Breakout Session 5:

“T2T-omics” at scale: Improving our understanding of human genetic variation using AnVIL

Professor Michael Schatz (Moderator)
Bloomberg Distinguished Professor of Computer Science and Biology,
Johns Hopkins University

“T2T-omics” at scale: Improving our understanding of human genetic variation using AnVIL

Michael Schatz
Johns Hopkins University

2024 NIH/ODSS Cloud Program PI Meeting
January 18, 2024

T2T Powered by AnVIL!

TELOMERE-TO-TELOMERE CONSORTIUM

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

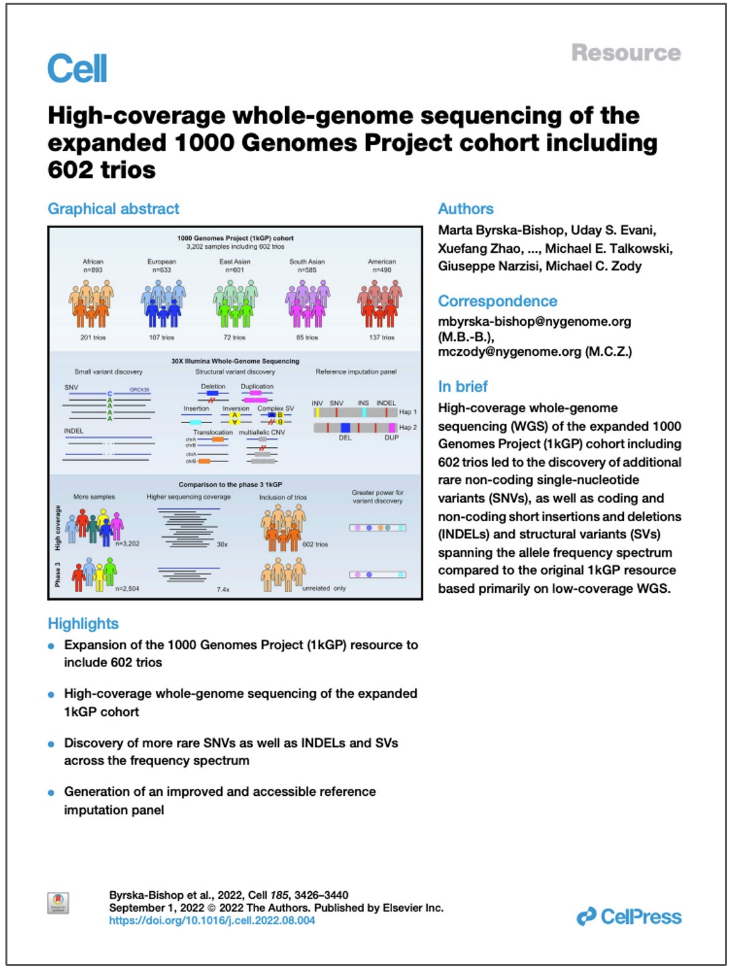
Science

$15
1 APRIL 2022
SPECIAL ISSUE
science.org

AAAS

FILLING THE GAPS

Closing in on a complete human genome p. 42



Cell

Resource

High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios

Graphical abstract

Highlights

- Expansion of the 1000 Genomes Project (1kGP) resource to include 602 trios
- High-coverage whole-genome sequencing of the expanded 1kGP cohort
- Discovery of more rare SNVs as well as INDELs and SVs across the frequency spectrum
- Generation of an improved and accessible reference imputation panel

Authors

Marta Byrska-Bishop, Uday S. Evani, Xuefang Zhao, ..., Michael E. Talkowski, Giuseppe Narzisi, Michael C. Zody

Correspondence

mbyrska-bishop@nygenome.org (M.B.-B.), mczody@nygenome.org (M.C.Z.)

In brief

High-coverage whole-genome sequencing (WGS) of the expanded 1000 Genomes Project (1kGP) cohort including 602 trios led to the discovery of additional rare non-coding single-nucleotide variants (SNVs), as well as coding and non-coding short insertions and deletions (INDELs) and structural variants (SVs) spanning the allele frequency spectrum compared to the original 1kGP resource based primarily on low-coverage WGS.

