

# Cloud-scale genomics

Michael Schatz

Sept 19, 2022

Lecture 6: Applied Comparative Genomics



# Assignment 2: Genome Assembly

## Due Monday Sept 19 by 11:59pm

The screenshot shows a GitHub repository page for 'appliedgenomics2022'. The repository has 2 forks and 9 contributors. The README.md file contains the assignment details:

```
mschatz: update assignment to use conda ✓
```

Latest commit 4e2554c 10 minutes ago History

1 contributor

156 lines (96 select) 8 kB

### Assignment 2: Genome Assembly

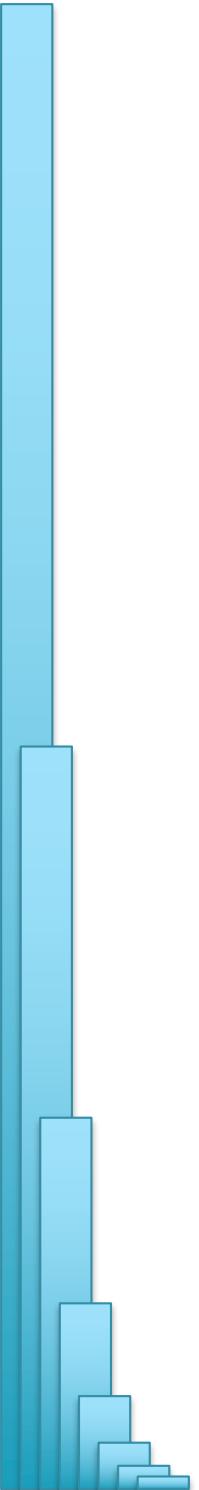
Assignment Date: Monday, September 12, 2022  
Due Date: Monday, September 19, 2022 @ 11:59pm

#### Assignment Overview

In this assignment, you are given a set of unassembled reads from a mysterious pathogen that contains a secret message encoded somewhere in the genome. The secret message will be recognizable as a novel insertion of sequence not found in the reference. Your task is to assess the quality of the reads, assemble the genome, identify, and decode the secret message. If all goes well the secret message should decode into a recognizable english text; otherwise doublecheck your coordinates and try again. As a reminder, any questions about the assignment should be posted to Piazza.

For this assignment, we recommend you install and run the tools using bioconda. There are some tips below in the Resources section. Alternatively, you can try running the tools using Docker. Docker is a powerful containerization tool to make software easier to distribute. This will

<https://github.com/schatzlab/appliedgenomics2022/tree/main/assignments/assignment2>  
Check Piazza for questions!



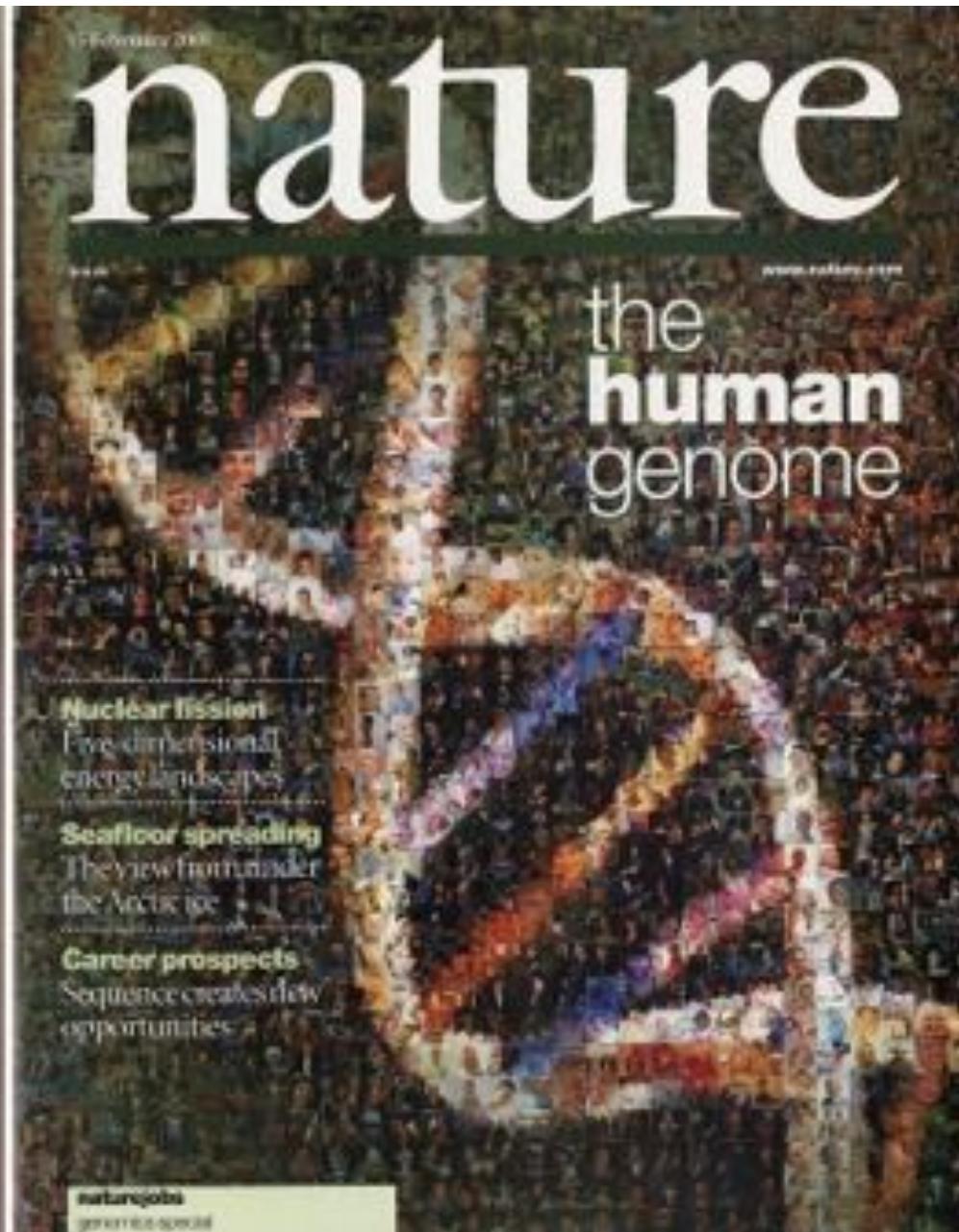
# Wednesday's lecture

How many of you know?

- Hash Table
- Suffix Array
- FM Index
- Dynamic Programming
- Edit Distance
- Learned Index Structure



**The Sequence of the Human Genome**  
Venter et al.  
Science 291, pp 1304-1351 (2001)



**Initial sequencing and analysis of the human genome**  
International Human Genome Sequencing Consortium  
Nature 409, pp 860-921 (2001)

# Single Molecule Long Read Sequencing

PacBio  
Sequel II

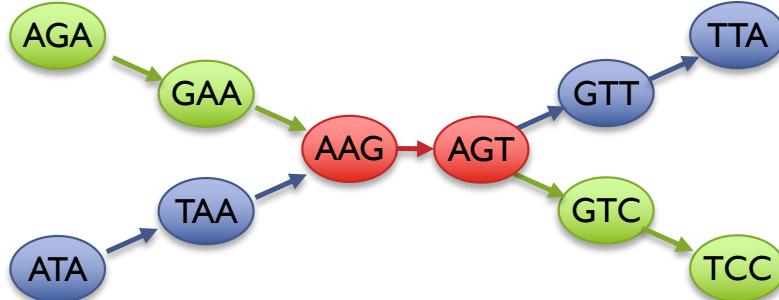


Oxford Nanopore  
PromethION



# Two Paradigms for Assembly

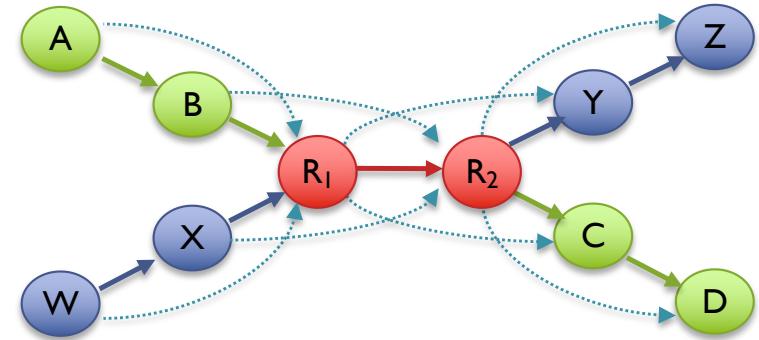
## de Bruijn Graph



Short read assemblers

- Repeats depends on word length
- Read coherency, placements lost
- Robust to high coverage

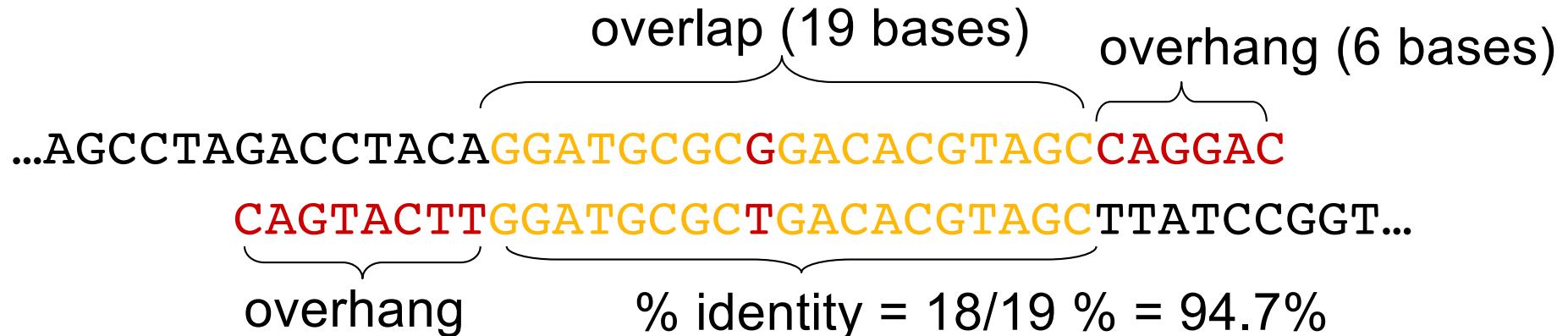
## Overlap Graph



Long read assemblers

- Repeats depends on read length
- Read coherency, placements kept
- Tangled by high coverage

# Overlap between two sequences



**overlap** - region of similarity between regions  
**overhang** - un-aligned ends of the sequences

The assembler screens merges based on:

- length of overlap
- % identity in overlap region
- maximum overhang size.

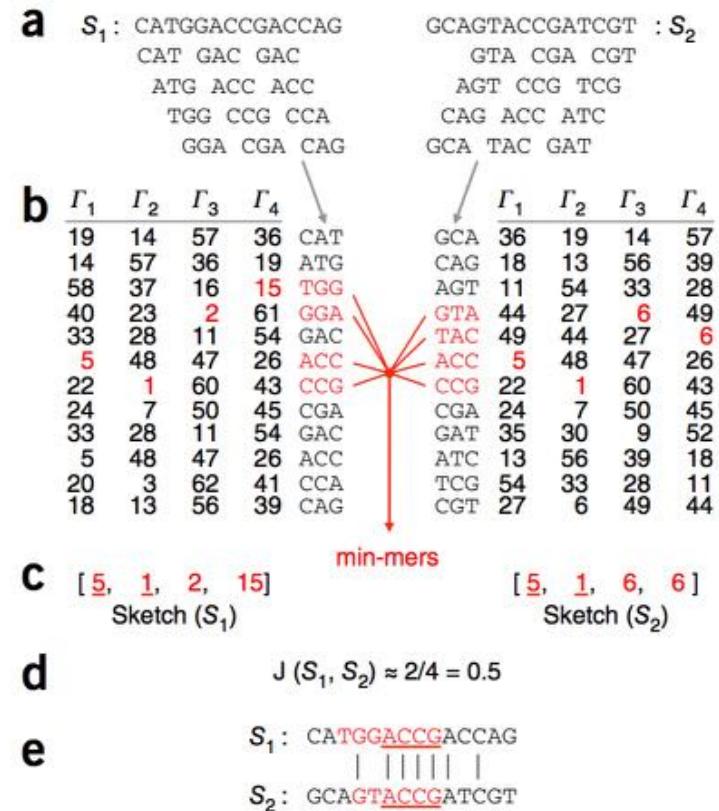
[How do we compute the overlap?]

[Do we really want to do all-vs-all?]

