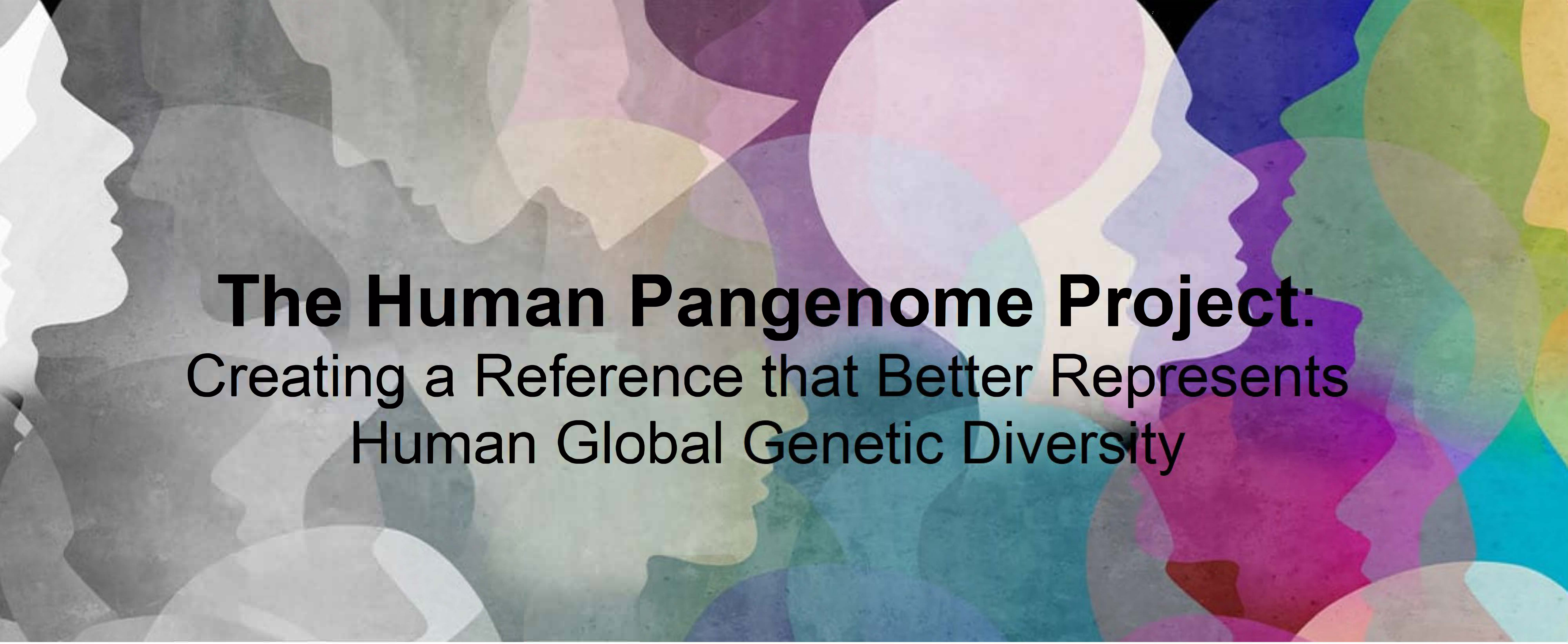


The Human Pangenome Project:

Creating a Reference that Better Represents Human Global Genetic Diversity

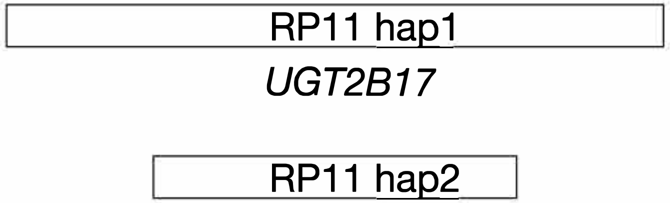
TELOMERE-TO-TELOMERE CONSORTIUM

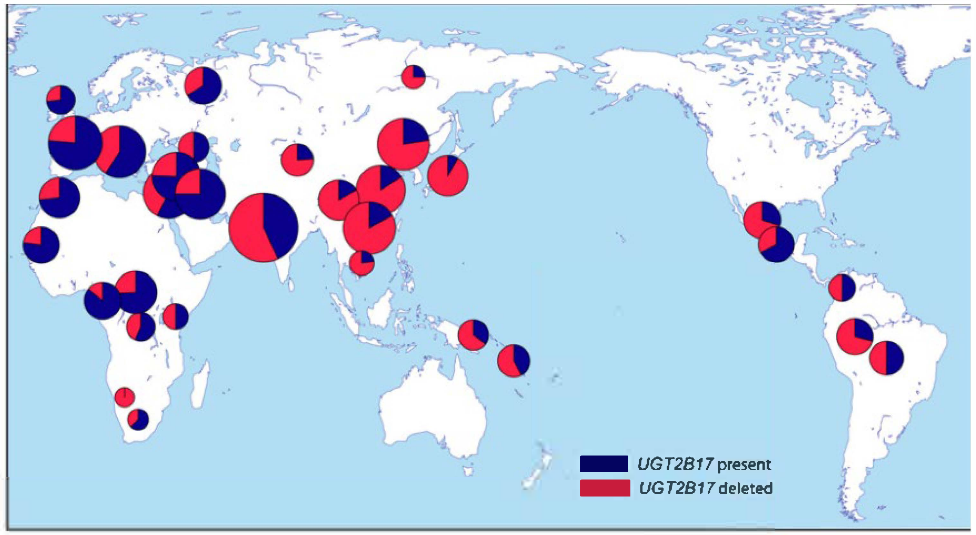
TOWARDS A COMPLETE REFERENCE OF HUMAN GENOME DIVERSITY

HUMAN PANGENOME

Deanna Church
Presenting

The current human reference fails to faithfully represent a single individual genome much less the genomes of a global population.





Xue et al, 2008

- **The human reference genome** is a foundational resource in human genetics and like most technology-driven resources, is overdue for an upgrade.

- The current structure is a **linear monoploid representation containing mixed haplotypes with too many gaps and errors. Additionally, the underlying sequence is predominately from a single individual.**

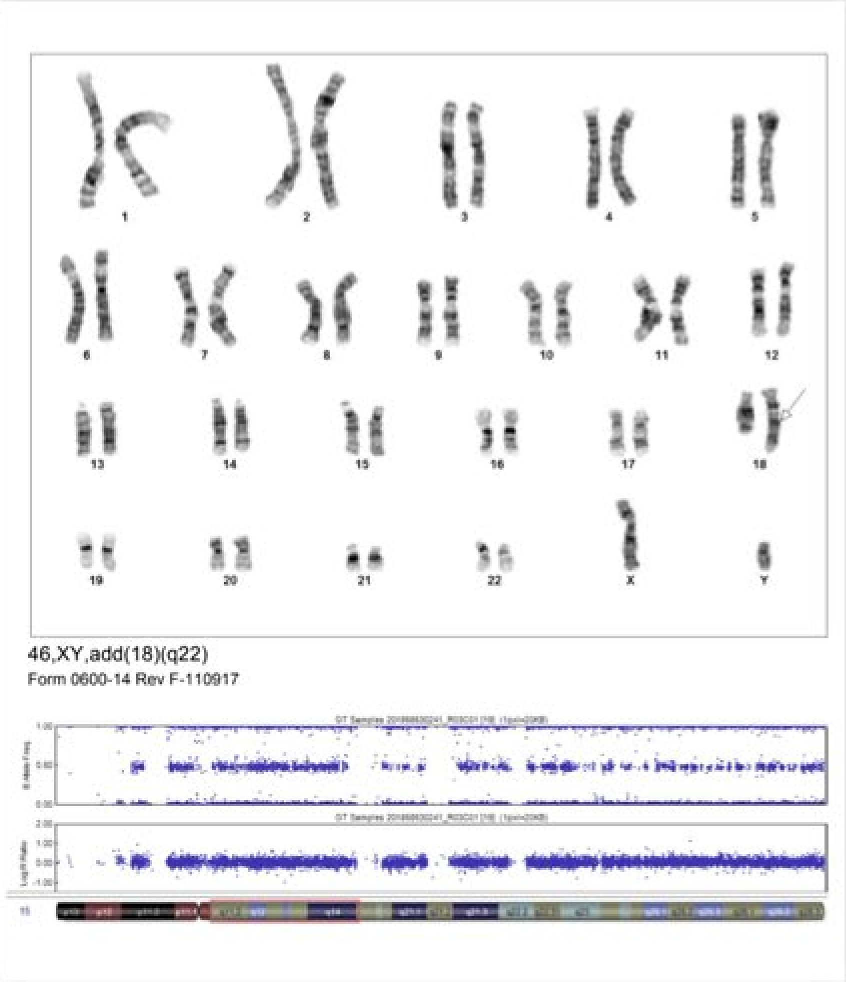
- **Mapping limitations of short reads and inherent reference biases** means we have missed more than 70% of structural variants in traditional whole-genome sequencing studies

Human Pangenome Reference Consortium

- Improve representation of **global genomic diversity** (>350 diverse diploid references)
- **Prioritizing quality**: we aim to release a complete (T2T) and comprehensive map of genome variation
- **Develop a new, non-linear reference data structure** and foster an innovative ecosystem of pangenomic tools
- Outreach, Education and Implementation