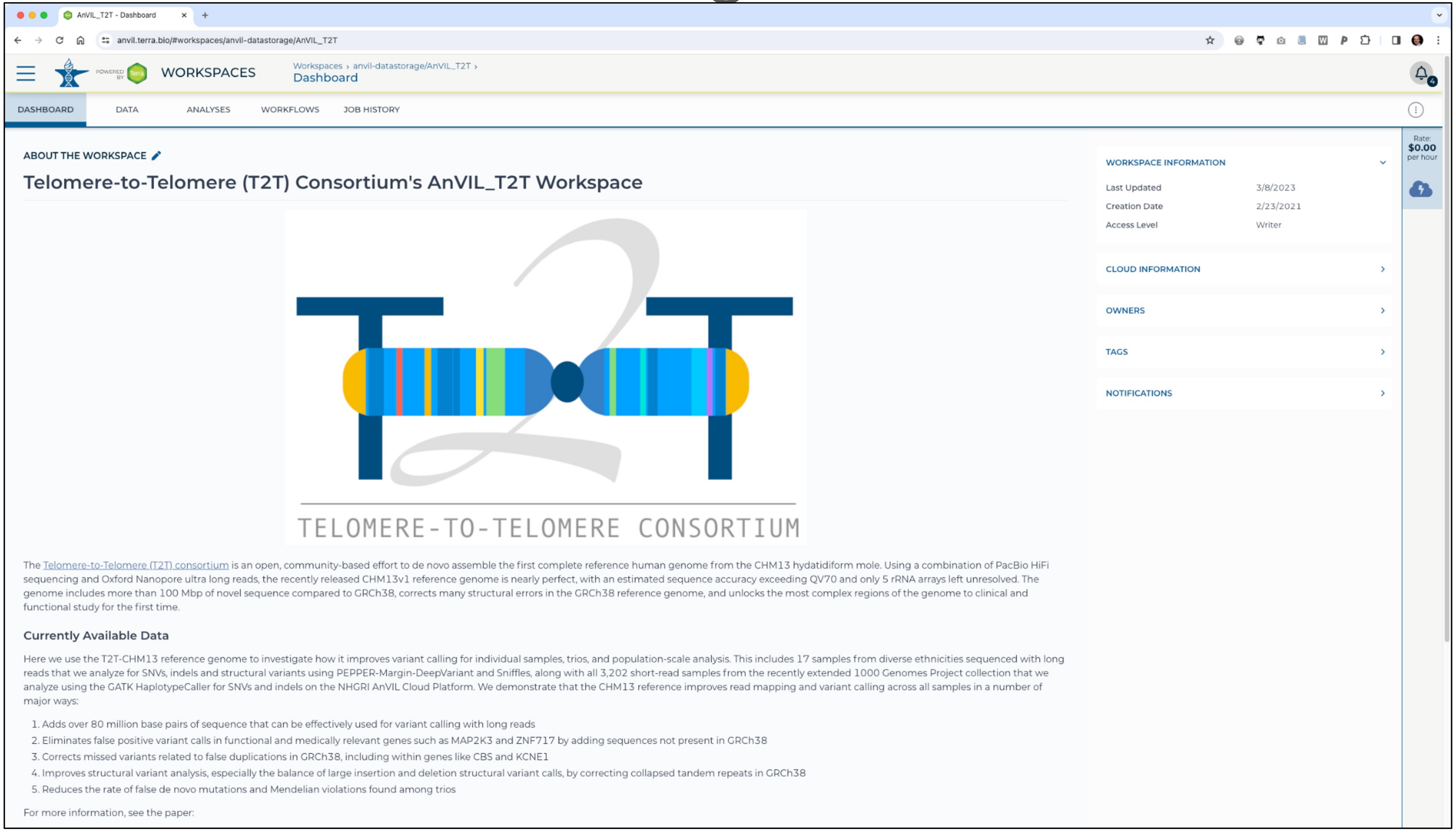
Byrska-Bishop et al., 2022, Cell *185*, 3426–3440
September 1, 2022 © 2022 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cell.2022.08.004

CellPress

3202 samples from 26 populations

3202 samples x 30Gb = 96Tb input data

T2T on AnVIL


AnVIL_T2T - Dashboard
anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_T2T
POWERED BY
WORKSPACES
Workspaces › anvil-datastorage/AnVIL_T2T ›
Dashboard
DASHBOARD
DATA
ANALYSES
WORKFLOWS
JOB HISTORY
Rate:
$0.00
per hour
TELOMERE-TO-TELOMERE CONSORTIUM

ABOUT THE WORKSPACE

Telomere-to-Telomere (T2T) Consortium's AnVIL_T2T Workspace

WORKSPACE INFORMATION

Last Updated	3/8/2023
Creation Date	2/23/2021
Access Level	Writer

CLOUD INFORMATION ›

OWNERS ›

TAGS ›

NOTIFICATIONS ›

The Telomere-to-Telomere (T2T) consortium is an open, community-based effort to de novo assemble the first complete reference human genome from the CHM13 hydatidiform mole. Using a combination of PacBio HiFi sequencing and Oxford Nanopore ultra long reads, the recently released CHM13v1 reference genome is nearly perfect, with an estimated sequence accuracy exceeding QV70 and only 5 rRNA arrays left unresolved. The genome includes more than 100 Mbp of novel sequence compared to GRCh38, corrects many structural errors in the GRCh38 reference genome, and unlocks the most complex regions of the genome to clinical and functional study for the first time.

