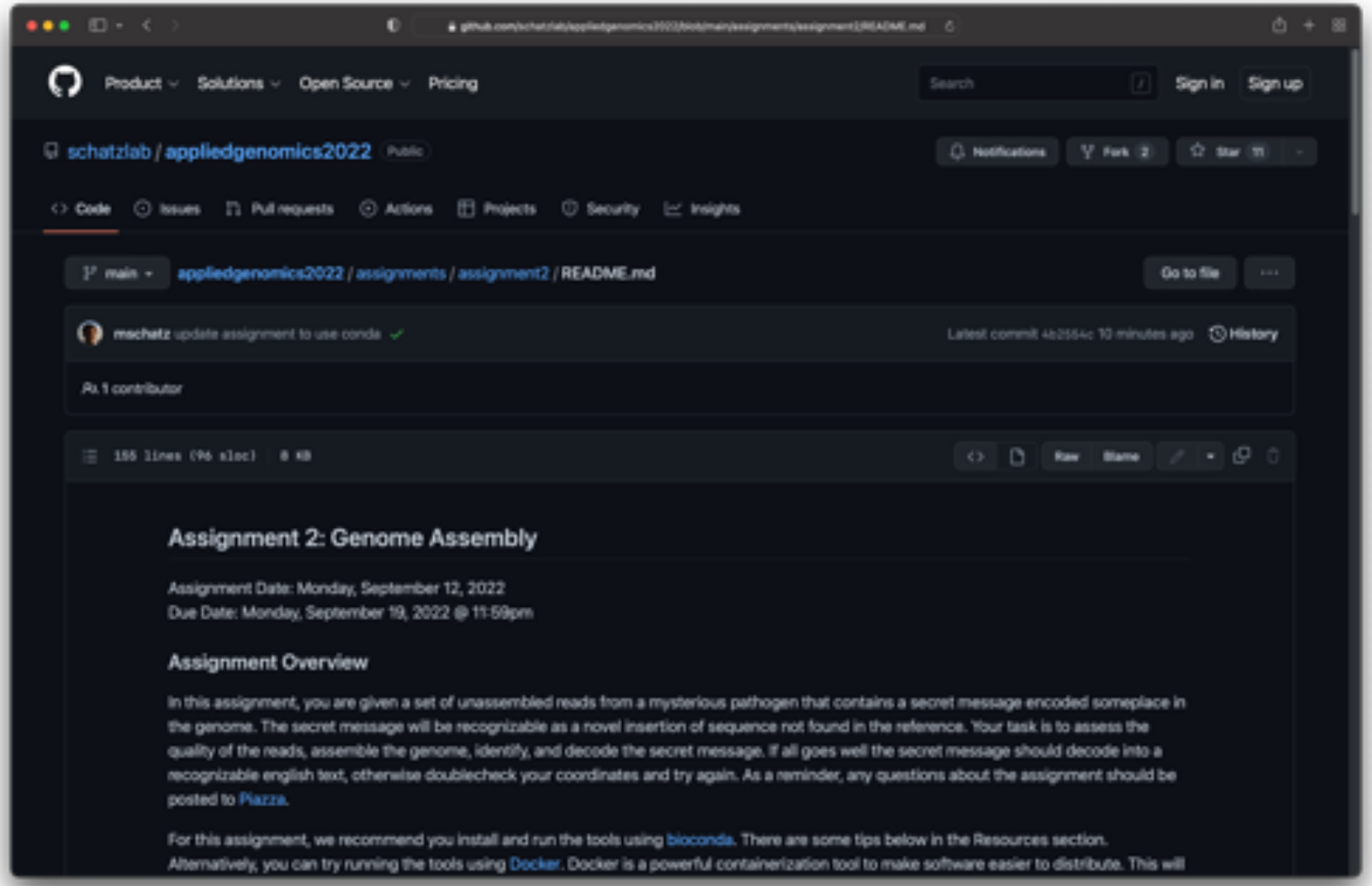
Cloud-scale genomics

Michael Schatz

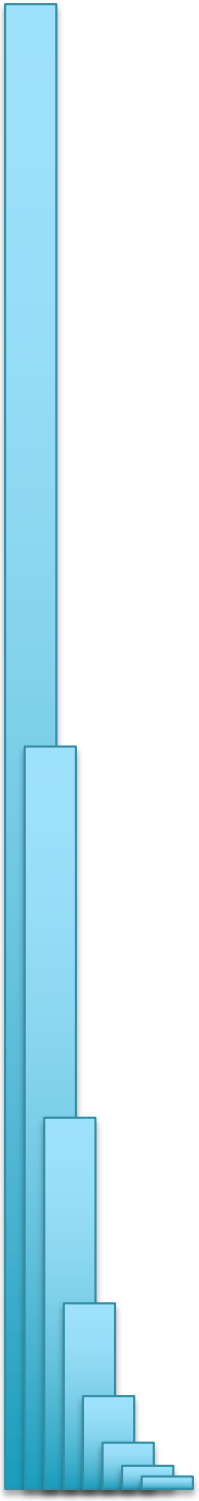
Sept 19, 2022

Lecture 6: Applied Comparative Genomics

Assignment 2: Genome Assembly
Due Monday Sept 19 by 11:59pm



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

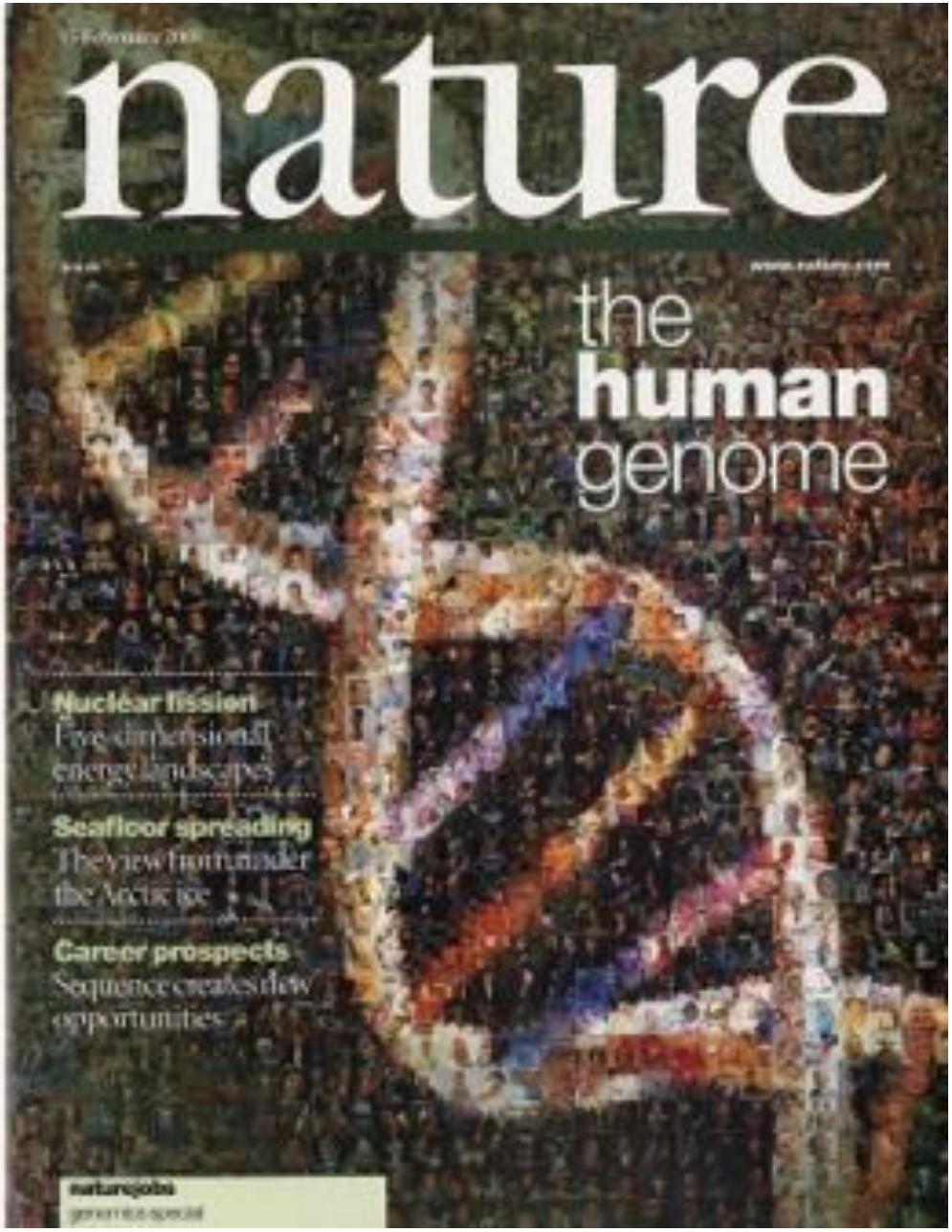


Wednesday's lecture

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)