# Very fast approximate overlapping

Maybe we don't need to compute the exact identity of the overlap region, just approximate it

- If two reads overlap, they should share many of the same kmers: Their Jaccard coefficient should be high:  $|\text{intersection}| / |\text{union}|$
- But tracking all of the kmers for a read is a lot of overhead
- Instead, compare the “sketch” of the reads: a small fraction of kmers carefully chosen
- LSH: Find the sketch by applying N hash functions to the kmers, and keeping the minimum hash values reported from each ( $N=4$  in example)
- This forms a nice “random” sample of the reads, and the Jaccard coefficient is a good approximation of the sequence similarity



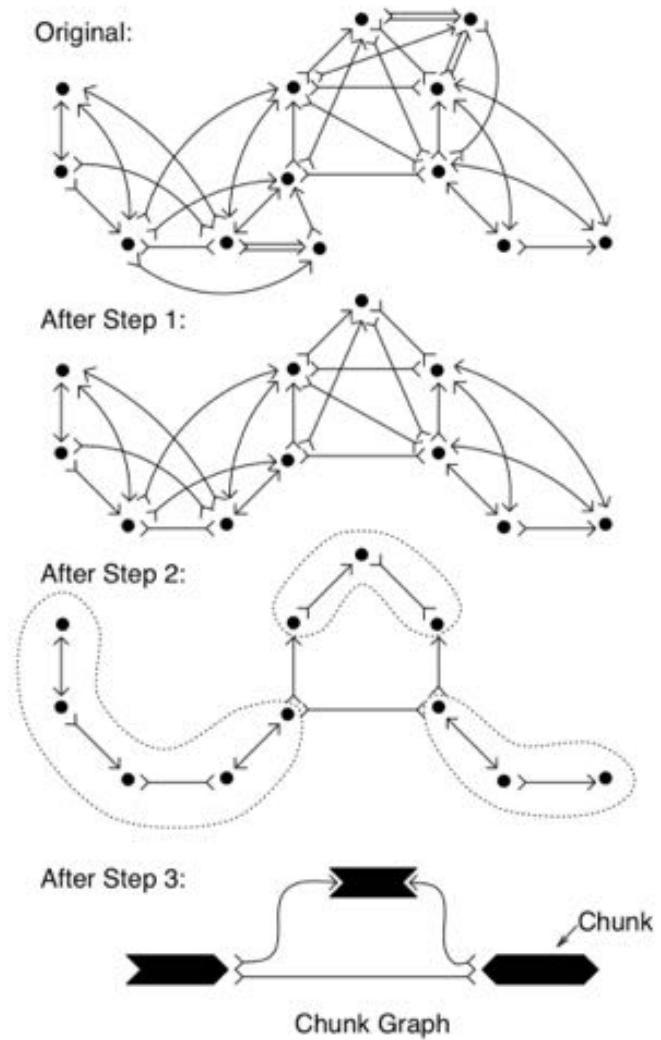
# Unitigging: Pruning the Overlap Graph

The overlap graph has many redundant edges:

- If the average coverage is D, we should expect D overlaps at the beginning of the read, and D at the end

Transform the graph to simplify the assembly problem (without changing the valid solutions):

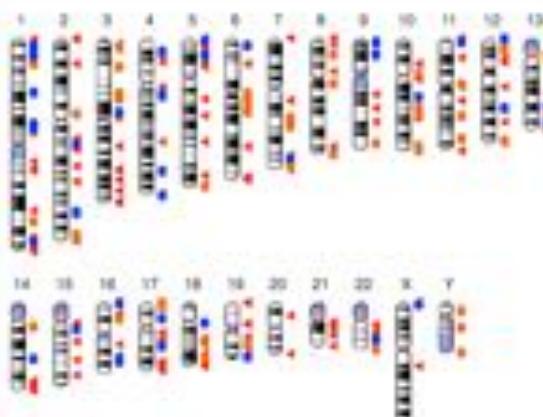
1. **Contained reads removal:** Short reads that are substrings of longer reads don't advance the assembly, remove those nodes and all of the edges
2. **Transitive edge removal:** If A  $\rightarrow$  B, and B  $\rightarrow$  C, remove the transitive edge A  $\rightarrow$  C
3. **“Chunkification”:** Linear subgraphs define uniquely assemblable segments: “unitigs”



Towards Simplifying and Accurately Formulating Fragment Assembly  
Myers (1995) J Comput Biol. Summer;2(2):275-90.

## Human Genome Overview

Information about the continuing improvement of the human genome



- Region containing alternate loci
- Region containing fix patches
- Region containing novel patches

Karyogram of the latest human assembly, GRCh38.p11

The GRC is working hard to provide the best possible assembly by both generating multiple representations (alternatives) for each chromosome and by allowing users to search for variants represented by a single path. Additionally, we are now providing a reference genome that allows users who are interested in a specific locus to quickly find the best representation. This will also allow users who need chromosome coordinate information to easily find it.

### Download data:

- GRCh38.p11 (latest minor release) FTP
- GRCh38 (latest major release) FTP
- Genome regions under review FTP
- Current Tiling Path Files (TPFs)

Transitioning to GRCh38? Try the NCBI Remap tool to find the assembly alignments used by the GRC.

### Next assembly update

The next assembly update (GRCh38.p12) will be released in June 2017.

[GRCh38.p11](#)   [GRCh37.p13](#)   [GRCh37](#)

## GRCh38.p11

Release date: June 14, 2017

Release type: minor

Release notes: GRCh38.p11 is the eleventh patch release for the GRCh38 reference assembly. No chromosome coordinate changes were made. The total number of patch scaffolds is now: 64 FIX and 59 NOVEL.

Assembly accessions: GenBank: [GCA\\_000001405.26](#), RefSeq: [GCF\\_000001405.37](#)

### Pseudoautosomal regions

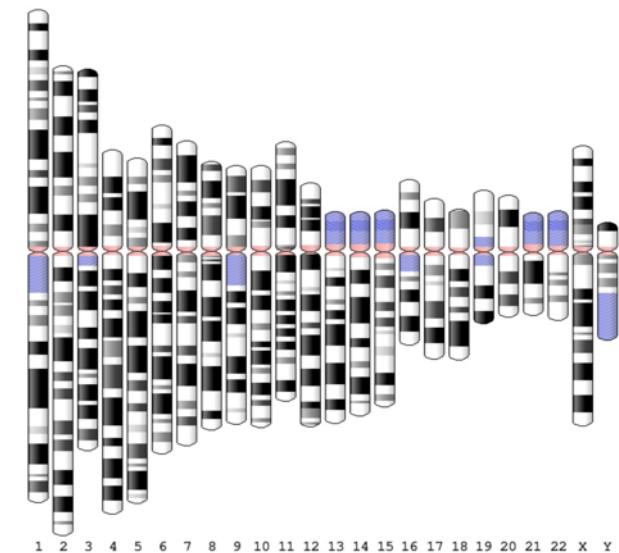
Name	Chr	Start	Stop
PARH1	X	10,001	2,781,479
PARH2	X	155,701,363	156,030,895
PARH1	Y	10,001	2,781,479
PARH2	Y	56,887,903	57,217,415



# Finishing the human genome

## 238Mbp is missing or incorrect

- Centromeres and telomeres
- Segmentally duplicated genes
- Tandem gene arrays (e.g. rDNAs)
- And an unknown number of errors...



## Why does it matter?

Variation in these regions is unexplored

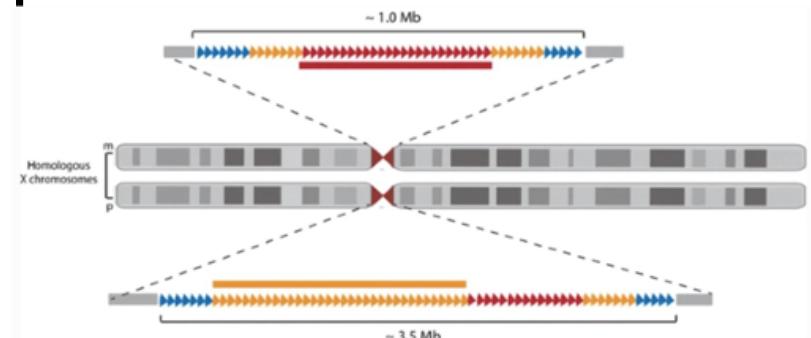
Functional studies need sequence

Reference gaps lead to artifacts

We don't know what we don't

## Why has it taken so long?

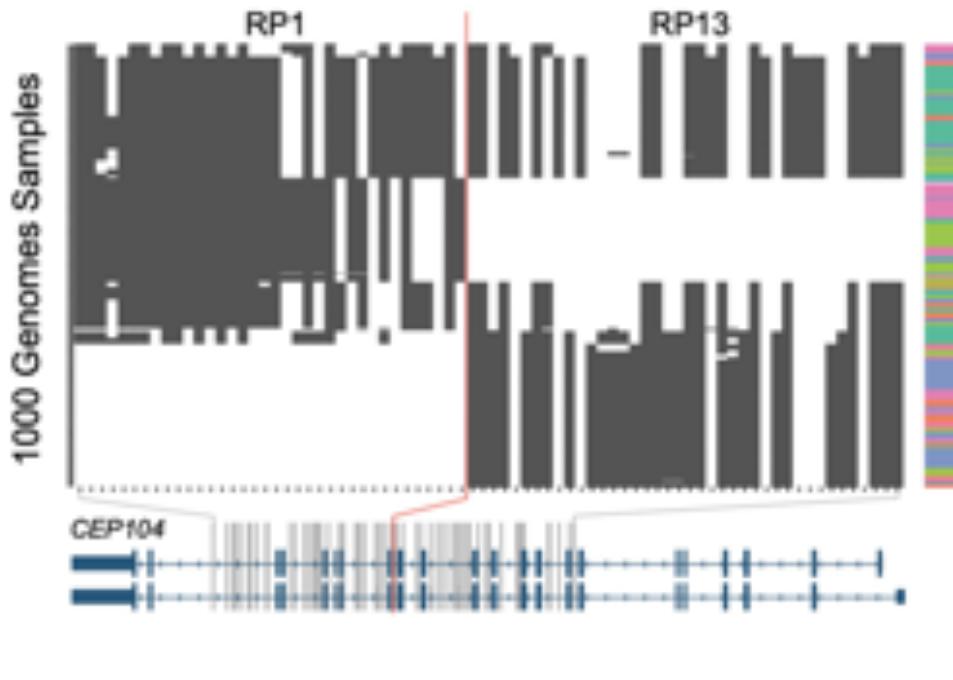
Repeats, repeats, repeats...



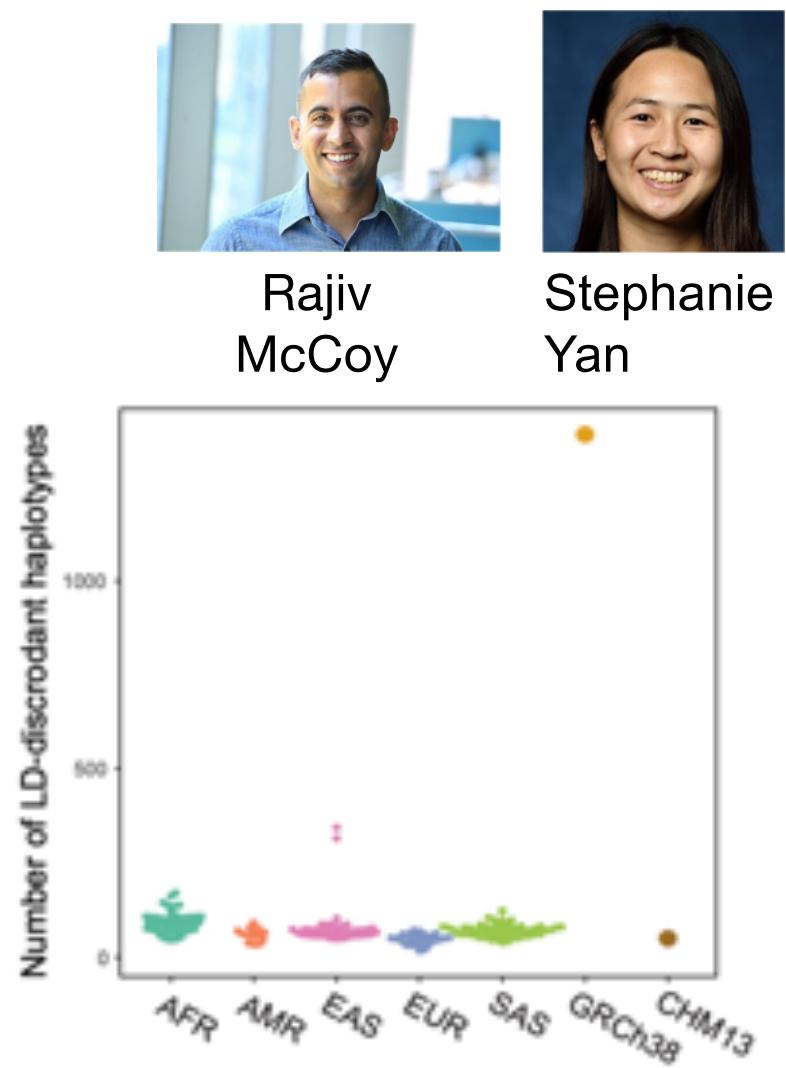
Miga 2015

# Local Ancestry Comparisons

C



D



Rajiv  
McCoy



Stephanie  
Yan

# Let's finish a human genome



T2T Working Group Home · Technology · Data · CHM13 Cell Line · Remaining Challenges · Who We Are · Join Us · Q

A karyotype image showing three rows of human chromosomes against a black background. The chromosomes are color-coded by chromosome group (e.g., pink for group 1, green for group 2, etc.).

The Telomere-to-Telomere (T2T) consortium is an open, community-based effort to generate the first complete assembly of a human genome.

