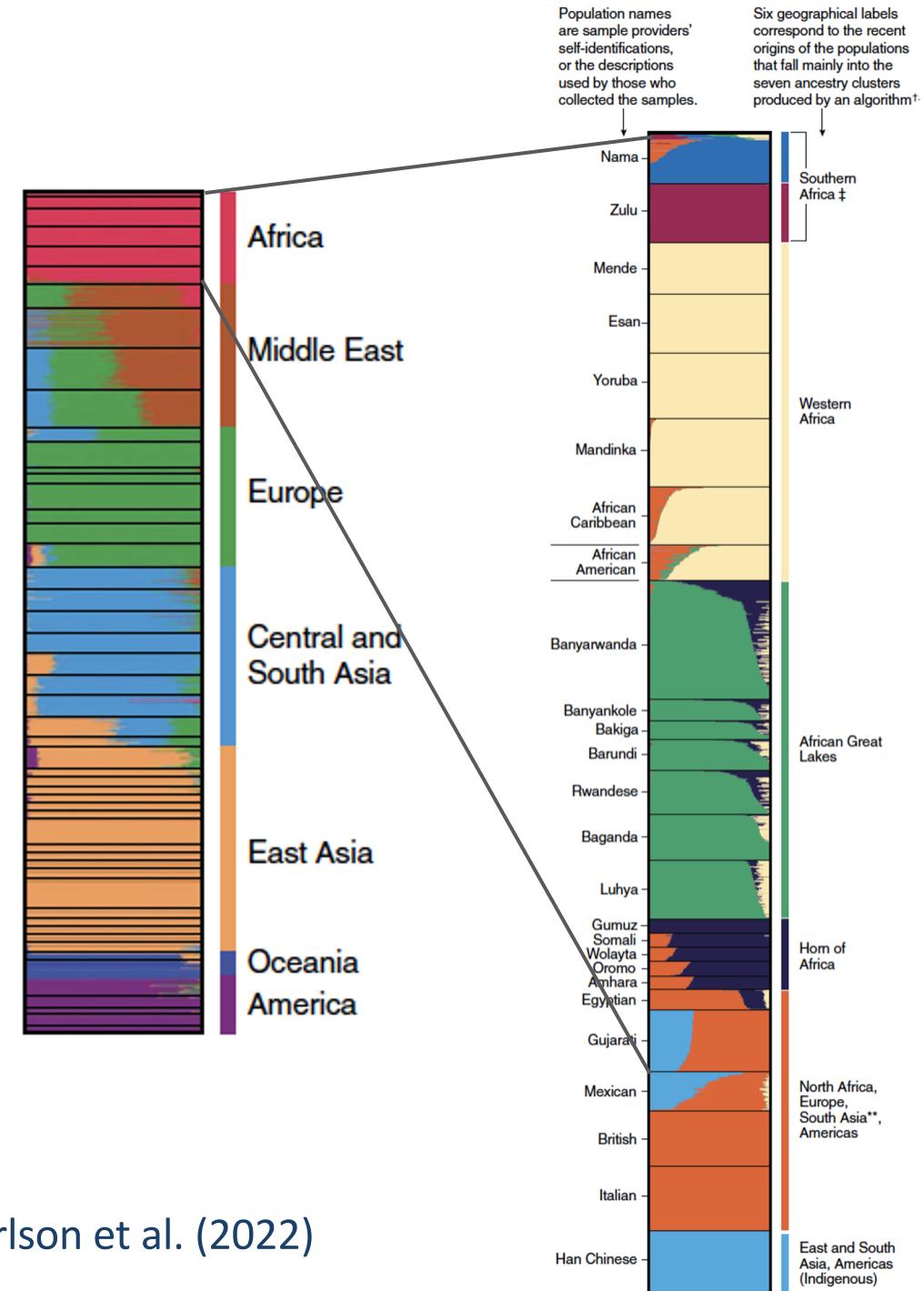


1000 Genomes data alone is insufficient to fully represent genomic diversity within the human population



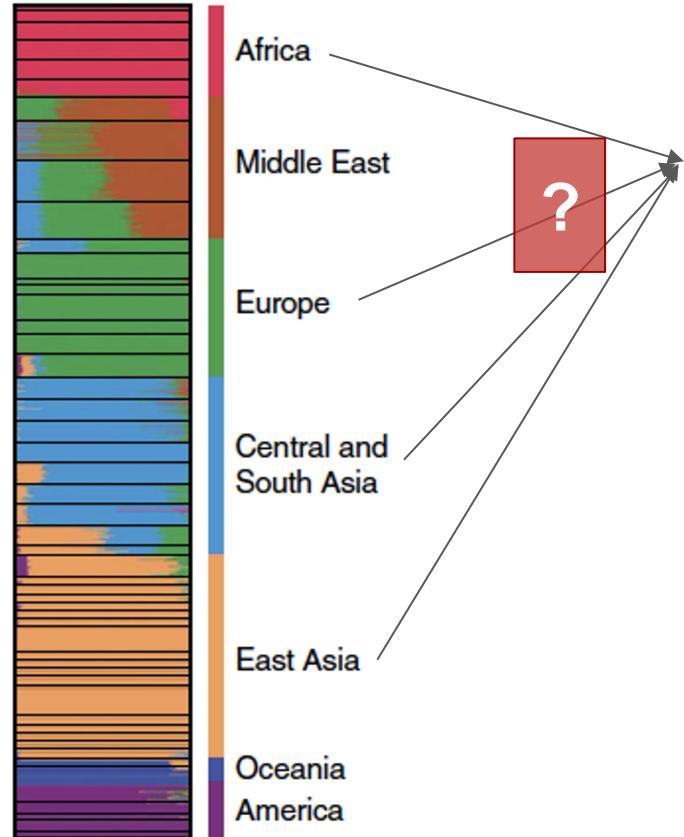
**Two randomly selected
'African' individuals
have greater genetic
variation between them
than two randomly
selected individuals
from different continents**

B Populations from Africa sampled: 13.5%

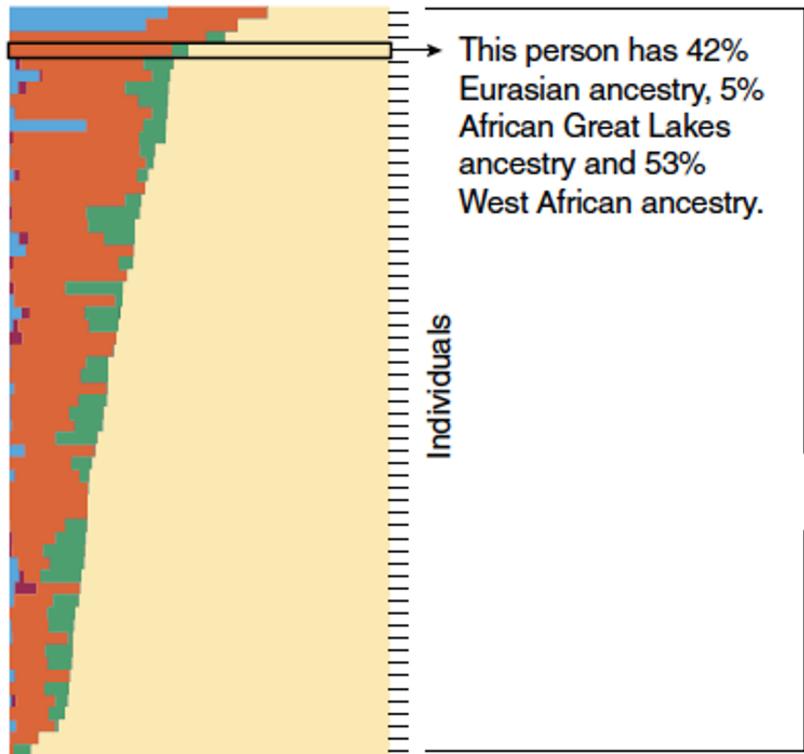


Why use a pangenome graph vs. population-specific references?

B Populations from Africa sampled: 13.5%

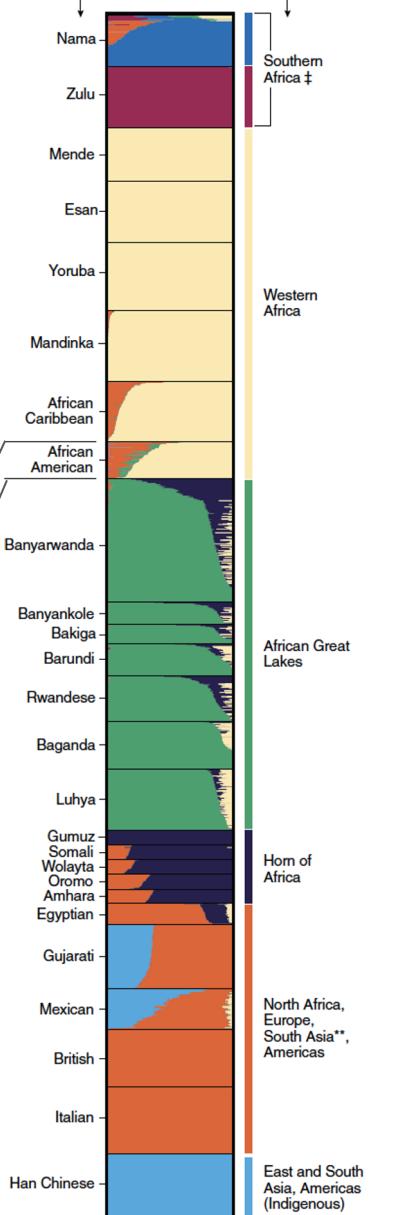


A Individual genomes
Each horizontal bar corresponds to the genome of a single person, in this case from a population identified or self-identified as African American.



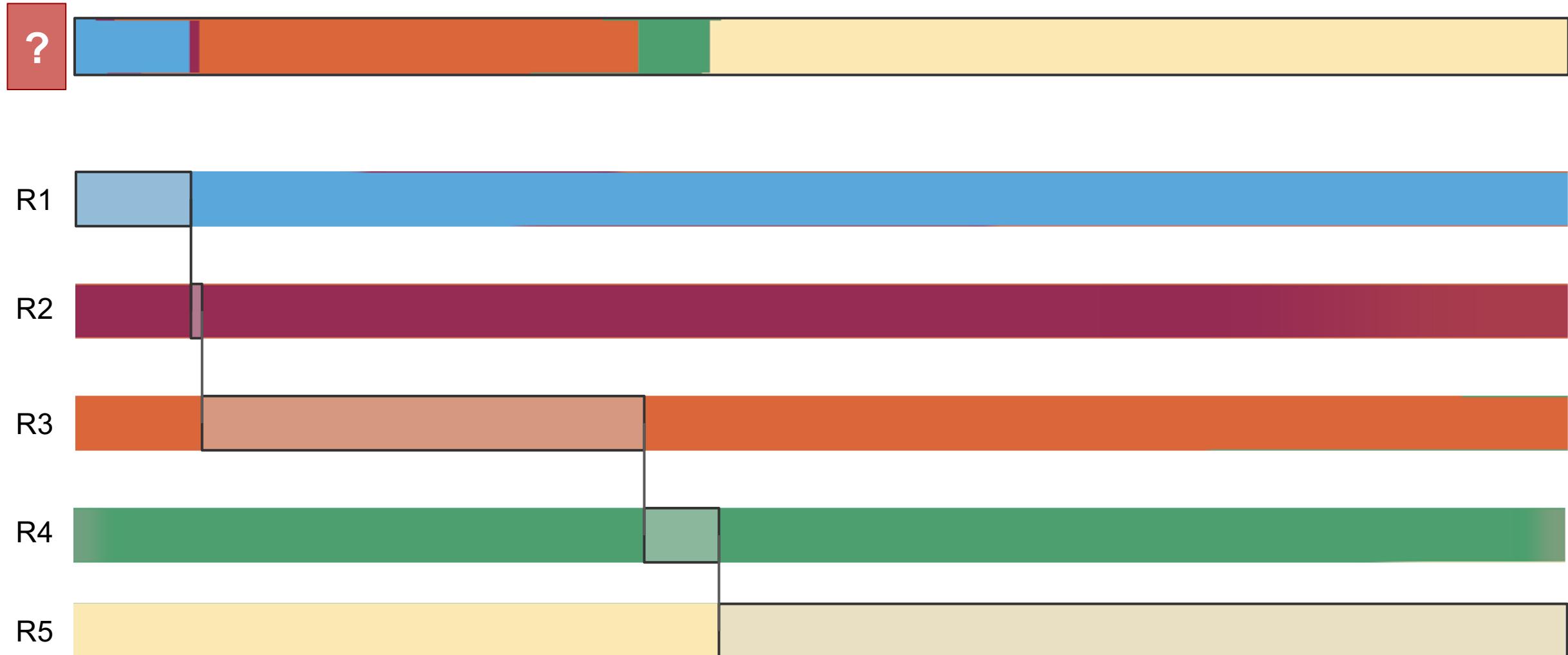
C Populations from Africa sampled: 85% (unpublished)
Adjusting the sampling and using an African-centric data set creates a more representative view.

Population names are sample providers' self-identifications, or the descriptions used by those who collected the samples.
Six geographical labels correspond to the recent origins of the populations that fall mainly into the seven ancestry clusters produced by an algorithm^t.

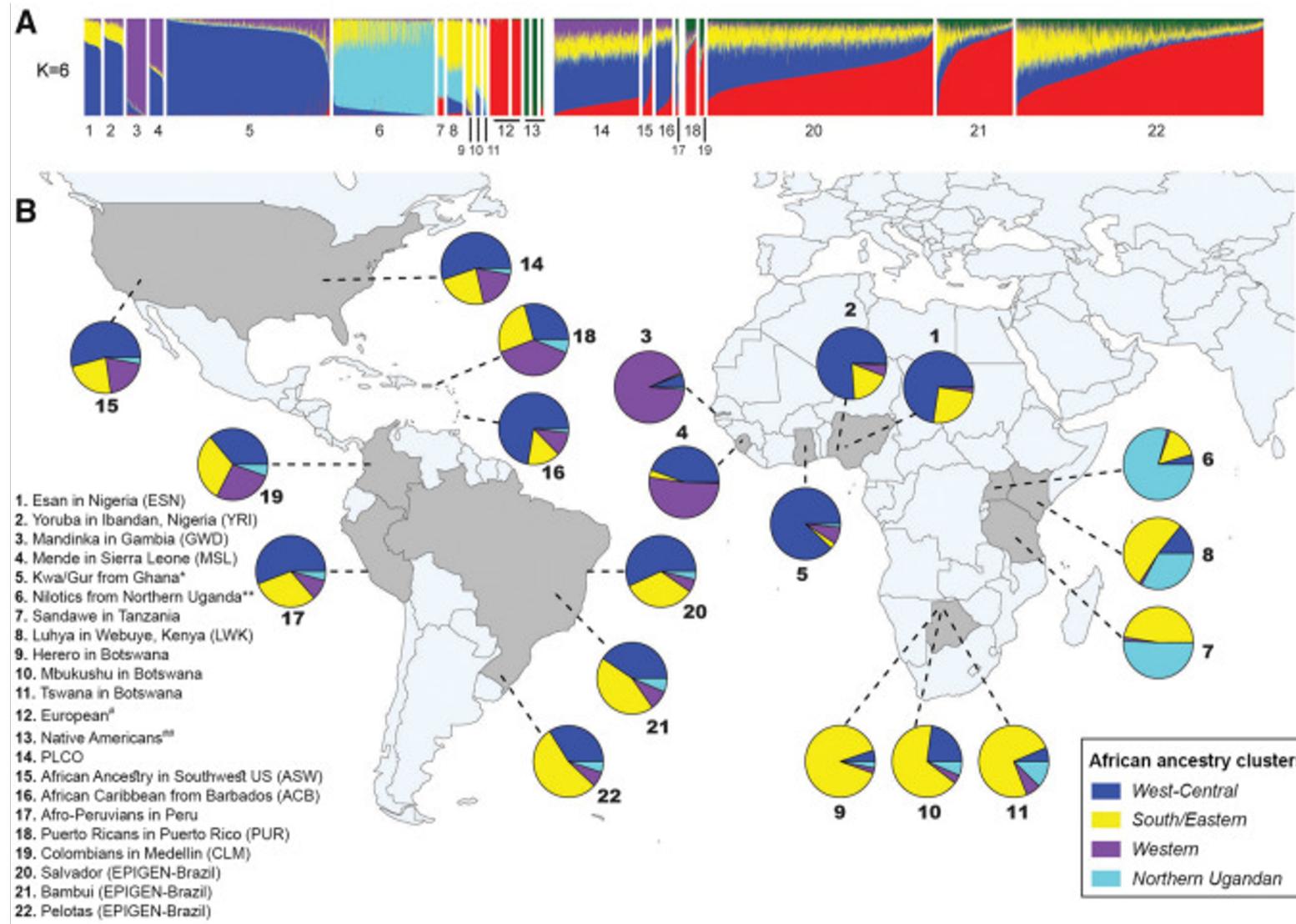


Why use a *pangenome graph* vs. population-specific references?