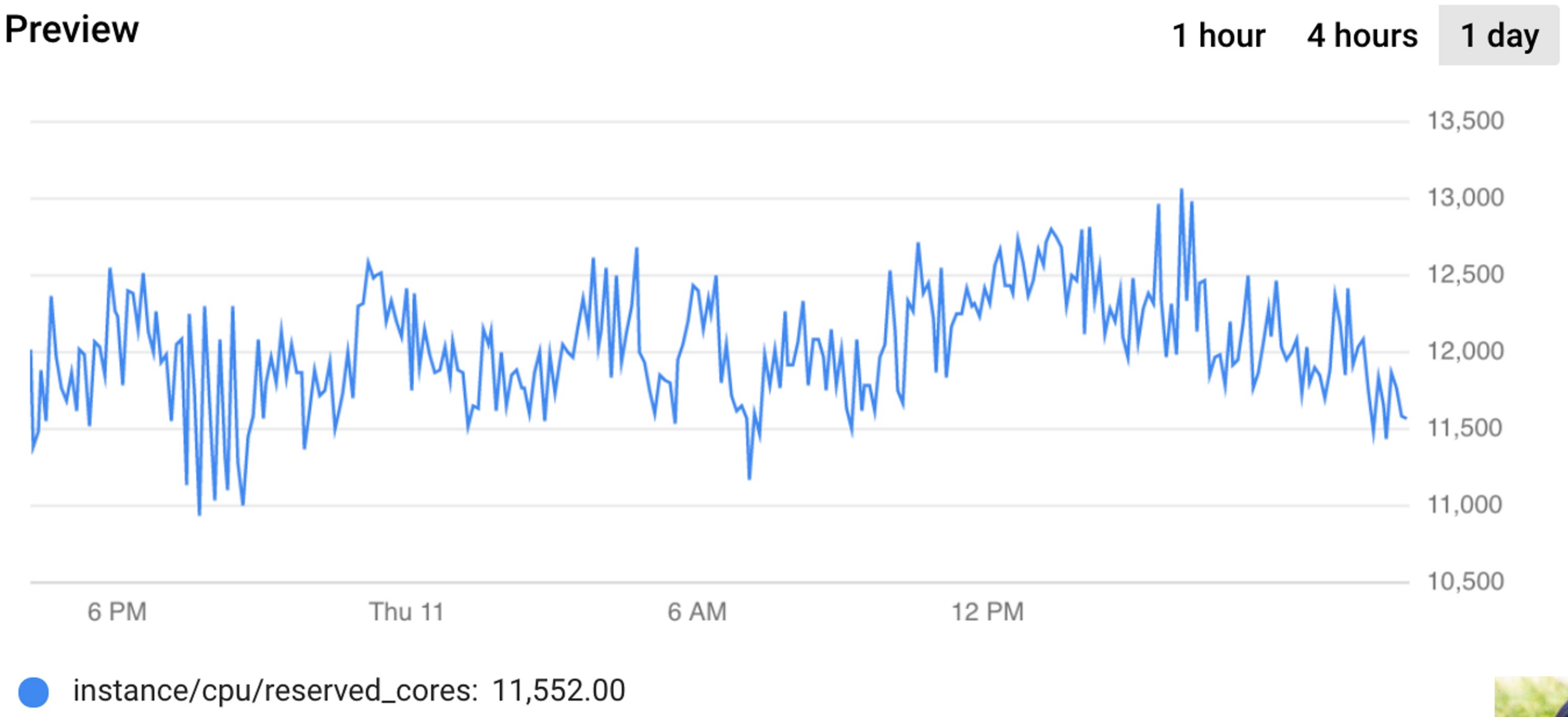
Currently Available Data

Here we use the T2T-CHM13 reference genome to investigate how it improves variant calling for individual samples, trios, and population-scale analysis. This includes 17 samples from diverse ethnicities sequenced with long reads that we analyze for SNVs, indels and structural variants using PEPPER-Margin-DeepVariant and Sniffles, along with all 3,202 short-read samples from the recently extended 1000 Genomes Project collection that we analyze using the GATK HaplotypeCaller for SNVs and indels on the NHGRI AnVIL Cloud Platform. We demonstrate that the CHM13 reference improves read mapping and variant calling across all samples in a number of major ways:

1. Adds over 80 million base pairs of sequence that can be effectively used for variant calling with long reads
2. Eliminates false positive variant calls in functional and medically relevant genes such as MAP2K3 and ZNF717 by adding sequences not present in GRCh38
3. Corrects missed variants related to false duplications in GRCh38, including within genes like CBS and KCNE1
4. Improves structural variant analysis, especially the balance of large insertion and deletion structural variant calls, by correcting collapsed tandem repeats in GRCh38
5. Reduces the rate of false de novo mutations and Mendelian violations found among trios

For more information, see the paper:

Core usage over 24 hours

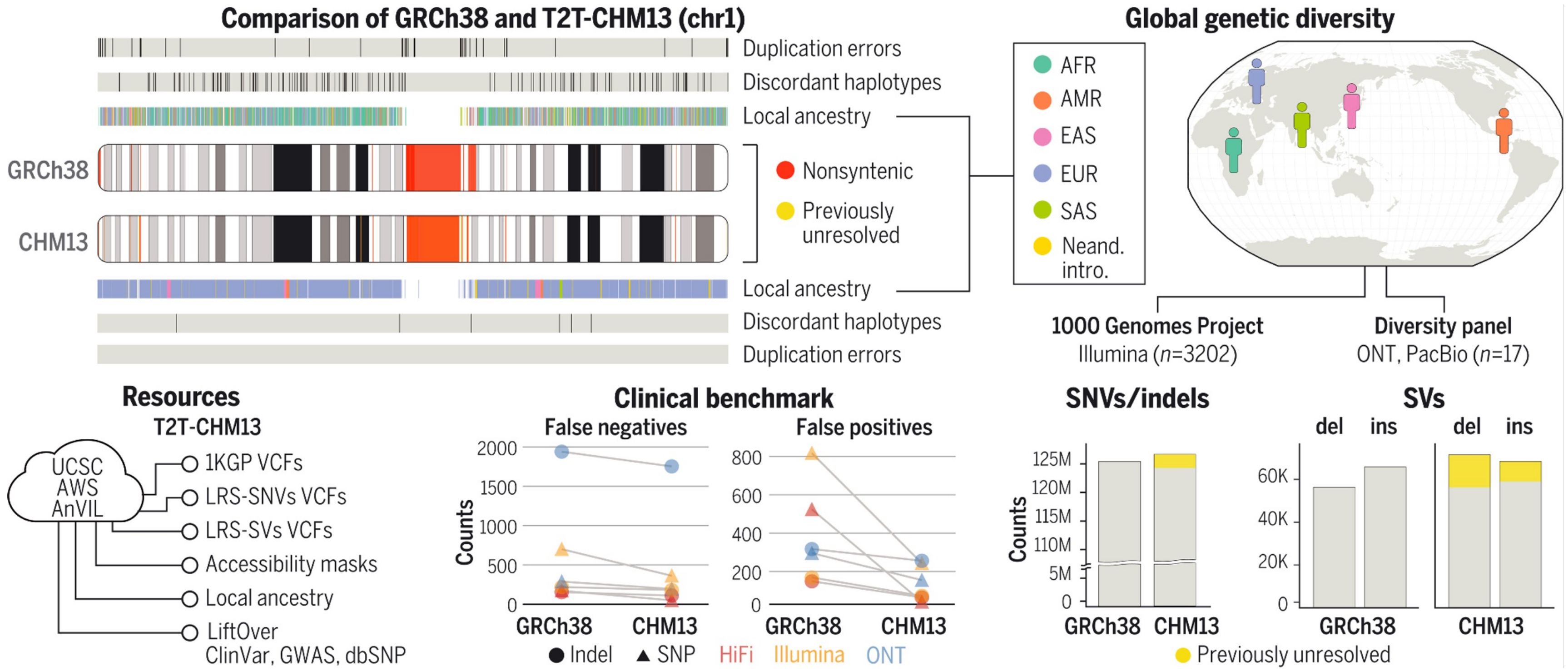


● instance/cpu/reserved_cores: 11,552.00

https://dockstore.org/workflows/github.com/schatzlab/t2t-variants/T2T_alignment

Samantha Zarate

T2T Genomes Powered by AnVIL



Sergey Aganezov

Stephanie Yan

Daniela Soto

Melanie Kirsche

Samantha Zarate

https://anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_T2T

A complete reference genome improves analysis of human genetic variation

Aganezov, S*, Yan, SM*, Soto, DC*, Kirsche, M*, Zarate, S*, *et al.* (2022) *Science*. doi: 10.1126/science.abl3533