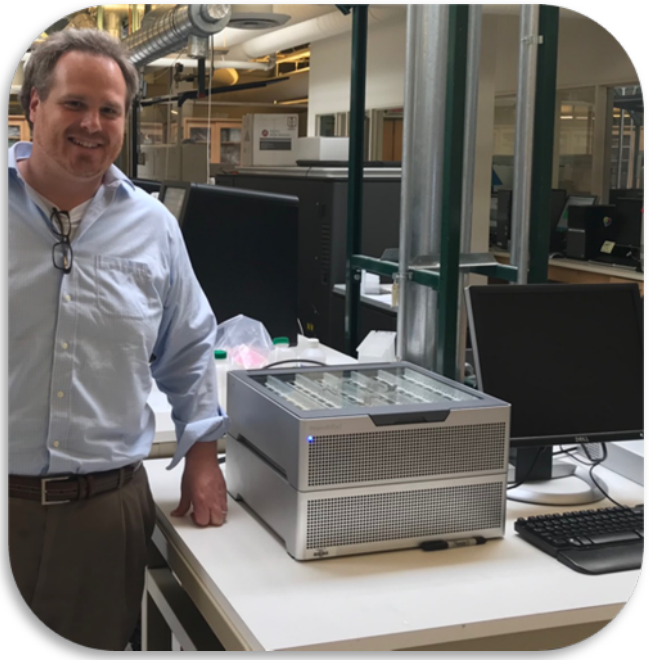


Initial sequencing and analysis of the human genome
International Human Genome Sequencing Consortium
Nature 409, pp 860–921 (2001)

Single Molecule Long Read Sequencing

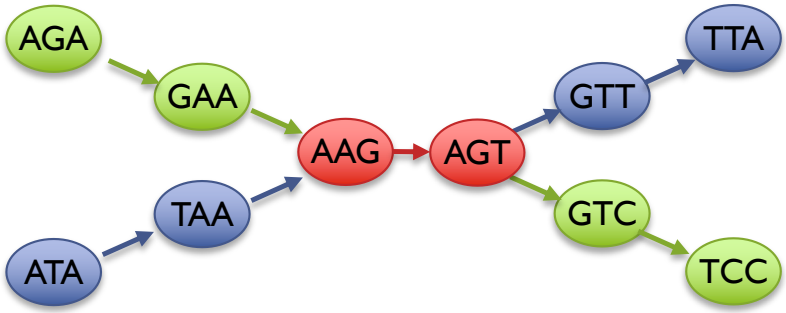
PacBio
Sequel II

Oxford Nanopore
PromethION



Two Paradigms for Assembly

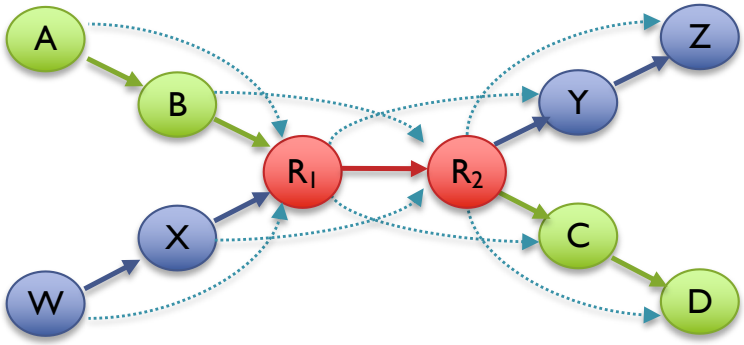
de Bruijn Graph



Short read assemblers

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

Overlap Graph

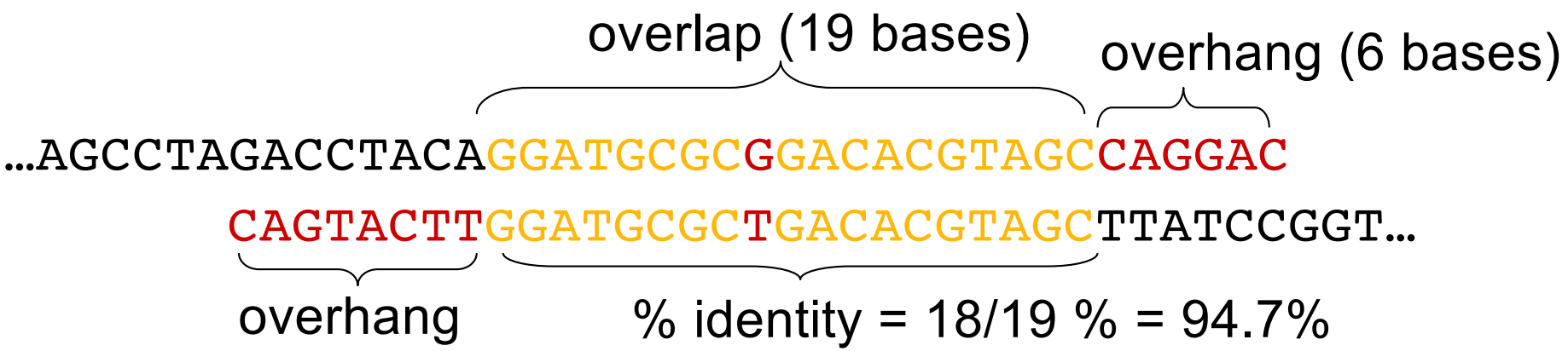


Long read assemblers

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

Assembly of Large Genomes using Second Generation Sequencing
Schatz MC, Delcher AL, Salzberg SL (2010) *Genome Research*. 20:1165-1173.

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.

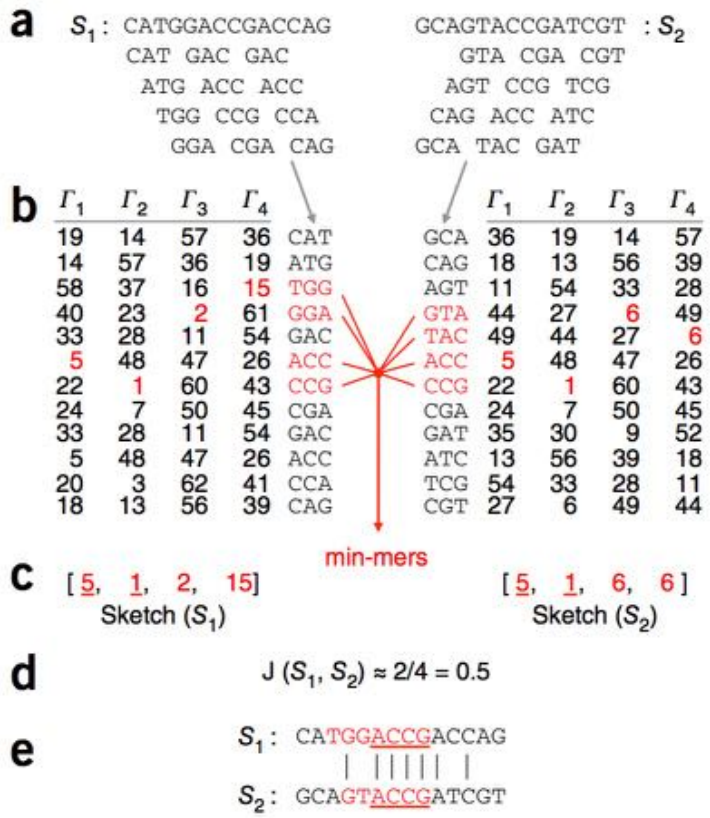
[How do we compute the overlap?]