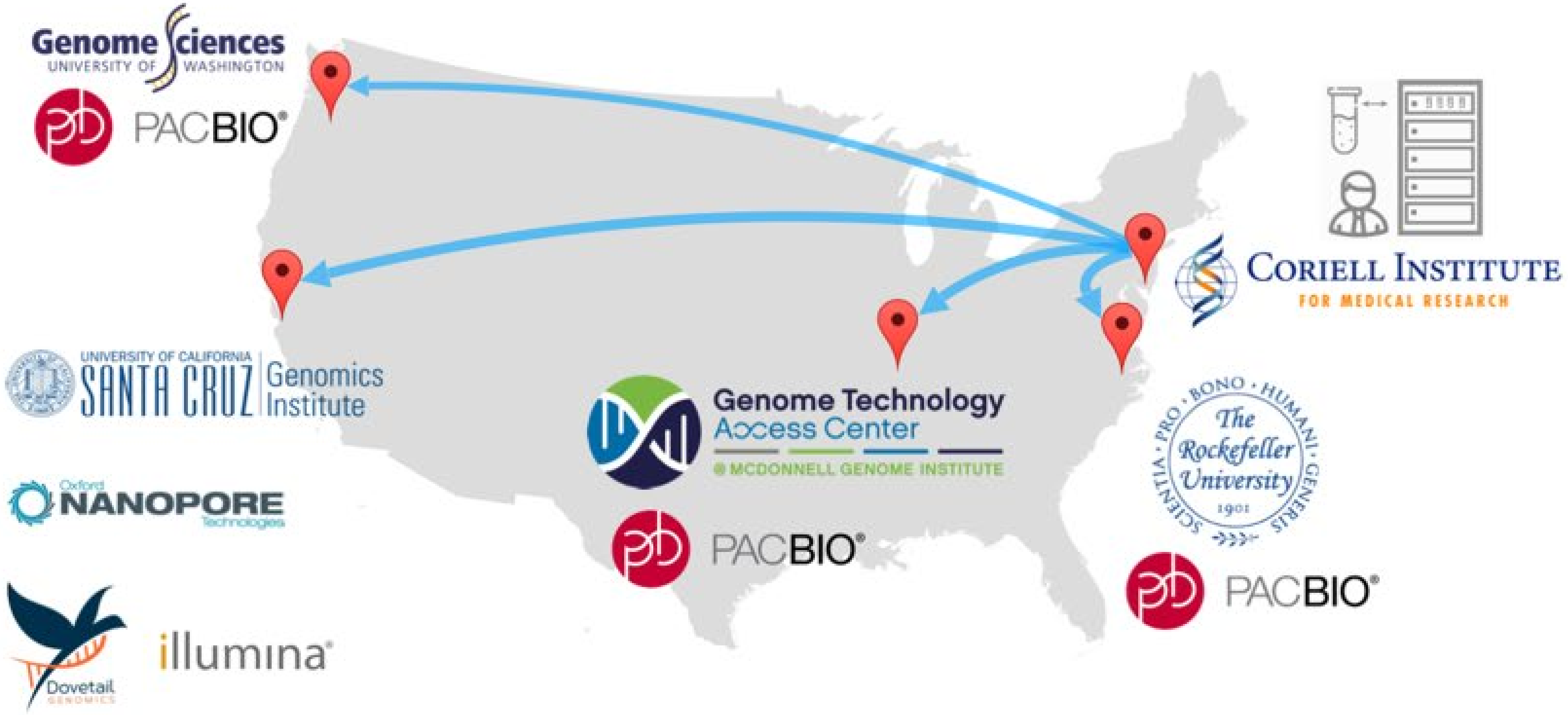
Multi-Center Sequencing Technology and Production: Optimized for Efficiency and Quality

Cell line stability/QC



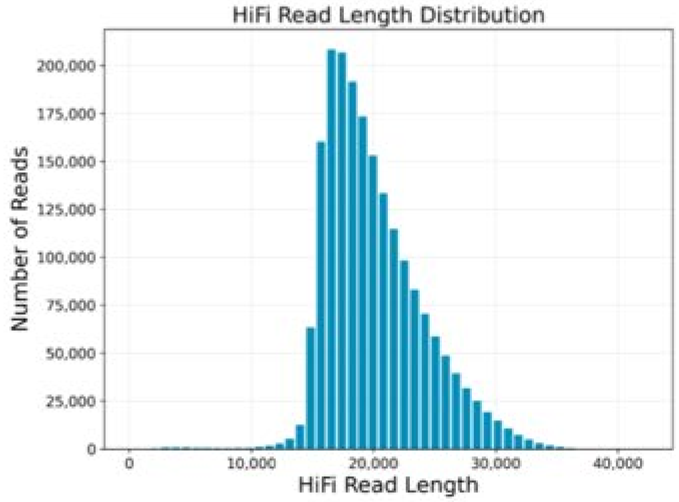
46,XY,add(18)(q22)

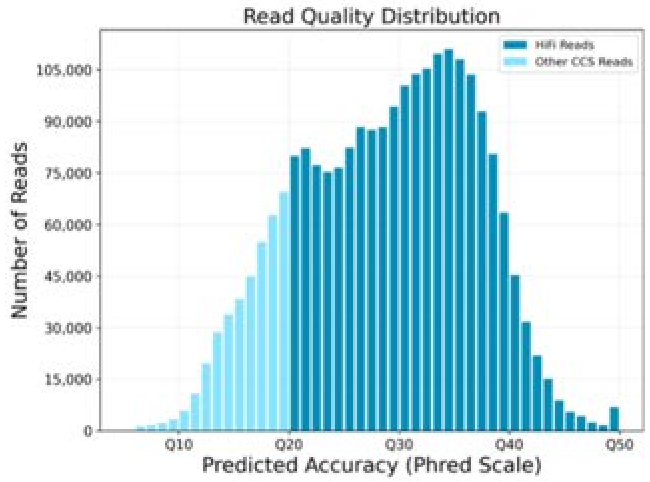
Form 0600-14 Rev F-110917

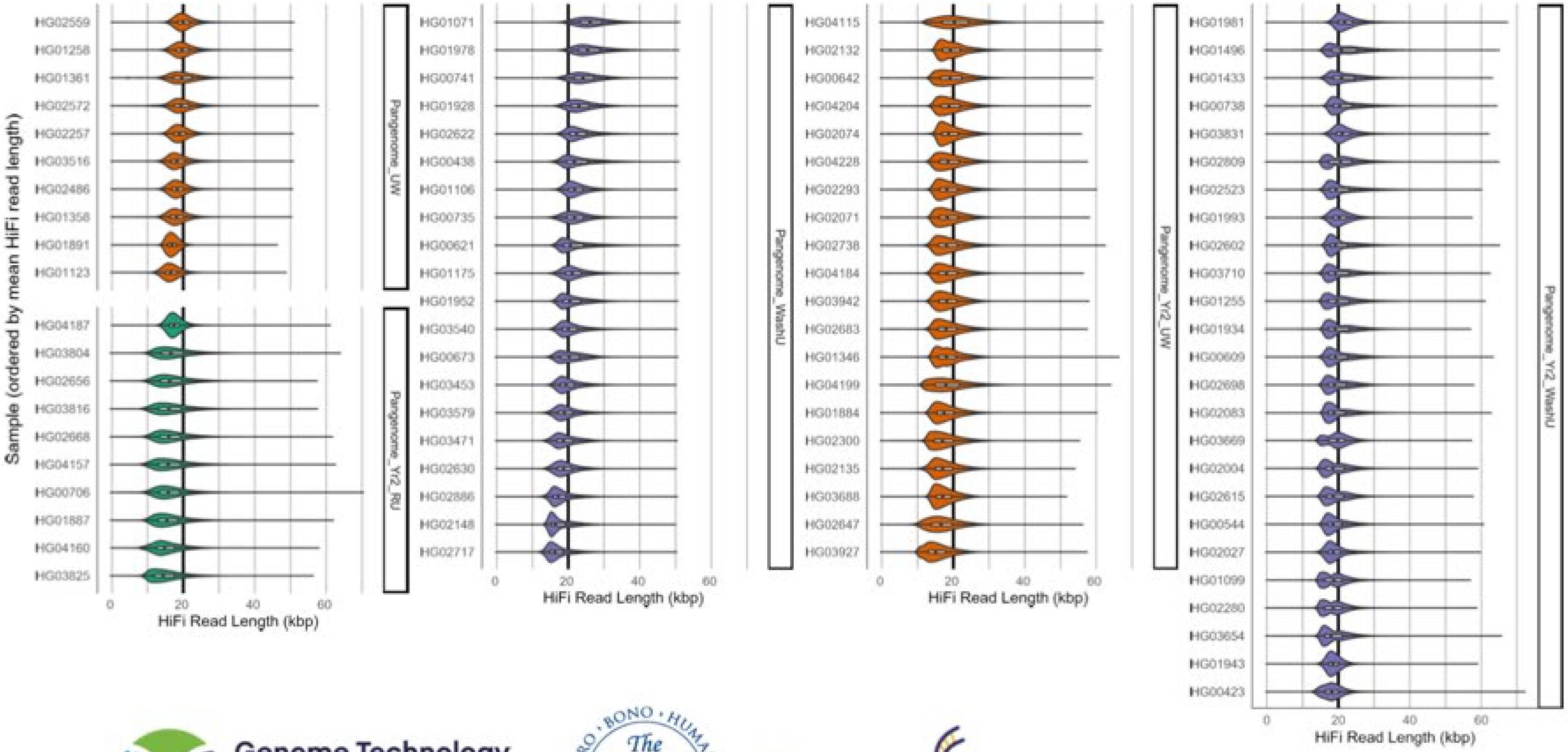

Genome Sciences
UNIVERSITY OF WASHINGTON
PACBIO®
UNIVERSITY OF CALIFORNIA
SANTA CRUZ Genomics Institute
Oxford NANOPORE Technologies
Dovetail GENOMICS
illumina®
Genome Technology Access Center
@ MCDONNELL GENOME INSTITUTE
PACBIO®
CORIELL INSTITUTE
FOR MEDICAL RESEARCH
The Rockefeller University
PRO BONO HUMANI GENERIS SCIENTIA
1901
PACBIO®