Earth's heart of iron begins to yield its secrets p. 18

Microglia in chronic pain recovery and relapse pp. 33 & 86

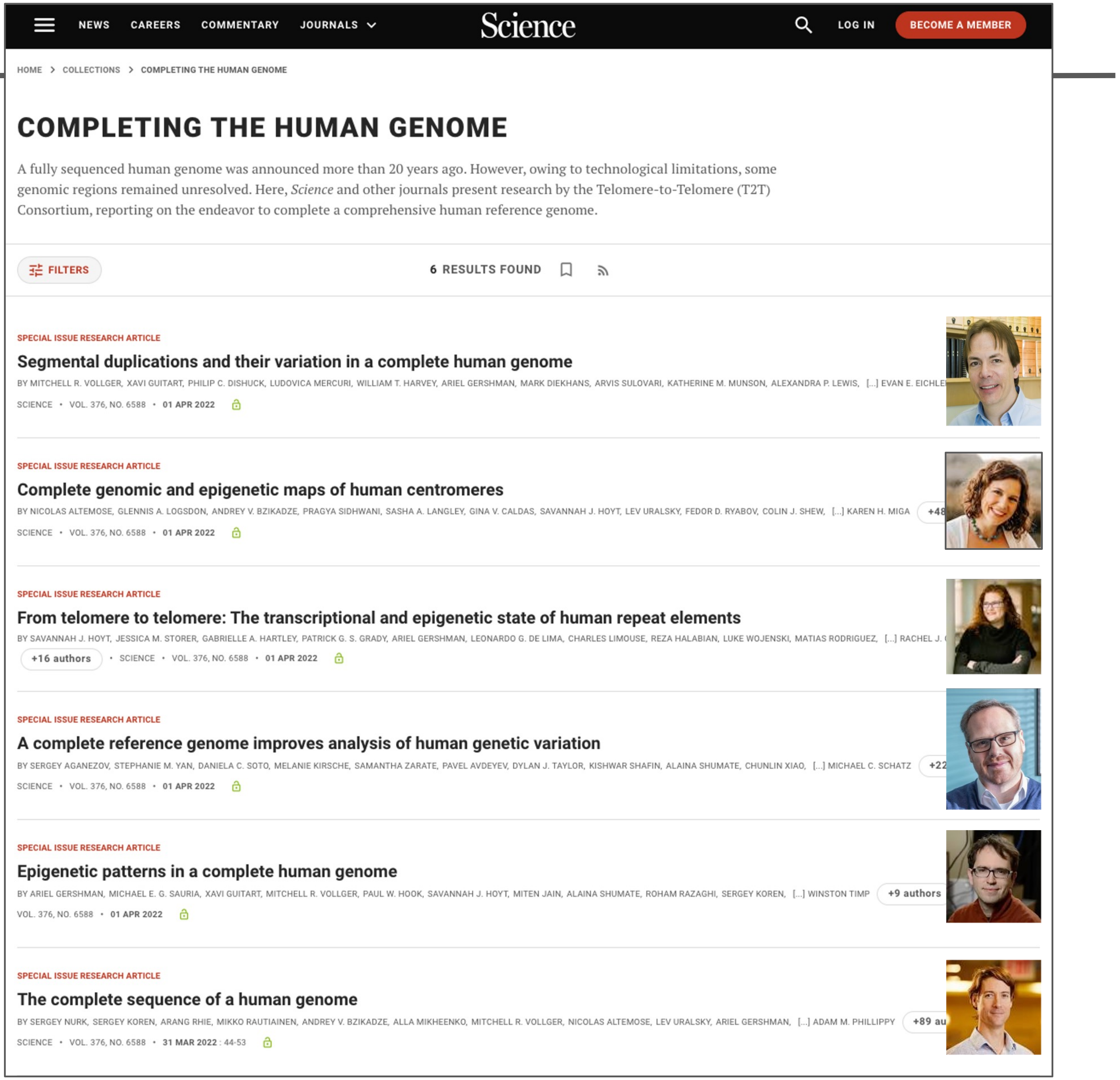
Particle acceleration in a nova explosion p. 77

Science

$15
1 APRIL 2022
SPECIAL ISSUE
science.org
AAAS

FILLING THE GAPS

Closing in on a complete human genome p. 42



COMPLETING THE HUMAN GENOME

A fully sequenced human genome was announced more than 20 years ago. However, owing to technological limitations, some genomic regions remained unresolved. Here, *Science* and other journals present research by the Telomere-to-Telomere (T2T) Consortium, reporting on the endeavor to complete a comprehensive human reference genome.

SPECIAL ISSUE RESEARCH ARTICLE

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. EICHLE

SCIENCE • VOL. 376, NO. 6588 • 01 APR 2022

SPECIAL ISSUE RESEARCH ARTICLE

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

SCIENCE • VOL. 376, NO. 6588 • 01 APR 2022

SPECIAL ISSUE RESEARCH ARTICLE

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.