[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: |intersection| / |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



Assembling large genomes with single-molecule sequencing and locality-sensitive hashing
Berlin et al (2015) *Nature Biotechnology*

Unitigging: Pruning the Overlap Graph

The overlap graph has many redundant edges:

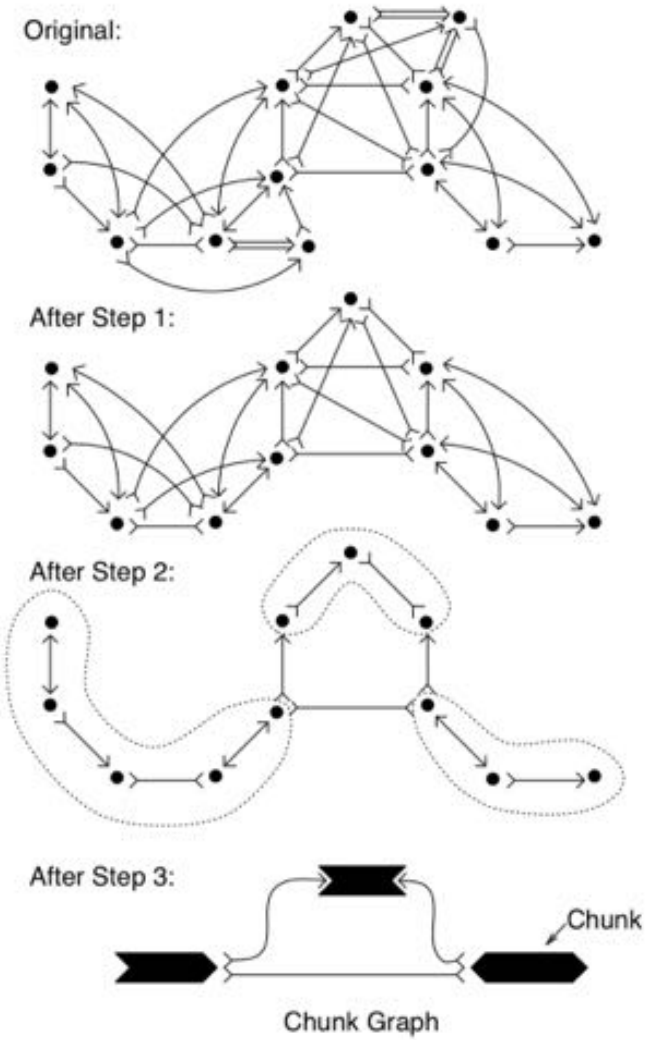
- If the average coverage is D, we should expect D overlaps at the beginning of the read, and D at the end

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 -> B, and B->C, remove the transitive edge A->C

3. ***"Chunkification":*** Linear subgraphs define uniquely assemblible segments: "unitigs"



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

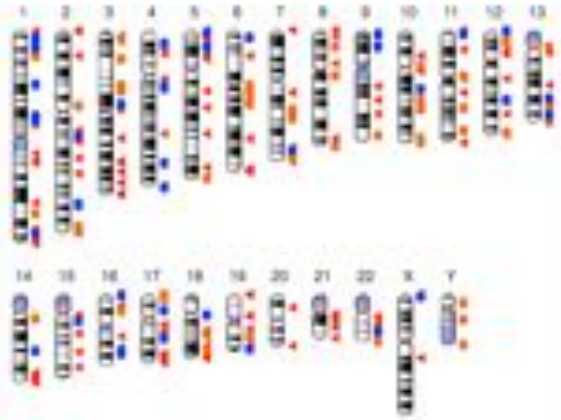
Genome Reference Consortium

GRC Home | Data | Help | Report an Issue | Contact Us | Credits | Curators Only

Human Overview | Human Genome Issues | Human Assembly Data

Human Genome Overview

Information about the continuing improvement of the human genome



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

Ideogram of the latest human assembly, GRCh38.p11

The GRC is working hard to provide the best possib
by both generating multiple representations (alterna
represented by a single path. Additionally, we are re
allows users who are interested in a specific locus o
affecting users who need chromosome coordinate s

Download data:

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

Transitioning to GRCh38? Try the NCBI Remapp
assembly alignments used by the GRC.

Next assembly update
The next assembly update (GRCh38.p12) will be

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 coordi
of patch scaffolds is now: 64 FIX and 59 NOVEL.
Assembly accessions: GenBank: GCA_000001405.26 , RefSeq: GCF_000001405.37

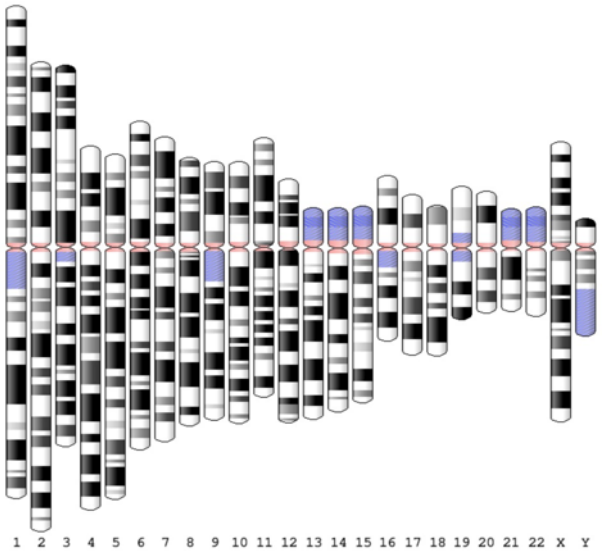
Pseudoautosomal regions

Name	Chr	Start	Stop
PAR#1	X	10,001	2,781,479
PAR#2	X	155,701,383	156,030,895
PAR#1	Y	10,001	2,781,479
PAR#2	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

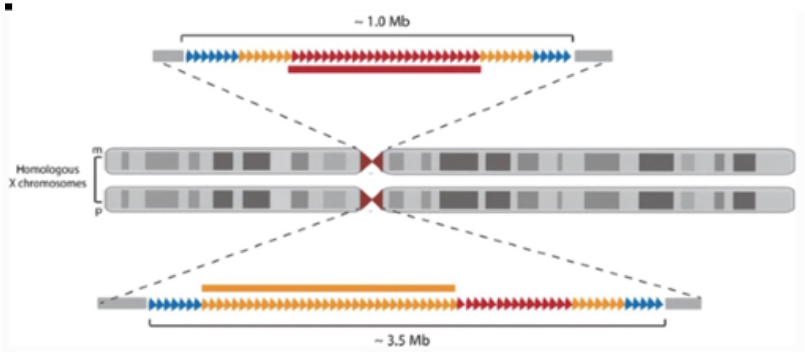
Functional studies need sequence

Reference gaps lead to artifacts

We don't know what we don't

Why has it taken so long?