Optimized and Consistent Long-Read HiFi Production

PACBIO®


HiFi Read Length Distribution


Read Quality Distribution



Genome Technology Access Center @ MCDONNELL GENOME INSTITUTE 🟪

The Rockefeller University 1901 PRO BONO HUMANI GENERIS SCIENTIA 🟩

Genome Sciences UNIVERSITY OF WASHINGTON 🟧

Forefront of ultra-long read sequencing innovation

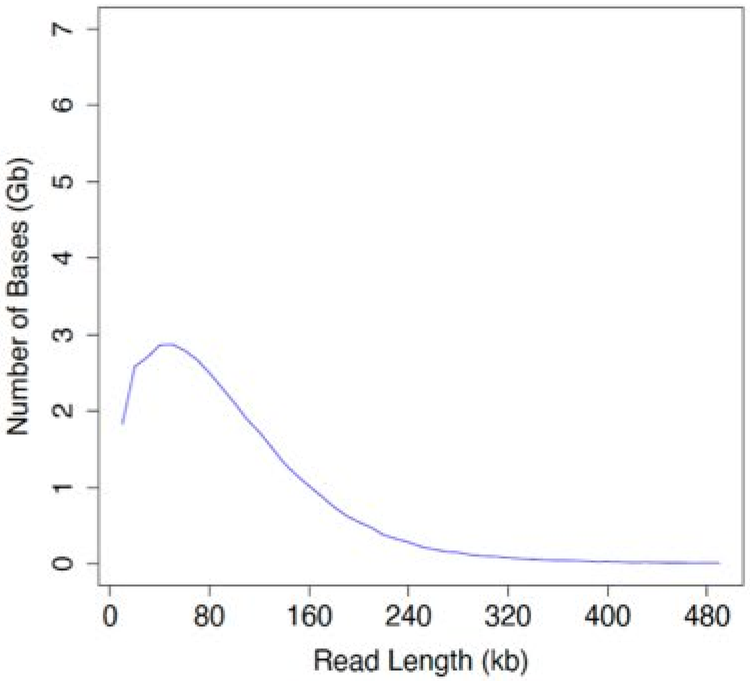
UC SANTA CRUZ | Genomics Institute

Oxford NANOPORE Technologies

circulomics

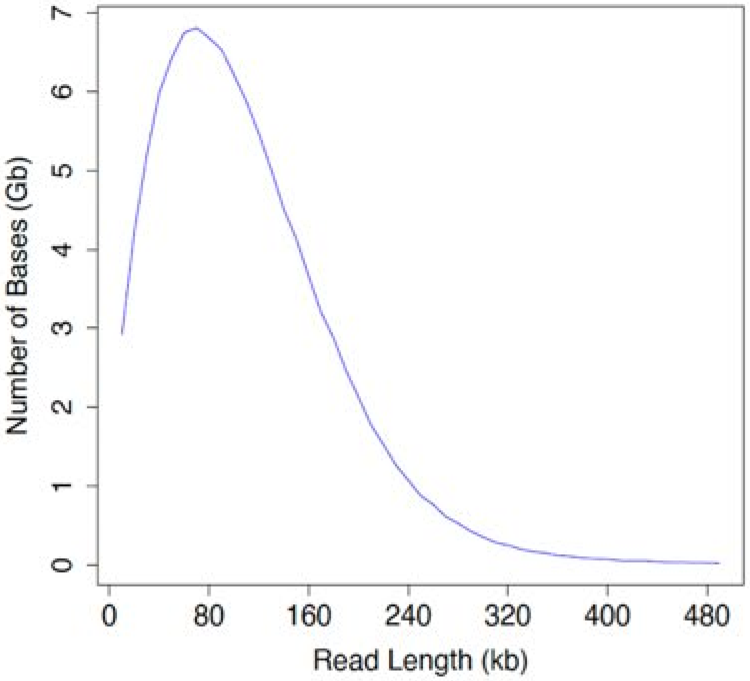
2020

Read N50 ~76 kb
~9X coverage per flow cell
~3.5X coverage in 100 kb+



2021

Read N50 ~73 kb
~30X coverage per flow cell
~9X coverage in 100 kb+

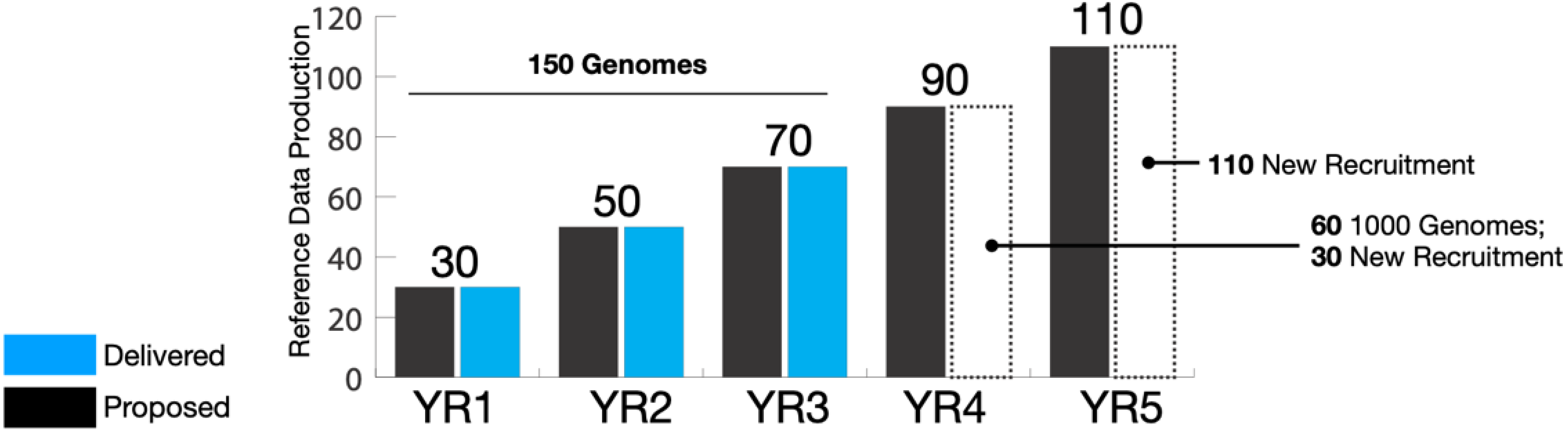


3x increase in throughput

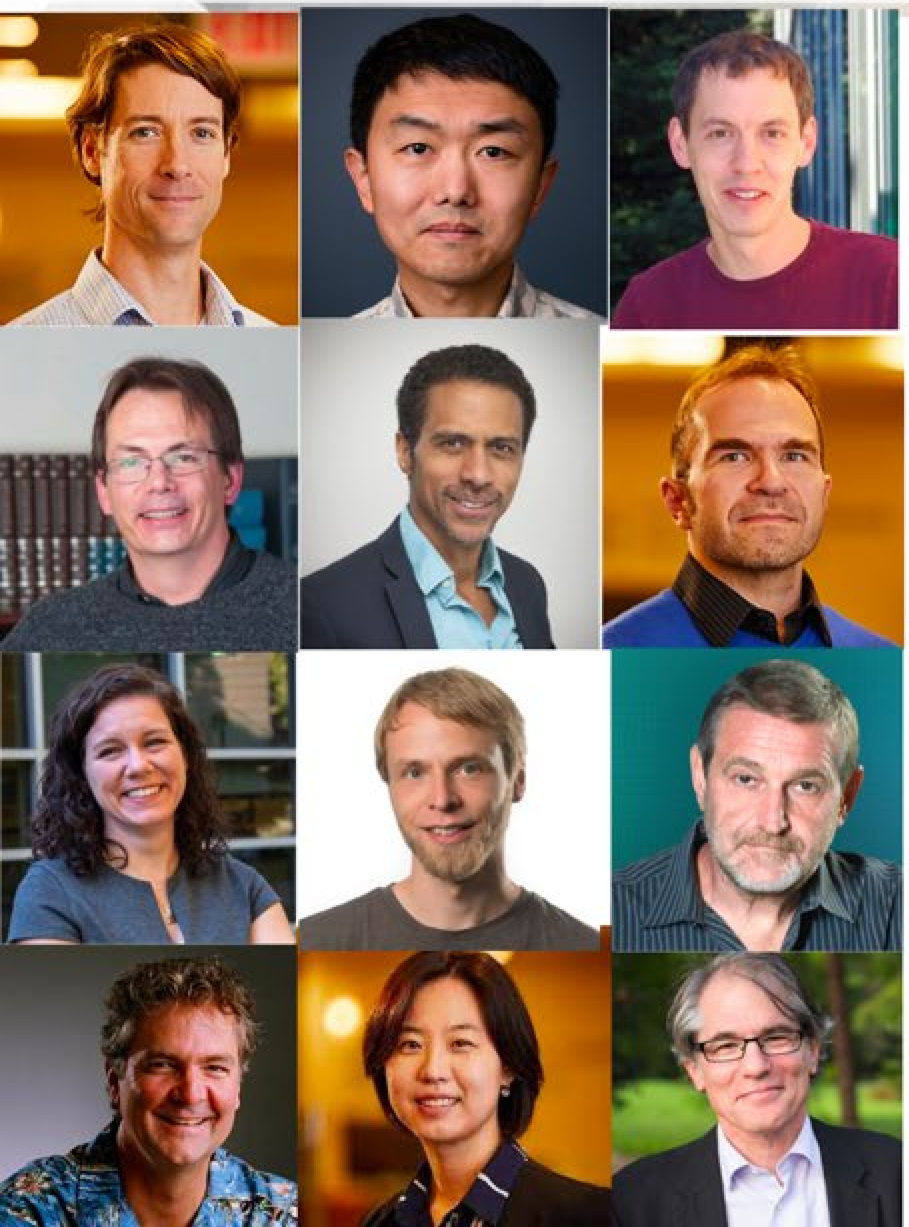
Overview

Population Sampling and Representation (Phase 1): We are representing >99.9% common SNVs (1%) in the 1000 Genomes lines. (**Phase 2)** We are in position for perspective recruitment of remaining 150 individuals (BioMe)