“African American” individual in the United States



Why use a pangenome graph vs. population-specific references?

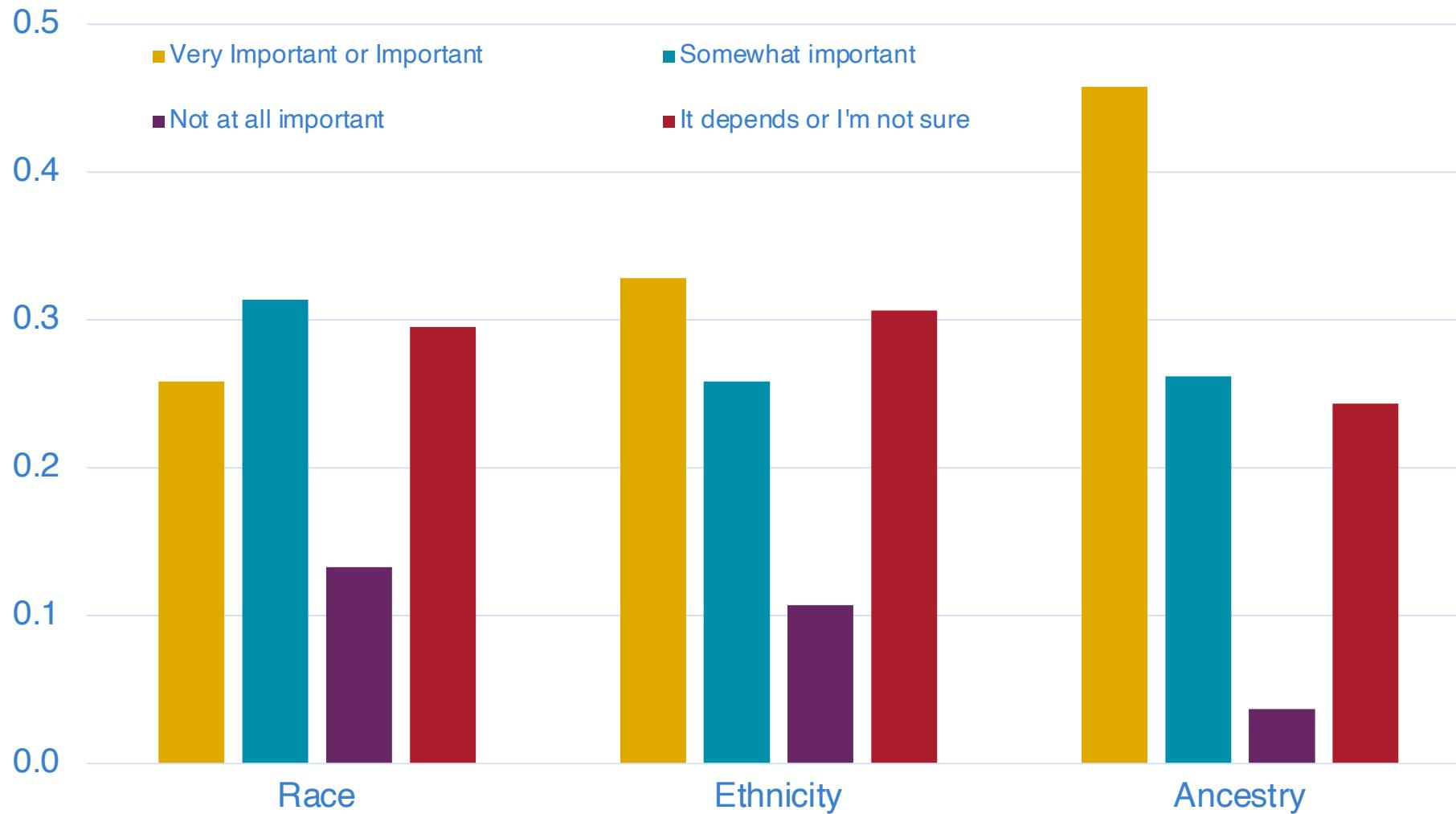


Gouveia et al. (2020)

- Genetic ancestries are continuous and relative to reference data (circular logic)
- Most human genomic variation is shared globally across all continents/populations
- Allele frequencies vary gradually across geography and time

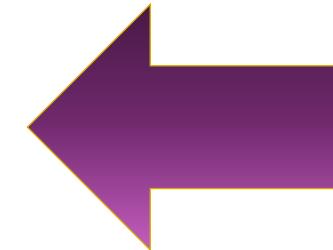
How important are race, ethnicity, and ancestry in the clinical interpretation of genetic variants? (N=271)

Popejoy et al., 2020



If information about a patient's race, ethnicity, or ancestry is used in any aspect of your clinical work, from where are these data obtained? (N=216)

Source of Data	No. Respondents	% Respondents
Race and/or ethnicity information is reported directly to me by the patient	204	94.4%
Data are obtained or provided from the patient's medical record	87	40.3%
Race and/or ethnicity are recorded by another care provider, possibly without asking the patient directly	39	18.1%
Genetic ancestry is inferred from the patient's DNA sample	10	4.6%
I do not use any of these measures in my work	5	2.3%
Other (Please specify)	6	2.8%
I'm not sure	4	1.9%



Popejoy et al., 2020

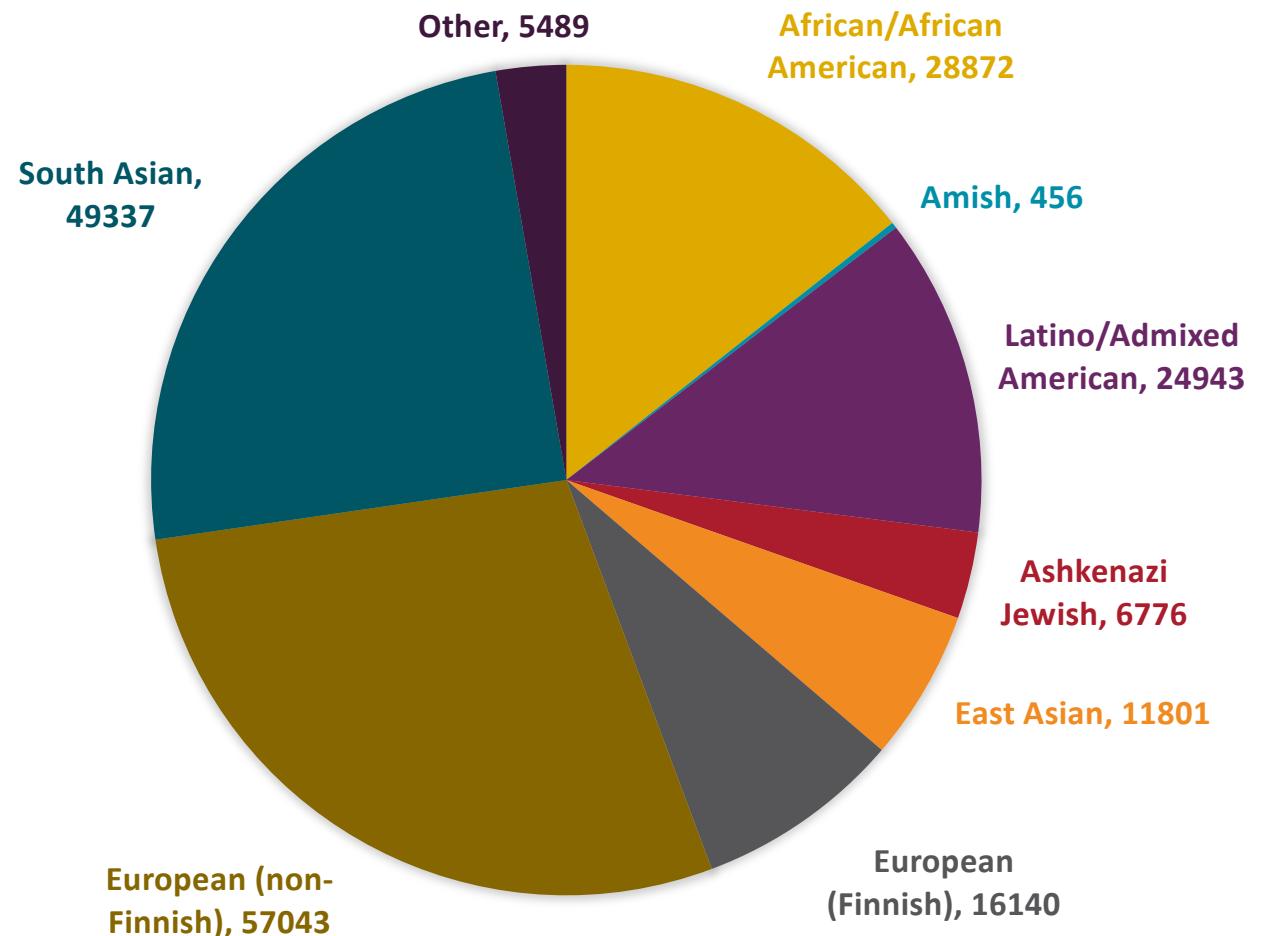


ClinGen
Clinical Genome Resource

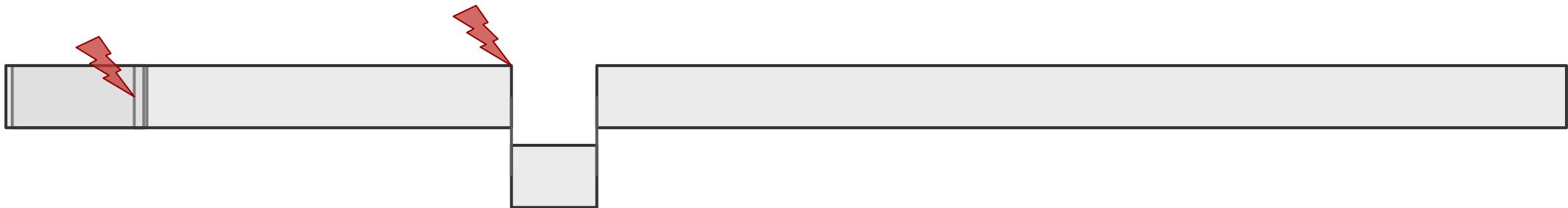
Populations represented in gnomAD allele frequency data

- African/African American
- Amish
- Ashkenazi Jewish
- East Asian
 - Japanese
 - Korean
 - Other East Asian
- European (Finnish)
- European (non-Finnish)
 - Southern European
 - North-western European
 - Other non-Finnish European
 - Bulgarian
 - Swedish
 - Estonian
- Latino/Admixed American
- Other
- South Asian

<https://gnomad.broadinstitute.org/>



Using The Human Reference [pan]Genome in Clinical Genetics (Discussion)

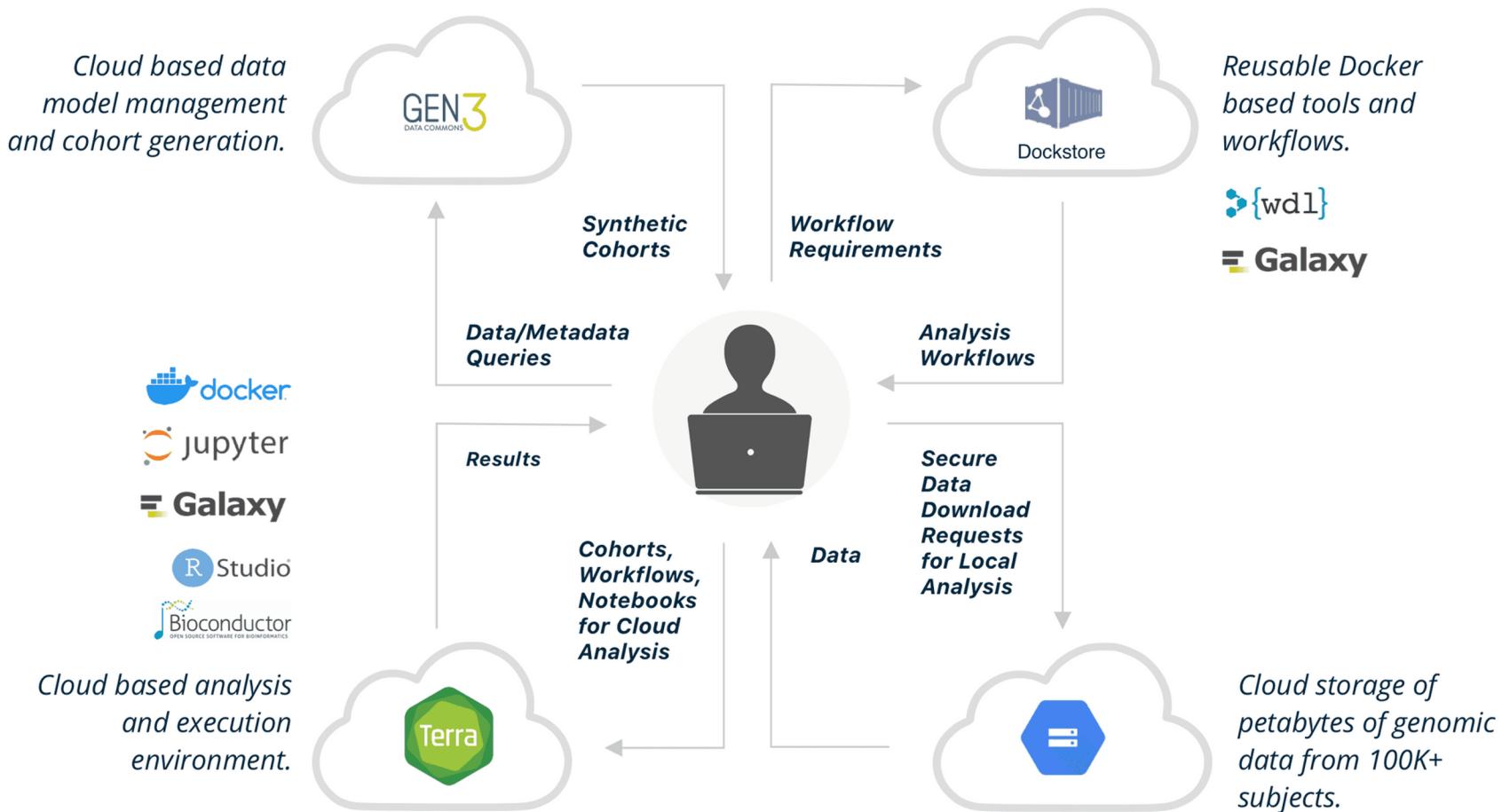


**What is exciting or intriguing to you about the development
of a graph-based pangenome reference, specifically as it
relates to your work in clinical genetics?**

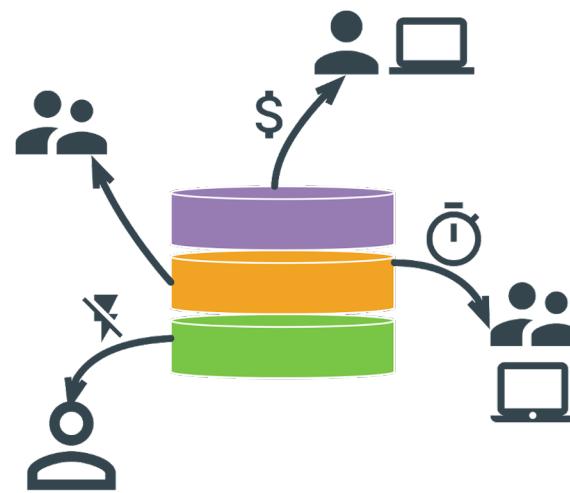
What questions or concerns do you have about the development of a human pangenome reference, and its implications for clinical genetics?