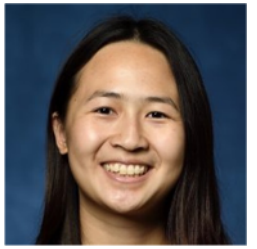
Repeats, repeats, repeats…



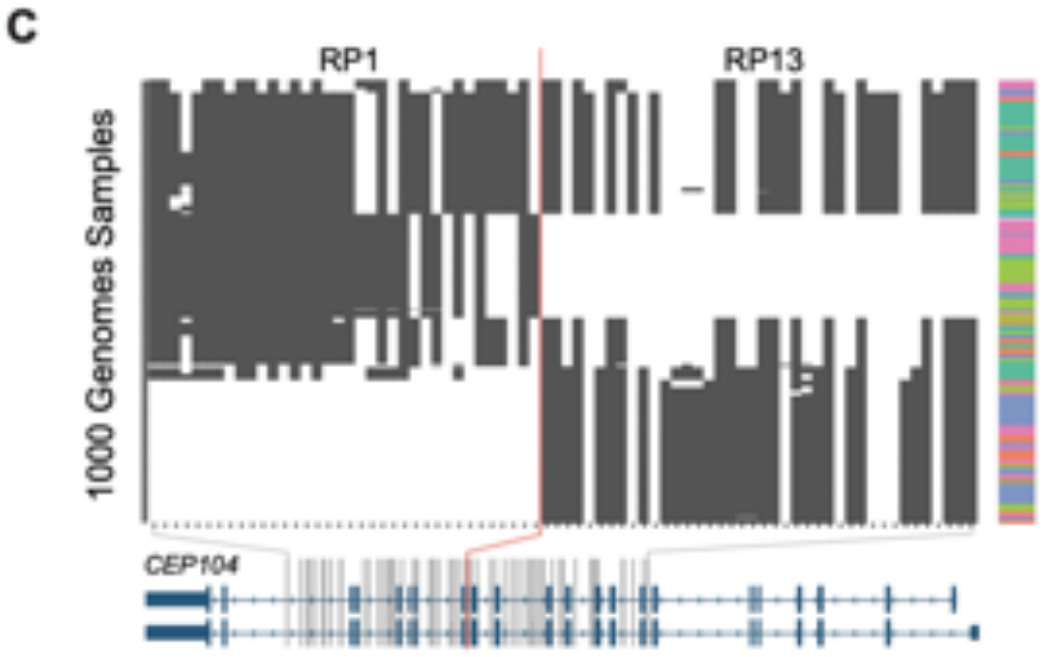
Miga 2015

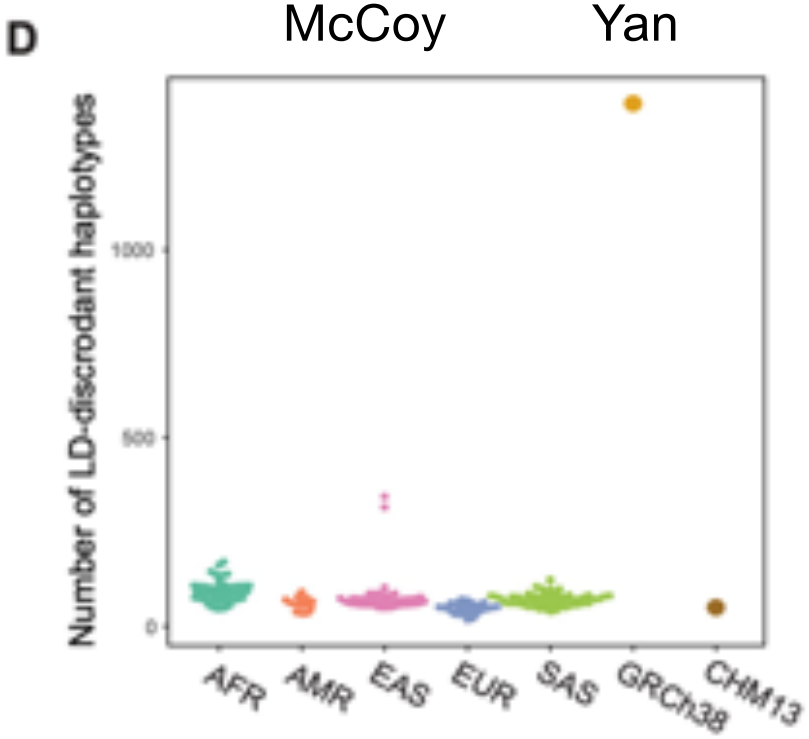
Local Ancestry Comparisons

Rajiv McCoy

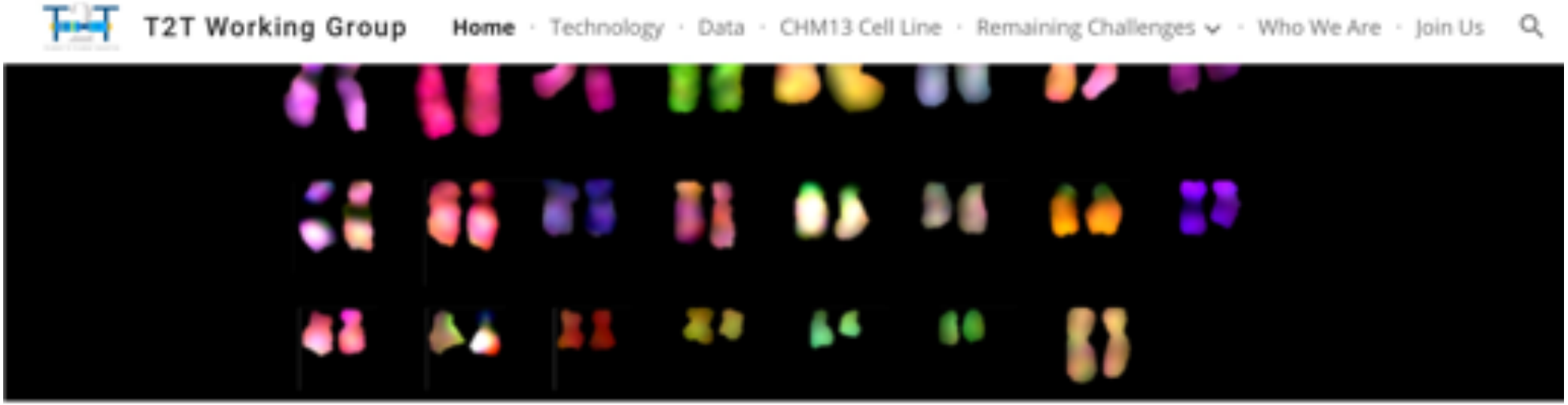


Stephanie Yan





Let's finish a human genome

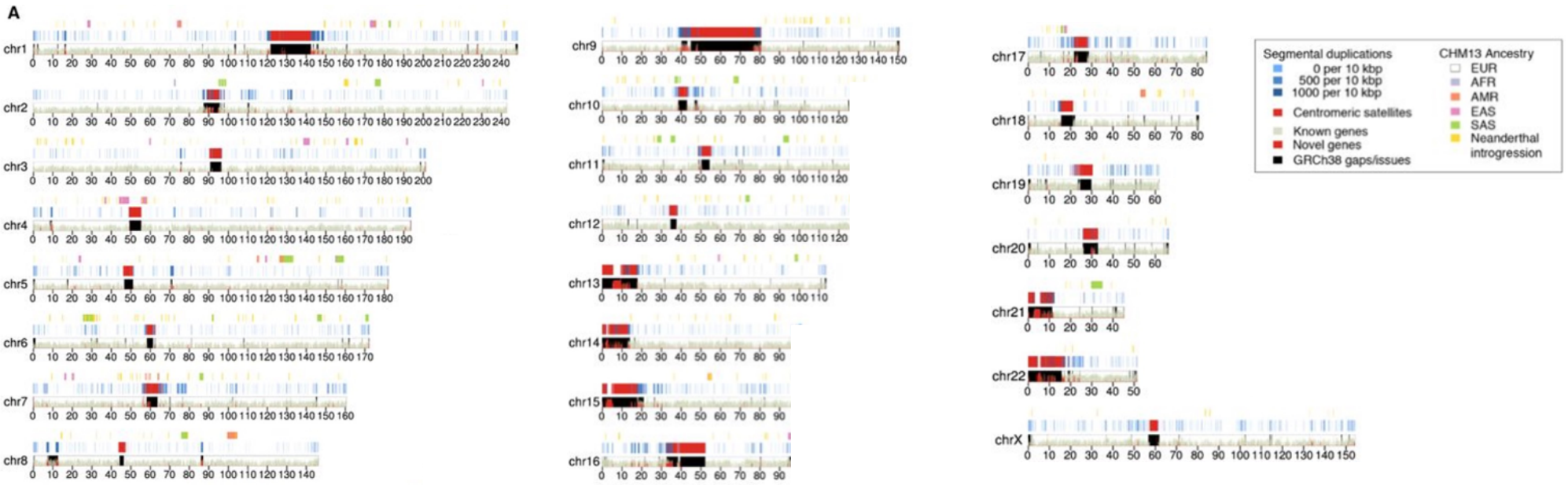
The Telomere-to-Telomere (T2T) consortium is an open, community-based effort to generate the first complete assembly of a human genome.

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

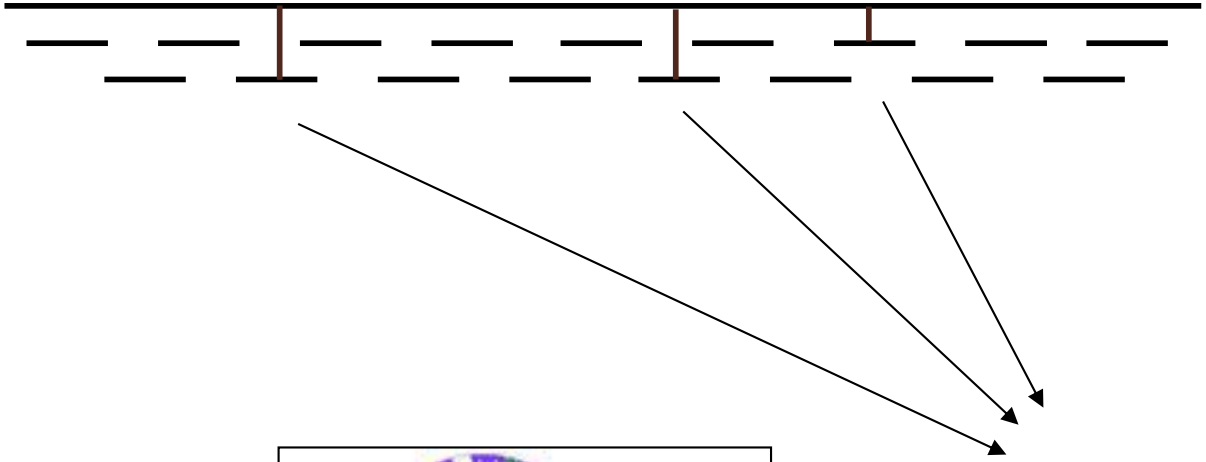


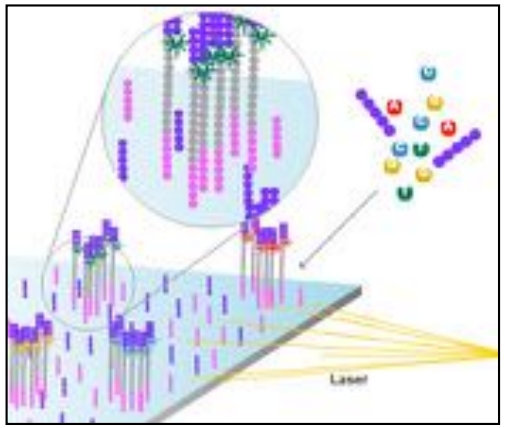
CHM13v1.1 genome size is **3.057 Gbp with zero Ns**
Every chromosome is telomere-to-telomere, quality estimated >Q70
~190 Mbp (~8%) of new sequence vs. GRCh38, fixes thousands of errors

(Nurk et al. Science, 2022)

Personal Genomics

How does your genome compare to the reference?