Sequence Technology and Production: Highly efficient multi-center production effort, automated assembly and quality assessment



Innovation in Long Read Assembly Methods



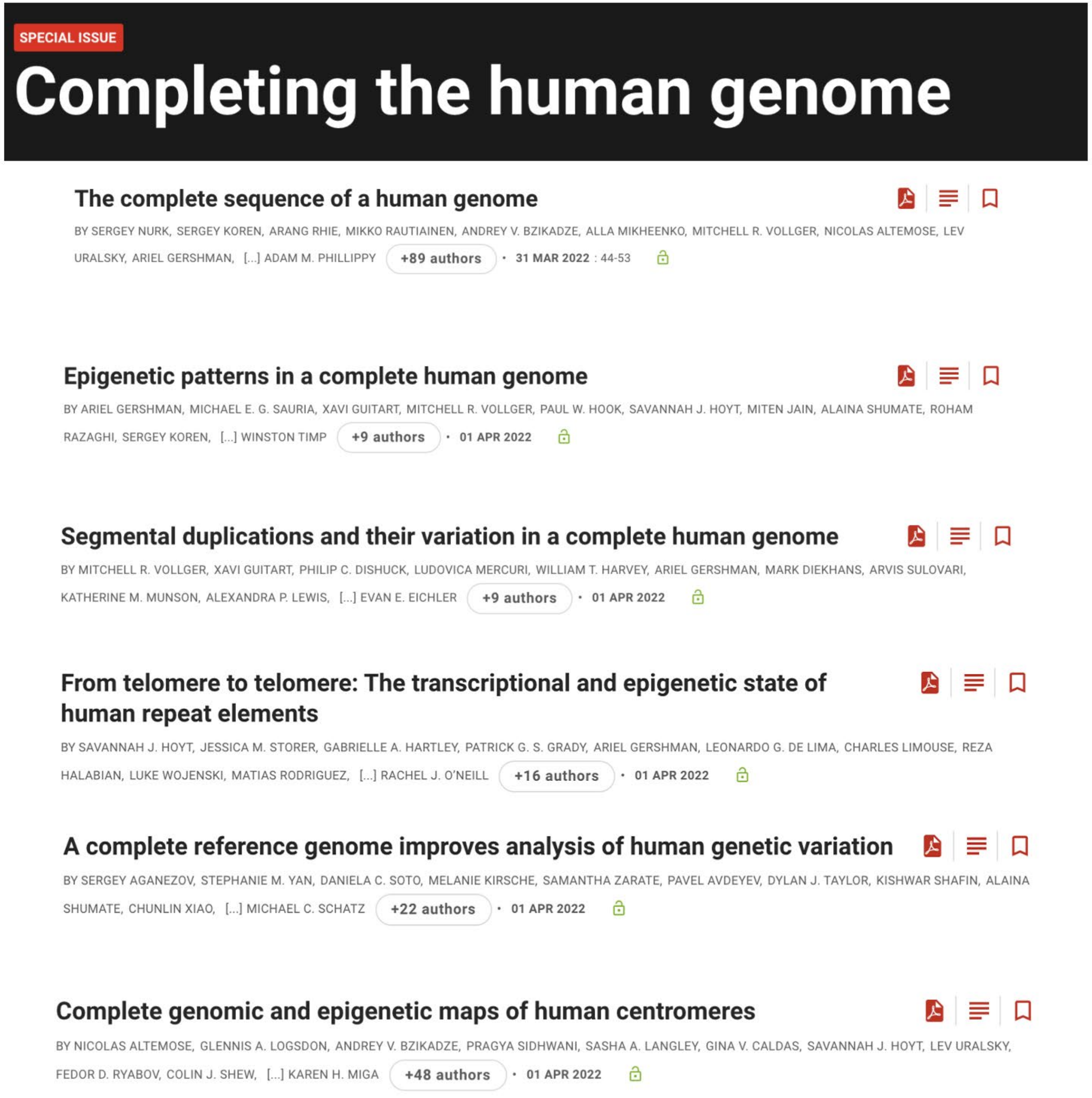
We have assembled the leaders in long-read assembly methods, with an emphasis in researchers involved in finishing and repeat assemblies.

T2T
TELOMERE-TO-TELOMERE CONSORTIUM

VGP VERTEBRATE GENOMES PROJECT
A PROJECT OF THE G10K CONSORTIUM

HGSVC

Earth's heart of iron begins to yield its secrets p. 18
Microglia in chronic pain recovery and relapse pp. 33 & 86
Particle acceleration in a nova explosion p. 77
Science
$15
1 APRIL 2022
SPECIAL ISSUE
science.org
AAAS
FILLING THE GAPS
Closing in on a complete human genome p. 42


SPECIAL ISSUE

Completing the human genome

The complete sequence of a human genome

BY SERGEY NURK, SERGEY KOREN, ARANG RHIE, MIKKO RAUTIAINEN, ANDREY V. BZIKADZE, ALLA MIKHEENKO, MITCHELL R. VOLLGER, NICOLAS ALTEMOSE, LEV URALSKY, ARIEL GERSHMAN, [...] ADAM M. PHILLIPPY +89 authors • 31 MAR 2022 : 44-53

Epigenetic patterns in a complete human genome

BY ARIEL GERSHMAN, MICHAEL E. G. SAURIA, XAVI GUITART, MITCHELL R. VOLLGER, PAUL W. HOOK, SAVANNAH J. HOYT, MITEN JAIN, ALAINA SHUMATE, ROHAM RAZAGHI, SERGEY KOREN, [...] WINSTON TIMP +9 authors • 01 APR 2022

Segmental duplications and their variation in a complete human genome

BY MITCHELL R. VOLLGER, XAVI GUITART, PHILIP C. DISHUCK, LUDOVICA MERCURI, WILLIAM T. HARVEY, ARIEL GERSHMAN, MARK DIEKHANS, ARVIS SULOVARI, KATHERINE M. MUNSON, ALEXANDRA P. LEWIS, [...] EVAN E. EICHLER +9 authors • 01 APR 2022

From telomere to telomere: The transcriptional and epigenetic state of human repeat elements

BY SAVANNAH J. HOYT, JESSICA M. STORER, GABRIELLE A. HARTLEY, PATRICK G. S. GRADY, ARIEL GERSHMAN, LEONARDO G. DE LIMA, CHARLES LIMOUSE, REZA HALABIAN, LUKE WOJENSKI, MATIAS RODRIGUEZ, [...] RACHEL J. O'NEILL +16 authors • 01 APR 2022

A complete reference genome improves analysis of human genetic variation

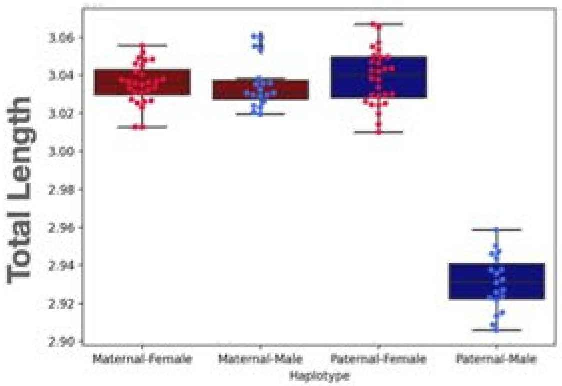
BY SERGEY AGANEZOV, STEPHANIE M. YAN, DANIELA C. SOTO, MELANIE KIRSCHE, SAMANTHA ZARATE, PAVEL AVDEYEV, DYLAN J. TAYLOR, KISHWAR SHAFIN, ALAINA SHUMATE, CHUNLIN XIAO, [...] MICHAEL C. SCHATZ +22 authors • 01 APR 2022

Complete genomic and epigenetic maps of human centromeres

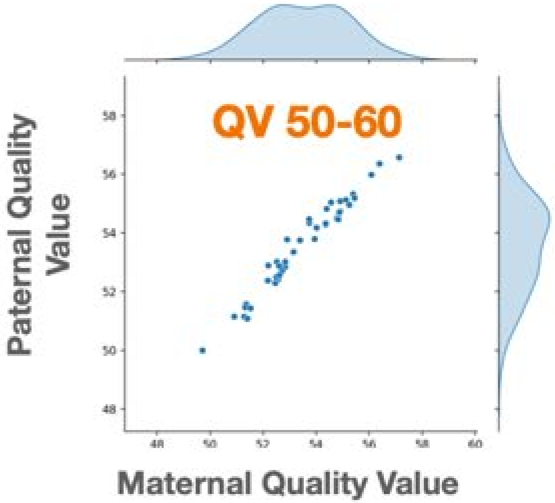
BY NICOLAS ALTEMOSE, GLENNIS A. LOGSDON, ANDREY V. BZIKADZE, PRAGYA SIDHWANI, SASHA A. LANGLEY, GINA V. CALDAS, SAVANNAH J. HOYT, LEV URALSKY, FEDOR D. RYABOV, COLIN J. SHEW, [...] KAREN H. MIGA +48 authors • 01 APR 2022