CHM13 homozygous 46,XX cell line from Urvashi Surti, Pitt; SKY karyotype from Jennifer Gerton, Stowers



# CHM13 assembly graph

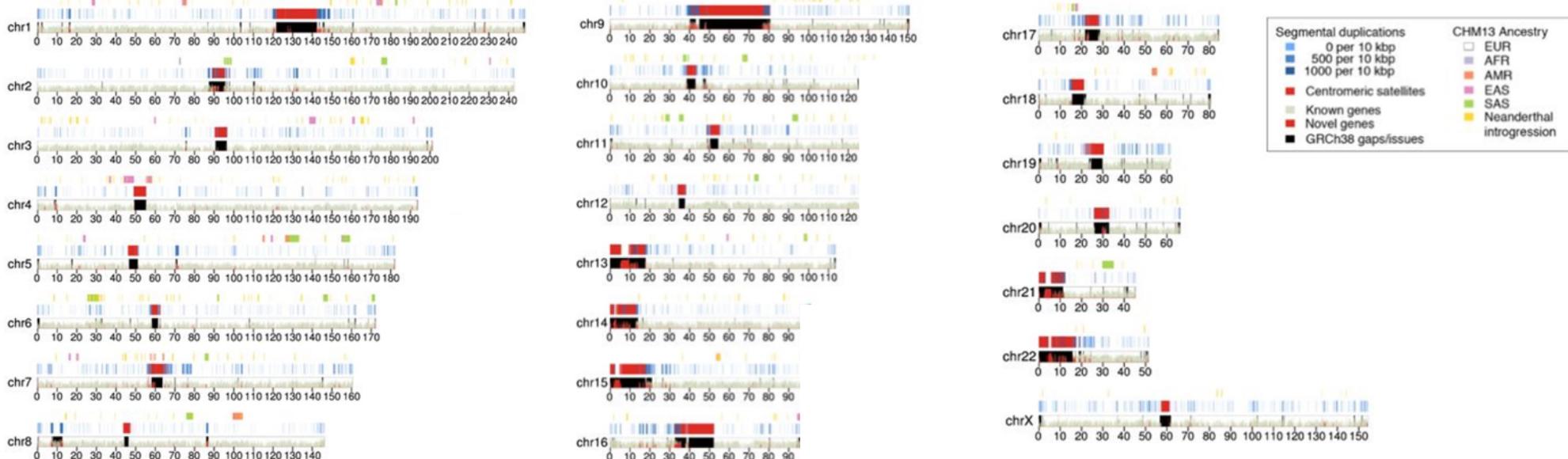


**HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads.**

Nurk et al. *Genome Research* (2020)

# The complete sequence of a human genome

A

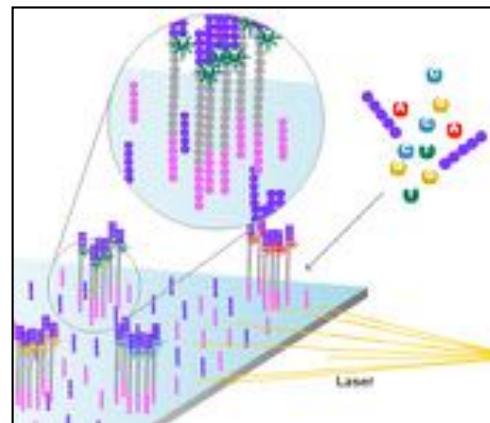
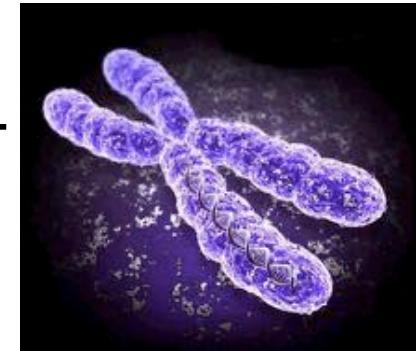
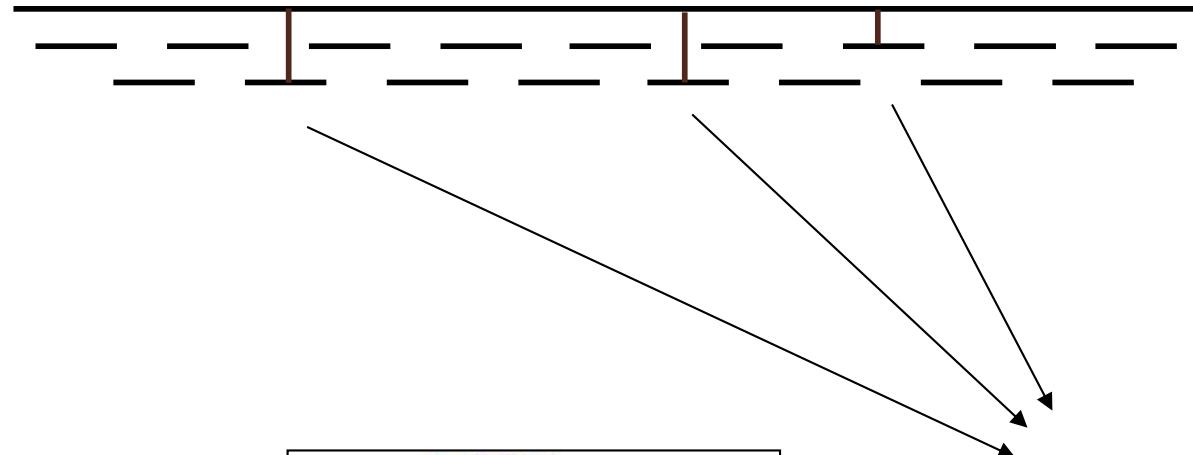


CHM13v1.1 genome size is **3.057 Gbp with zero Ns**  
Every chromosome is telomere-to-telomere, quality estimated >Q70  
~190 Mbp (~8%) of new sequence vs. GRCh38, fixes thousands of errors

(Nurk et al. Science, 2022)

# Personal Genomics

How does your genome compare to the reference?

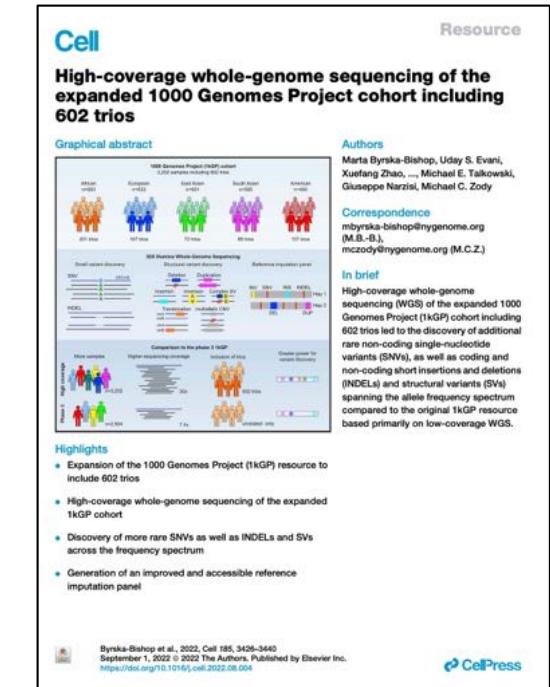


Heart Disease

Cancer

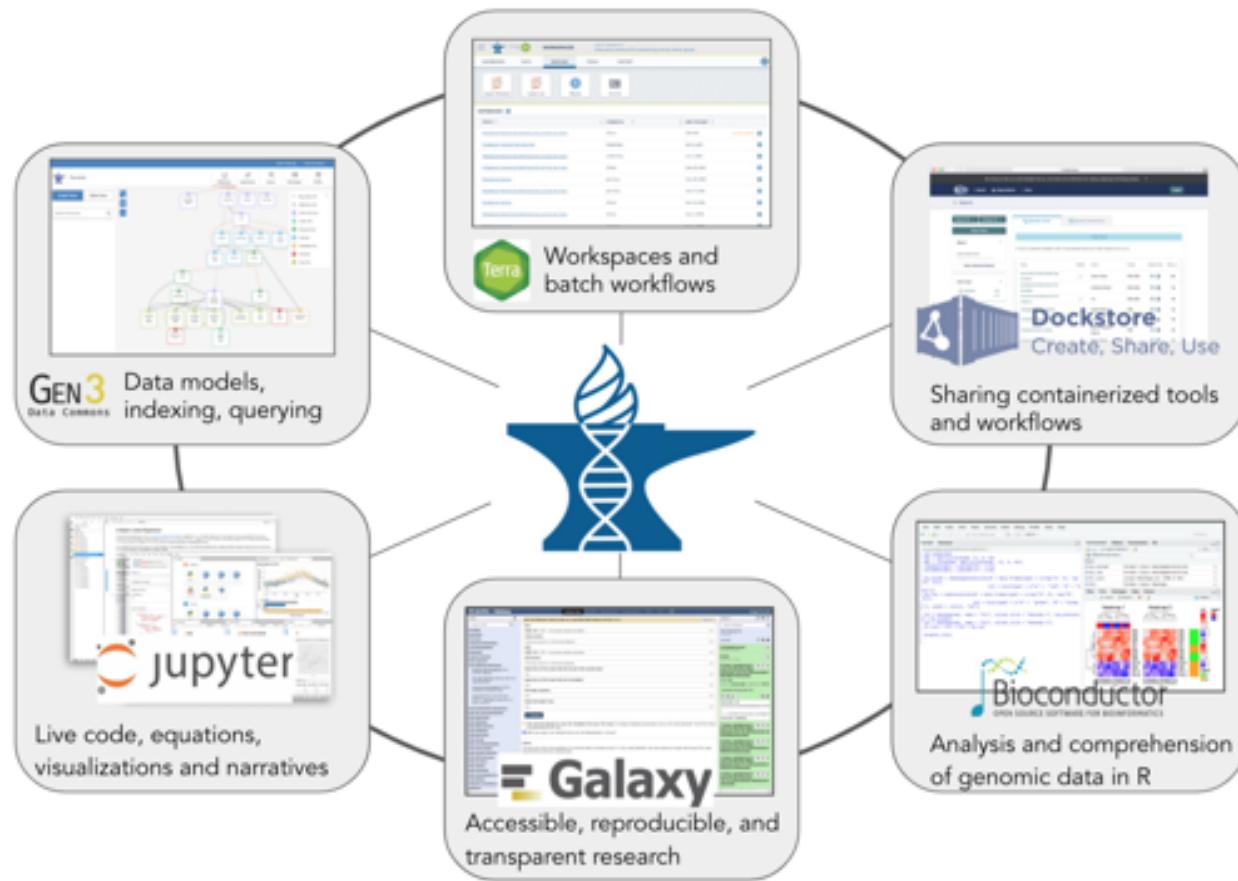
Presidential smile

# T2T Variants: 1000 Genomes Project



26 populations from 5 superpopulations (continental regions)  
3202 samples (2504 core genomes + 698 offspring)

3202 samples x 30Gb = 96Tb input data | >5Pb of intermediate data | >>1M core hours

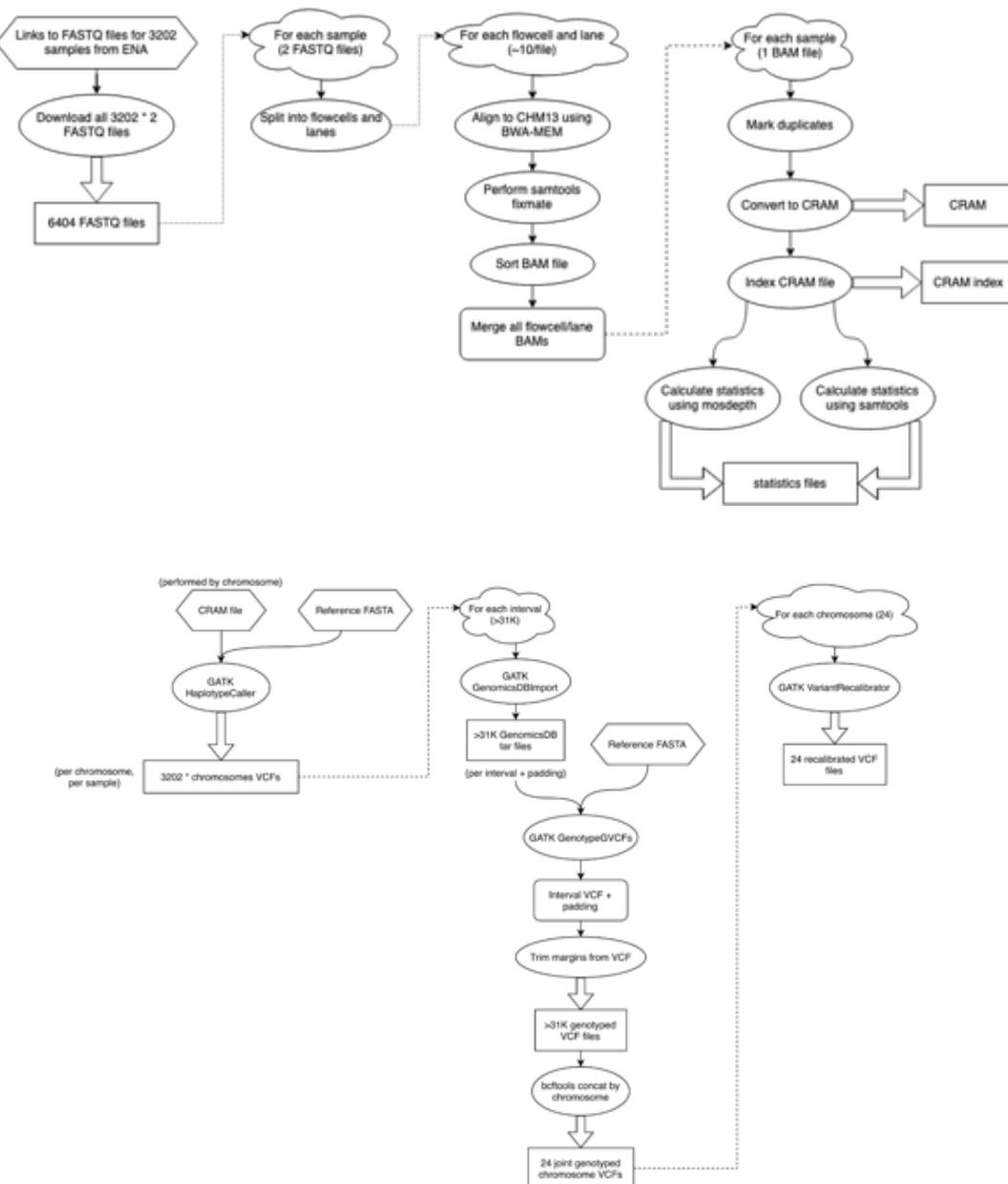


**Inverting the model of genomics data sharing with the  
NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL)**  
Schatz, Philippakis et al. (2022) *Cell Genomics*. doi: <https://doi.org/10.1016/j.xgen.2021.100085>