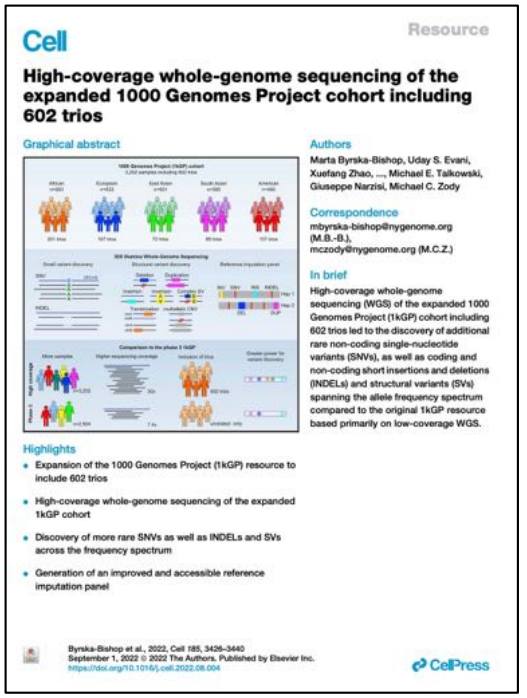


Heart Disease

Cancer

Presidential smile

T2T Variants: 1000 Genomes Project



Cell

Resource

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

Graphical abstract

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.

Highlights

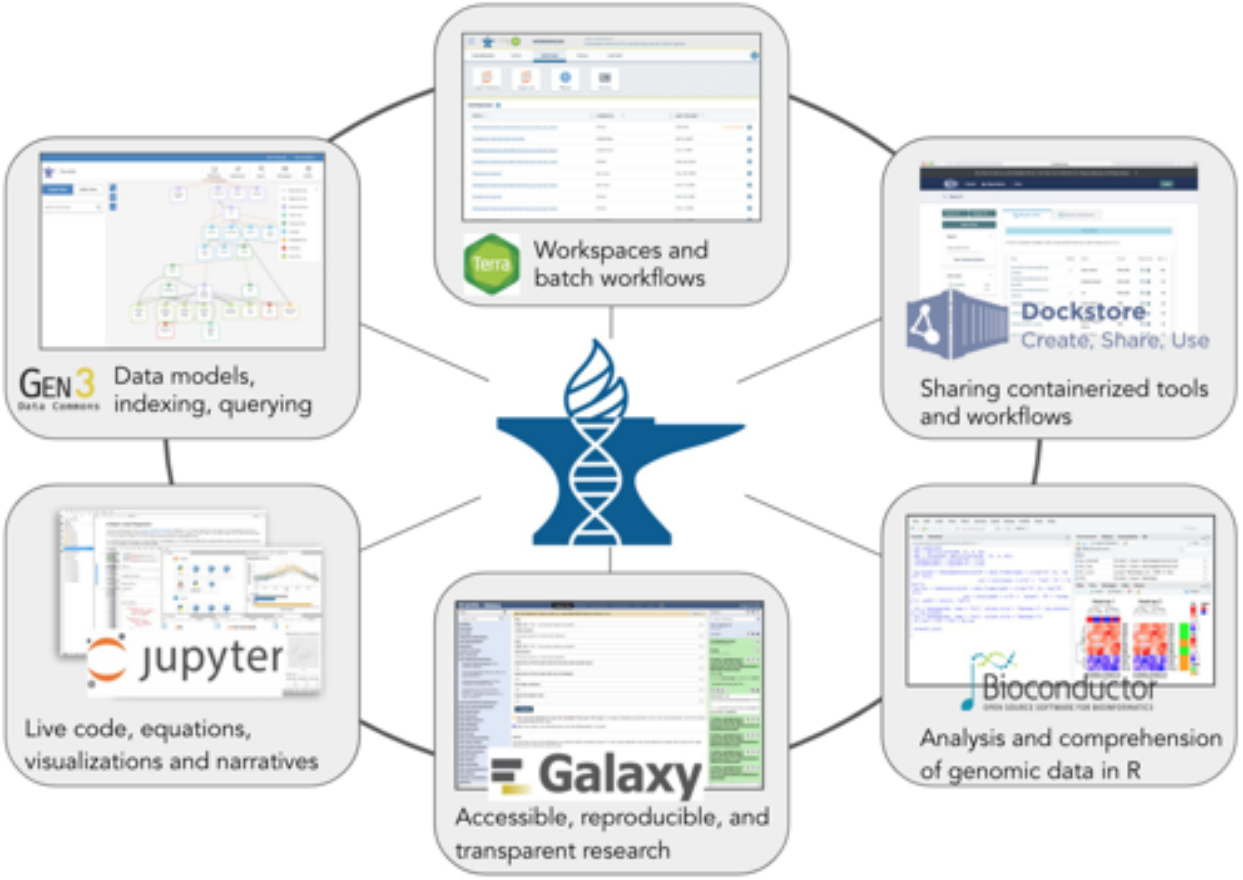
- Expansion of the 1000 Genomes Project (1kGP) resource to include 602 trios
- High-coverage whole-genome sequencing of the expanded 1kGP cohort
- Discovery of more rare SNVs as well as INDELs and SVs across the frequency spectrum
- Generation of an improved and accessible reference imputation panel

Byrska-Bishop et al., 2022, Cell 185, 3426–3440
September 1, 2022 © 2022 The Authors. Published by Elsevier Inc.

CellPress

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

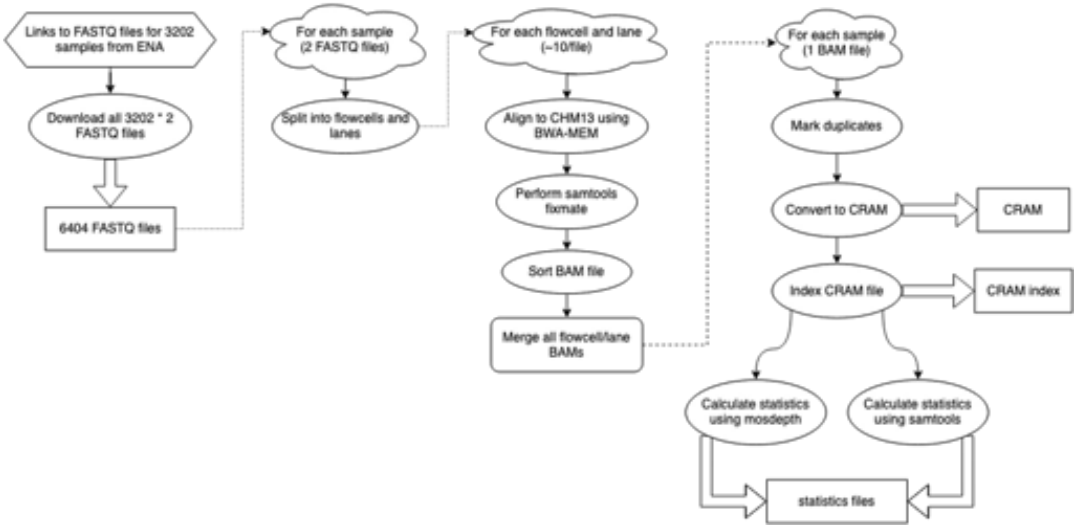


Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL)