Automated assembly standards: High-quality References

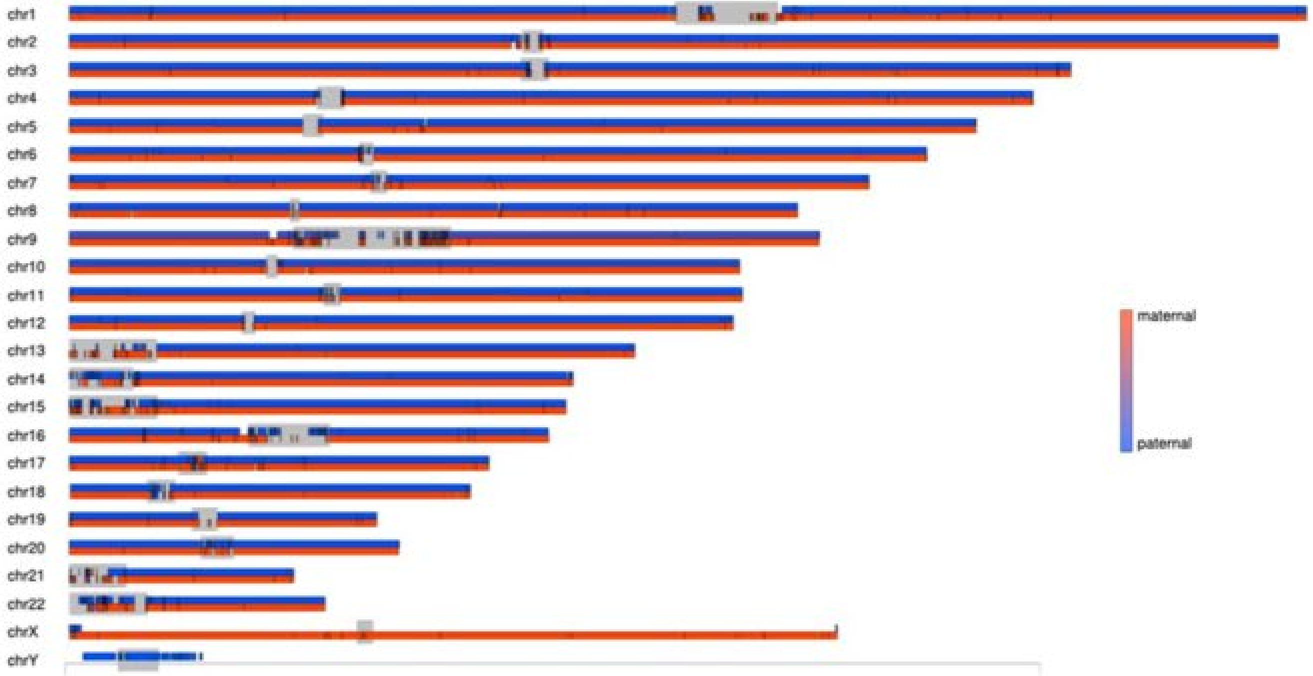
We tested the current best practices in sequencing technologies and automated assembly algorithms on one human sample, HG002, an openly- consented Ashkenazi individual from the Personal Genome Project



T2T-CHM13
3.055 Gb

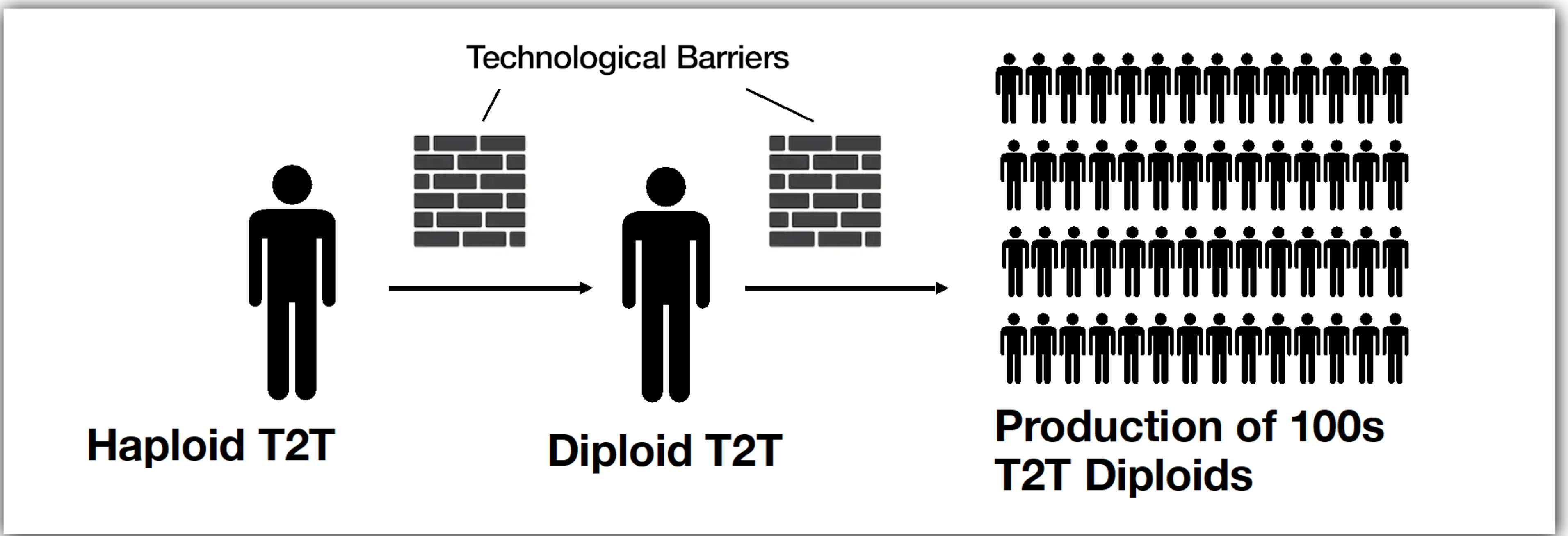


QV 50-60



Near complete chromosome scaffolds between HPRC-HG002 maternal and paternal assemblies.

One genome is not enough….



Technological Barriers

Haploid T2T

Diploid T2T

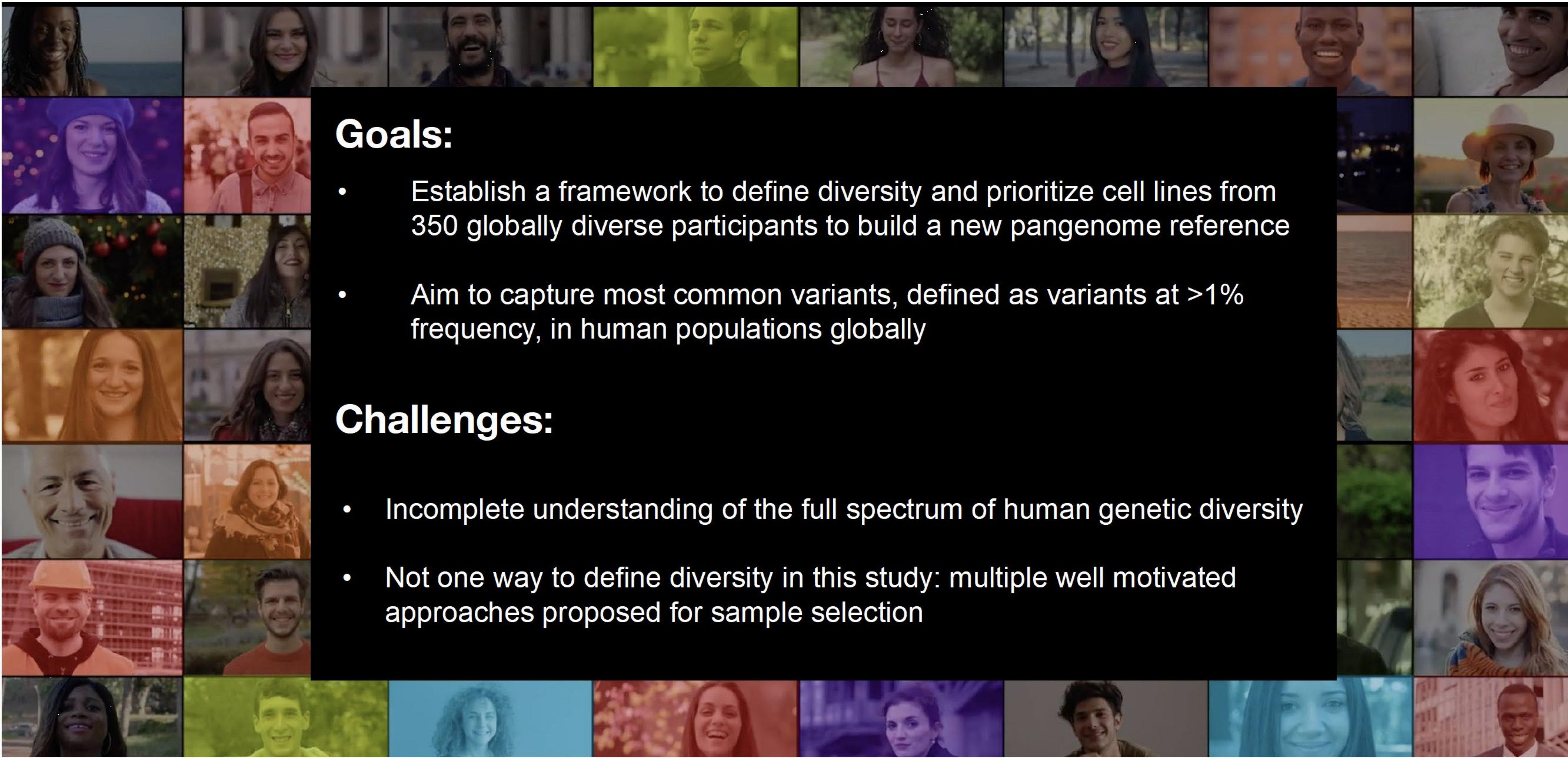
Production of 100s T2T Diploids

TELOMERE-TO-TELOMERE CONSORTIUM

TOWARDS A COMPLETE REFERENCE OF HUMAN GENOME DIVERSITY

HUMAN PANGENOME

Population Sampling and Representation



Goals:

- Establish a framework to define diversity and prioritize cell lines from 350 globally diverse participants to build a new pangenome reference

- Aim to capture most common variants, defined as variants at >1% frequency, in human populations globally

Challenges:

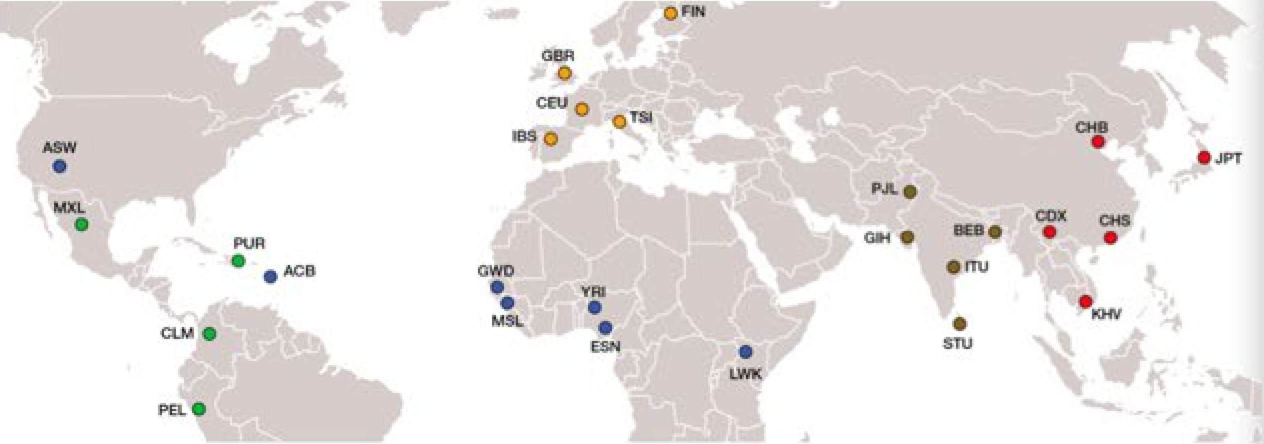
- Incomplete understanding of the full spectrum of human genetic diversity

- Not one way to define diversity in this study: multiple well motivated approaches proposed for sample selection

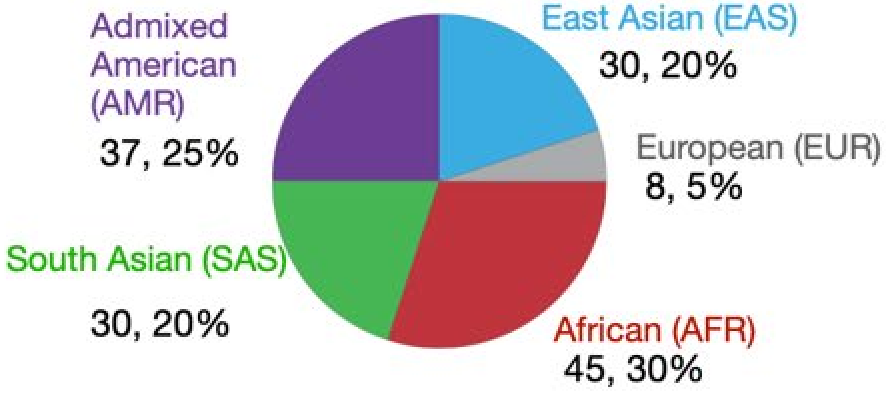
Phase 1 HPRC Sample Selection

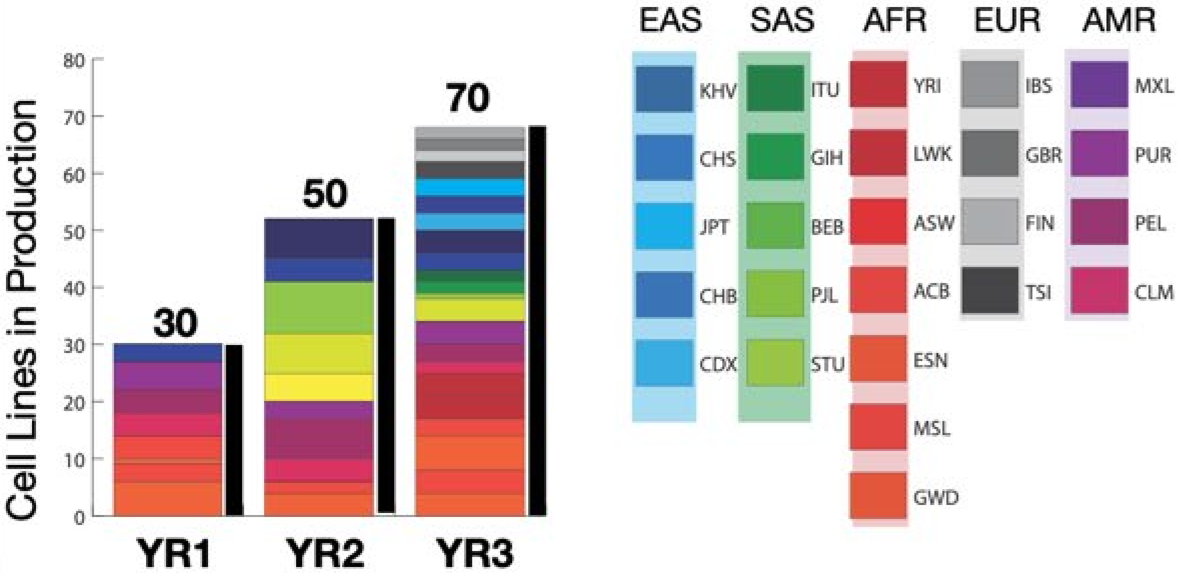
A THOUSAND GENOMES
Pilot studies prepare the way for population-scale gene sequencing

- ☑ Cover genetic and geographic diversity
- ☑ Availability of low passage cell lines
- ☑ Availability of trios/parental data (YR1-2)



Phase 1 Production: 150 total lines



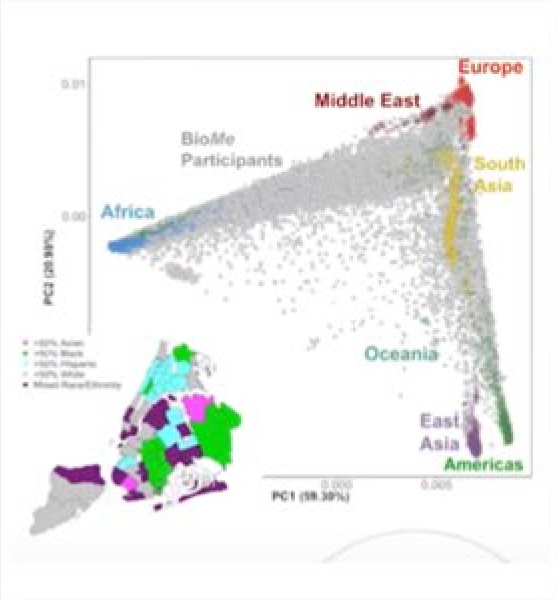


Mount Sinai The Charles Bronfman Institute for Personalized Medicine

Bio*Me* Biobank

Your history is our future

Eimear Kenny, PI



Mount Sinai's BioMe BioBank:

>70,000 participants from 160 countries, with sequencing data for more than 40,000 participants.