# 3202 Genomes to go...



Samantha Zarate



```
task: samtoolsStats {
    input {
        File inputCram
        File cramIndex
        File targetRef
        String sampleName
    }

    command <><
        samtools stats -r "~{targetRef}" \
            --reference "~{targetRef}" \
            -@ "${inproc}" \
            "~{inputCram}" > "~{sampleName}.samtools.stats.txt"
    >>>

    Int diskGb = ceil(2.0 * size(inputCram, "G"))

    runtime {
        docker : "szarate/t2t_variants:v0.0.2"
        disks : "local-disk ${diskGb} SSD"
        memory: "12G"
        cpu : 16
        preemptible: 3
        maxRetries: 3
    }

    output {
        File stats = "~{sampleName}.samtools.stats.txt"
    }
}
```



# Core usage over 24 hours

Preview

1 hour 4 hours 1 day



- instance/cpu/reserved\_cores: 11,552.00



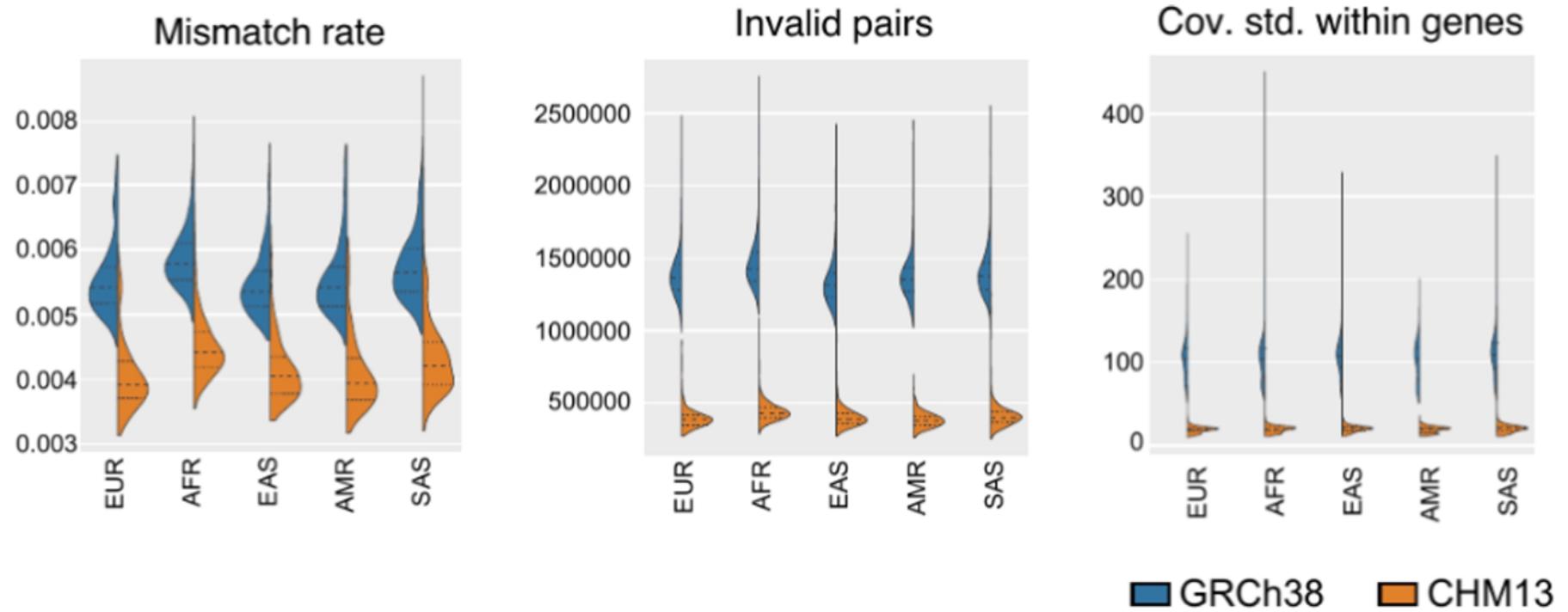
# AnVIL Data Table

Screenshot of the AnVIL Data Table interface showing a large dataset.

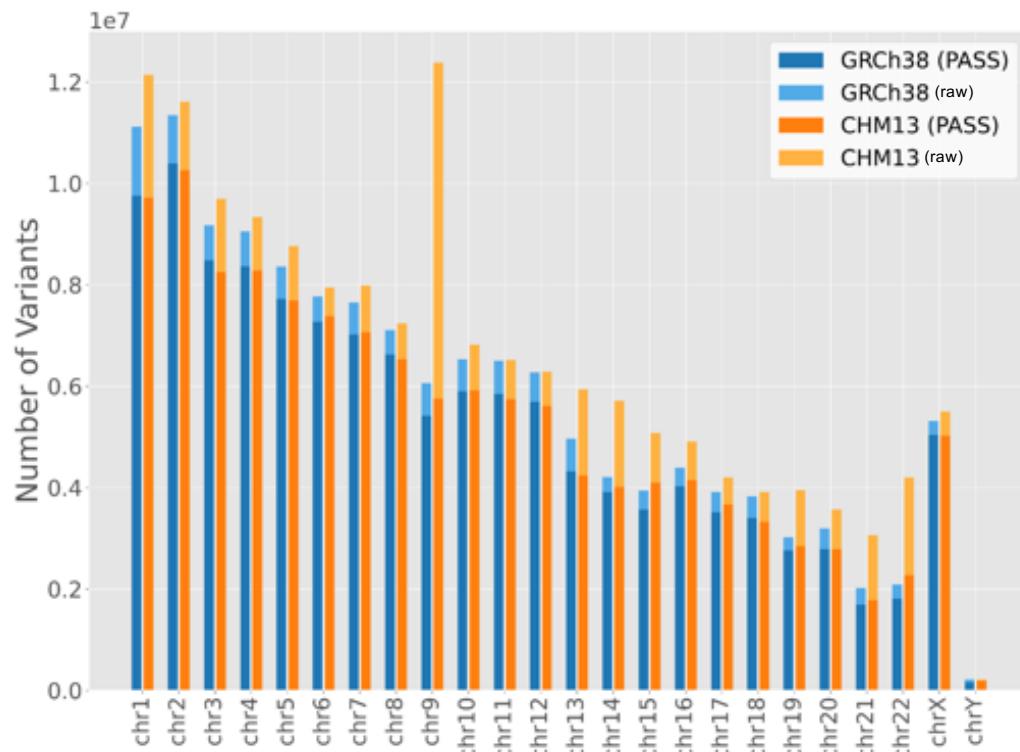
The interface includes a top navigation bar with tabs for Home, Workspaces, Data, Notebooks, Workflows, and Job history. The Data tab is selected. A sidebar on the left lists categories: ANALYSIS (1 item), REFERENCE DATA (1 item), OTHER DATA (1 item), WORKSPACE DATA (1 item), and FILES (1 item). The main area displays a table titled "anv.sample\_1" with 20 rows and 8 columns. The columns are labeled: ID, name, sample\_id, mousepm\_grossular, mousepm\_regions\_axial, mousepm\_regions\_axial\_val, mousepm\_regions\_axial\_val, mousepm\_regions\_axial\_val, mousepm\_axial.

ID	name	sample_id	mousepm_grossular	mousepm_regions_axial	mousepm_regions_axial_val	mousepm_regions_axial_val	mousepm_axial
HQ0001	HQ0001sample	HQ0001sample	HQ0001mousepm_grossular	HQ0001regions_axial	HQ0001regions_axial_val	HQ0001regions_axial_val	HQ0001_axial
HQ0002	HQ0002sample	HQ0002sample	HQ0002mousepm_grossular	HQ0002regions_axial	HQ0002regions_axial_val	HQ0002regions_axial_val	HQ0002_axial
HQ0003	HQ0003sample	HQ0003sample	HQ0003mousepm_grossular	HQ0003regions_axial	HQ0003regions_axial_val	HQ0003regions_axial_val	HQ0003_axial
HQ0004	HQ0004sample	HQ0004sample	HQ0004mousepm_grossular	HQ0004regions_axial	HQ0004regions_axial_val	HQ0004regions_axial_val	HQ0004_axial
HQ0005	HQ0005sample	HQ0005sample	HQ0005mousepm_grossular	HQ0005regions_axial	HQ0005regions_axial_val	HQ0005regions_axial_val	HQ0005_axial
HQ0006	HQ0006sample	HQ0006sample	HQ0006mousepm_grossular	HQ0006regions_axial	HQ0006regions_axial_val	HQ0006regions_axial_val	HQ0006_axial
HQ0007	HQ0007sample	HQ0007sample	HQ0007mousepm_grossular	HQ0007regions_axial	HQ0007regions_axial_val	HQ0007regions_axial_val	HQ0007_axial
HQ0008	HQ0008sample	HQ0008sample	HQ0008mousepm_grossular	HQ0008regions_axial	HQ0008regions_axial_val	HQ0008regions_axial_val	HQ0008_axial
HQ0009	HQ0009sample	HQ0009sample	HQ0009mousepm_grossular	HQ0009regions_axial	HQ0009regions_axial_val	HQ0009regions_axial_val	HQ0009_axial
HQ0010	HQ0010sample	HQ0010sample	HQ0010mousepm_grossular	HQ0010regions_axial	HQ0010regions_axial_val	HQ0010regions_axial_val	HQ0010_axial
HQ0011	HQ0011sample	HQ0011sample	HQ0011mousepm_grossular	HQ0011regions_axial	HQ0011regions_axial_val	HQ0011regions_axial_val	HQ0011_axial
HQ0012	HQ0012sample	HQ0012sample	HQ0012mousepm_grossular	HQ0012regions_axial	HQ0012regions_axial_val	HQ0012regions_axial_val	HQ0012_axial
HQ0013	HQ0013sample	HQ0013sample	HQ0013mousepm_grossular	HQ0013regions_axial	HQ0013regions_axial_val	HQ0013regions_axial_val	HQ0013_axial
HQ0014	HQ0014sample	HQ0014sample	HQ0014mousepm_grossular	HQ0014regions_axial	HQ0014regions_axial_val	HQ0014regions_axial_val	HQ0014_axial
HQ0015	HQ0015sample	HQ0015sample	HQ0015mousepm_grossular	HQ0015regions_axial	HQ0015regions_axial_val	HQ0015regions_axial_val	HQ0015_axial
HQ0016	HQ0016sample	HQ0016sample	HQ0016mousepm_grossular	HQ0016regions_axial	HQ0016regions_axial_val	HQ0016regions_axial_val	HQ0016_axial
HQ0017	HQ0017sample	HQ0017sample	HQ0017mousepm_grossular	HQ0017regions_axial	HQ0017regions_axial_val	HQ0017regions_axial_val	HQ0017_axial
HQ0018	HQ0018sample	HQ0018sample	HQ0018mousepm_grossular	HQ0018regions_axial	HQ0018regions_axial_val	HQ0018regions_axial_val	HQ0018_axial
HQ0019	HQ0019sample	HQ0019sample	HQ0019mousepm_grossular	HQ0019regions_axial	HQ0019regions_axial_val	HQ0019regions_axial_val	HQ0019_axial
HQ0020	HQ0020sample	HQ0020sample	HQ0020mousepm_grossular	HQ0020regions_axial	HQ0020regions_axial_val	HQ0020regions_axial_val	HQ0020_axial
HQ0021	HQ0021sample	HQ0021sample	HQ0021mousepm_grossular	HQ0021regions_axial	HQ0021regions_axial_val	HQ0021regions_axial_val	HQ0021_axial
HQ0022	HQ0022sample	HQ0022sample	HQ0022mousepm_grossular	HQ0022regions_axial	HQ0022regions_axial_val	HQ0022regions_axial_val	HQ0022_axial
HQ0023	HQ0023sample	HQ0023sample	HQ0023mousepm_grossular	HQ0023regions_axial	HQ0023regions_axial_val	HQ0023regions_axial_val	HQ0023_axial
HQ0024	HQ0024sample	HQ0024sample	HQ0024mousepm_grossular	HQ0024regions_axial	HQ0024regions_axial_val	HQ0024regions_axial_val	HQ0024_axial
HQ0025	HQ0025sample	HQ0025sample	HQ0025mousepm_grossular	HQ0025regions_axial	HQ0025regions_axial_val	HQ0025regions_axial_val	HQ0025_axial
HQ0026	HQ0026sample	HQ0026sample	HQ0026mousepm_grossular	HQ0026regions_axial	HQ0026regions_axial_val	HQ0026regions_axial_val	HQ0026_axial
HQ0027	HQ0027sample	HQ0027sample	HQ0027mousepm_grossular	HQ0027regions_axial	HQ0027regions_axial_val	HQ0027regions_axial_val	HQ0027_axial
HQ0028	HQ0028sample	HQ0028sample	HQ0028mousepm_grossular	HQ0028regions_axial	HQ0028regions_axial_val	HQ0028regions_axial_val	HQ0028_axial
HQ0029	HQ0029sample	HQ0029sample	HQ0029mousepm_grossular	HQ0029regions_axial	HQ0029regions_axial_val	HQ0029regions_axial_val	HQ0029_axial
HQ0030	HQ0030sample	HQ0030sample	HQ0030mousepm_grossular	HQ0030regions_axial	HQ0030regions_axial_val	HQ0030regions_axial_val	HQ0030_axial
HQ0031	HQ0031sample	HQ0031sample	HQ0031mousepm_grossular	HQ0031regions_axial	HQ0031regions_axial_val	HQ0031regions_axial_val	HQ0031_axial
HQ0032	HQ0032sample	HQ0032sample	HQ0032mousepm_grossular	HQ0032regions_axial	HQ0032regions_axial_val	HQ0032regions_axial_val	HQ0032_axial
HQ0033	HQ0033sample	HQ0033sample	HQ0033mousepm_grossular	HQ0033regions_axial	HQ0033regions_axial_val	HQ0033regions_axial_val	HQ0033_axial
HQ0034	HQ0034sample	HQ0034sample	HQ0034mousepm_grossular	HQ0034regions_axial	HQ0034regions_axial_val	HQ0034regions_axial_val	HQ0034_axial
HQ0035	HQ0035sample	HQ0035sample	HQ0035mousepm_grossular	HQ0035regions_axial	HQ0035regions_axial_val	HQ0035regions_axial_val	HQ0035_axial
HQ0036	HQ0036sample	HQ0036sample	HQ0036mousepm_grossular	HQ0036regions_axial	HQ0036regions_axial_val	HQ0036regions_axial_val	HQ0036_axial
HQ0037	HQ0037sample	HQ0037sample	HQ0037mousepm_grossular	HQ0037regions_axial	HQ0037regions_axial_val	HQ0037regions_axial_val	HQ0037_axial
HQ0038	HQ0038sample	HQ0038sample	HQ0038mousepm_grossular	HQ0038regions_axial	HQ0038regions_axial_val	HQ0038regions_axial_val	HQ0038_axial