+16 authors • SCIENCE • VOL. 376, NO. 6588 • 01 APR 2022

SPECIAL ISSUE RESEARCH ARTICLE

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

SCIENCE • VOL. 376, NO. 6588 • 01 APR 2022

SPECIAL ISSUE RESEARCH ARTICLE

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

VOL. 376, NO. 6588 • 01 APR 2022

SPECIAL ISSUE RESEARCH ARTICLE

The complete sequence of a human genome

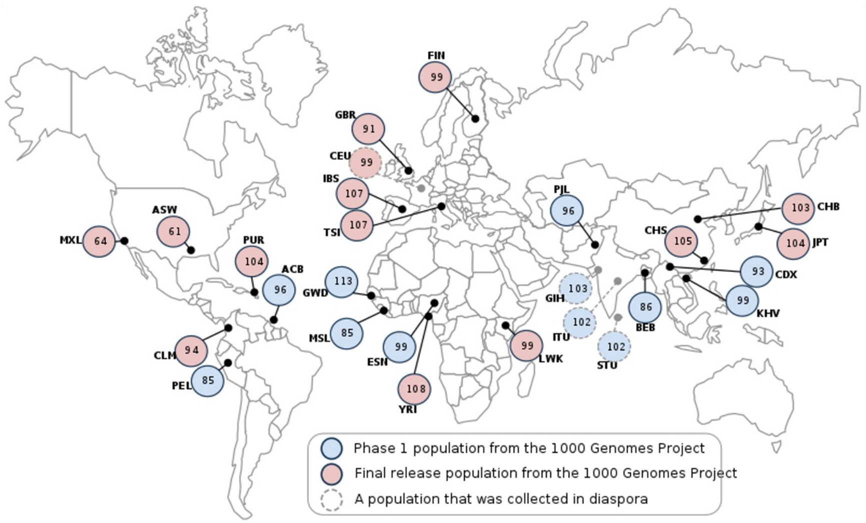
BY SERGEY NURK, SERGEY KOREN, ARANG RHIE, MIKKO RAUTIAINEN, ANDREY V. BZIKADZE, ALLA MIKHEENKO, MITCHELL R. VOLLGER, NICOLAS ALTEMOSE, LEV URALSKY, ARIEL GERSHMAN, [...] ADAM M. PHILLIPPY +89 au

SCIENCE • VOL. 376, NO. 6588 • 31 MAR 2022 : 44-53

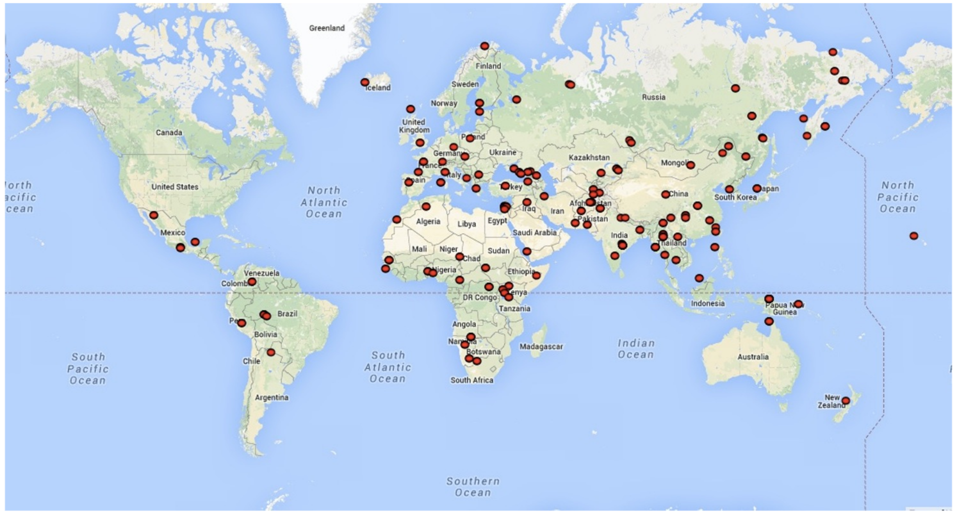
T2T-chrY: Human variation across 156 populations

1000 Genomes Project (1KGP)
3,202 samples from 26 populations

Simons Genome Diversity Project (SGDP)
279 open access samples from 130 populations



(Byrska-Bishop et al., Cell, 2022)



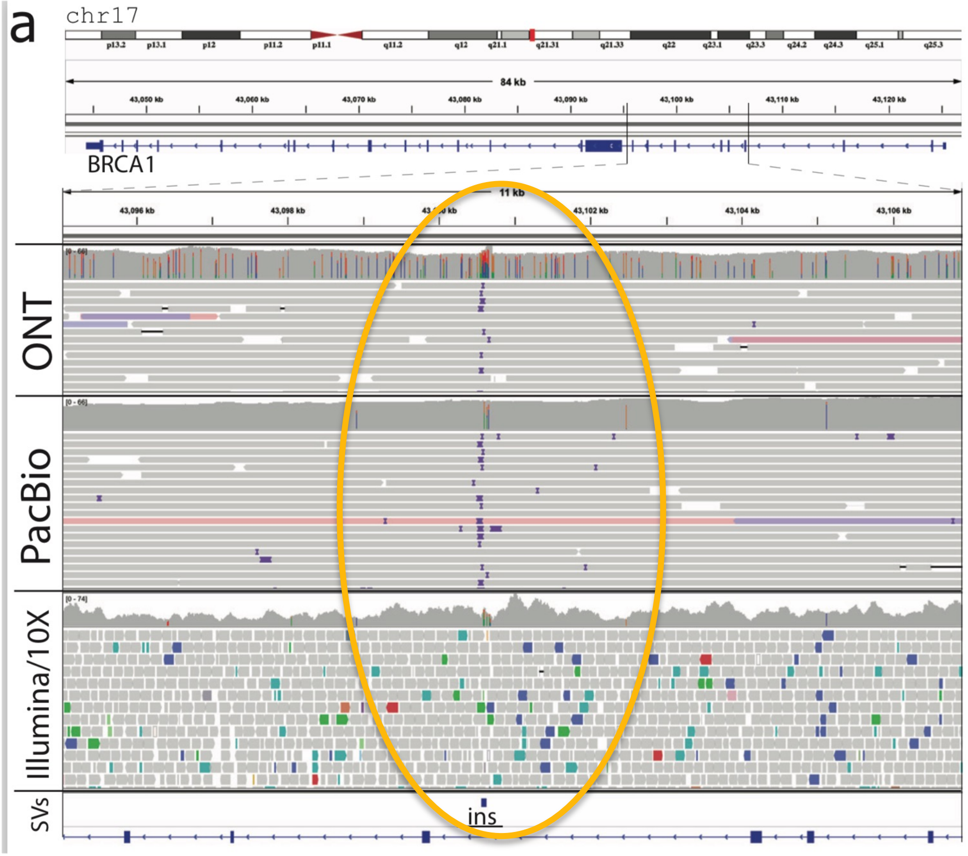
(Mallick et al., Nature, 2016)