Washington University, St. Louis Recruitment Center:

New recruitment from African American communities

Phase 2: Prioritization of New Participants

- **Model 1:** Maximizing common variant diversity

Leveraging sequence data, iteratively select samples that maximize common variant coverage in an out-of-sample dataset

- **Model 2**: Maximizing genetic divergence

Using sequence/array data, plot PCA and select participants to represent the continuum of genetic dissimilarity in principle component (PC) space

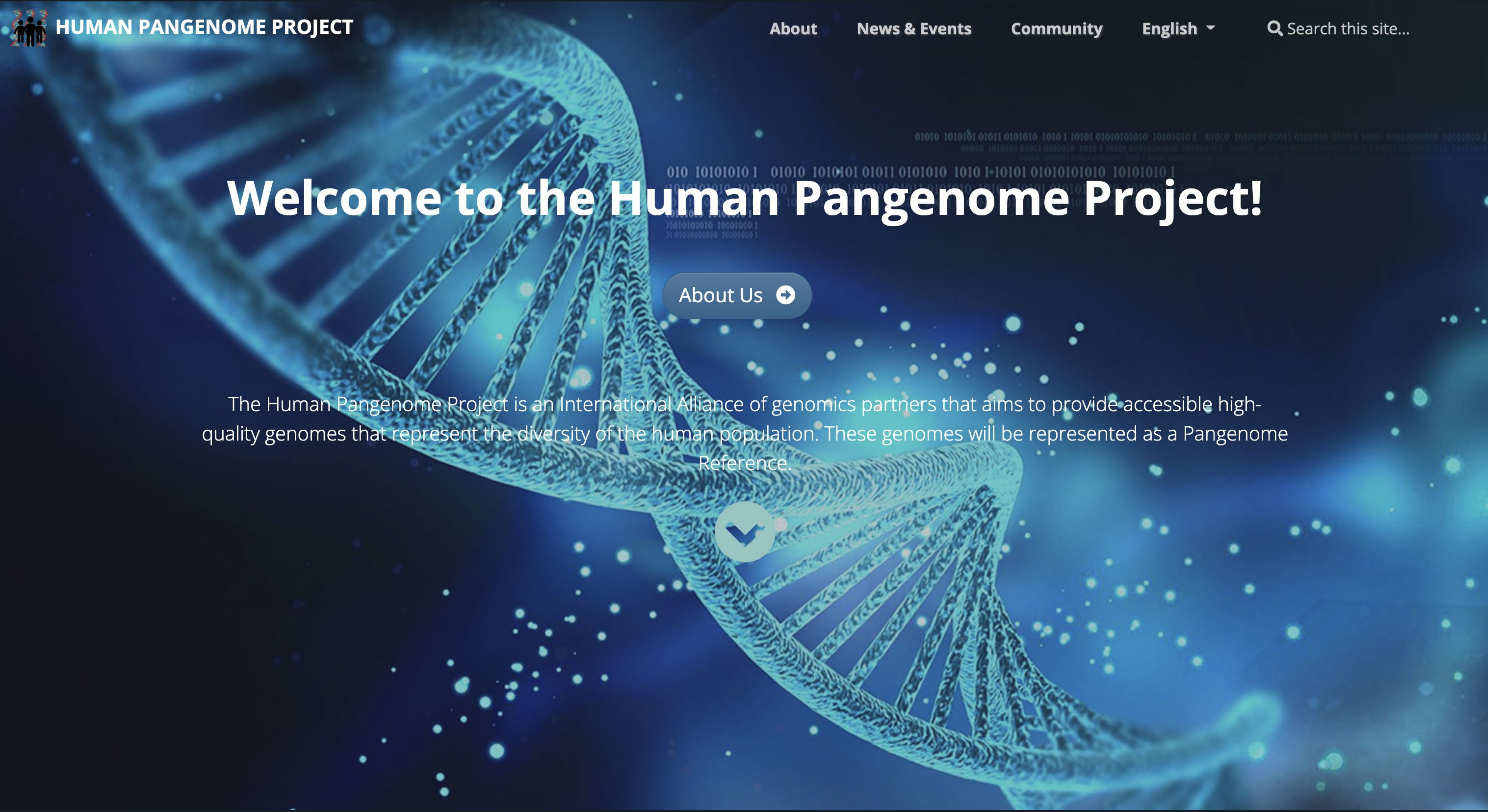
- **Model 3**: Targeting underrepresented populations based on self-reporting/ geographical data

Using country-of-origin data, designate countries/subcontinental regions of interest to the project, and select participants from those regions

Partnership with ELSI Team for Consent and Outreach

Embedded Ethics Team

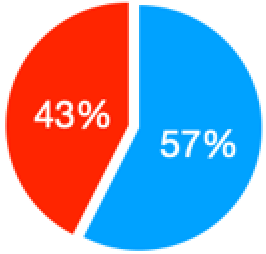
- Formal Review of Consent Language (compatibility with 1000 genomes consent)
- Design and review of outreach materials for prospective recruitment
- Alignment with internal ethics/genomic review board at BioMe
- Review of the use of external IRB (BRANY)



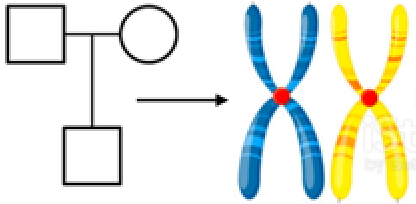
Welcome to the Human Pangenome Project!

The Human Pangenome Project is an International Alliance of genomics partners that aims to provide accessible high-quality genomes that represent the diversity of the human population. These genomes will be represented as a Pangenome Reference.

Improving Population Sampling and Representation



Continue reference production of 1000 genomes (200/350)

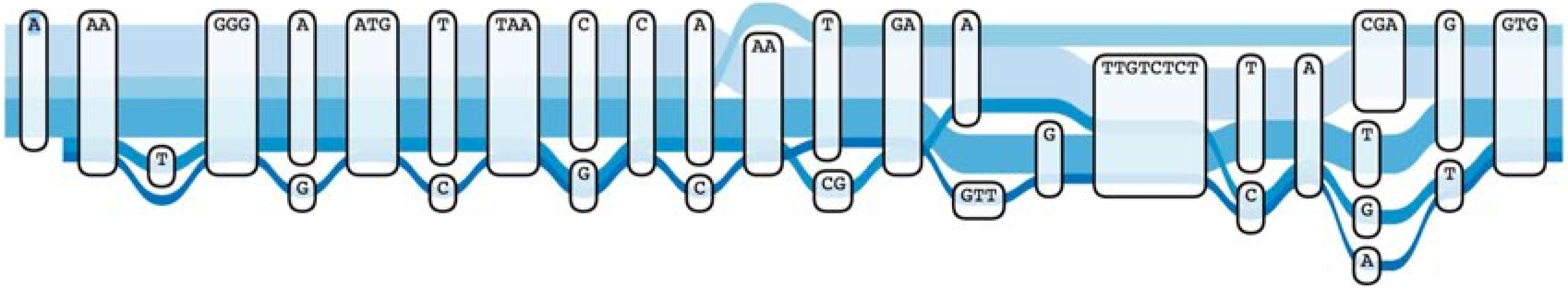


Move away from dependency on trios (priority based on genomic diversity) Establish new assembly methods that do not rely on parental data

Launch new recruitment efforts and establish 100 new LCLs (NHGRI Collection: Human Pangenome) outside of 1000 Genome Cell lines

Establish new collaborations and international partnerships

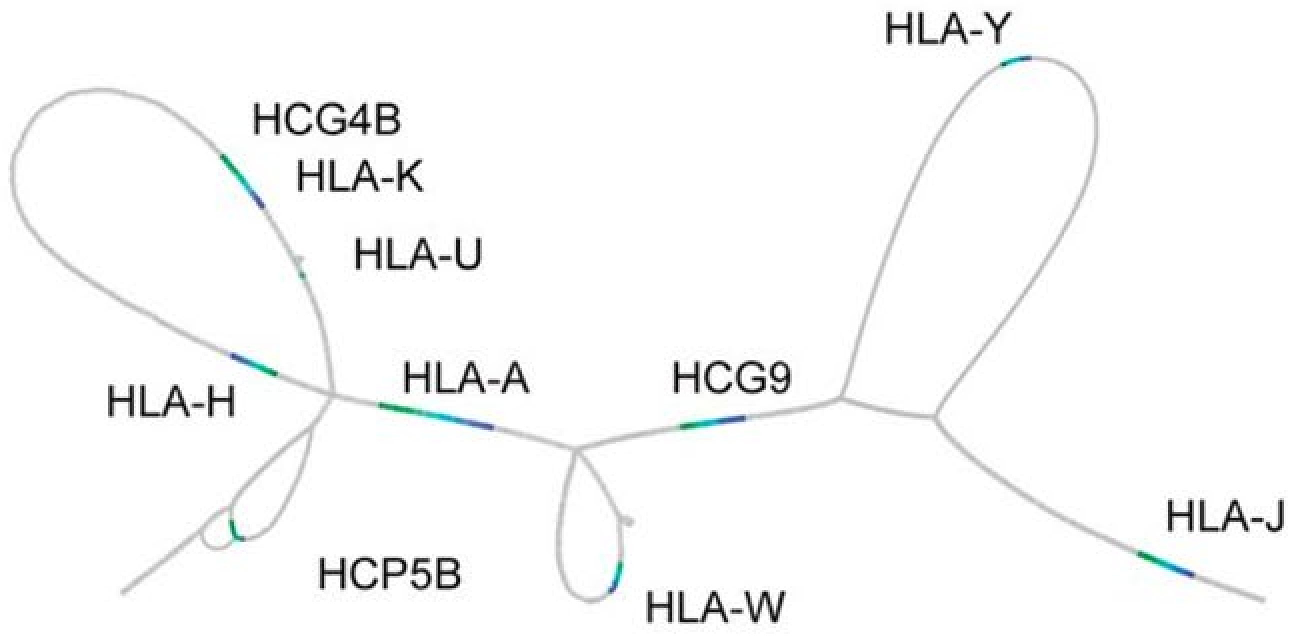
The Human Pangenome



Alignments of high-quality assemblies were performed using three different methods, pioneered by our team:

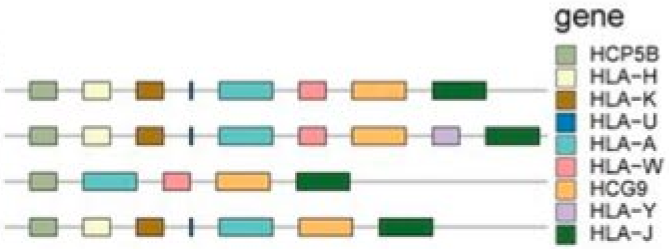
Minigraph (Li et al., 2020),
Minigraph-Cactus (MC)
PanGenome Graph Builder (PGGB).