# 1000G Mapping on T2T-CHM13

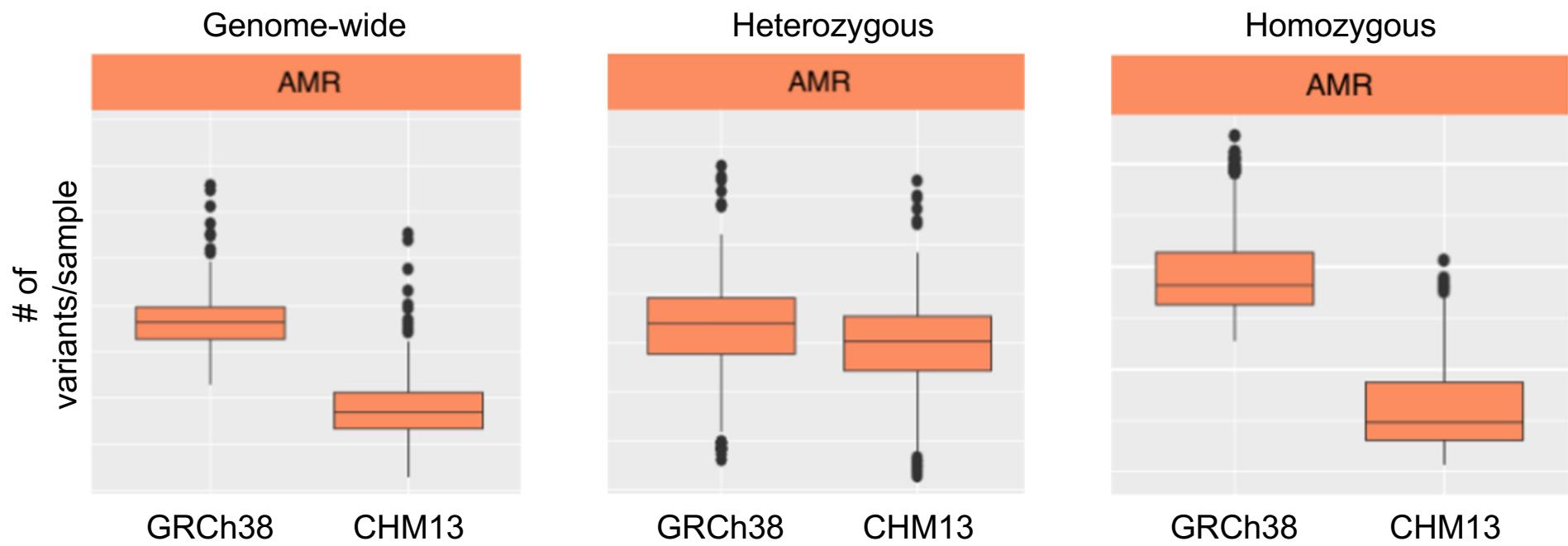


# Across 1000G, Many More Variants Found Using CHM13

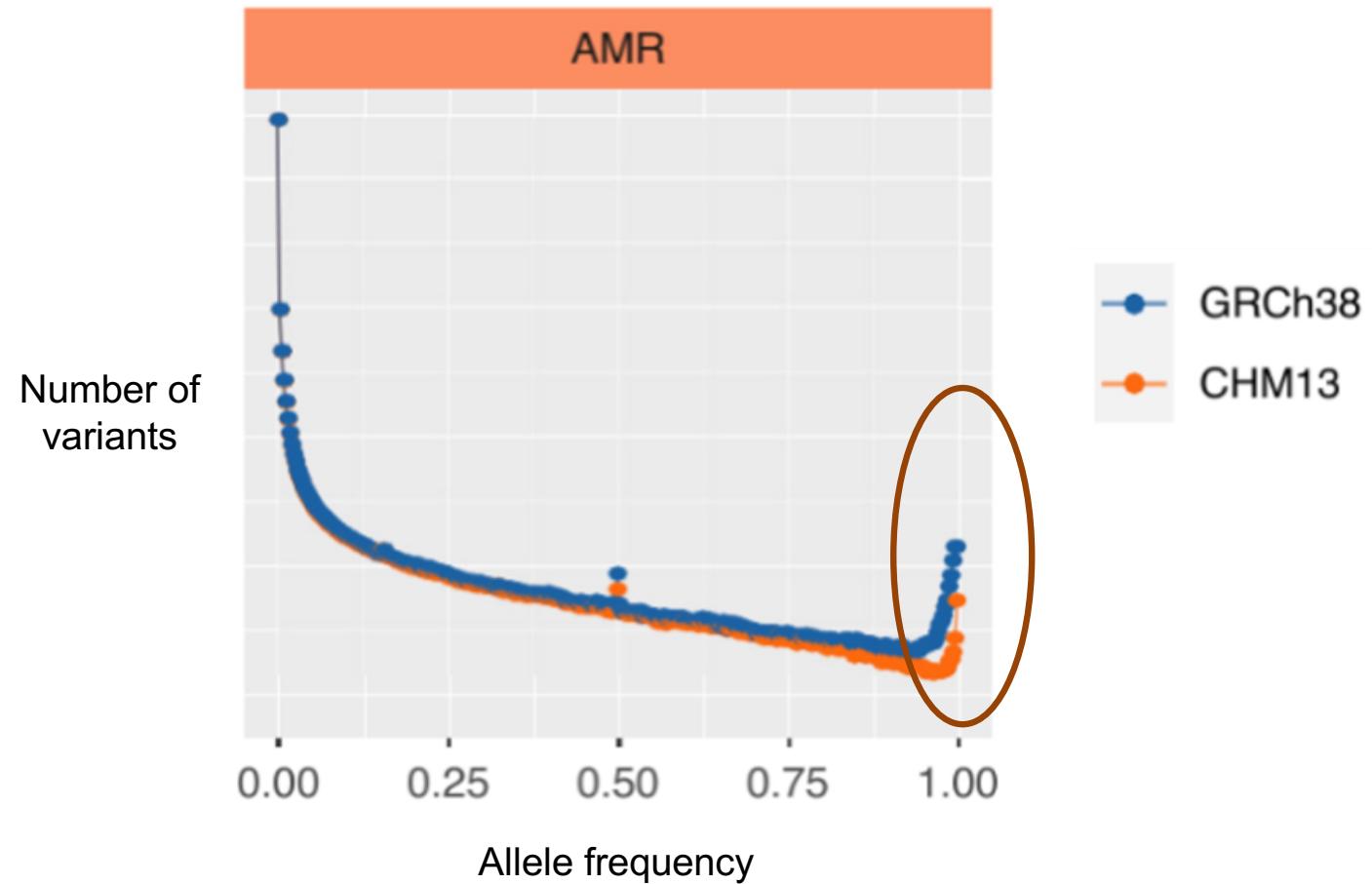


	GRCh38	CHM13
# PASS variants	125,484,020	126,591,489

# 1000G Per-Sample Variant Counts on T2T-CHM13



# Explaining Decreased Per-Sample Count

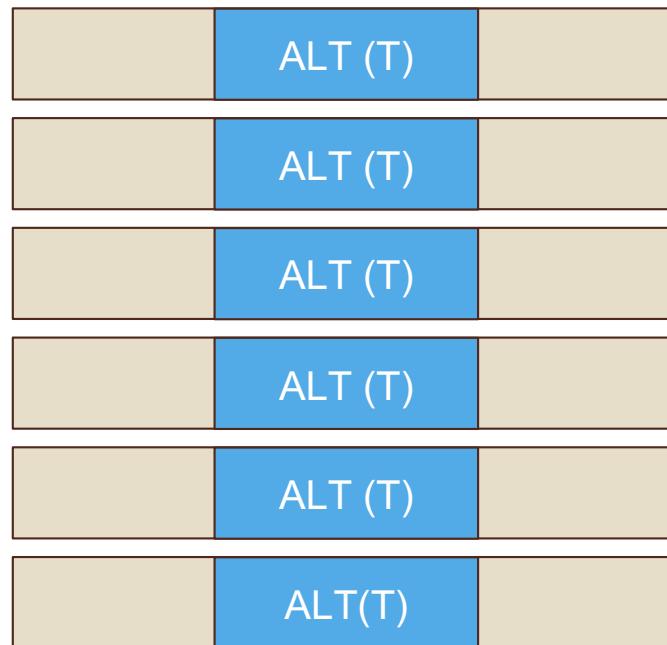


# Allele Frequency = I: Reference error / private variant

GRCh38



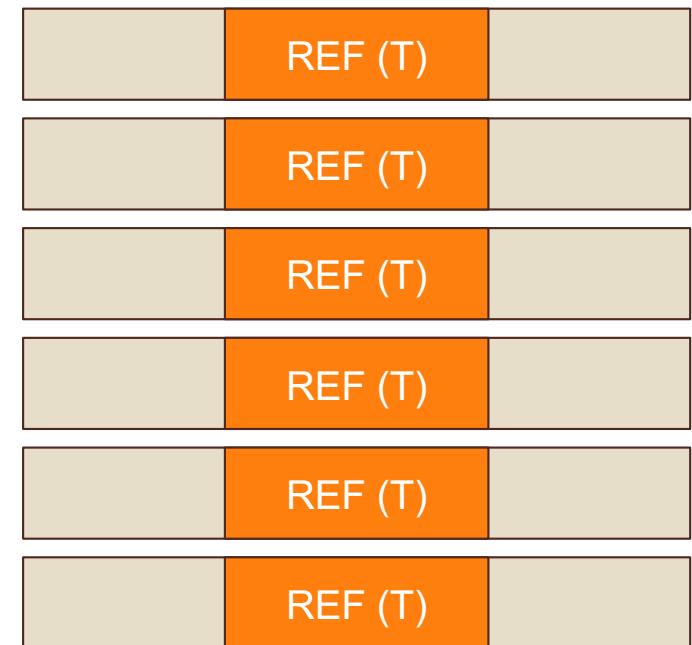
Samples



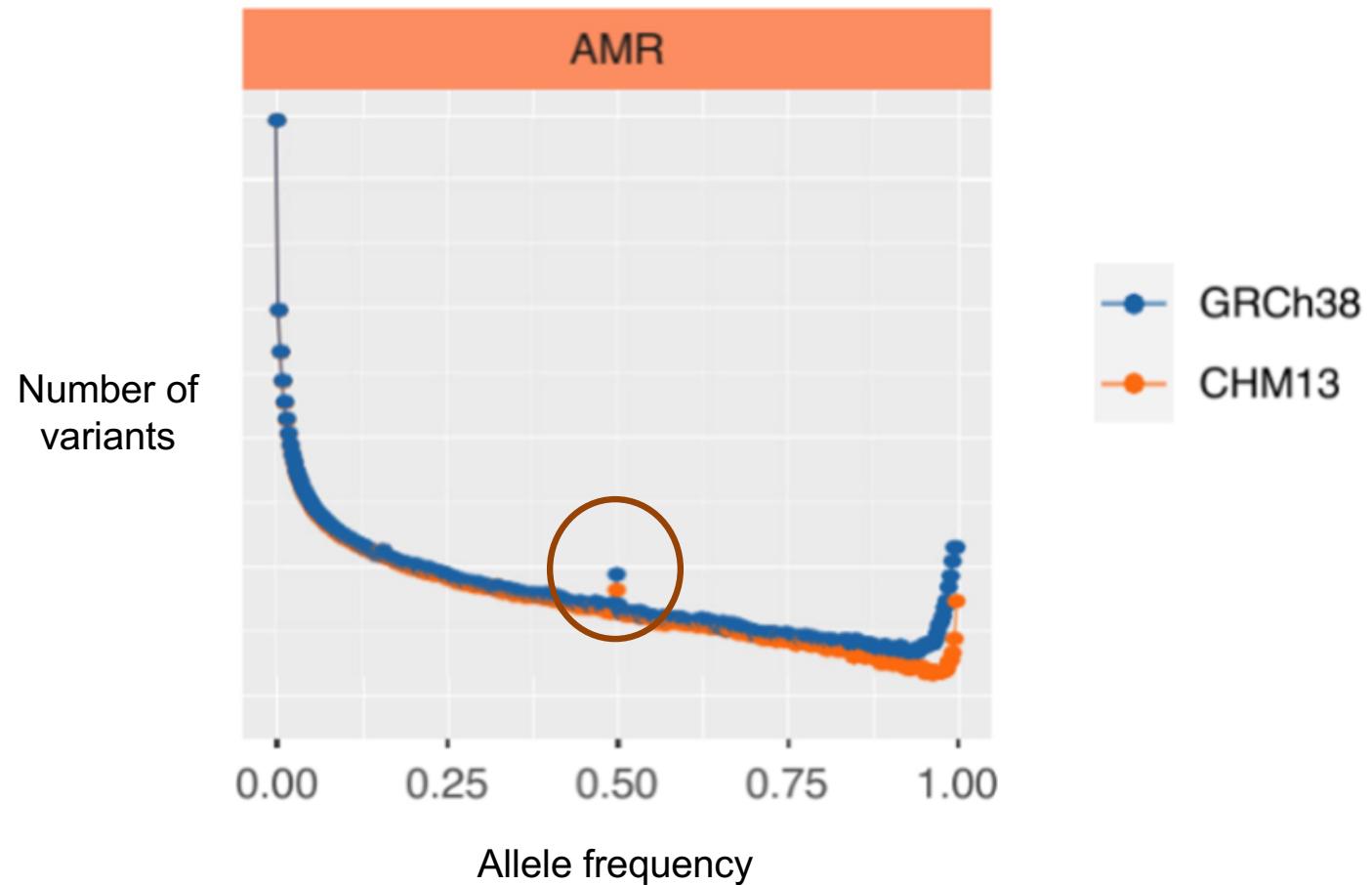
CHM13



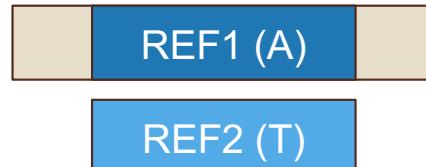
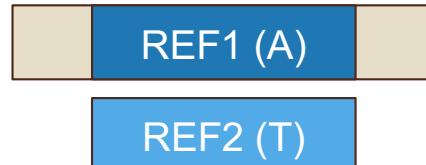
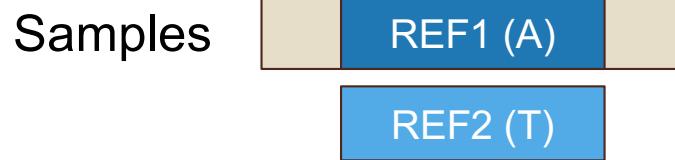
Samples



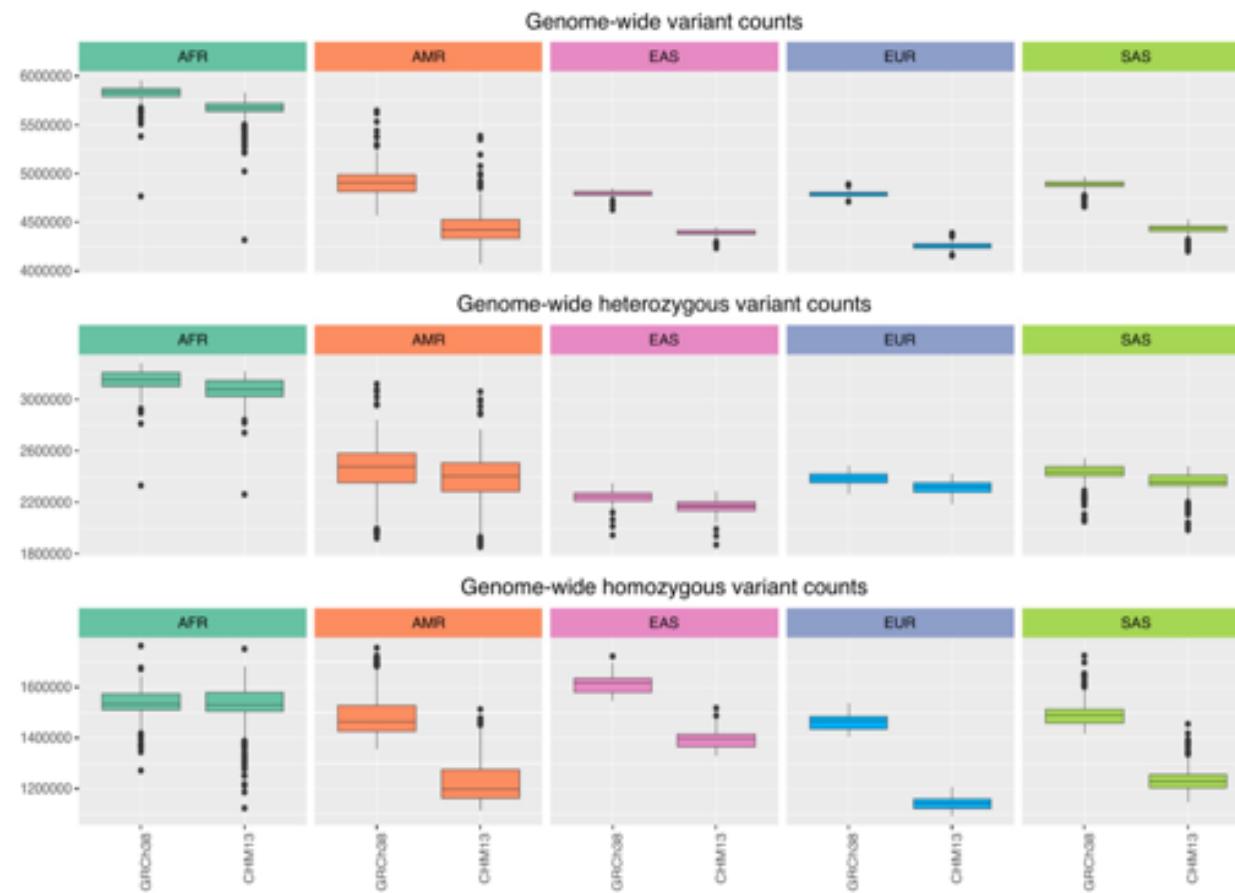