Overview of AnVIL and HPRC Resources

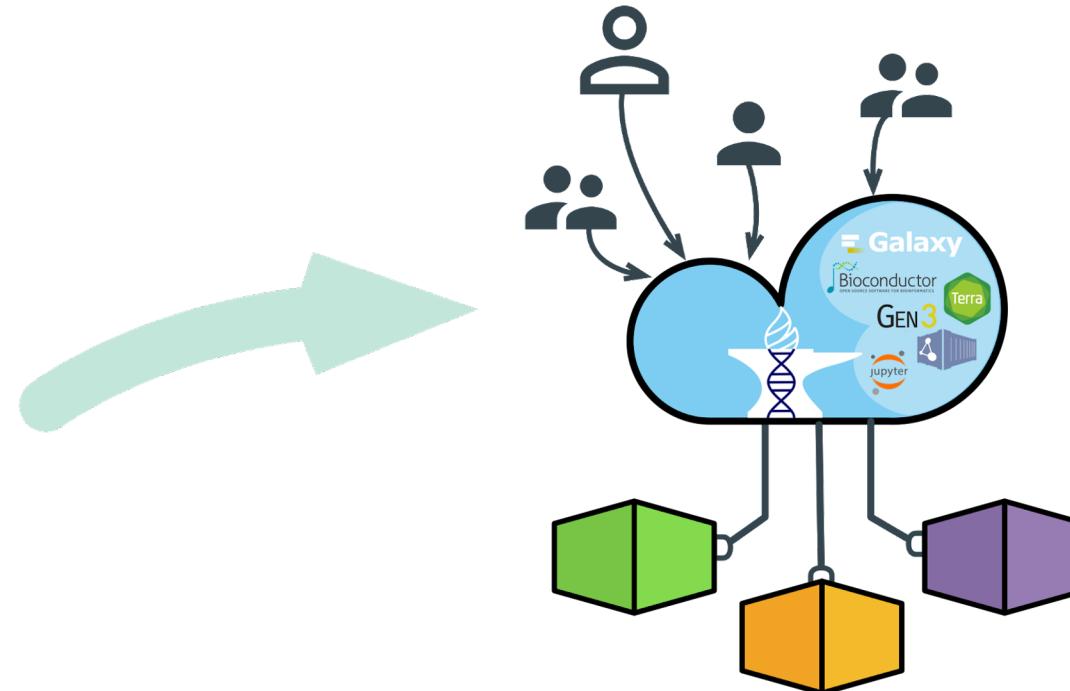
AnVIL: Ecosystem Overview



AnVIL: Inverting The Model Of Genomic Data Sharing



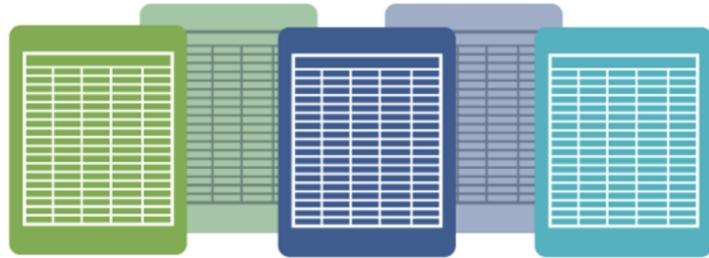
Traditional: Bring data to the researcher



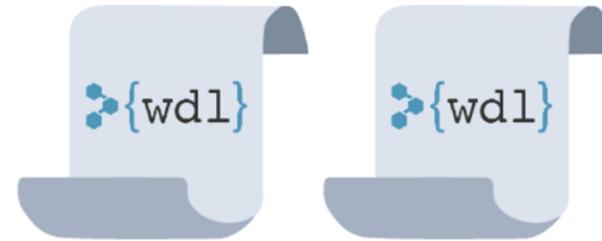
Goal: Bring researcher to the data

Terra: Cloud Data Storage & Processing

Data Tables



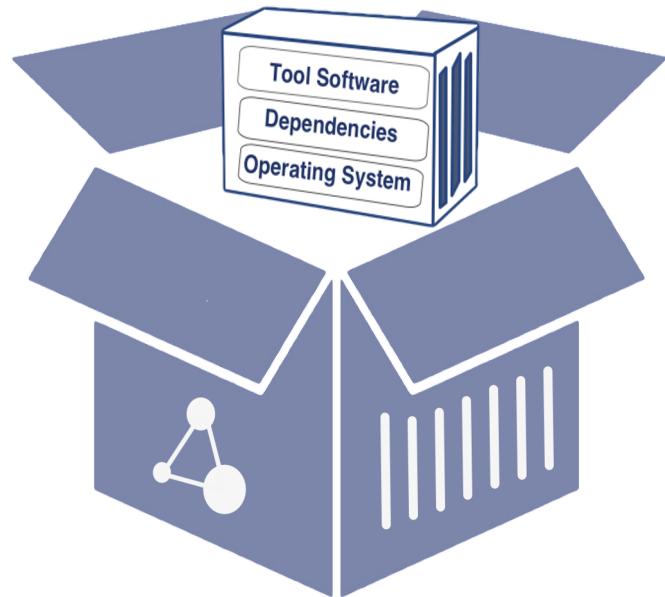
Workflows



Analyses



Dockstore: AnVIL's Repository For Workflows



Launch with

dockstore.org

AnVIL Hosts HPRC Data

- Community can directly access and analyze data in the HPRC data workspace
- When you work in AnVIL or GCP, you don't pay to use the data (no need to copy or host individually)

The screenshot shows the AnVIL Workspaces interface. At the top, there's a header with the Terra logo, a 'POWERED BY' badge, and a 'BETA' badge next to 'WORKSPACES'. The URL 'Workspaces > anvil-datastorage/AnVIL_HPRC > Data' is visible. On the right, there are icons for 'Cloud Environment' (Stopped <\$0.01/hr) and a gear. Below the header, there are tabs for 'DASHBOARD', 'DATA' (which is selected), 'NOTEBOOKS', 'WORKFLOWS', and 'JOB HISTORY'. A search bar and a refresh button are also present.

The main area is a table titled 'TABLES'. It has a sidebar on the left with categories: 'assembly_sample' (47), 'minigraph' (2), 'minigraph_cactus' (6), 'participant' (47), 'pggb' (1), 'sample' (57), 'REFERENCE DATA' (with a plus sign), and 'OTHER DATA'. The table itself has columns: a checkbox, 'assembl...', 'mat_chm13_aln_bai', 'mat_chm13_aln_bam', 'mat_fasta', 'mat_grch38_aln_bai', and a settings icon. There are download and copy buttons at the top of the table. The table lists several rows corresponding to the categories in the sidebar, each with a unique identifier and multiple links to different data files.

	assembl...	mat_chm13_aln_bai	mat_chm13_aln_bam	mat_fasta	mat_grch38_aln_bai	
assembly_sample	(47)					
minigraph	(2)	HG002	HG002.maternal.CHM13Y_EBV.ba...	HG002.maternal.CHM13Y_EBV.bam	HG002.maternal.f1_assembly_v2.g...	HG002.maternal.GRCh38_no_alt.ba...
minigraph_cactus	(6)	HG00438	HG00438.maternal.CHM13Y_EBV.ba...	HG00438.maternal.CHM13Y_EBV.b...	HG00438.maternal.f1_assembly_v2...	HG00438.maternal.GRCh38_no_alt...
participant	(47)	HG005	HG005.maternal.CHM13Y_EBV.ba...	HG005.maternal.CHM13Y_EBV.bam	HG005.maternal.f1_assembly_v2.g...	HG005.maternal.GRCh38_no_alt.ba...
pggb	(1)	HG00621	HG00621.maternal.CHM13Y_EBV.ba...	HG00621.maternal.CHM13Y_EBV.b...	HG00621.maternal.f1_assembly_v2...	HG00621.maternal.GRCh38_no_alt...
sample	(57)	HG00673	HG00673.maternal.CHM13Y_EBV.ba...	HG00673.maternal.CHM13Y_EBV.b...	HG00673.maternal.f1_assembly_v2...	HG00673.maternal.GRCh38_no_alt...
REFERENCE DATA	(+)	HG00733	HG00733.maternal.CHM13Y_EBV.ba...	HG00733.maternal.CHM13Y_EBV.b...	HG00733.maternal.f1_assembly_v2...	HG00733.maternal.GRCh38_no_alt...
OTHER DATA		HG00735	HG00735.maternal.CHM13Y_EBV.ba...	HG00735.maternal.CHM13Y_EBV.b...	HG00735.maternal.f1_assembly_v2...	HG00735.maternal.GRCh38_no_alt...

https://anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_HPRC

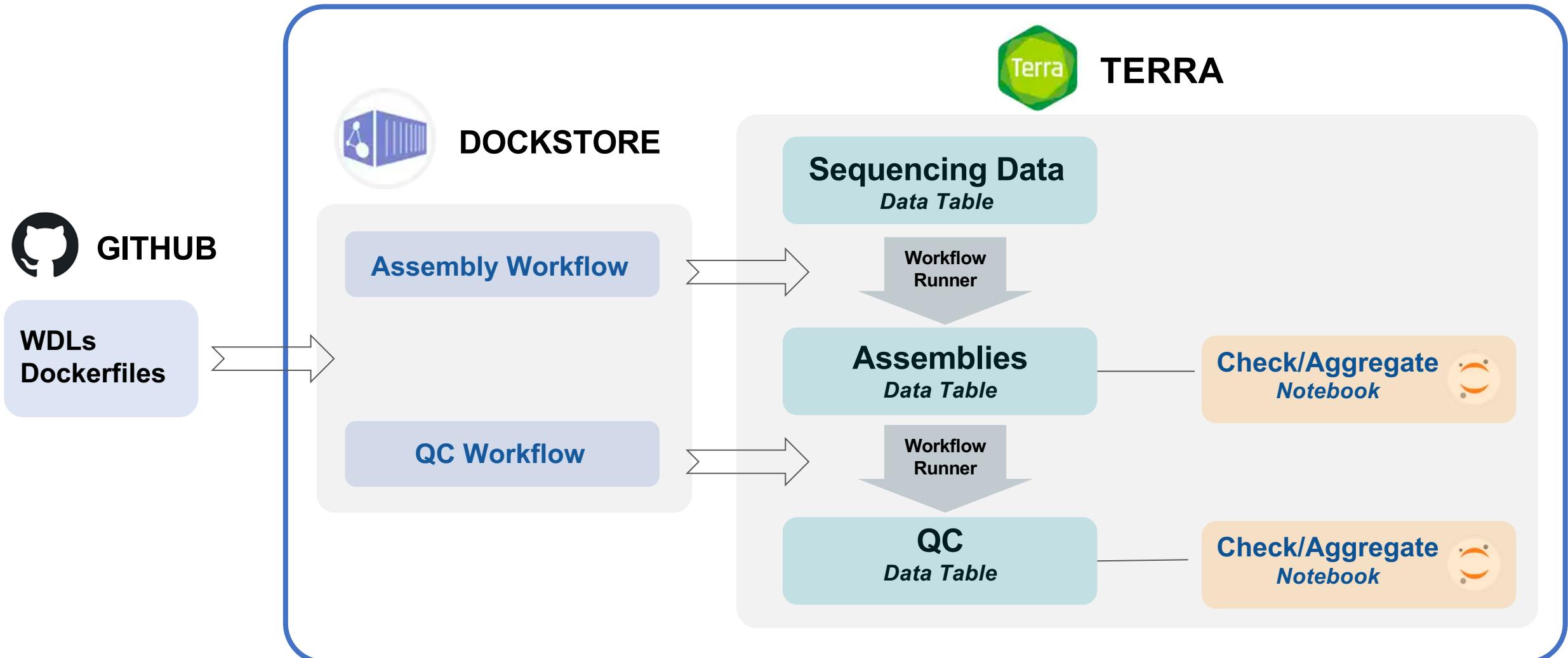
HPRC Shares Workflows In Dockstore

- Consortia share production workflows with the community using GitHub and Dockstore
- Leverage a container-based workflow language (WDL)
- Easily launch from Dockstore to Terra

The screenshot shows the Dockstore organization page for the Human Pangenome Reference Consortium. The top navigation bar includes links for Dockstore, Search, Organizations, About, Docs, Forum, Login, and Register. The main header for the organization is "Human Pangenome Reference Consortium" with a subtitle "Assembly and analysis workflows for generating Human Pangenome References". Below the header, there are sections for Collections (3), Members (3), and Updates (10). A large callout box highlights three specific workflow categories: "Hifiasm Assembly Workflows" (1 Workflow), "Assembly QC" (10 Workflows), and "T2T CHM13 Short-Read Small-Variant Calling". Each category has a "View" button. To the right, there are sections for "About the Organization" (link to https://humanpangenome.org/), "About" (describing the consortium's role in developing the human pangenome reference assembly), and "Open-Access Data" (mentioning open-access data availability on AnVIL and AWS).

<https://dockstore.org/organizations/HumanPangenome>

HPRC Created Year 1 Assemblies In AnVIL



HPRC's Data Flow



illumina®

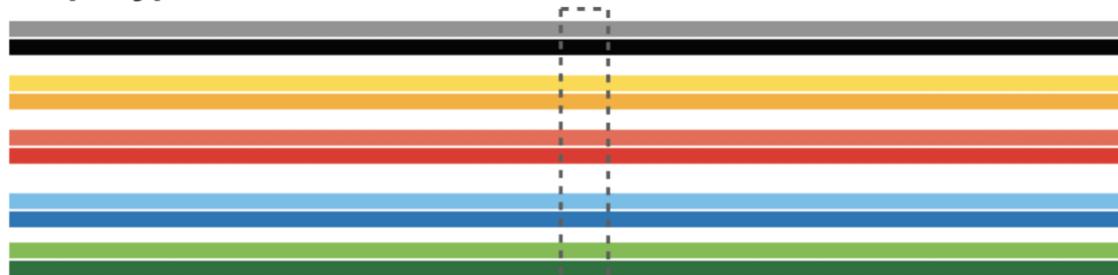
Oxford
NANOPORE
Technologies

PACBIO®

Release

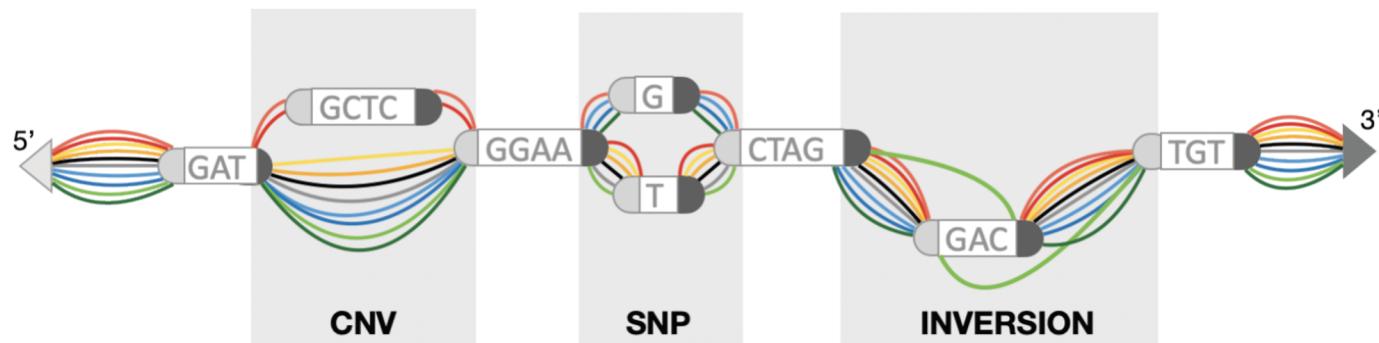
Data/
Workflows

Haplotype-Phased Assemblies:



Release

Data/
Workflows



Release

Data/
Workflows

HPRC Sequencing Data

The AnVIL

Public workspace (requires AnVIL account)



GitHub

Repository with data indexes for GCP and AWS files



Sequence Read Archive

BioProject with INSDC copies of raw sequencing data



HPRC Assemblies



The AnVIL

Public workspace (requires AnVIL account)



GitHub

Assembly repository with indexes for AWS & GCP files



Genbank

BioProject for HPRC assemblies



UCSC Genome Browser

Browser hub with year 1 assemblies & annotations



ENSEMBL

Project page containing assemblies & gene annotations

HPRC Pangenomes

The AnVIL

Public workspace (requires AnVIL account)