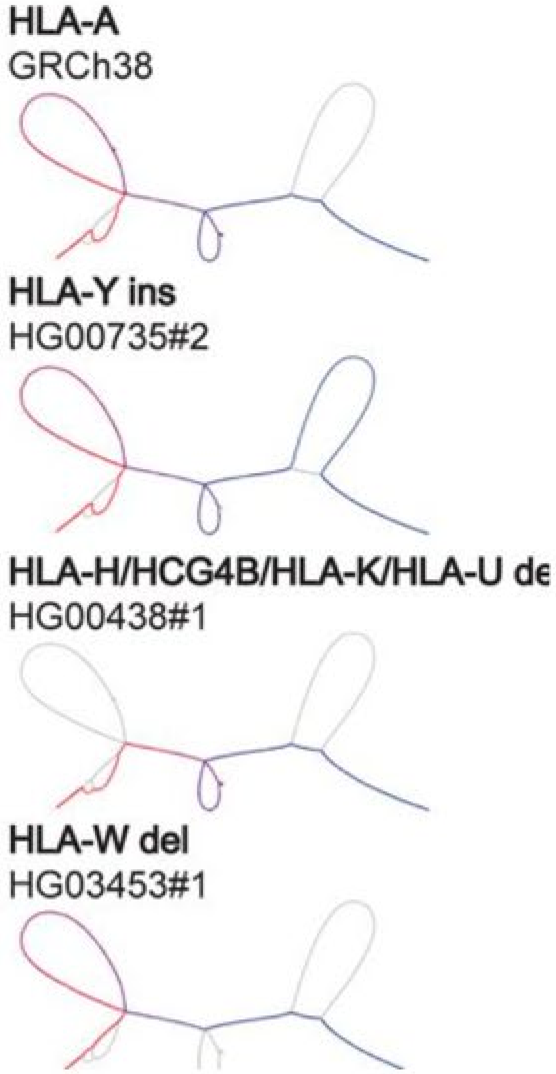
HLA-A: Aligning Complex Pangenome Loci



F

Count	Frequency	Haplotype name
57	0.63	HLA-A
25	0.28	HLA-Y ins
7	0.08	HLA-H/HCG4B/HLA-K/HLA-U del
1	0.01	HLA-W del





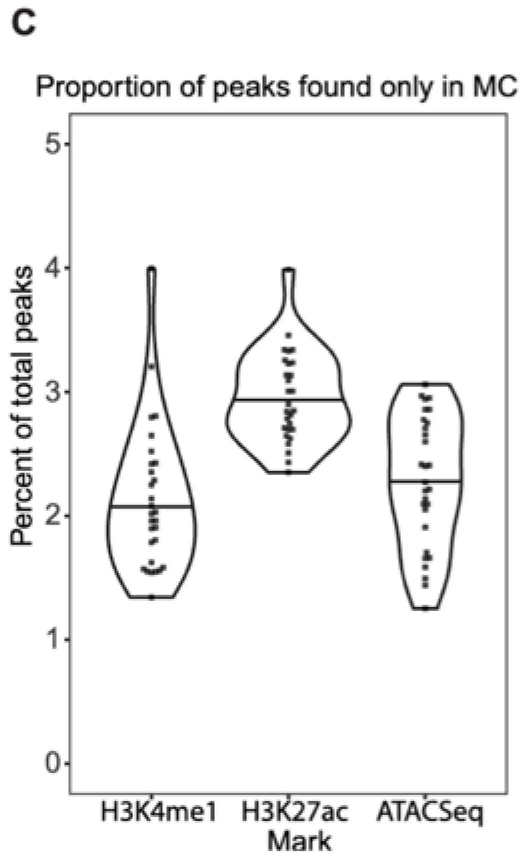
HLA-A
GRCh38

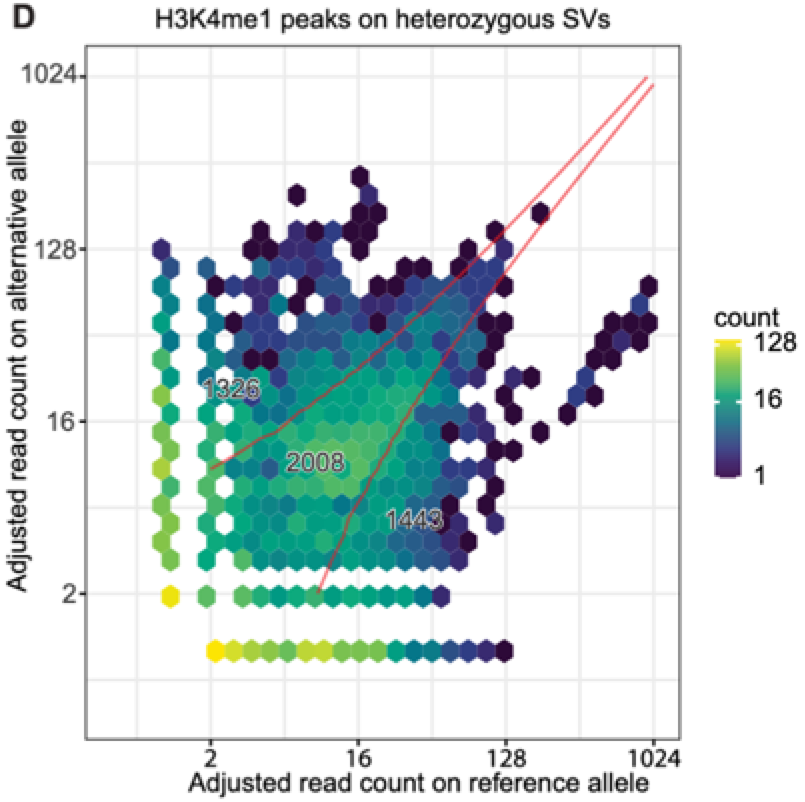
HLA-Y ins
HG00735#2

HLA-H/HCG4B/HLA-K/HLA-U de
HG00438#1

HLA-W del
HG03453#1

Applications





A pangenome approach captures annotation missing in a linear reference

A pangenome approach allows annotation at heterozygous structural variant sites

Remaining challenges

- Sample selection and resource development
    - Ensuring this project is truly international in scope
    - Ensuring that we common (> 1% MAF) variation is included
    - Development of usable cell lines (iPSCs, LCLs, etc) for experimental work
- Pangenome implementation
    - Will one graph representation rule them all, or do we need different graphs for different applications?
- Pangenome adoption
    - In a world where many people still use GRCh37, how do we encourage adoption of this resource?

Acknowledgements

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

David Haussler
Benedict Paten
Karen Miga
Ed Green
Mark Akeson
Adam Novack

Miten Jain
Hugh Olsen
Erik Garrison
Marina Haukness
Mark Akeson
Jean Monlong
Glenn Hickey

Adirna Fuller
Tony Tsung Yu Lu
Xian Chang
Trevor Pesout
Ryan Lorig-Roach
Charles Markello
Jonas Sibbesen

Melissa Meredith
Kishwar Shafin
Jouni Siren
Jordan Eizenga
Beth Sheets
Julian Lucas
Brian Hannafious

THE GENOME INSTITUTE at Washington University

Ting Wang
Lucinda Fulton
Sarah Cody
Robert Fulton
Heather Lawson

Wen-Wei Liao
Nathan Stitziel
Milinn Kremitizki
Haley Abel
Eddie Belter

Derek Albracht
Chad Tomlinson
Allison Regier
Chris Markovic
Tina Lindsay

EMBL-EBI

Paul Flicek
Susan Fairley
Daniel Zerbino

Icahn School of Medicine at Mount Sinai

Eimear Kenny
Vimi Desai

NIH National Human Genome Research Institute

Adam Felsenfeld
Mike Smith
Carolyn Hutter
Taylorlyn Stephan
Heidi Sofia
Nicole Lockhart
Sara Currin

Adam Phillippy
Sergey Koren
Arang Rhie
Chirag Jain
Baergen Schultz
Kris Wetterstrand

Genome Sciences UNIVERSITY OF WASHINGTON

Evan Eichler
Katy Munson
Mitchell Vollger
Pete Audano
David Porubksy
Arvis Sulovari

CBG Max Planck Institute of Molecular Cell Biology and Genetics

Gene Myers

Yale University

Ira Hall
Wen-Wei Liao
Shuangjia Lu

wellcome sanger institute

Kerstin Howe

UNIVERSITY OF CAMBRIDGE

Richard Durbin

NCBI National Center for Biotechnology Information

Valerie Schneider
Terence Murphy
Paul Kitts
Chunlin Xiao
Francoise Thibaud-Nissent

Barbara Koenig (UCSF)
Nanibaa' Garrison (UCLA)
Bob Cook-Deegan (ASU)
Alice Popejoy (UC Davis)

HARVARD MEDICAL SCHOOL

Heng Li
Shilpa Garg
Haoyu Cheng
Xiaowen Feng

AnVIL

CORIELL INSTITUTE FOR MEDICAL RESEARCH
DECODING THE GENOME

Alissa Resch
Brittany Kerr
Brittney Martinez
Ellen Kellly

THE ROCKEFELLER UNIVERSITY
Science for the benefit of humanity

Erich Jarvis
Olivier Fedrigo
Giulio Formenti
Sadye Paez
Lauren Shalmiyev

TOWARDS A COMPLETE REFERENCE OF HUMAN GENOME DIVERSITY

HUMAN PANGENOME

NIH/NHGRI Funding:
U01HG010971, U41HG010972

NIST

Justin Zook

Company Partnerships

aws
Google DeepVariant
illumina
ARIMA GENOMICS
Dovetail GENOMICS
PACBIO
Oxford NANOPORE Technologies
HUDSONALPHA INSTITUTE FOR BIOTECHNOLOGY