# Explaining Decreased Per-Sample Count



# Allele Frequency $\approx$ 0.5: Collapsed duplication



# 1000G Per-Sample Variant Counts on T2T-CHM13



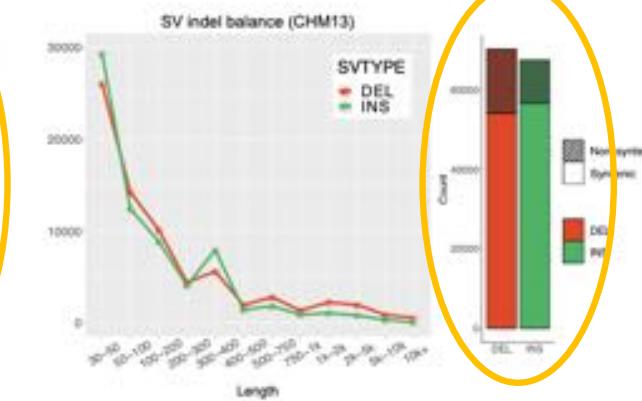
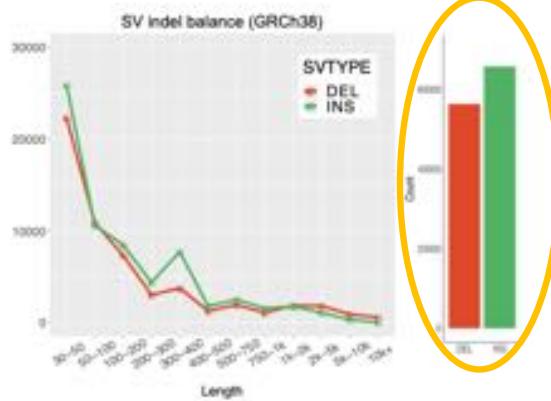
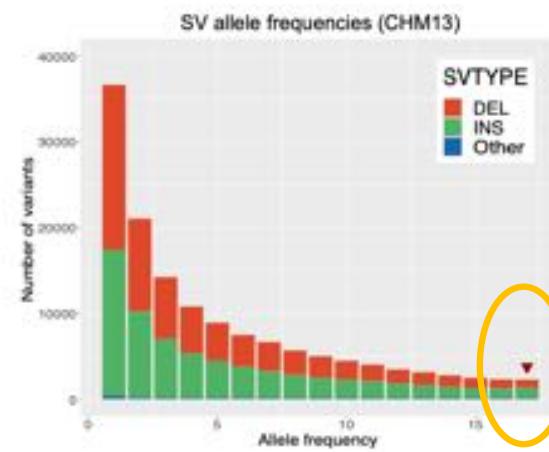
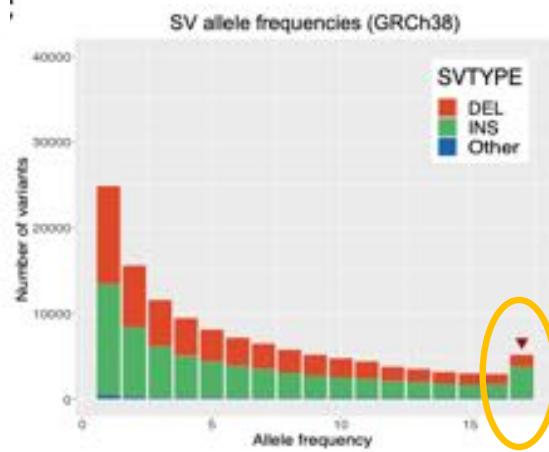
# Long-Read Analysis with T2T-CHM13



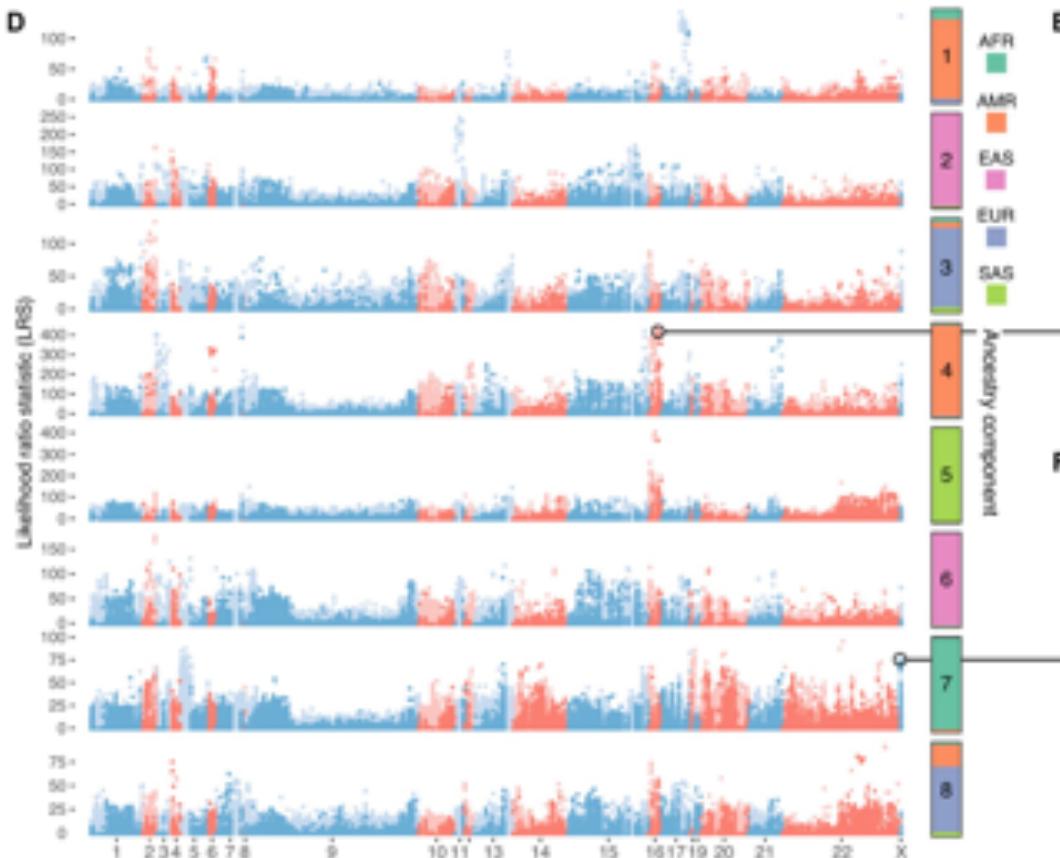
Melanie  
Kirsche



Sergey  
Aganezov



# Variants in Newly Resolved Regions

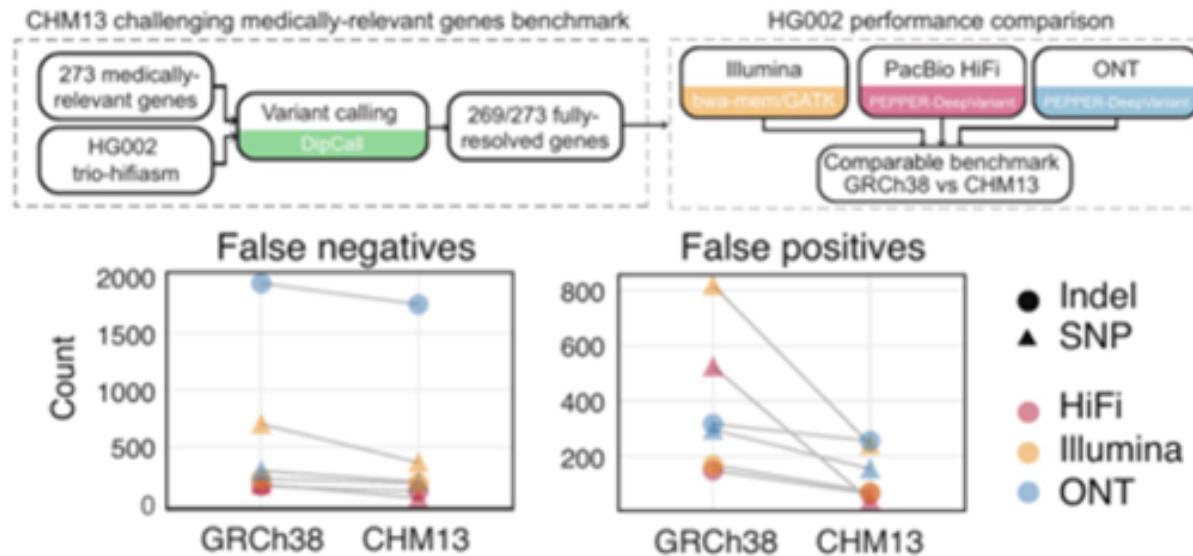


Stephanie  
Yan



Rajiv McCoy

# T2T-CHM13 Improves Clinical Genomics Variant Calling



Danny Miller



Daniela Soto



Megan Dennis



Justin Zook



Fritz Sedlazeck

For more information, see: (Wagner, J et al., NBT, 2022)