GitHub

Repository with data indexes for GCP and AWS files



European Nucleotide Archive

BioProject with pangenomes



First Data Release

sequencing data & QC metrics
for the first 30 samples*



30 HiFi
(30x, 17-20kb)



30 ONT
Ultra-Long
(~6x 100 kb+)



30 Hi-C
(Omni-C, ~60X)



30 Bionano Maps
(N50>250kb,
~100X coverage)



60 Parental
Datasets
(30x, 150 bp PE)



10 Strand-Seq
single-cell
libraries

Open Data Sharing and
Cloud-based Data Management



S3://human-pangenomics



AnVIL_HPRC



Dockstore
Human
Pangenome



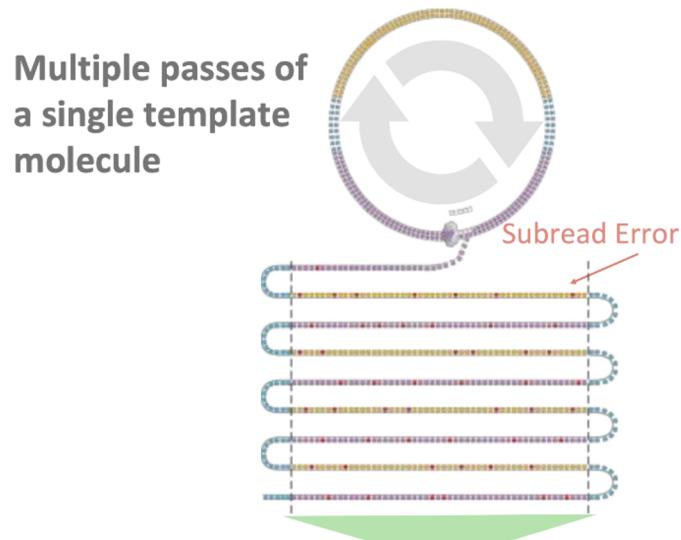
GitHub
Human-
Pangenomics



Data Production Has An Emphasis On Long Reads

Advances in Long-Read Sequencing

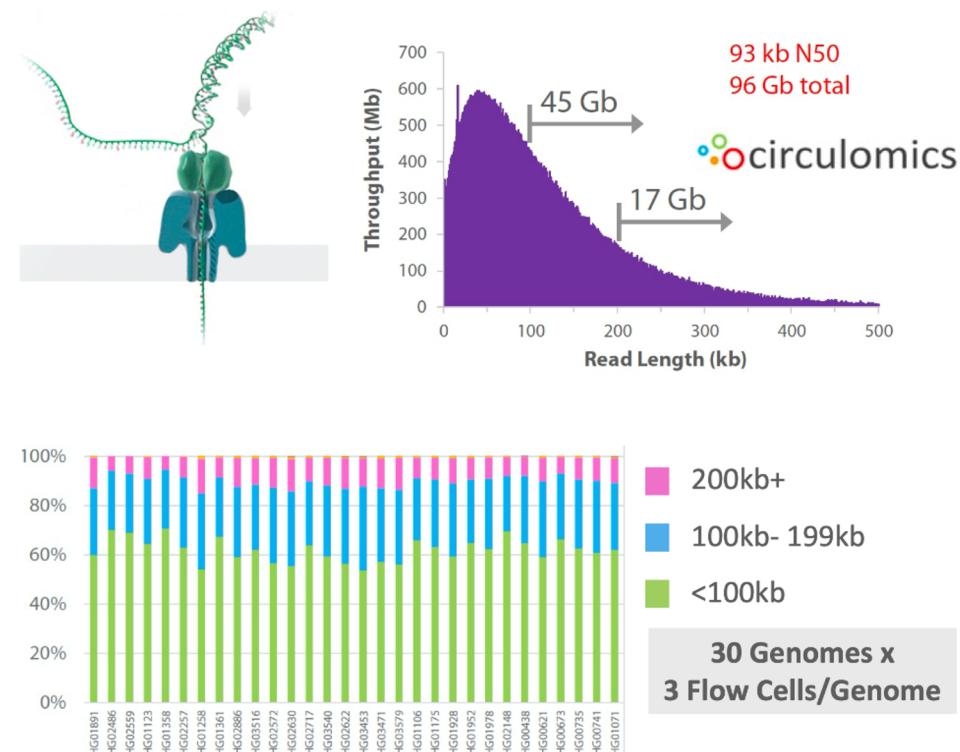
PacBio High Fidelity (HiFi) Data



99.9% Consensus Read Accuracy
(35-40x Coverage >Q20 HiFi Reads; 18-20kb)

Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Wenger et al. *Nature Biotechnology* (2019)

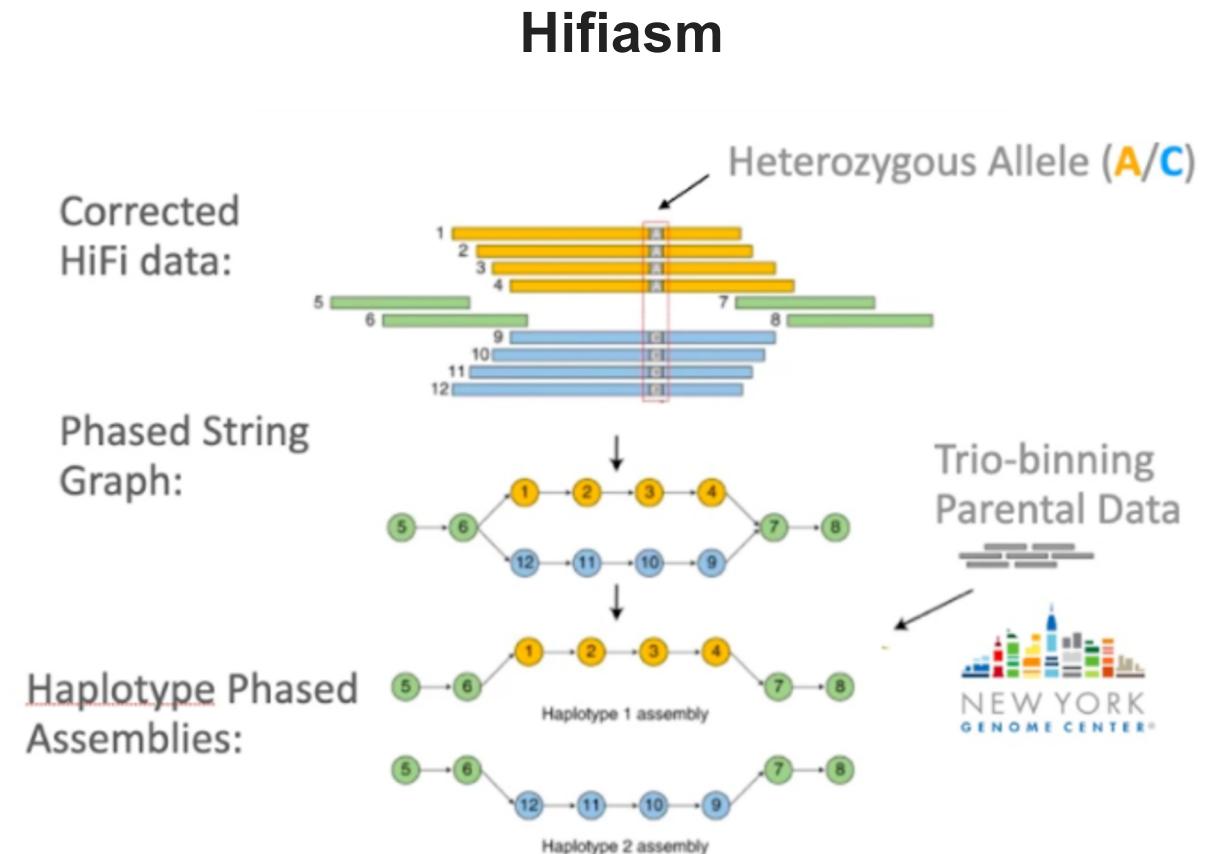
ONT Ultra-Long (UL) Data



Nanopore sequencing and assembly of a human genome with ultra-long reads.
Jain et al. *Nature Biotechnology* (2018)

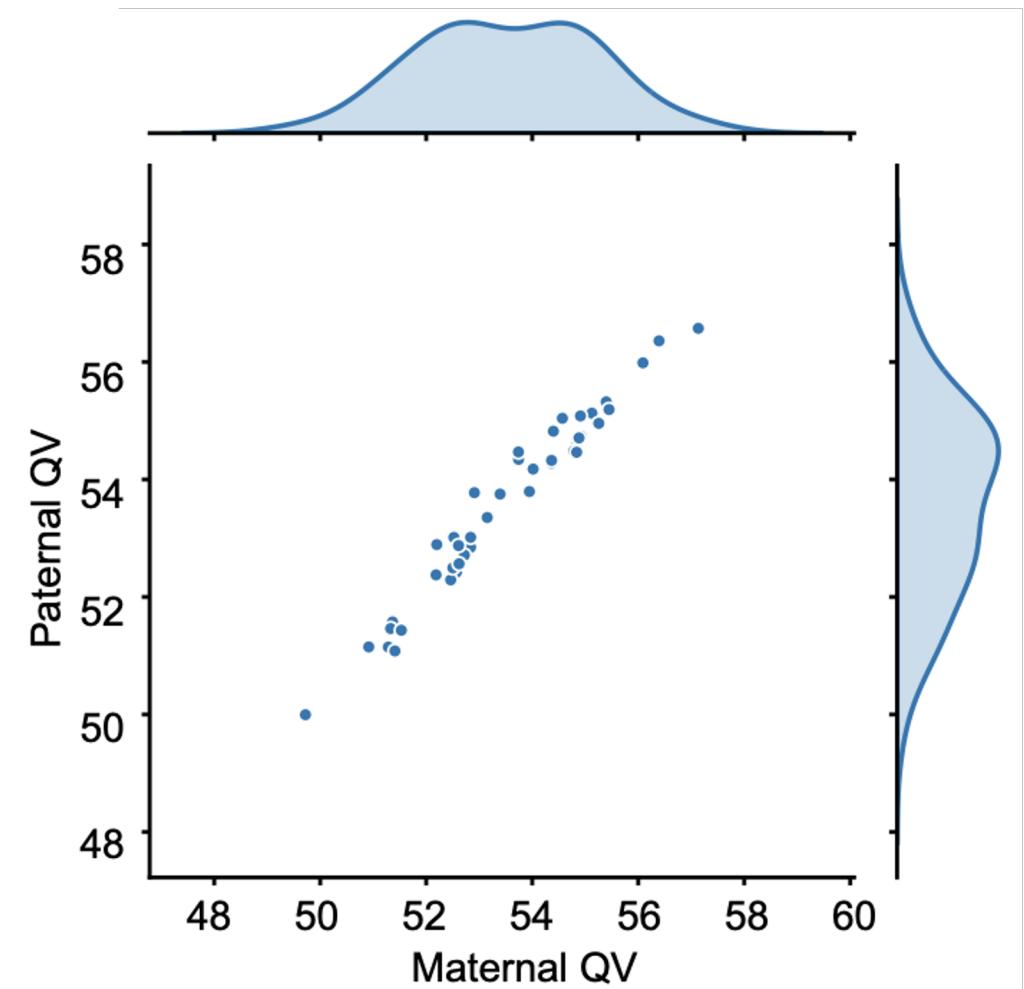
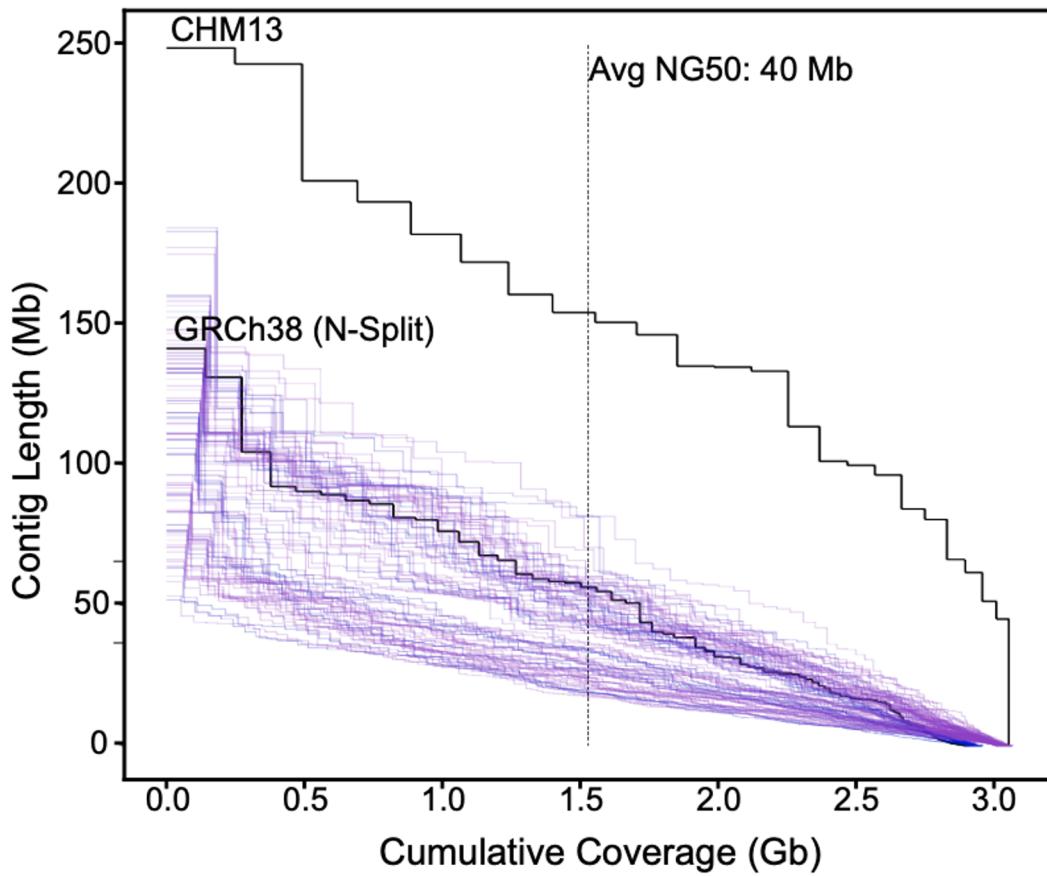
Assembly Strategy

- Needed a high-quality, automated workflow to create contig-level diploid assemblies
- Held a bake-off
 - Jarvis, et al. *Nature* 2022
- Developing methods for automated QC, scaffolding, T2T, & non-trio assembly



Cheng, Haoyu, et al. *Nature Methods* (2021)