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

3202 Genomes to go…

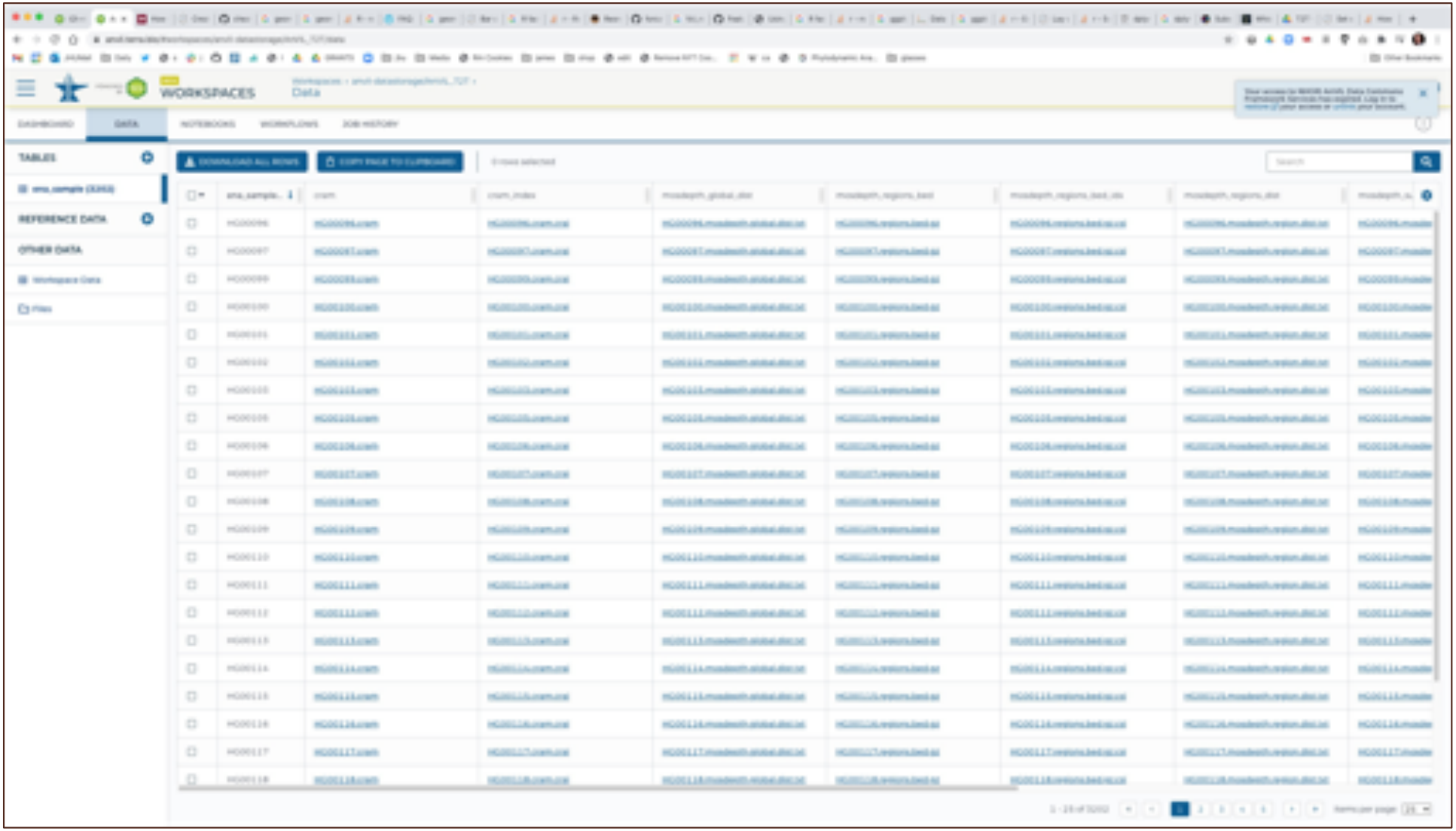
Samantha Zarate



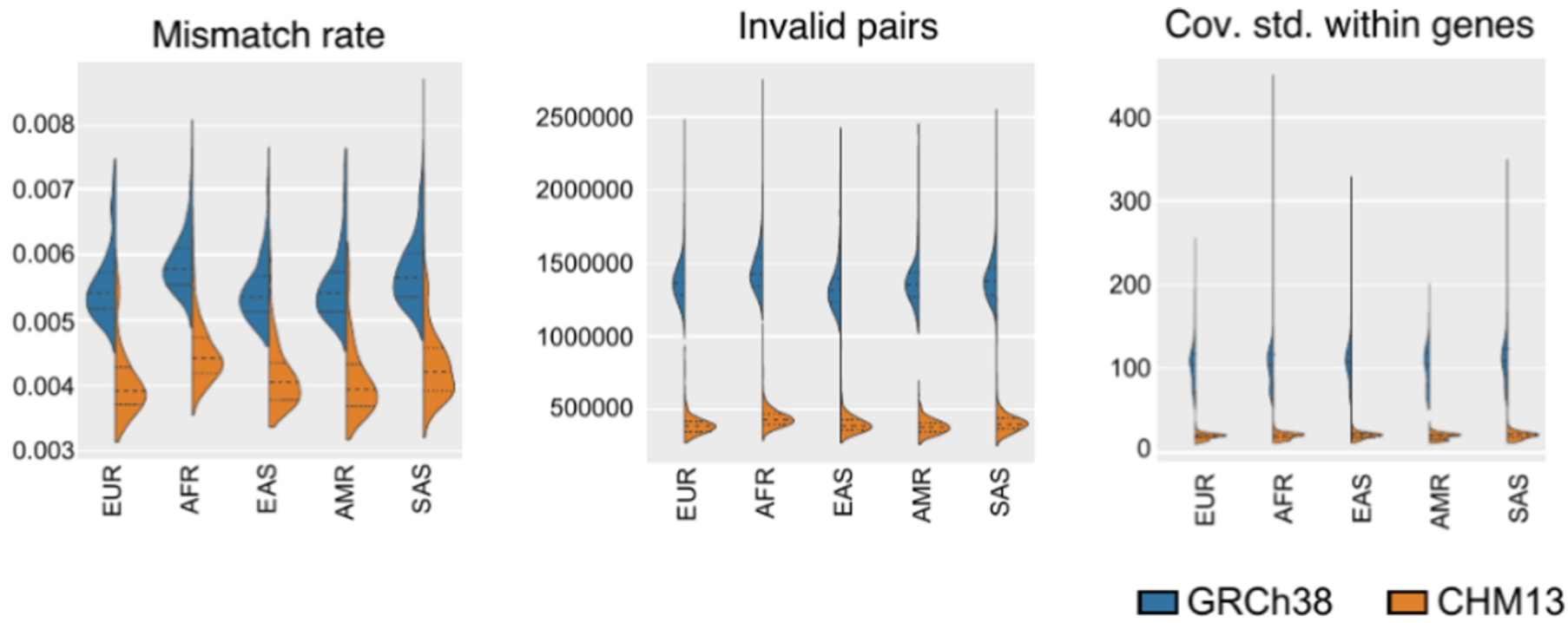

```
task samtoolsStats {
    input {
        File inputCram
        File cramIndex
        File targetRef
        String sampleName
    }

    command <<<
        samtools stats -r "~{targetRef}" \