The complete sequence of a human Y chromosome
Rhie *et al.* (2023) *Nature. https://doi.org/10.1038/s41586-023-06457-y*

Stephen Hwang

Dylan Taylor

Hidden Variants in Breast Cancer Genes



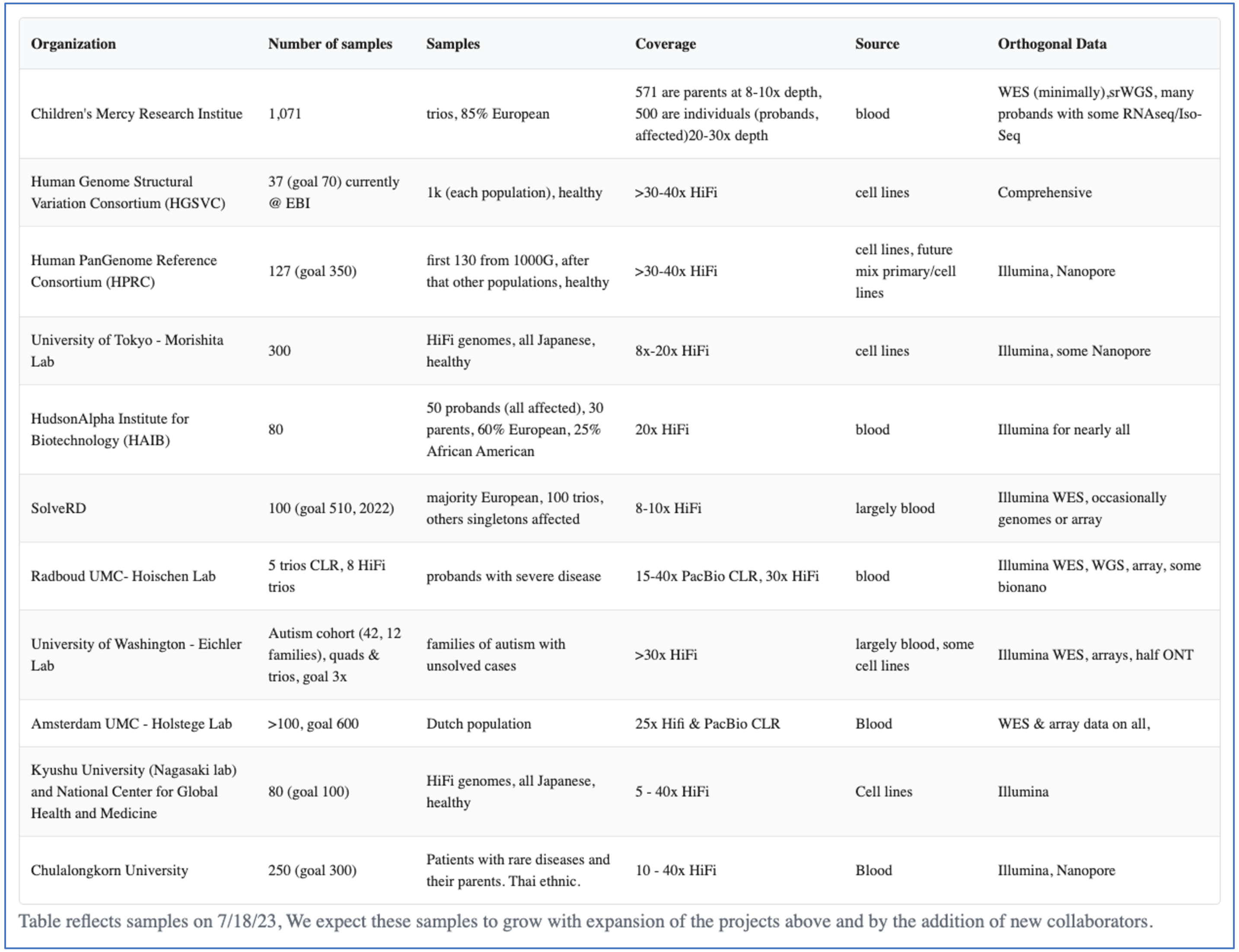
Thanks to long reads we can now robustly detect entirely new types of variation

…

But how can we identify those variants with clinical & functional impact?