High-Quality Diploid Assemblies



How To Use Pangenomes To Improve (Clinical) Variant Calling

Pangenome Creation From Assemblies

TCGGACTTGCACCAGCAATGGAGATGCTCAGCGTTAGATACGCCGAATTAATGGACAA **haplotype 1**

TCGGACTTGCACCAGCAATGGAGATGCTCAGTGTAGATACGCCGAATTAATGGACAA **haplotype 2**

TCGGACTTGCACCAGCAATGGA-----TAATGGACAA **haplotype 3**

Pangenome Creation From Assemblies

TCGGACTTGCACCAGCAATGGAGATGCTCAGCGTTAGATACGCCGAATTAATGGACAA

TCGGACTTGCACCAGCAATGGAGATGCTCAGTGTAGATACGCCGAATTAATGGACAA

TCGGACTTGCACCAGCAATGGA-----TAATGGACAA

Pangenome Creation From Assemblies

GATGCTCAGCGTTAGATAACGCCGAAT
TCGGACTTGCGACCAGCAATGGATTAATGGACAA

GATGCTCAGTGTAGATAACGCCGAAT
TCGGACTTGCGACCAGCAATGGATTAATGGACAA

TCGGACTTGCGACCAGCAATGG-----TAATGGACAA

Pangenome Creation From Assemblies

GATGCTCAGCGTTAGATA CGCCGAAT
GATGCTCAGTGTAGATA CGCCGAAT
TCGGACTTGCGACCAGCAATGGA TAATGGACAA

Pangenome Creation From Assemblies

TCGGACTTGCGACCAGCAATGGA

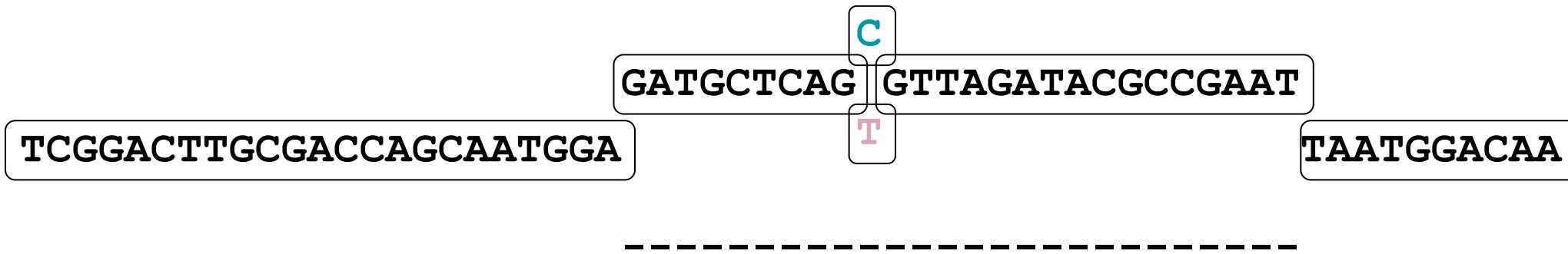
GATGCTCAG C GTTAGATA CGCCGAAT
GATGCTCAG T GTTAGATA CGCCGAAT

TAATGGACAA

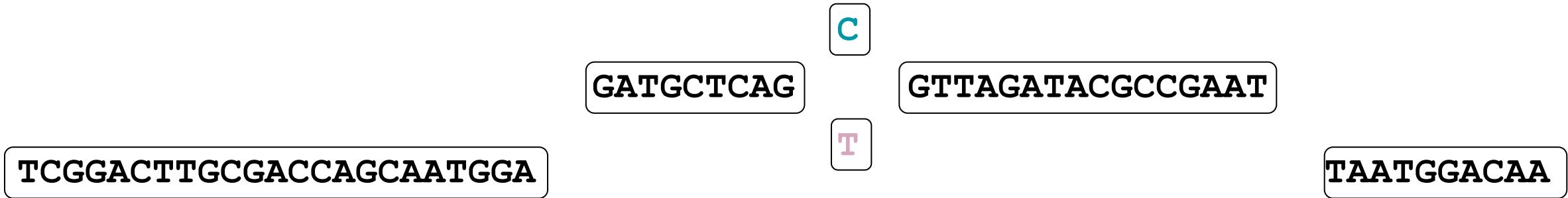
Pangenome Creation From Assemblies

GATGCTCAG GTTAGATAACGCCGAAT
TCGGACTTGCGACCAGCAATGGAA

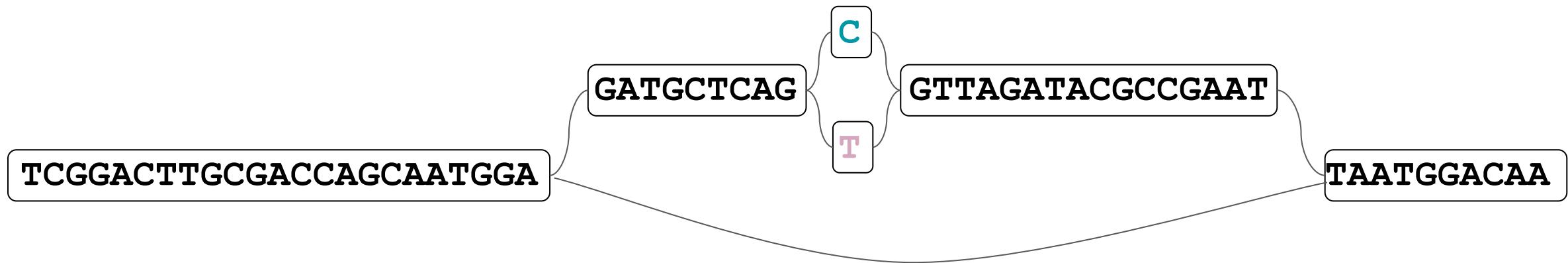
Pangenome Creation From Assemblies



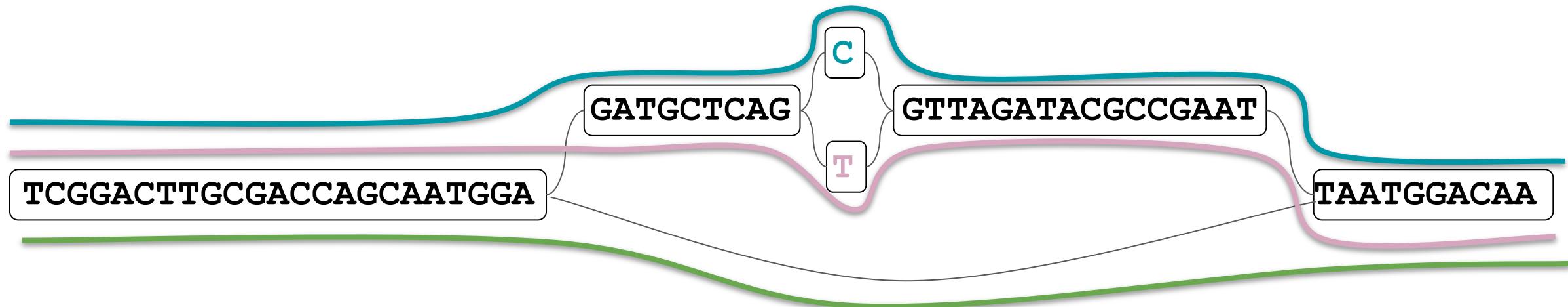
Pangenome Creation From Assemblies



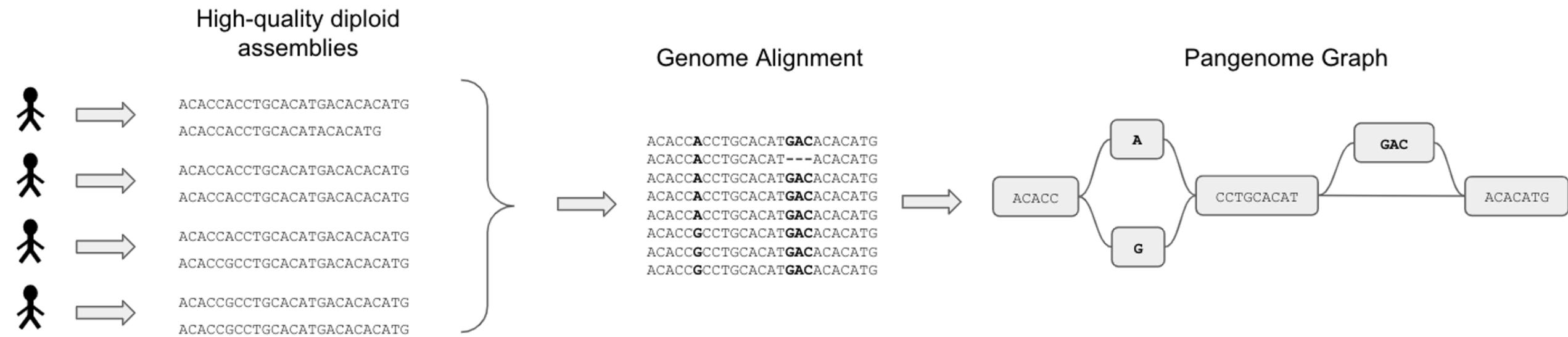
Pangenome Creation From Assemblies



Pangenome Creation From Assemblies



Pangenome: Conceptual Approach To Creation



Pangenome: Three Approaches

- Pangenome team currently has “Freeze1” pangenomes
 - The best pangenomes we can create now with current data and tools
 - 90 haplotypes were included (includes GRCh38 and CHM13)

	Minigraph	Minigraph/CACTUS	PGGB
Short Read Mapping	Untested	Yes (fast)	Untested
Long Read Mapping	Yes (fastest)	Yes	Yes
Assembly Mapping	Yes (direct)	Untested	Yes (via injection)

Pangenome: Three Approaches

- Generalization of minimap2
- Iterative construction
- SV only



	Minigraph	Minigraph/CACTUS	PGGB
Short Read Mapping	Untested	Yes (fast)	Untested
Long Read Mapping	Yes (fastest)	Yes	Yes
Assembly Mapping	Yes (direct)	Untested	Yes (via injection)

Pangenome: Three Approaches

- Adds base-level alignments to minigraph
- Omits centromeric variation



	Minigraph	Minigraph/CACTUS	PGGB
Short Read Mapping	Untested	Yes (fast)	Untested
Long Read Mapping	Yes (fastest)	Yes	Yes
Assembly Mapping	Yes (direct)	Untested	Yes (via injection)

Pangenome: Three Approaches

- Constructed with all-to-all pairwise mapping
- Contains centromeric variation



	Minigraph	Minigraph/CACTUS	PGGB
Short Read Mapping	Untested	Yes (fast)	Untested
Long Read Mapping	Yes (fastest)	Yes	Yes
Assembly Mapping	Yes (direct)	Untested	Yes (via injection)

Pangenome: How You Can Use A Pangenome

1

Calling SNV/InDels From Short Reads
VG-Giraffe

2

Short RNA Mapping
VG

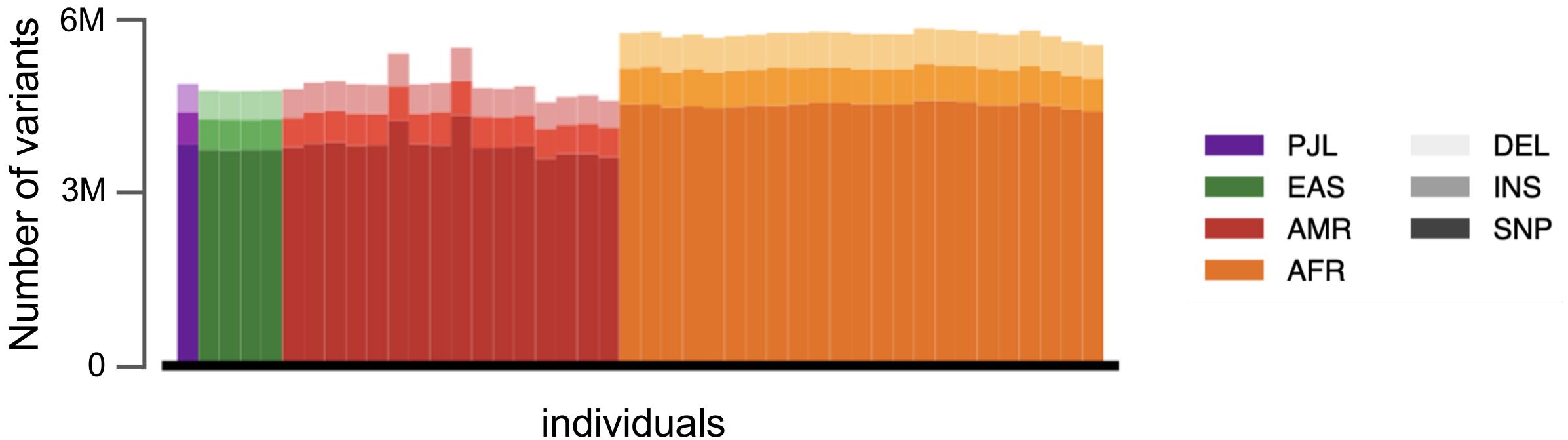
3

Comparative Genomics
Create browser hub (MC)
Liftover with Comparative Analysis Toolkit

4

Calling Structural Variants From Short Reads
Pangenie

Pangenome: How You Can Use A Pangenome



Credit: Wen-Wei Liao

You Can Do This Too!

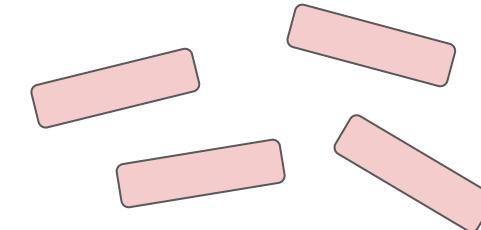
- Navigate to Demo Workspace:

<https://github.com/human-pangenomics/hprc-tutorials>

Later (after workshop): Google “Sign Up For Terra”

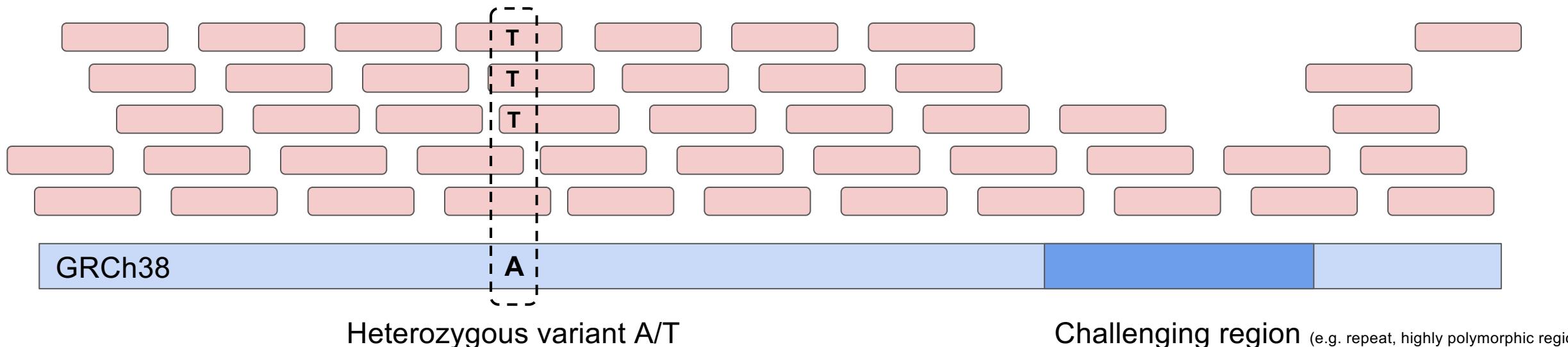
- FYI: need a Gmail account
- Use Demo Workspace (link above) to replicate analysis

Variant Calling From Short Sequencing Reads



Sequencing reads mapped to a reference genome.

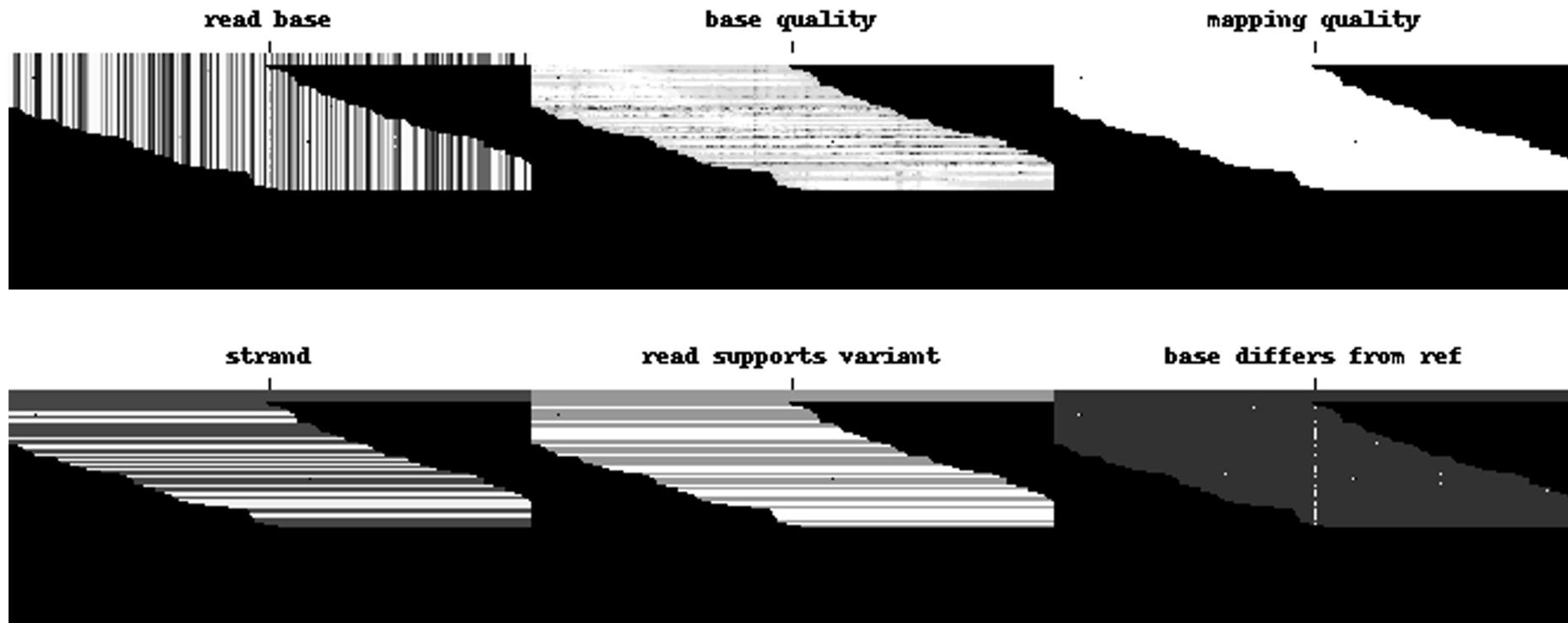
Variants called from the differences between reads and reference.



Unmapped when sequence not in the reference

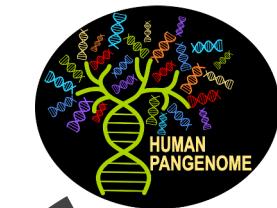
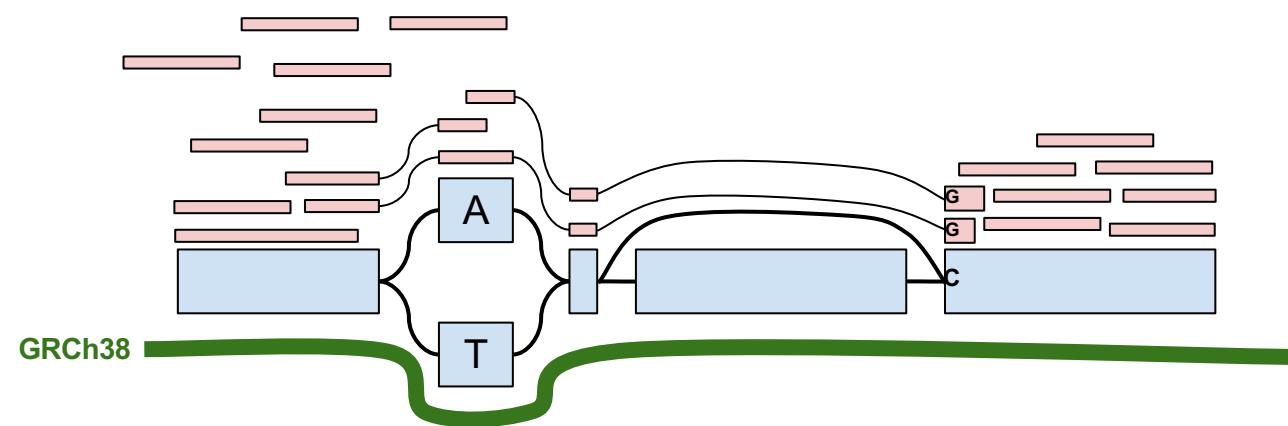
DeepVariant

Machine learning approach to predict genotype at a candidate site from read pile-up images.