            --reference "~{targetRef}" \
            -@ "$(nproc)" \
            "~{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

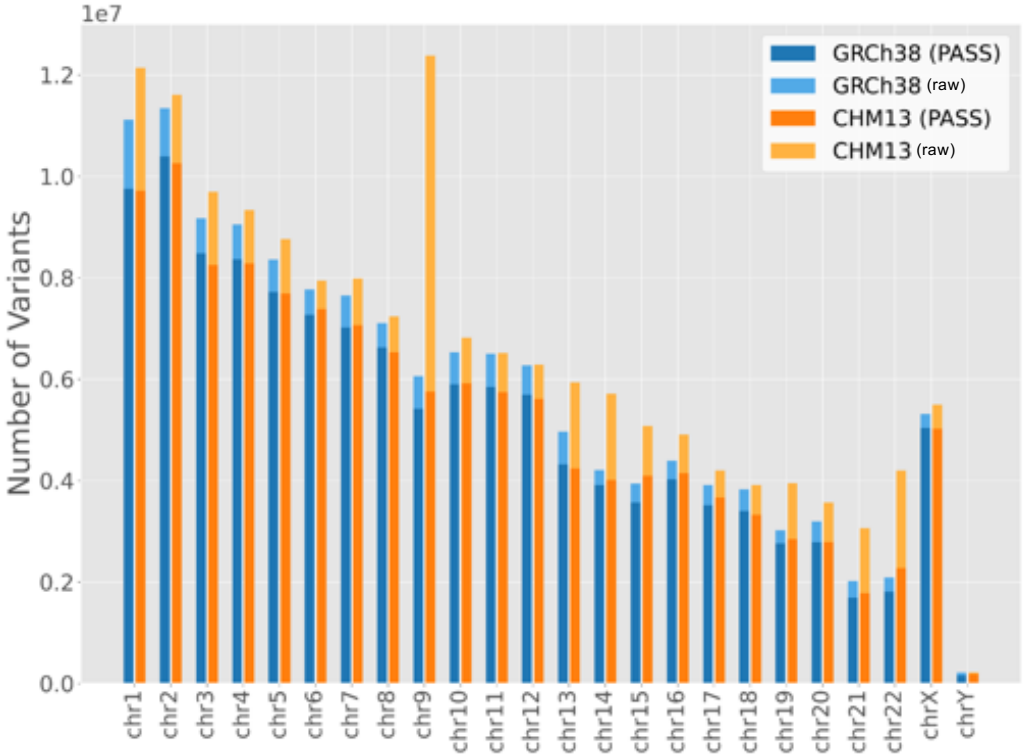
AnVIL Data Table



1000G Mapping on T2T-CHM13

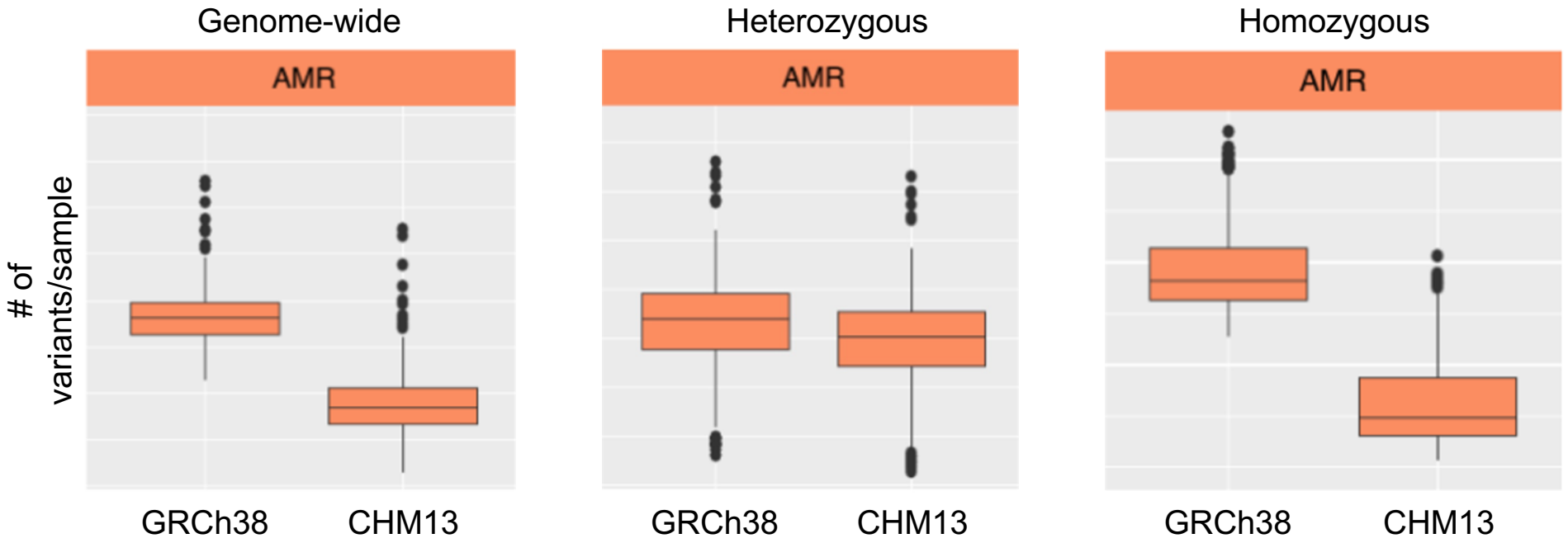


Across 1000G, Many More Variants Found Using CHM13