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

**Science**

**\$15**  
1 APRIL 2022  
**SPECIAL ISSUE**  
[science.org](http://science.org)

**Micoglia in chronic pain recovery and relapse** pp. 33 & 86

**Particle acceleration in a nova explosion** p. 77

**FILLING THE GAPS**  
Closing in on a complete human genome p. 42

Science

LOG IN [BECOME A MEMBER](#)

HOME > COLLECTIONS > COMPLETING THE HUMAN GENOME

## 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.

6 RESULTS FOUND

**SPECIAL ISSUE RESEARCH ARTICLE**  
**Segmental duplications and their variation in a complete human genome**  
BY MITCHELL R. VOLLMER, XAVI GUERRAT, PHILIP C. DISHONIK, LUDOVICA MERCURI, WILLIAM T. HARVEY, ARIEL GERSHMAN, MARK DEDKOWSKI, ARVIS SULOVARI, KATHERINE M. MUNSON, ALEXANDRA P. LEWIS, [...] EVAN E. SCHLESINGER

SCIENCE • VOL. 376, NO. 6588 • 9 APRIL 2022

**SPECIAL ISSUE RESEARCH ARTICLE**  
**Complete genomic and epigenetic maps of human centromeres**  
BY NICOLAS ALTEMOSSE, GLENNIA A. LOGGISON, ANDREY V. BZKADZE, PRAGYA SIDHWANI, SASHA A. LINGLEY, DINA V. CALDAIS, SAVANNAH J. HOYT, LEV URALSKY, FEDOR B. RABINOVICH, COLIN J. SHIER, [...] KAREN H. MIGA

SCIENCE • VOL. 376, NO. 6588 • 9 APRIL 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, LEONARD G. DE UMA, CHARLES LIMOUSE, REZZA HALABIAN, LUKE WOJENSKI, MATIAS RODRIGUEZ, [...] RACHEL J. HARRIS

+16 authors SCIENCE • VOL. 376, NO. 6588 • 9 APR 2022

**SPECIAL ISSUE RESEARCH ARTICLE**  
**A complete reference genome improves analysis of human genetic variation**  
BY SERGEY AGAMENZOV, STEPHANIE M. YAN, DANIEL K. SOTO, MELANIE KIRSCH, SAMANTHA ZAKATE, PAVEL KVOYEVICH, DYLAN J. TAYLOR, KISHWAR SHAFIN, ALAINA SHUMATE, CHUNLIN XIAO, [...] MICHAEL C. SCHATTNER

SCIENCE • VOL. 376, NO. 6588 • 9 APR 2022

**SPECIAL ISSUE RESEARCH ARTICLE**  
**Epigenetic patterns in a complete human genome**  
BY ARIEL GERSHMAN, MICHAEL E. G. SAURA, XAVI GUERRAT, MITCHELL R. VOLLMER, PAUL W. HOOK, SAVANNAH J. HOYT, MITEN JAIN, ALAINA SHUMATE, ROHAM RAZNAKH, SERGEY KOREN, [...] WINSTON TIPPMAN

SCIENCE • VOL. 376, NO. 6588 • 9 APR 2022

**SPECIAL ISSUE RESEARCH ARTICLE**  
**The complete sequence of a human genome**  
BY SERGEY KUREK, SERGEY KOREN, ARVIND RHEE, MIKKO RAUTIÄRÖ, ANDREY V. BZKADZE, ALLA MIKHAELENKO, MITCHELL R. VOLLMER, NICOLAS ALTEMOSSE, LEV URALSKY, ARIEL GERSHMAN, [...] ADAM M. PHILLIPPI

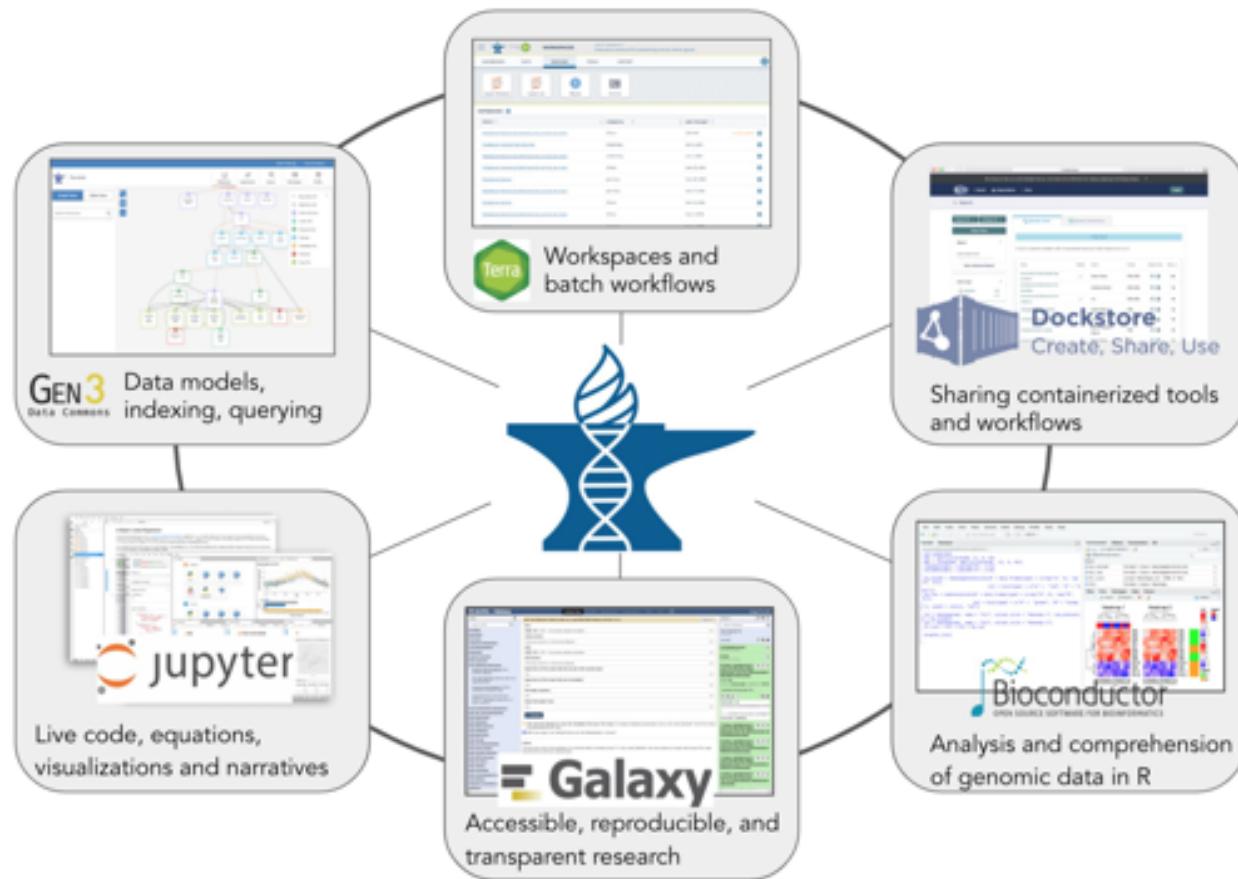
SCIENCE • VOL. 376, NO. 6588 • 31 MAR 2022 • 64-53

# T2T Team @ Santa Cruz, August 2022



# T2T Team @ TIME100, June 2022





**Inverting the model of genomics data sharing with the  
NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL)**  
Schatz, Philippakis et al. (2022) *Cell Genomics*. doi: <https://doi.org/10.1016/j.xgen.2021.100085>

# What is the AnVIL?

Scalable and interoperable computing resource for the genomics scientific community

- **Cloud-based infrastructure**
  - Highly elastic; shared analysis and computing environment
- **Data access and security**
  - Genomic datasets, phenotypes and metadata
  - Large datasets generated by NHGRI programs, as well as other initiatives / agencies
  - dbGaP Authenticated sharing of primary and derived datasets
- **Collaborative computing environment for datasets and analysis workflows**
  - Storage, scalable analytics, data visualization
  - Security, training & outreach, with new models of data access
  - ...for both users with limited computational expertise and sophisticated data scientist users

The screenshot shows the homepage of the AnVIL project at [anvilproject.org](https://anvilproject.org). The header features the AnVIL logo and navigation links for About, Data, Tools, Training, News, Events, FAQ, Contact, and NCPI. The main title "Migrate Your Genomic Analysis Workflows to the Cloud" is prominently displayed. Below the title, a subtitle reads: "Analyze large, open & controlled-access genomic datasets with familiar tools and reproducible workflows in a secure cloud-based computing environment." Two columns of cards provide information about Terra, GEM3, Dockstore, and NIH Cloud Platform Interoperability. A summary section at the bottom right highlights 5 consortia, 100 cohorts, 75K subjects, 90K samples, and 1.1PB of size.

<https://anvilproject.org>

Schatz, Philippakis et al. (2022) *Cell Genomics*  
doi: <https://doi.org/10.1016/j.xgen.2021.100085>



# AnVIL Data Dashboard

Extensive unrestricted and protected  
data sets already available within  
AnVIL

- 484 cohorts (CCDG, CMG,  
GTEx, 1000G, eMerge)
- 614k subjects
- 4.7Pb and rapidly growing
- Open access (e.g. 1000G), dbGaP  
authenticated (e.g. GTex) and  
consortium authenticated (e.g.  
CCDG) options available

Consortium	Cohorts	Samples	Participants	Size (TB)
1000 Genomes	1	3,202	3,202	72.98
CCDG	374	422,780	422,504	2,809.31
CMG	72	24,401	21,642	130.03
CMH	5	3,098	3,086	97.56
Convergent Neuroscience	4	2,547	2,548	6.91
CSER	6	2,498	1,471	160.81
eMERGE	3	114,118	114,118	73.43
GTEx	6	31,302	2,031	324.78
HPRC	1	57	47	260.23
PAGE	5	903	903	21.99
T2T	1	0	3,202	572.10
Unspecified	1	368	368	0.60
WGSFO1	5	9,588	9,575	177.36
Totals	484	614,860	584,687	4,708.08

<https://anvilproject.org/data>



# AnVIL is “renting computers”



WORKSPACES

Workspaces > hoffman-ava/Bioconductor-Workflow-DESeq... E.1\_DESeq2Analysis.ipynb

PREVIEW (READ-ONLY) EDIT PLAYGROUND MODE

### Introduction

This vignette will walk you through how to examine results from a DESeq2 analysis. The output data should have been saved to the bucket in the previous vignette [DESeq2 Analysis](#).

### Installation

Instructions for installing packages necessary for this notebook are given in [An Overview of AnVIL BulkRNASeq](#). Refer to that vignette for installation steps.

Load the packages to be used in this notebook:

```
In [1]: # Load packages
suppressPackageStartupMessages({
  library(DESeq2)
  library(ggplot2)
})
```

### Load the DESeq results:

```
In [2]: # Move the result saved in the bucket to the compute workspace
AnVIL::avfiles_restore(source = "DESeq_result.RData")

# Load the results
dds <- readRDS("DESeq_result.RData")
dds
```

Copying gs://fc-70e013b-f0f4-456c-8eb6-83ab2553af18/DESeq\_result.RData...
/ [0/1 files] 0.0 B/ 22.4 MiB] 0% Done
/ [0/1 files] 264.0 KiB/ 22.4 MiB] 1% Done
= [1/1 files] 22.4 MiB/ 22.4 MiB] 100% Done

## Standard Computing

- You buy a laptop one time
- You get that one laptop
- You pay little per use

## Cloud computing

- You use any web browser
- You rent the computers
- You pay per hour/gigabyte/etc.