Giraffe/DeepVariant

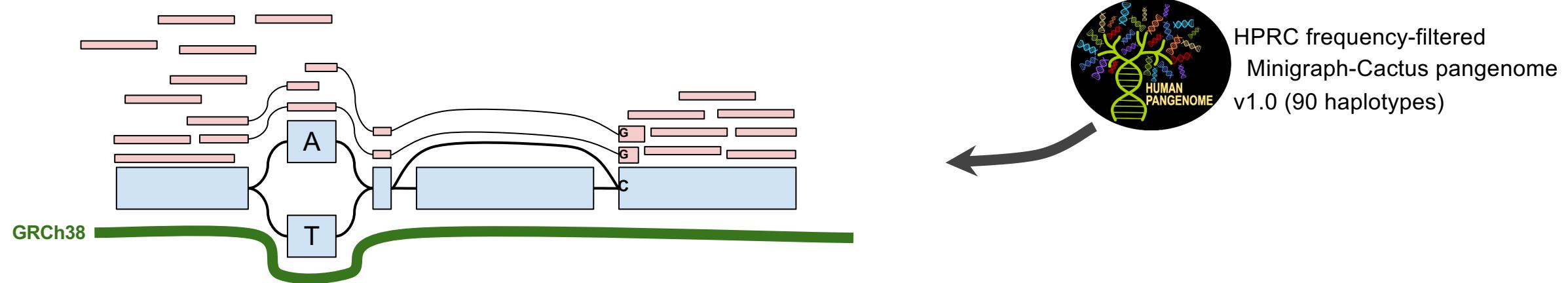
1. Short sequencing reads mapped to the pangenome with *vg giraffe*



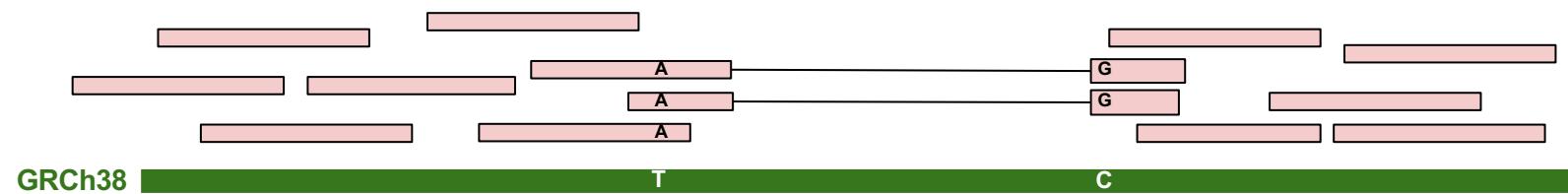
HPRC frequency-filtered
Minigraph-Cactus pangenome
v1.0 (90 haplotypes)

Giraffe/DeepVariant

1. Short sequencing reads mapped to the pangenome with *vg giraffe*

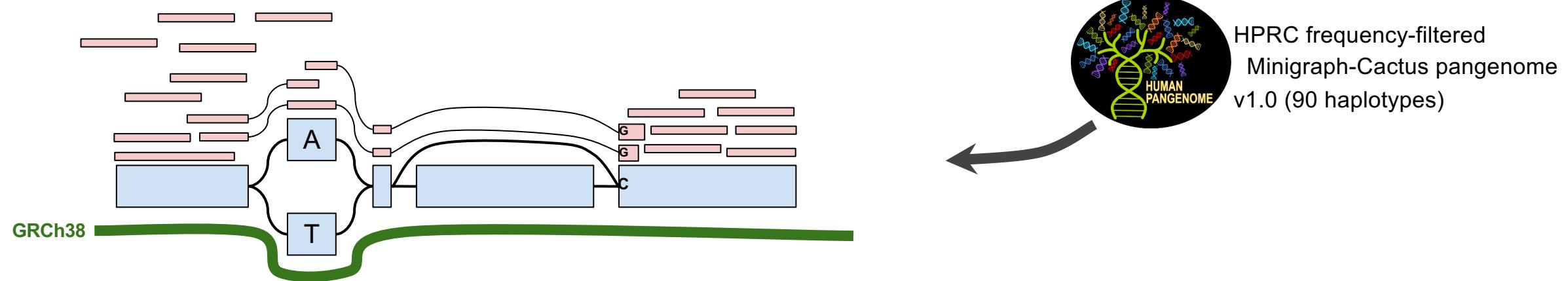


2. Aligned reads projected to GRCh38 with *vg surject*

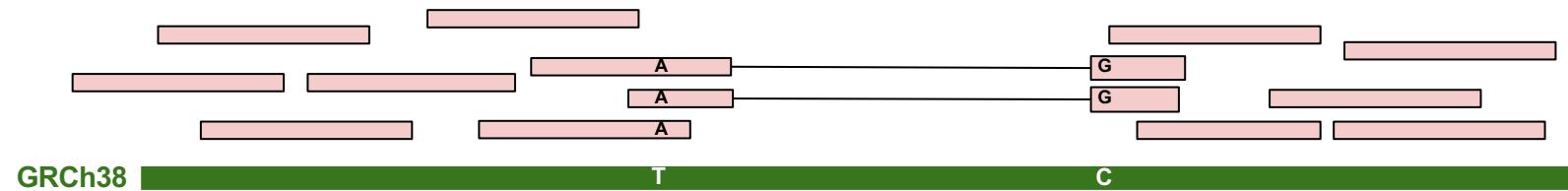


Giraffe/DeepVariant

1. Short sequencing reads mapped to the pangenome with *vg giraffe*

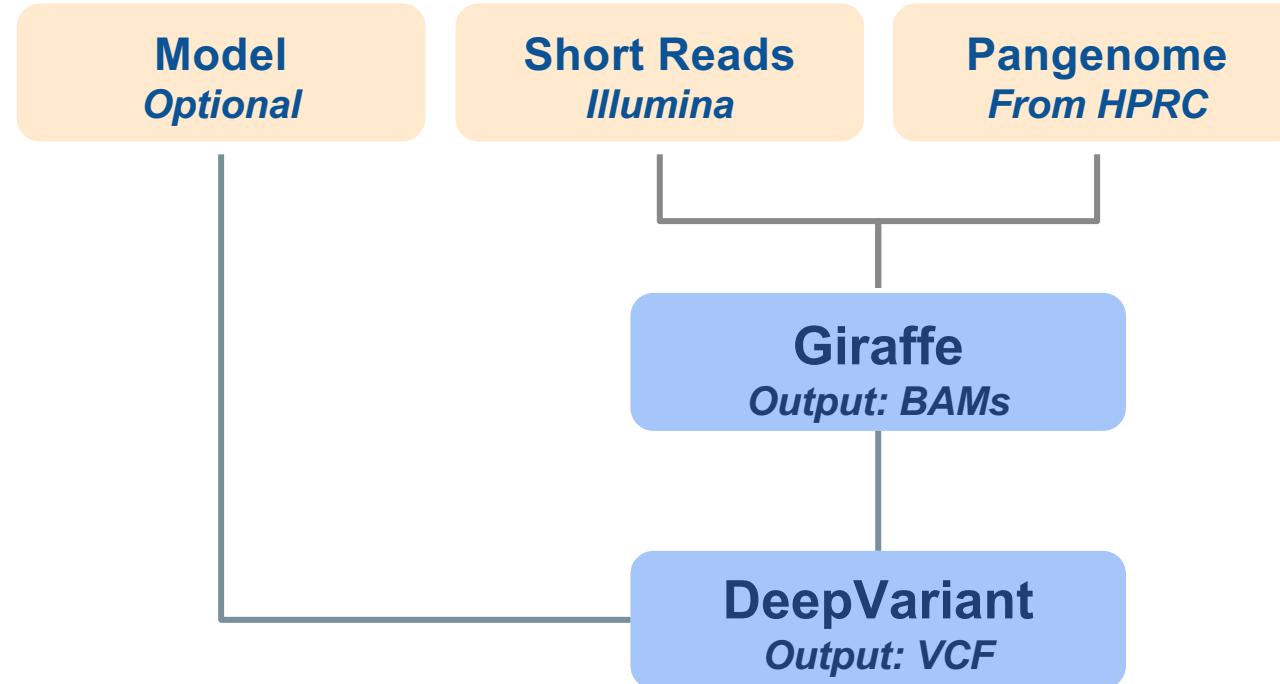


2. Aligned reads projected to GRCh38 with *vg surject*

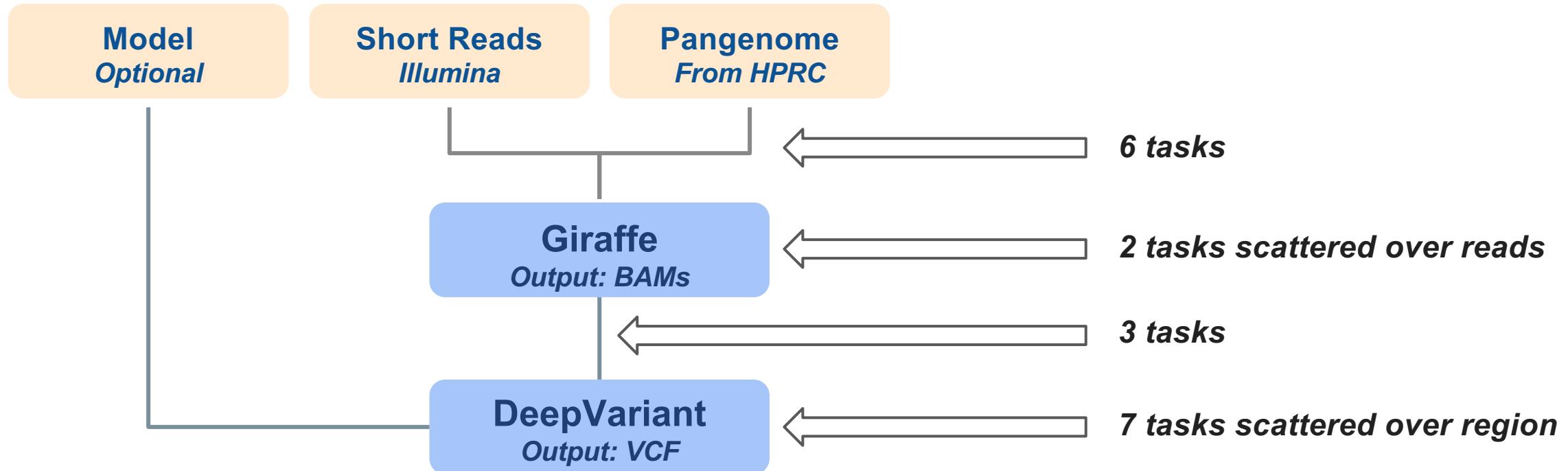


3. Variants called with DeepVariant
trained on surjected alignments, after
indel realignment.

Giraffe/DeepVariant Workflow (WDL)

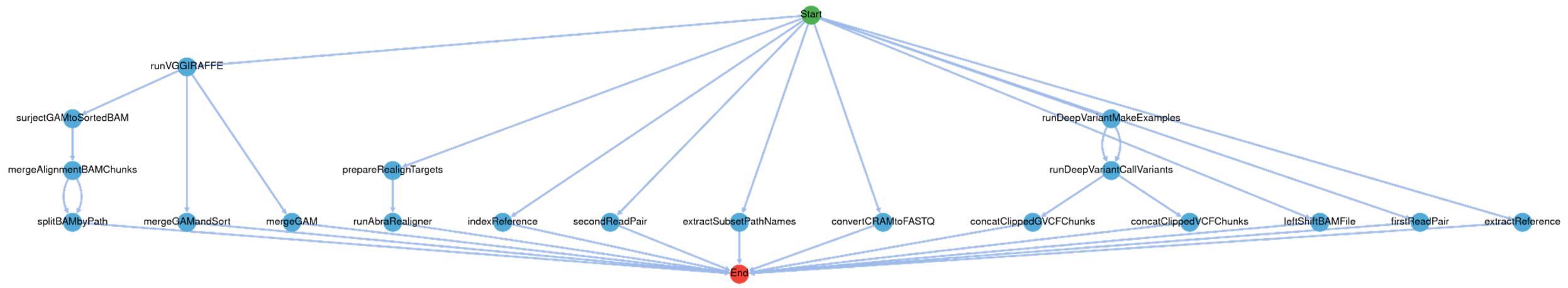


Giraffe/DeepVariant Workflow (WDL)



Giraffe/DeepVariant Workflow (WDL)

Can Be Treated By Users As One Step

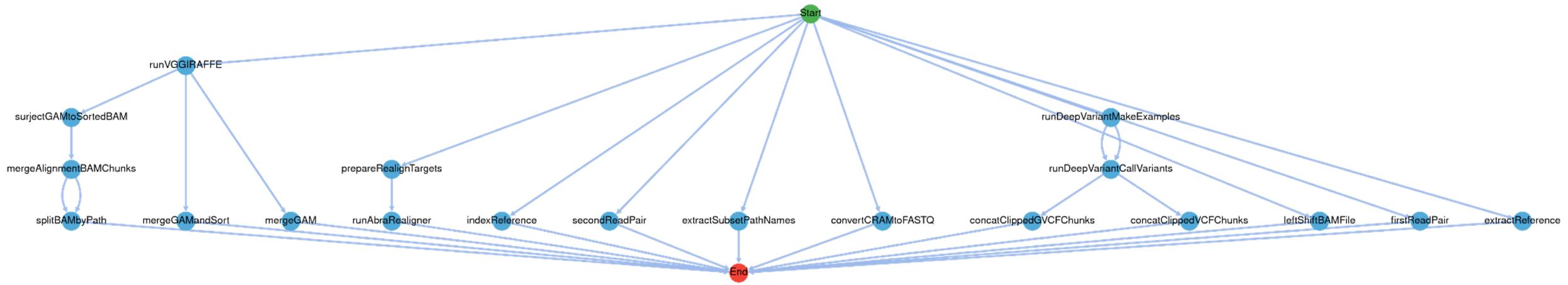


Motivation For Using Tools/Data In AnVIL

- Able to use pangenome in HPRC's workspace
- Easily find and import workflows from Dockstore
- Allows easy sharing of analysis results

Giraffe/DeepVariant Workflow (WDL)

Can Be Treated By Users As One Step



Motivation For Using Tools/Data In AnVIL

- Able to use pangenome in HPRC's workspace
- Easily find and import workflows from Dockstore
- Allows easy sharing of analysis results

Running workflows with a workflow runner in the cloud (using data stored centrally in the cloud) is an important tool for genomic analysis.

First Look At a Workspace On Terra

Tabs to describe data, run interactive notebooks, or run workflows

The screenshot shows the Terra workspace interface. At the top, there is a navigation bar with five tabs: DASHBOARD (highlighted in green), DATA, ANALYSES, WORKFLOWS, and JOB HISTORY. To the right of the tabs is a three-dot menu icon. Below the tabs, the main content area is divided into several sections:

- WORKSPACE INFORMATION**:
 - Last Updated: 10/29/2022
 - Creation Date: 10/29/2022
 - Workflow Submissions: 1
 - Access Level: Project Owner
- CLOUD INFORMATION**: (with a right arrow)
- OWNERS**: (with a right arrow)
- TAGS**: (with a right arrow)
- NOTIFICATIONS**: (with a right arrow)

On the far right, there is a vertical sidebar with the text "Rate: \$0.00 per hour" and a blue cloud icon with a lightning bolt.

Data Tab: Organizing The Data In Tables

The screenshot shows the Data Tab interface. At the top, there are tabs for DASHBOARD, DATA (which is selected and highlighted in green), ANALYSES, WORKFLOWS, and JOB HISTORY. Below the tabs is a toolbar with buttons for IMPORT DATA, EDIT, OPEN WITH..., EXPORT, SETTINGS, and ADVANCED SEARCH. A search bar shows "0 rows selected". On the right, there is a sidebar with a cloud icon and text indicating a rate of \$0.00 per hour.

A red arrow points to the "TABLES" section in the sidebar. The main area displays a table with the following columns: sample_id, sample, input_fastq_1, and input_fastq_2. One row is visible, showing HG002_chrl_2... under sample_id, HG002 under sample, and two file paths under the fastq columns.

	sample_id	sample	input_fastq_1	input_fastq_2
	HG002_chrl_2...	HG002	HG002.chr1_25053647_25685365.R1.fastq.gz	HG002.chr1_25053647_25685365.R2.fastq.gz

The sidebar also lists other sections: sample (1), sample_prerun (1), REFERENCE DATA (hg38), and OTHER DATA (Workspace Data, Files).

Data Tab: Organizing The Data In Tables

The screenshot shows the Data Tab interface with the following components:

- Header:** DASHBOARD, DATA (highlighted in green), ANALYSES, WORKFLOWS, JOB HISTORY.
- Toolbar:** IMPORT DATA, EDIT, OPEN WITH..., EXPORT, SETTINGS, 0 rows selected, ADVANCED SEARCH, Search, Rate: \$0.00 per hour.
- Tables Sidebar:** TABLES dropdown, Search all tables, sample (1), sample_prerun (1), REFERENCE DATA (hg38), OTHER DATA (Workspace Data, Files).
- Table View:** A table with columns: sample_id, sample, input_fastq_1, and input_fastq_2. It shows one row for HG002 with values: HG002_chr1_2..., HG002, [HG002.chr1_25053647_25685365.R1.fastq.gz](#), and [HG002.chr1_25053647_25685365.R2.fastq.gz](#).
- Annotations:** A box labeled "metadata" has an arrow pointing to the "sample" column. Two boxes labeled "paths to raw data" have arrows pointing to the "input_fastq_1" and "input_fastq_2" columns.