Comprehensive analysis of structural variants in breast cancer genomes using single molecule sequencing
Aganezov *et al.* (2020) Genome Research. doi:10.1101/gr.260497.119

CoLoRS: Consortium of Long Read Sequencing



Organization	Number of samples	Samples	Coverage	Source	Orthogonal Data
Children's Mercy Research Institue	1,071	trios, 85% European	571 are parents at 8-10x depth, 500 are individuals (probands, affected)20-30x depth	blood	WES (minimally),srWGS, many probands with some RNAseq/Iso-Seq
Human Genome Structural Variation Consortium (HGSVC)	37 (goal 70) currently @ EBI	1k (each population), healthy	>30-40x HiFi	cell lines	Comprehensive
Human PanGenome Reference Consortium (HPRC)	127 (goal 350)	first 130 from 1000G, after that other populations, healthy	>30-40x HiFi	cell lines, future mix primary/cell lines	Illumina, Nanopore
University of Tokyo - Morishita Lab	300	HiFi genomes, all Japanese, healthy	8x-20x HiFi	cell lines	Illumina, some Nanopore
HudsonAlpha Institute for Biotechnology (HAIB)	80	50 probands (all affected), 30 parents, 60% European, 25% African American	20x HiFi	blood	Illumina for nearly all
SolveRD	100 (goal 510, 2022)	majority European, 100 trios, others singletons affected	8-10x HiFi	largely blood	Illumina WES, occasionally genomes or array
Radboud UMC- Hoischen Lab	5 trios CLR, 8 HiFi trios	probands with severe disease	15-40x PacBio CLR, 30x HiFi	blood	Illumina WES, WGS, array, some bionano
University of Washington - Eichler Lab	Autism cohort (42, 12 families), quads & trios, goal 3x	families of autism with unsolved cases	>30x HiFi	largely blood, some cell lines	Illumina WES, arrays, half ONT
Amsterdam UMC - Holstege Lab	>100, goal 600	Dutch population	25x Hifi & PacBio CLR	Blood	WES & array data on all,
Kyushu University (Nagasaki lab) and National Center for Global Health and Medicine	80 (goal 100)	HiFi genomes, all Japanese, healthy	5 - 40x HiFi	Cell lines	Illumina
Chulalongkorn University	250 (goal 300)	Patients with rare diseases and their parents. Thai ethnic.	10 - 40x HiFi	Blood	Illumina, Nanopore

Table reflects samples on 7/18/23, We expect these samples to grow with expansion of the projects above and by the addition of new collaborators.

Open coalition of international researchers focused on cataloging all classes of variation using long-read whole genome sequencing.

- The goal is to provide variant frequency data for public use and as a resource to the global scientific and clinical research community
- Complements existing databases such as gnomAD
- Develop state-of-the-art pipelines, execute at individual sites or within the AnVIL cloud platform

>2195 samples and growing!

https://colorsdb.org/

TELOMERE-TO-TELOMERE CONSORTIUM

Google Health

DNAnexus

amazon web services

PACBIO®

Oxford NANOPORE Technologies

ARIMA GENOMICS

NIH

National Human Genome Research Institute

Acknowledgements

Schatz Lab

Enis Afghan
Ahmed Awan
Dannon Baker
Tyler Collins
Arun Das
John Davis
Sam Guerler
Katie Jenike
Sam Kovaka
Natalie Kucher
Qiuhui Li
Stephen Mosher

Matthew Nguyen
Bohan Ni
Alex Ostrovsky
Srividya Ramakrishnan
Michelle Savage
Michael Sauria
Margaret Starostik
Alex Sweeten
Jenn Vessio
Natalie Whitaker

T2T, AnVIL, & Galaxy Teams

Miga, Phillippy, Eichler, Nekrutenko, Goecks, Tan, Leek, Morgan, Carey, Philippakis *et al.*

CoLoRS

Eichler, Lake, Wenger, Korlach, Beck, Pastinen, Audano, Garimella, Schmutz, Chen *et al.*

JHU

Battle Lab
Klein Lab
Genetic Resources Core

Timp Lab

Carolina Montano
Jessica Hosea
Luke Morina

Stanford

Montgomery Lab
Ashley Lab

Mayo Clinic

Gloria Petersen / Sam Antwi

University of Toronto

Steven Gallinger

CoLoRS

National Human Genome Research Institute

NIH NATIONAL CANCER INSTITUTE

NSF

Bloomberg Professors

NIH National Institutes of Health

Office of Data Science Strategy

HVD 21: Telomere-to-Telomere Consortium Analyses on the NHGRI AnVIL

HVD 22: Long Read Variant Frequency Database on AnVIL (CoLoRS)

Thank you!

schatz-lab.org