# Cloud Architecture

- The cloud is built from several very large clusters of computers
  - Effectively infinite resources
  - High-end servers with many cores, many GB RAM, high speed networking, and exabytes of storage
- Computers run in a virtualized environment
  - Cloud providers subdivide large nodes into smaller instances
  - You are 100% protected from other users on the machine
  - You get to pick the operating system, all software installed

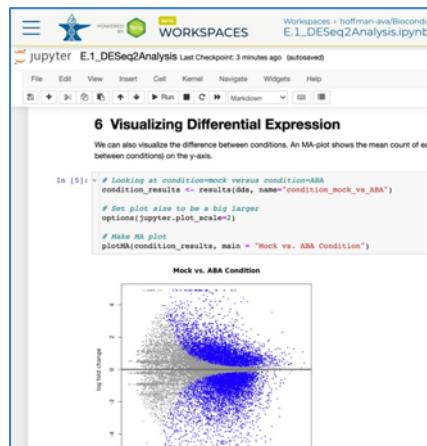


<https://www.google.com/about/datacenters/locations/>

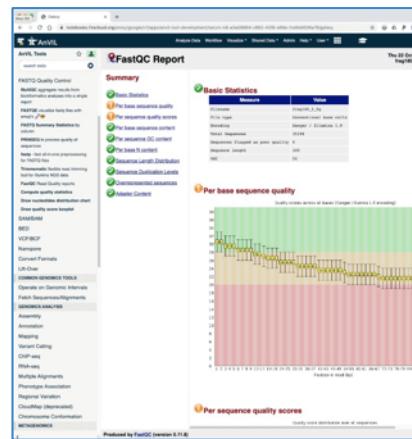


[https://en.wikipedia.org/wiki/Virtual\\_machine](https://en.wikipedia.org/wiki/Virtual_machine)

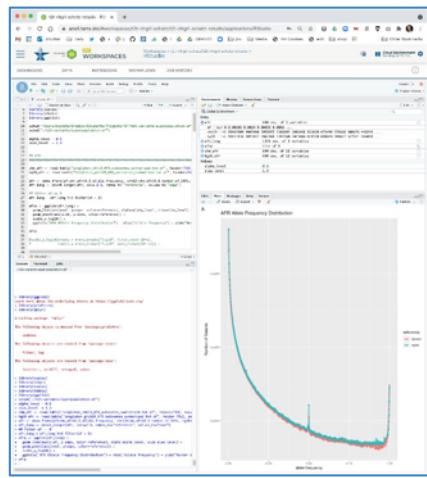
# AnVIL Analysis Platforms



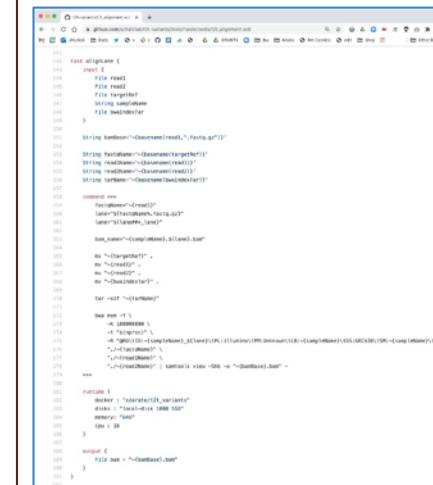
- + Code, text and plots in one document
- + Supports coding in Python or R
- Least scalable, not a complete IDE



- + Graphical interface for thousands of tools and workflows
- + Highly accessible and reproducible
- Tools must be preconfigured to use

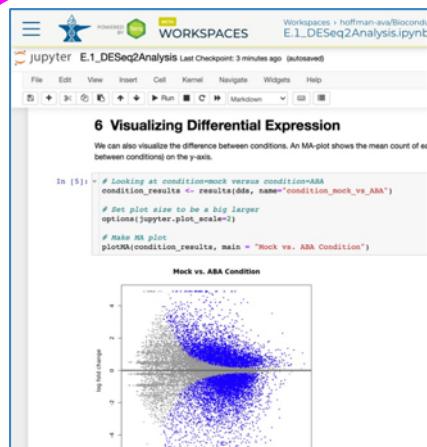


- + Feature rich IDE for programming in R
- + Rich statistics, ML, and visualizations
- Limited support for other programming languages



- + Extremely scalable and flexible
- Most technically demanding
- Unpredictable and potentially large costs

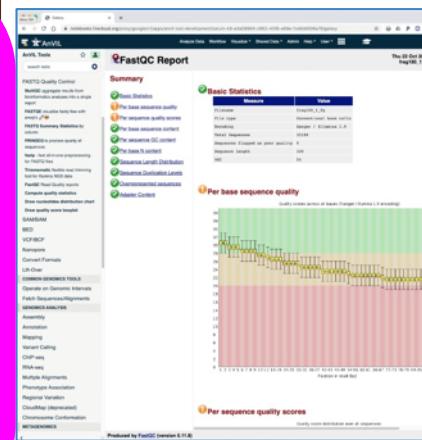
# AnVIL Analysis Platforms (coding)



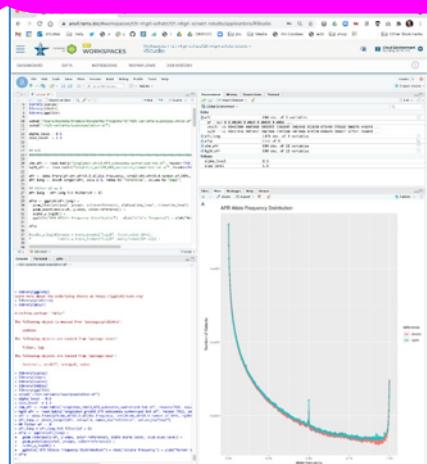
A screenshot of a Jupyter notebook titled "E\_1\_DESeq2Analysis.ipynb". The notebook contains Python code for visualizing differential expression. It includes a scatter plot titled "Mock vs. ABA Condition" showing log fold change versus negative log p-value.



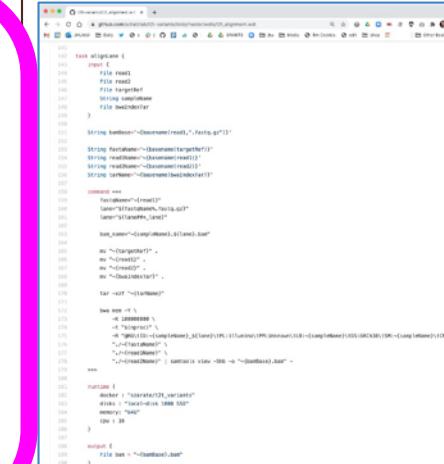
- + Code, text and plots in one document
- + Supports coding in Python or R
- Least scalable, not a complete IDE



- + Graphical interface for thousands of tools and workflows
- + Highly accessible and reproducible
- Tools must be preconfigured to use



- + Feature rich IDE for programming in R
- + Rich statistics, ML, and visualizations
- Limited support for other programming languages



A screenshot of a terminal window showing a WDL (Workflow Description Language) script and its execution logs. The logs show the flow of data from fastq files through various processing steps to a final BAM file.



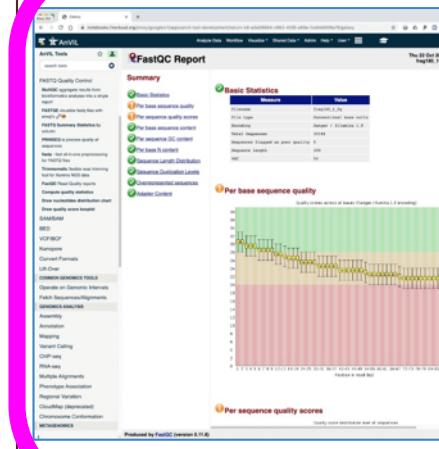
- + Extremely scalable and flexible
- Most technically demanding
- Unpredictable and potentially large costs

# AnVIL Analysis Platforms (GUI)

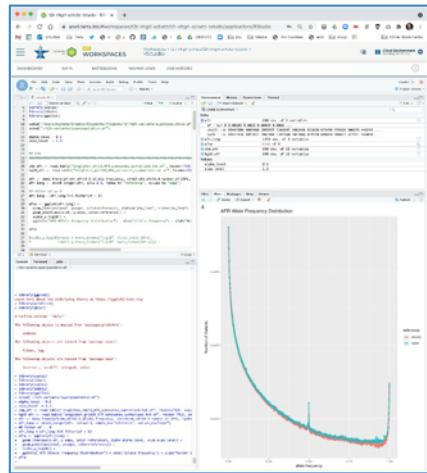
A screenshot of a Jupyter notebook titled "E\_1\_DESeq2Analysis.ipynb". The notebook shows code for visualizing differential expression, specifically comparing "Mock vs. ABA Condition". The code uses ggplot2 to create an MA-plot where the y-axis is "log fold change" and the x-axis is "neg log p-value".



- + Code, text and plots in one document
- + Supports coding in Python or R
- Least scalable, not a complete IDE



- + Graphical interface for thousands of tools and workflows
- + Highly accessible and reproducible
- Tools must be preconfigured to use



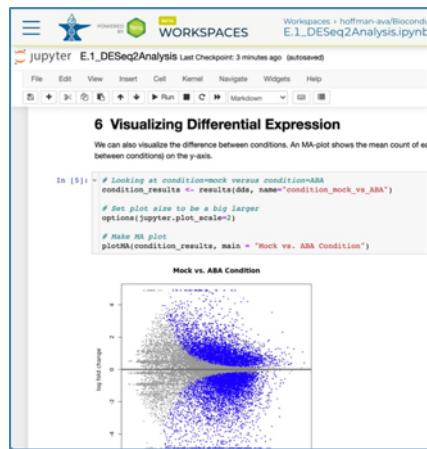
- + Feature rich IDE for programming in R
- + Rich statistics, ML, and visualizations
- Limited support for other programming languages

A screenshot of a WDL (Workflow Description Language) script. The script defines tasks for reading fastq files, filtering them, and then running a string sampler tool. It also includes sections for Docker container configuration and output handling.



- + Extremely scalable and flexible
- Most technically demanding
- Unpredictable and potentially large costs

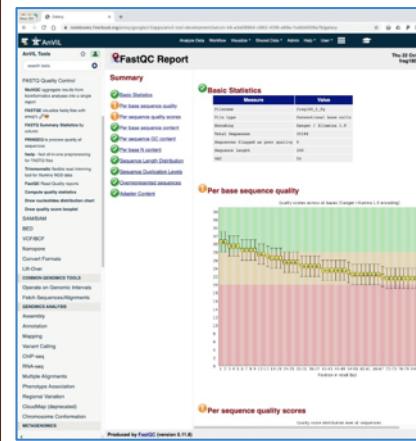
# AnVIL Analysis Platforms (workflows)



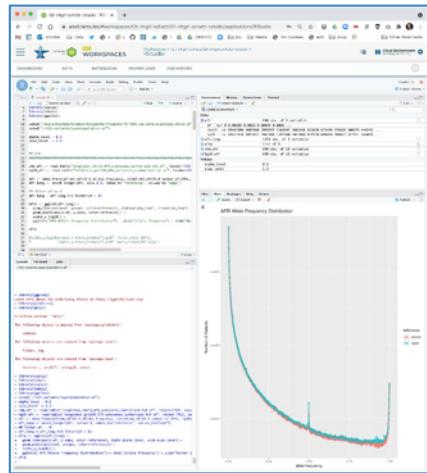
A screenshot of a Jupyter notebook titled "6 Visualizing Differential Expression". The code uses R to load a dataset, set plot size, and create an MA-plot comparing "Mock vs. ABA Condition". The plot shows log fold change on the y-axis and -log10(p-value) on the x-axis.



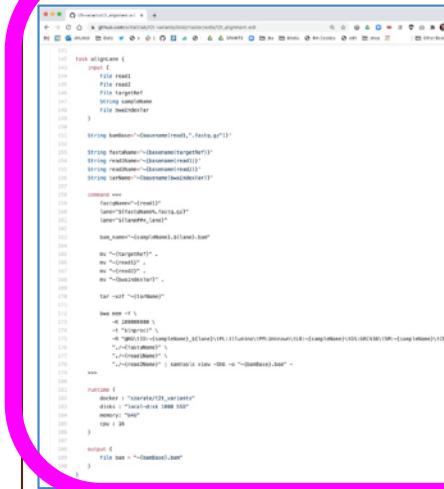
- + Code, text and plots in one document
- + Supports coding in Python or R
- Least scalable, not a complete IDE



- + Graphical interface for thousands of tools and workflows
- + Highly accessible and reproducible
- Tools must be preconfigured to use



- + Feature rich IDE for programming in R
- + Rich statistics, ML, and visualizations
- Limited support for other programming languages



A screenshot of a WDL (Workflow Description Language) workflow script. The script defines tasks for fastq processing, including filtering, quality trimming, and adapter removal. It also includes a task for running a pipeline and a final task for generating a report. The script uses various parameters and tool configurations.



- + Extremely scalable and flexible
- Most technically demanding
- Unpredictable and potentially large costs

# My perspective



## jupyter

- + Code, text and plots in one document
- + Supports coding in Python or R
- Least scalable, not a complete IDE



## = Galaxy

- + Graphical interface for thousands of tools and workflows
- + Highly accessible and reproducible
- Tools must be preconfigured to use



## R Studio®

- + Feature rich IDE for programming in R
- + Rich statistics, ML, and visualizations
- Limited support for other programming languages



## {wdl}

- + Extremely scalable and flexible
- Most technically demanding
- Unpredictable and potentially large costs

# Parallel Algorithm Spectrum

Embarrassingly Parallel



Each item is Independent

Loosely Coupled



Independent-Sync-Independent

Tightly Coupled



Constant Sync