Workflows Tab: Launch Analyses

The screenshot shows the 'WORKFLOWS' tab selected in a navigation bar. Below the header, there's a search bar labeled 'SEARCH WORKFLOWS' and a sorting dropdown set to 'Alphabetical'. To the right, there are two icons: a grid and a list. On the far right, a sidebar displays a rate of '\$0.00 per hour' next to a cloud icon.

WORKFLOWS

Find a Workflow [+ New](#)

GiraffeDeepVariantLite
V. giraffe-dv-dt-hprcy1
Source: Dockstore [More](#)

publicly available workflow imported from Dockstore

Configure Inputs Of The Workflow

DASHBOARD DATA ANALYSES WORKFLOWS JOB HISTORY

Back to list GiraffeDeepVariantLite

Version: giraffe-dv-dt-hprc... Rate: \$0.00 per hour

Source: github.com/vgteam/vg_wdl/GiraffeDeepVariantLite:giraffe-dv-dt-hprcy1

Synopsis:
No documentation provided

Run workflow with inputs defined by file paths
 Run workflow(s) with inputs defined by data table

Step 1
Select root entity type: sample

Step 2
SELECT DATA 1 selected sample

Use call caching i Delete intermediate outputs i Use reference disks i Retry with more memory i Ignore empty outputs i

SCRIPT •• INPUTS •• OUTPUTS •• RUN ANALYSIS

Hide optional inputs Download json | Drag or click to upload json | Clear inputs SEARCH INPUTS

Task name ↓	Variable	Type	Attribute
vgMultiMap	INPUT_READ_FILE_1	File	this.input_fastq_1 <small>[...]</small>
vgMultiMap	INPUT_READ_FILE_2	File	this.input_fastq_2 <small>[...]</small>
vgMultiMap	SAMPLE_NAME	String	this.sample <small>[...]</small>

Configure Inputs Of The Workflow

DASHBOARD DATA ANALYSES WORKFLOWS JOB HISTORY ⋮

← Back to list

GiraffeDeepVariantLite

Version: giraffe-dv-dt-hprc...

Source: github.com/vgteam/vg_wdl/GiraffeDeepVariantLite:giraffe-dv-dt-hprcy1

Synopsis:
No documentation provided

Run workflow with inputs defined by file paths
 Run workflow(s) with inputs defined by data table

Step 1
Select root entity type:
Step 2

1 selected sample

Use call caching i Delete intermediate outputs i Use reference disks i Retry with more memory i Ignore empty outputs i

SCRIPT ⋯ INPUTS ⋯ OUTPUTS ⋯ RUN ANALYSIS

Hide optional inputs

Download json | Drag or click to upload json | Clear inputs

SEARCH INPUTS

Task name ↓	Variable	Type	Attribute
vgMultiMap	INPUT_READ_FILE_1	File	<input type="text" value="this.input_fastq_1"/> <input style="width: 20px; height: 20px; border: 1px solid #ccc;" type="button" value="..."/>
vgMultiMap	INPUT_READ_FILE_2	File	<input type="text" value="this.input_fastq_2"/> <input style="width: 20px; height: 20px; border: 1px solid #ccc;" type="button" value="..."/>
vgMultiMap	SAMPLE_NAME	String	<input type="text" value="this.sample"/> <input style="width: 20px; height: 20px; border: 1px solid #ccc;" type="button" value="..."/>

Configure Inputs Of The Workflow

DASHBOARD DATA ANALYSES WORKFLOWS JOB HISTORY

Back to list GiraffeDeepVariantLite

Version: giraffe-dv-dt-hprc... Rate: \$0.00 per hour

Source: github.com/vgteam/vg_wdl/GiraffeDeepVariantLite:giraffe-dv-dt-hprcy1

Synopsis:
No documentation provided

Run workflow with inputs defined by file paths
 Run workflow(s) with inputs defined by data table

Step 1
Select root entity type: sample

Step 2
SELECT DATA 1 selected sample

Use call caching i Delete intermediate outputs i Use reference disks i Retry with more memory i Ignore empty outputs i

SCRIPT •• INPUTS •• OUTPUTS •• RUN ANALYSIS

Hide optional inputs Download json | Drag or click to upload json | Clear inputs SEARCH INPUTS

Task name ↓	Variable	Type	Attribute
vgMultiMap	INPUT_READ_FILE_1	File	this.input_fastq_1  
vgMultiMap	INPUT_READ_FILE_2	File	this.input_fastq_2  
vgMultiMap	SAMPLE_NAME	String	this.sample  

links to Table's columns

Configure The Outputs Of The Workflow

Select root entity type: sample 1 selected sample

Use call caching i Delete intermediate outputs i Use reference disks i Retry with more memory i Ignore empty outputs i

SCRIPT •• INPUTS •• |OUTPUTS| •• RUN ANALYSIS

Output files will be saved to
Files / submission unique ID / vgMultiMap / workflow unique ID

References to outputs will be written to
Tables / sample

Fill in the attributes below to add or update columns in your data table

Task name ↓	Variable	Type	Attribute Use defaults
vgMultiMap	output_gam	File	this.output_gam <small>[...]</small>
vgMultiMap	output_gam_index	File	this.output_gam_index <small>[...]</small>
vgMultiMap	output_gvcf	File	Optional <small>[...]</small>
vgMultiMap	output_gvcf_index	File	Optional <small>[...]</small>
vgMultiMap	output_vcf	File	this.output_vcf <small>[...]</small>
vgMultiMap	output_vcf_index	File	this.output_vcf_index <small>[...]</small>

Download json | Drag or click to upload json | Clear outputs

links to Table's columns

Job History Tab: Monitor Running/Prior Jobs

DASHBOARD DATA ANALYSES WORKFLOWS **JOB HISTORY** ⋮

Search Rate:
\$0.00 per hour

Submission (click for details)	Data entity	No. of Workflows	Status	Submitted ↑	Submission ID	Comment	Actions
HPRC2022/GiraffeDeepVariantLite Submitted by jmonlong@ucsc.edu	HG002_chr1_25053...	1	✓ Done	Oct 29, 2022 8:24 AM	9ba0378a-871f-4693-9fc5-356986ab3818	First workflow!!!	⋮

Job History Tab: Monitor Running/Prior Jobs

DASHBOARD DATA

Submission (click for detail)
HPRC2022/GiraffeDeepVa
Submitted by jmonlong@u

Job Manager SIGN OUT

vgMultiMap

ID: 8471552d-d76d-4bdb-9950-b9b619394491 workspace-id: d0e33d74-48f4-4480-ac0-98b7369157a4 submission-id: 9ba0378a-871f-4693-9fc5-356986ab38f8

Status: Succeeded

Tasks: 14 succeeded, 0 failed, 0 currently being processed

Submitted: Oct 29, 2022

Started: Oct 29, 2022

Ended: Oct 29, 2022 (0h 4m)

LIST VIEW INPUTS OUTPUTS LABELS TIMING DIAGRAM

Task Name	Status	Start	Duration	Inputs	Outputs	Links	Attempts
firstReadPair		Oct 29, 2022	0h 0m				1
secondReadPair		Oct 29, 2022	0h 0m				1
<u>runVGGIRAFFE</u>		Oct 29, 2022	0h 0m				
<u>surjectGAMtoSortedBAM</u>		Oct 29, 2022	0h 0m				
mergeGAMandSort		Oct 29, 2022	0h 0m				1
mergeAlignmentBAMchu...		Oct 29, 2022	0h 0m				1
splitBAMbyPath		Oct 29, 2022	0h 0m				1
<u>leftShiftBAMFile</u>		Oct 29, 2022	0h 0m				

Rate: \$0.00 per hour

Results Are Also Organized in Data Tables

The screenshot shows a user interface for managing data tables. The top navigation bar includes tabs for DASHBOARD, DATA (which is selected and highlighted in green), ANALYSES, WORKFLOWS, and JOB HISTORY. Below the navigation is a toolbar with buttons for IMPORT DATA, EDIT, OPEN WITH..., EXPORT, SETTINGS, and ADVANCED SEARCH. A search bar is also present. On the right side, there is a sidebar with a cloud icon and text indicating a rate of \$0.00 per hour.

The main area displays a table titled "sample" with one row. The columns are labeled: sample_id, output_gam, output_gam_index, and output_vcf. The "output_gam" column contains the value "HG002.sorted.gam". The "output_gam_index" column contains the value "HG002.sorted.gam.gai". The "output_vcf" column contains the value "HG002.vcf.gz". A "Search all tables" input field is located below the table.

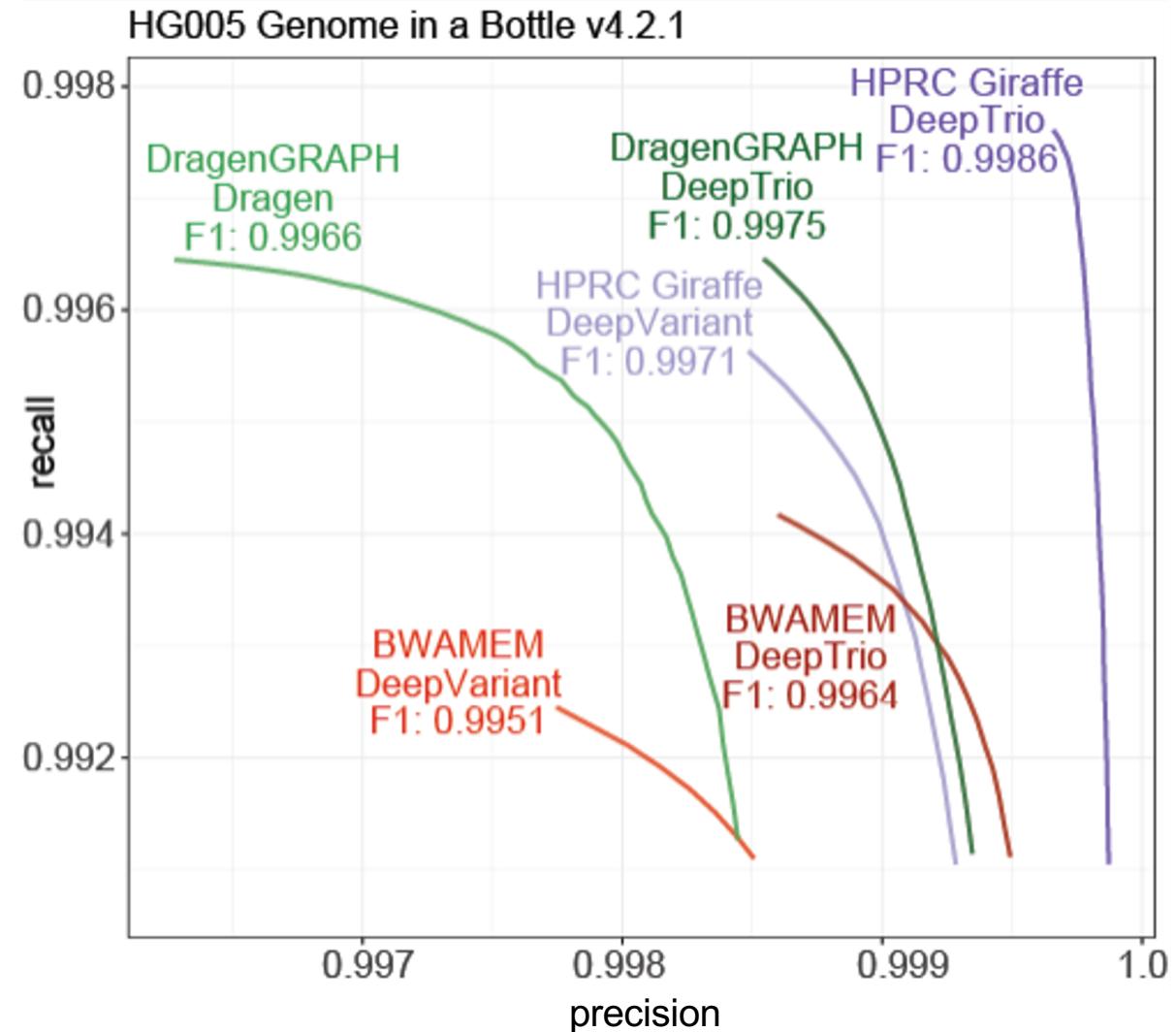
A vertical sidebar on the left lists categories: TABLES, REFERENCE DATA, and OTHER DATA. Under TABLES, there is a list item "sample (1)". Under REFERENCE DATA, there is "hg38". Under OTHER DATA, there are "Workspace Data" and "Files".

Three curved arrows point from a callout box labeled "new columns" to the "output_gam", "output_gam_index", and "output_vcf" columns in the table.

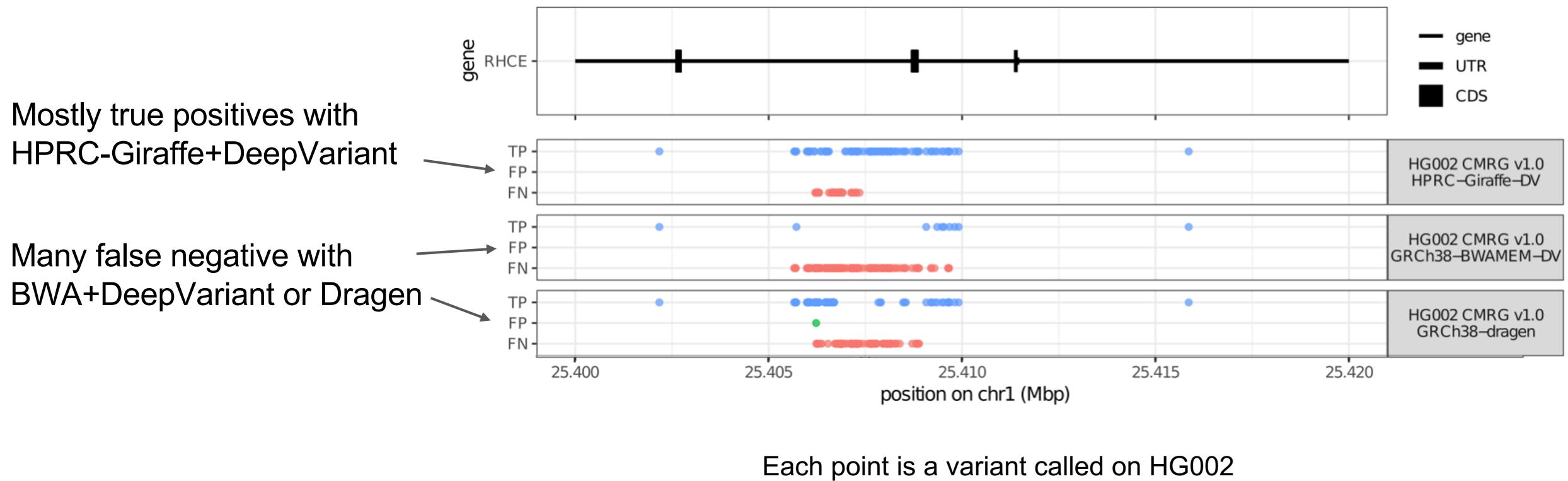
	sample_id	output_gam	output_gam_index	output_vcf
	HG002_chr1_2... tq.gz	HG002.sorted.gam	HG002.sorted.gam.gai	HG002.vcf.gz

Giraffe/DeepVariant Improves Variant Calling

On average, 34% less errors compared to the linear reference approach (GRCh38-BWA-DeepVariant)



RHCE: A Challenging, Medically-Relevant Gene



Complex Variation Around RHCE

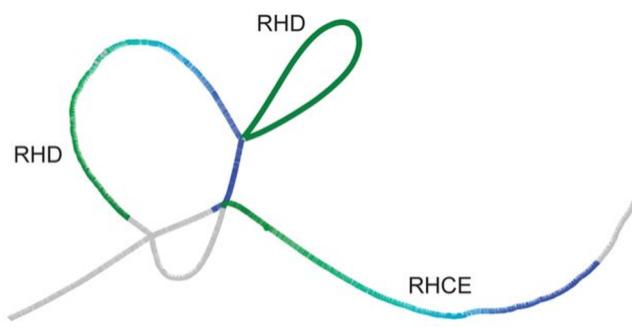
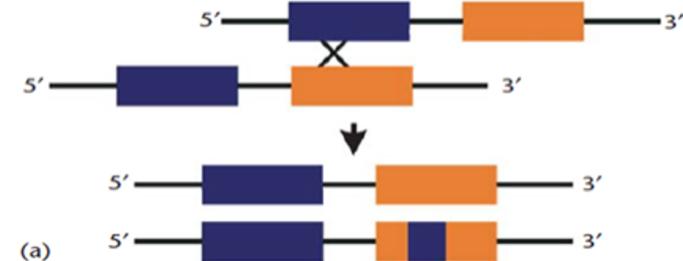


Image from [Wikipedia](#)



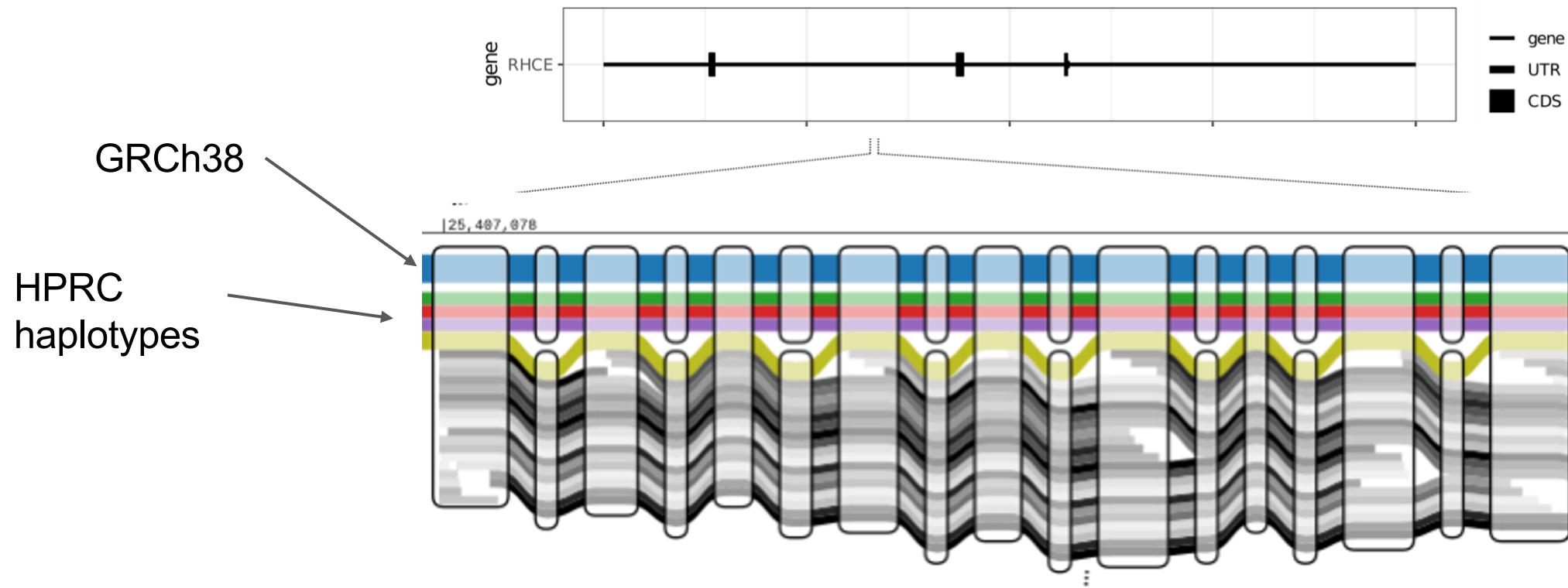
Gene conversion event

GRCh38

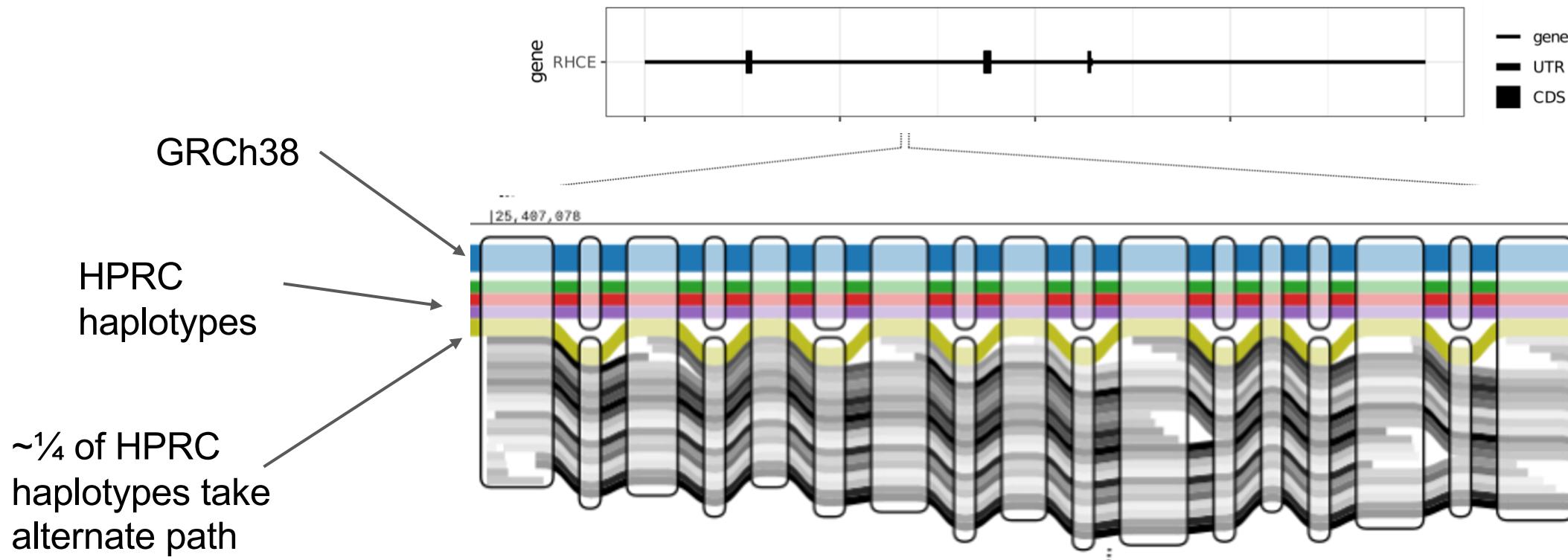
HG002

	Count	Frequency	Haplotype name	gene
	43	0.48	RHD;RHCE	RHD RHCE TMEM50A
	22	0.24	RHD;RHCE-RHD(2)-RHCE	RHD RHCE TMEM50A
	15	0.17	RHCE	RHD RHCE TMEM50A
	3	0.03	RHD-RHCE(2-3)-RHD;RHCE	RHD RHCE TMEM50A
	2*	0.02	RHD-RHCE(8)-RHD;RHCE	RHD RHCE TMEM50A
	1	0.01	RHCE-RHD(2)-RHCE	RHD RHCE TMEM50A
	1*	0.01	RHD-RHCE(2-9)-RHD;RHCE	RHD RHCE TMEM50A
	1*	0.01	RHD;RHCE-RHD(9)-RHCE	RHD RHCE TMEM50A
	1*	0.01	RHD-RHCE(10);inv;RHCE-RHD(10)	RHD RHCE TMEM50A
	1*	0.01	RHD;RHD;RHCE-RHD(9)-RHCE	RHD RHCE TMEM50A

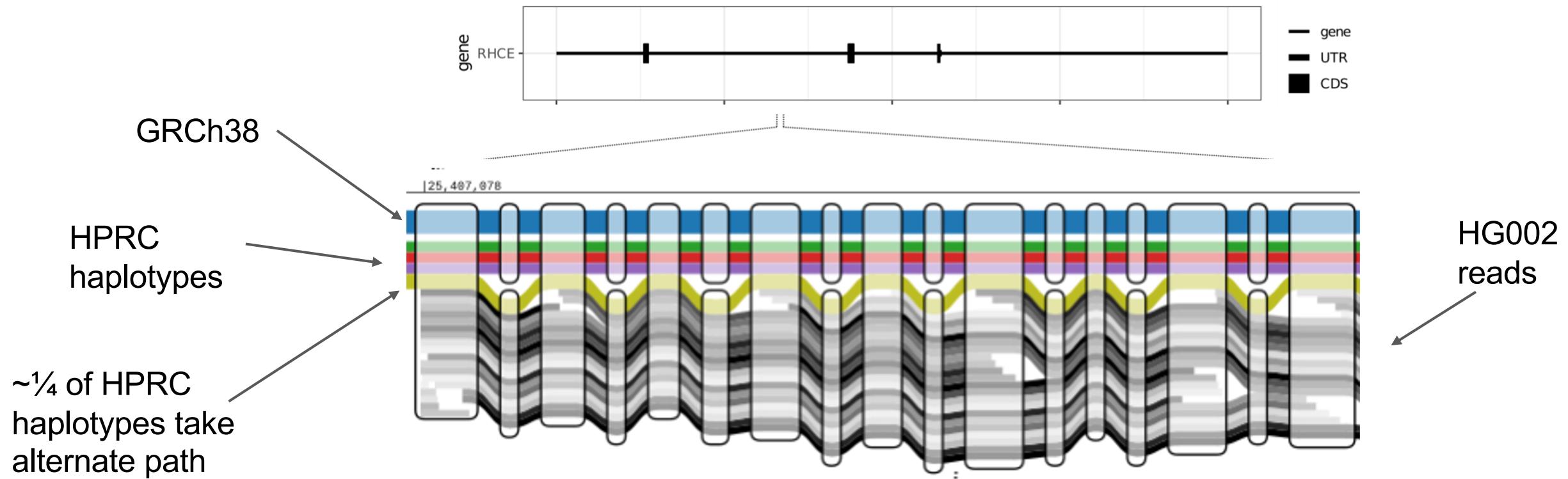
Small Reads Map To A Subset Of Haplotypes



Small Reads Map To A Subset Of Haplotypes



Small Reads Map To A Subset Of Haplotypes



Another example: VUS observed in healthy individuals

rs1553862732 <https://www.ncbi.nlm.nih.gov/snp/rs1553862732>

Current Build 156
Released September 21, 2022

Organism	<i>Homo sapiens</i>	Clinical Significance	Reported in ClinVar
Position	chr2:151600710 (GRCh38.p14) ?	Gene : Consequence	NEB : Missense Variant
Alleles	A>G	Publications	0 citations
Variation Type	SNV Single Nucleotide Variation	Genomic View	See rs on genome
Frequency	None		

Variant Details Clinical Significance HGVS Submissions History Publications Flanks

Allele: G (allele ID: 516380)

ClinVar Accession Disease Names Clinical Significance

RCV000641401.5 Nemaline myopathy 2 Uncertain-Significance

Frequencies <https://varsome.com/variant/hg38/rs1553862732>

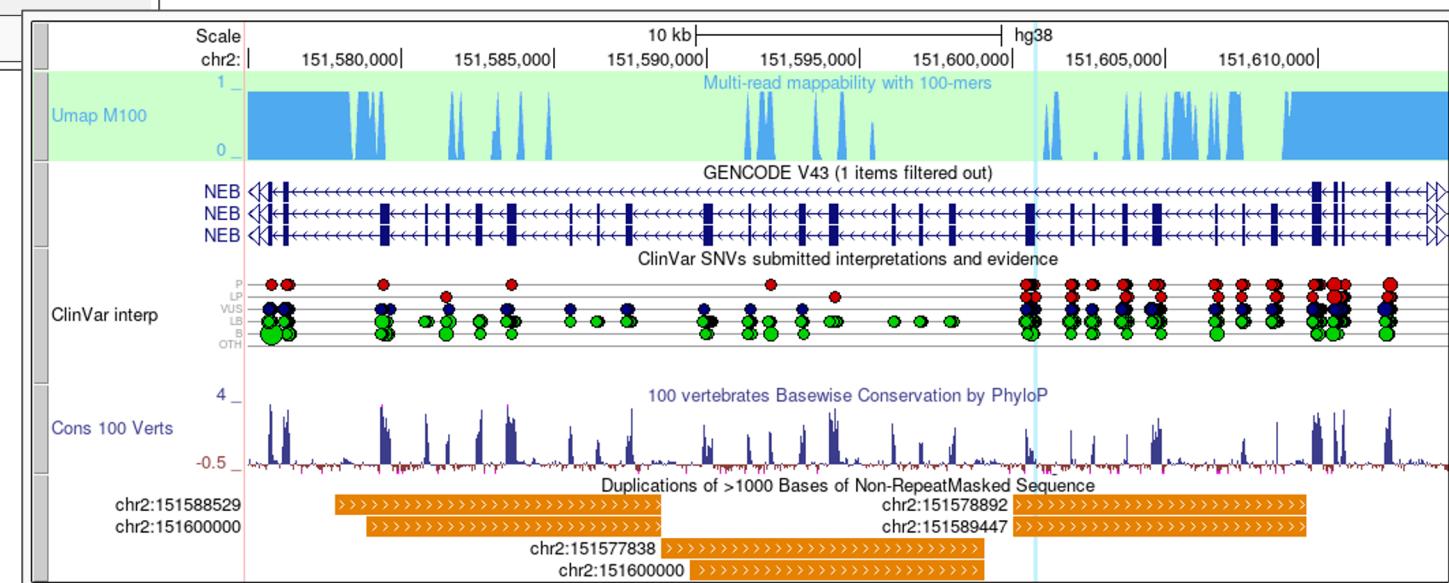
gnomAD Genomes

This variant does not have a gnomAD genomes entry, but its locus is covered in gnomAD genomes as follows.

gnomAD Genomes Coverage Version: 3.0

Coverage in gnomAD Genomes samples

Mean coverage	Median coverage	% of samples over 20x coverage
3.1	3	0.0076%



Hard region to map short reads to because of segmental duplications.

Found in 3 individuals of the 1000 Genomes cohort when using Giraffe-DeepVariant on the HPRC pangenome (AF ~0.002)

The Pangenome Reference For Clinical Variant Calling and Interpretation (Discussion)

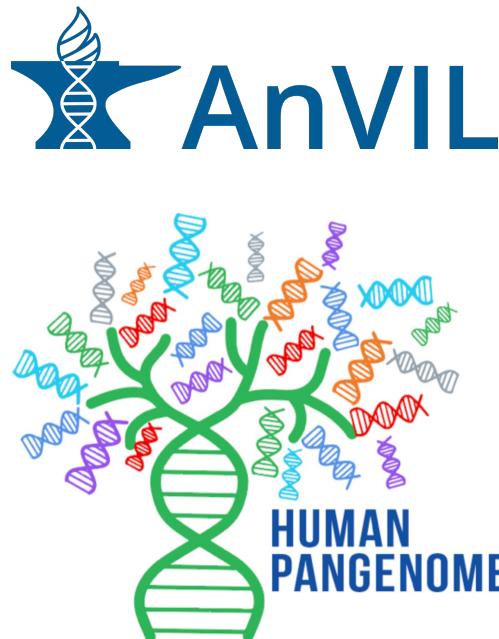


How might the adoption of a pangenome reference improve variant calling and/or interpretation?

What are the key barriers to adoption of a human pangenome reference in clinical genetics?

How might we design the pangenome reference tools to best serve the clinical genetics community?

Acknowledgements



HPRC Coordinating Center

Ting Wang Heather Lawson
Nathan Stitzel Lucinda Fulton
Robert Fulton Tina Lindsay
Sarah Cody Milinn Kremitsiki
Eddie Belter Derek Albracht
Haley Abel Chris Markovic
Wen-Wei Liao



David Haussler Benedict Paten
Karen Miga Ed Green Mark Akeson
Adam Novak Glenn Hickey Miten Jain
Hugh Olsen Erik Garrison Jean Monlong
Adirna Fuller Xian Chang Marina Haukness
Trevor Pesout Beth Sheets Tony Tsung Yu Lu
Jonas Sibbesen Julian Lucas Ryan Lorig-Roach
Jouni Siren Kishwar Shafin Charles Markello
Jordan Eizenga Melissa Meredith
Brian Hannafious



Peter Goodhand
Angela Page



National Human Genome Research Institute

Vimi Desai

Adam Felsenfeld
Mike Smith
Carolyn Hutter
Taylorlyn Stephan
Heidi Sofia

Adam Phillippy
Sergey Koren
Arang Rhie
Chirag Jain
Baergen Schultz



Richard Durbin



Justin Zook

HPRC-ELSI Working Group Members*



National Center for Biotechnology Information



Max Planck Institute of Molecular Cell Biology and Genetics
Gene Myers



Kerstin Howe



Ira Hal
Wen-Wei Liao
Shuangjia Lul



Heng Li
Shilpa Garg
Haoyu Cheng
Xiaowen Feng



Paul Flicek
Susan Fairley
Daniel Zerbino



We would like to acknowledge the National Genome Research Institute (NHGRI) for funding the following grants which are in support of creating the human pangenome reference: 1U41HG010972, 1U01HG010971, 1U01HG010961, 1U01HG010973, 1U01HG010963, and the Human Pangenome Reference Consortium (<https://humanpangenome.org/>)

Partnerships



*Robert Cook-Deegan, Nanibaa' Garrison, Barbara Koenig, Alice Popejoy, Heather Lawson, Vasiliki Rahimzadeh, Ann McCartney, Joe Yracheta, Jonathan Lotempio, Pearl O'Rourke, Jean McKewen, Shawneequa Callier, Pilar Ossorio, Mahsa Shabani, Leroy Hubert, Clement Odebamowo



✉ Respond at pollev.com/popejoy

Pangenome Workshop ACMG 2023

0 done

 **0 underway**

Start the presentation to see live content. For screen share software, share the entire screen. Get help at pollev.com/app