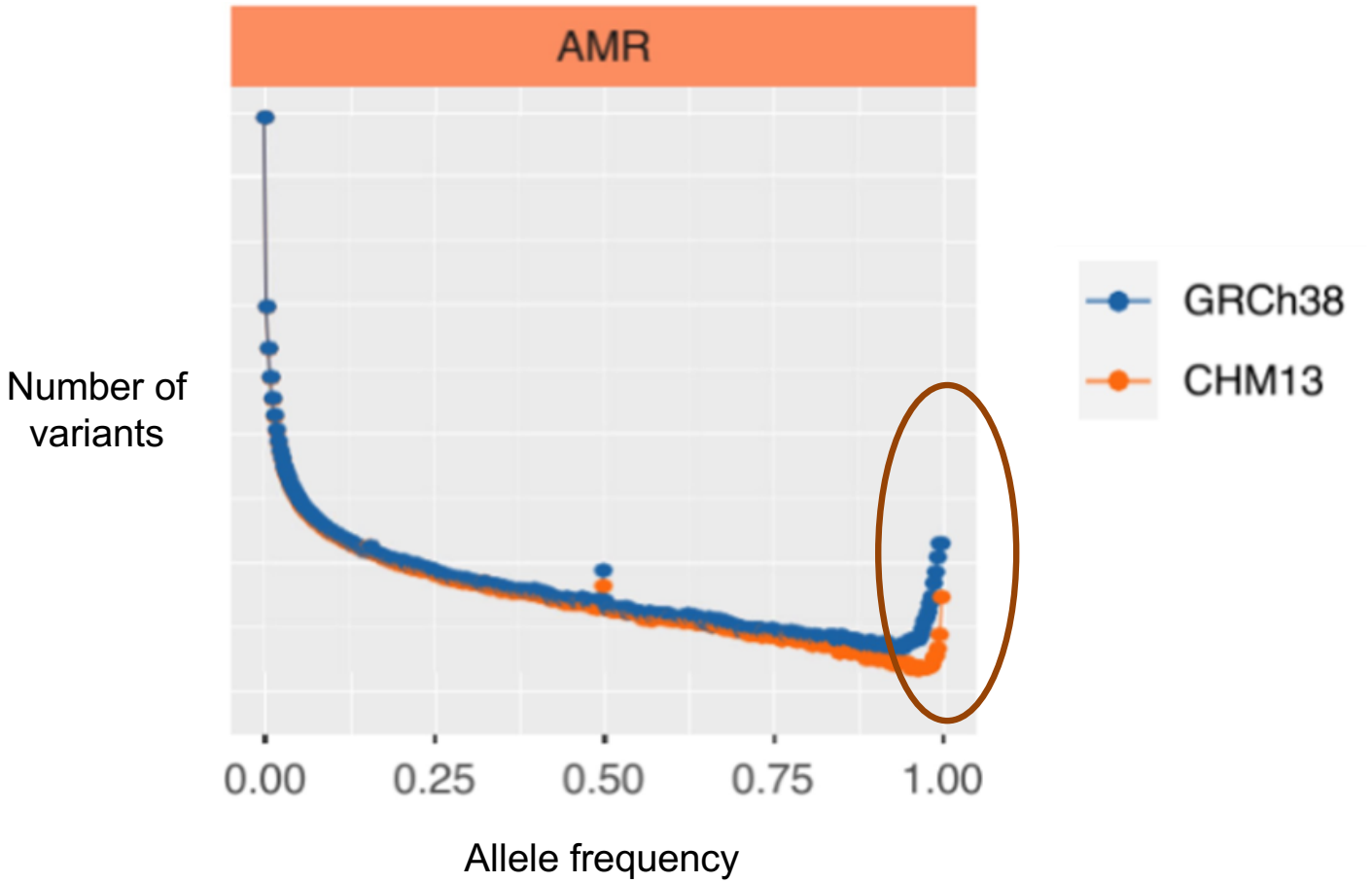


	GRCh38	CHM13
# PASS variants	125,484,020	126,591,489

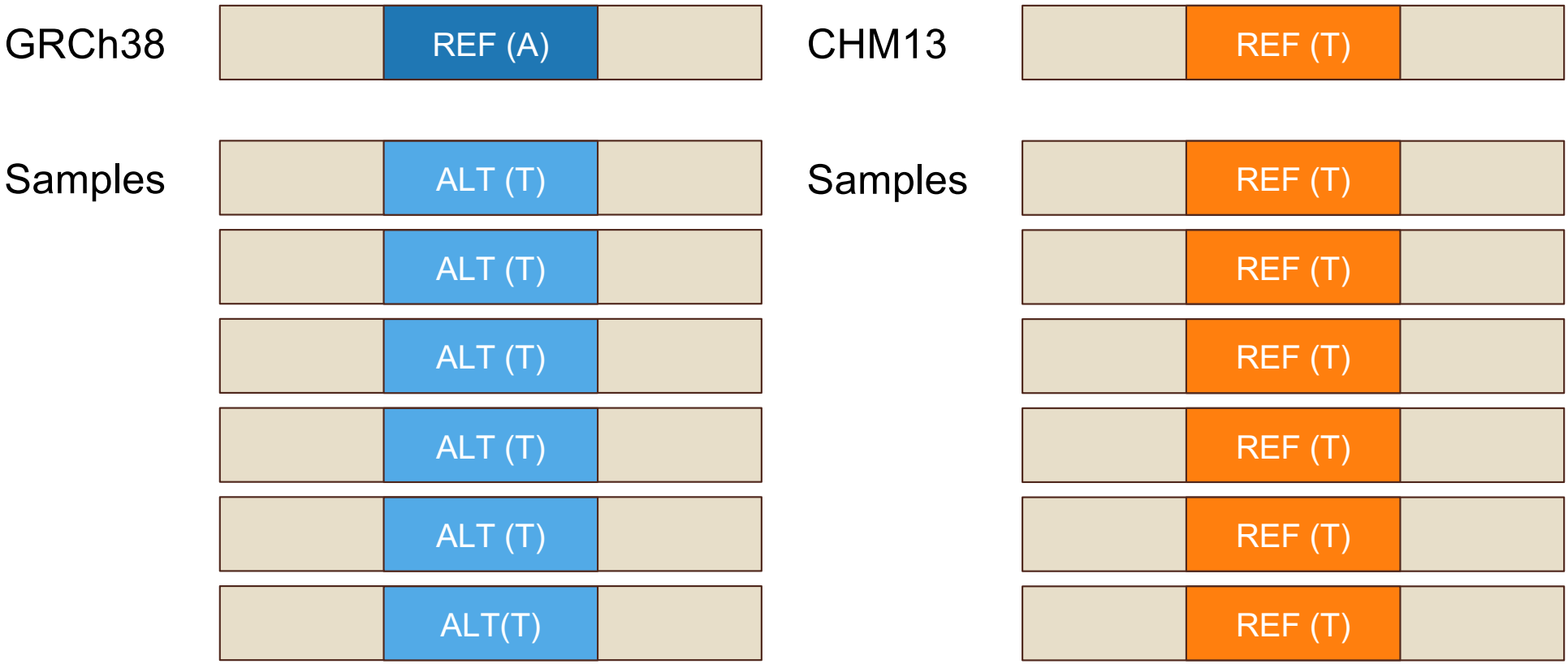
1000G Per-Sample Variant Counts on T2T-CHM13



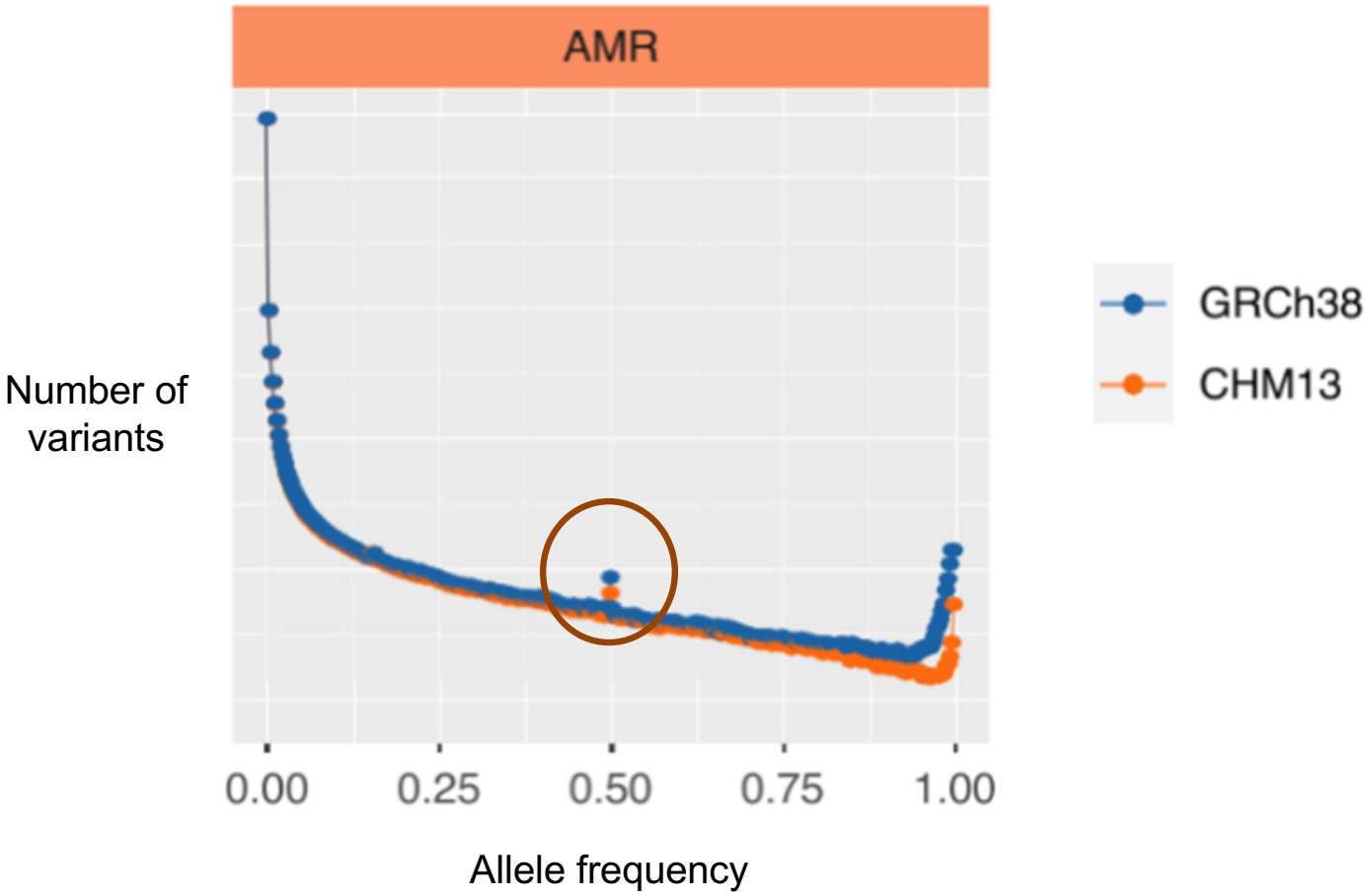
Explaining Decreased Per-Sample Count



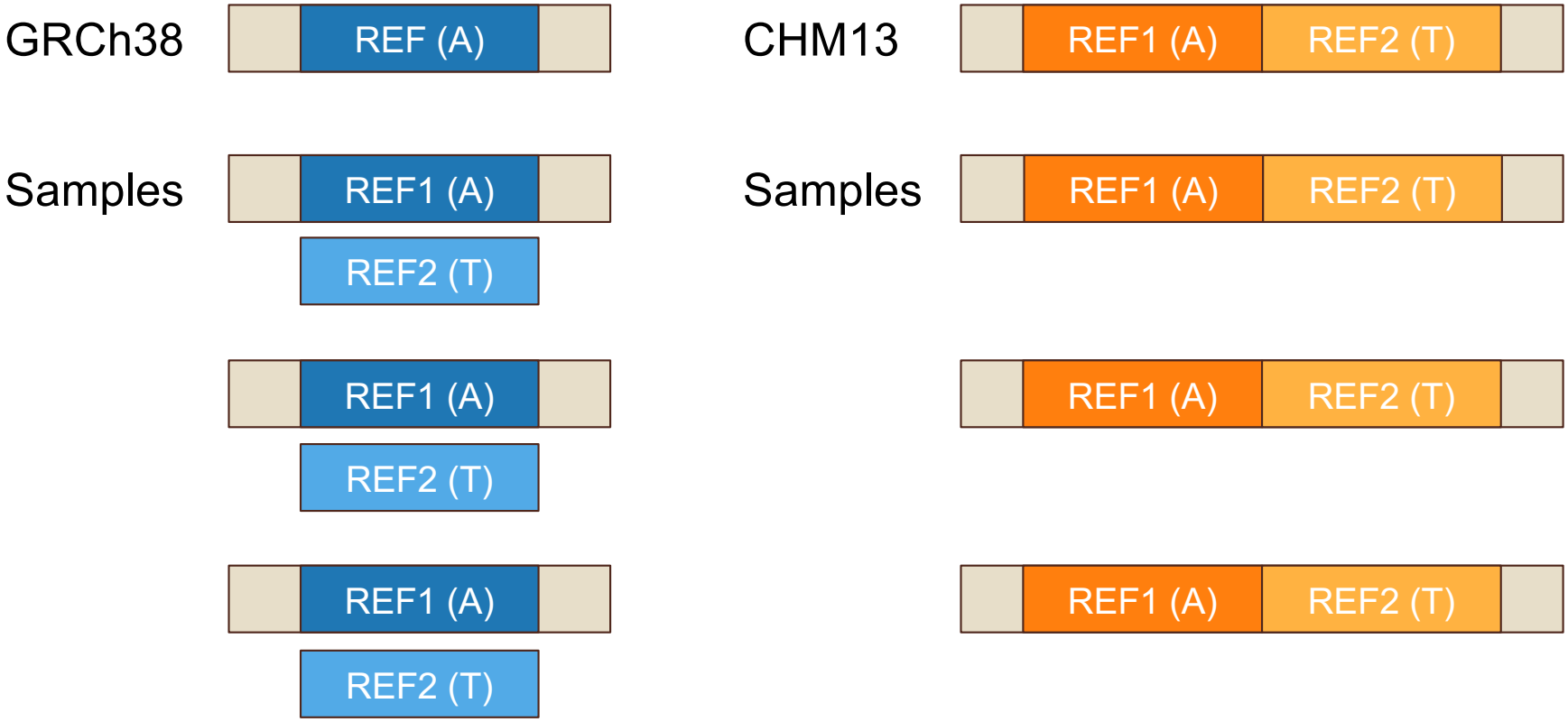
Allele Frequency = 1: Reference error / private variant



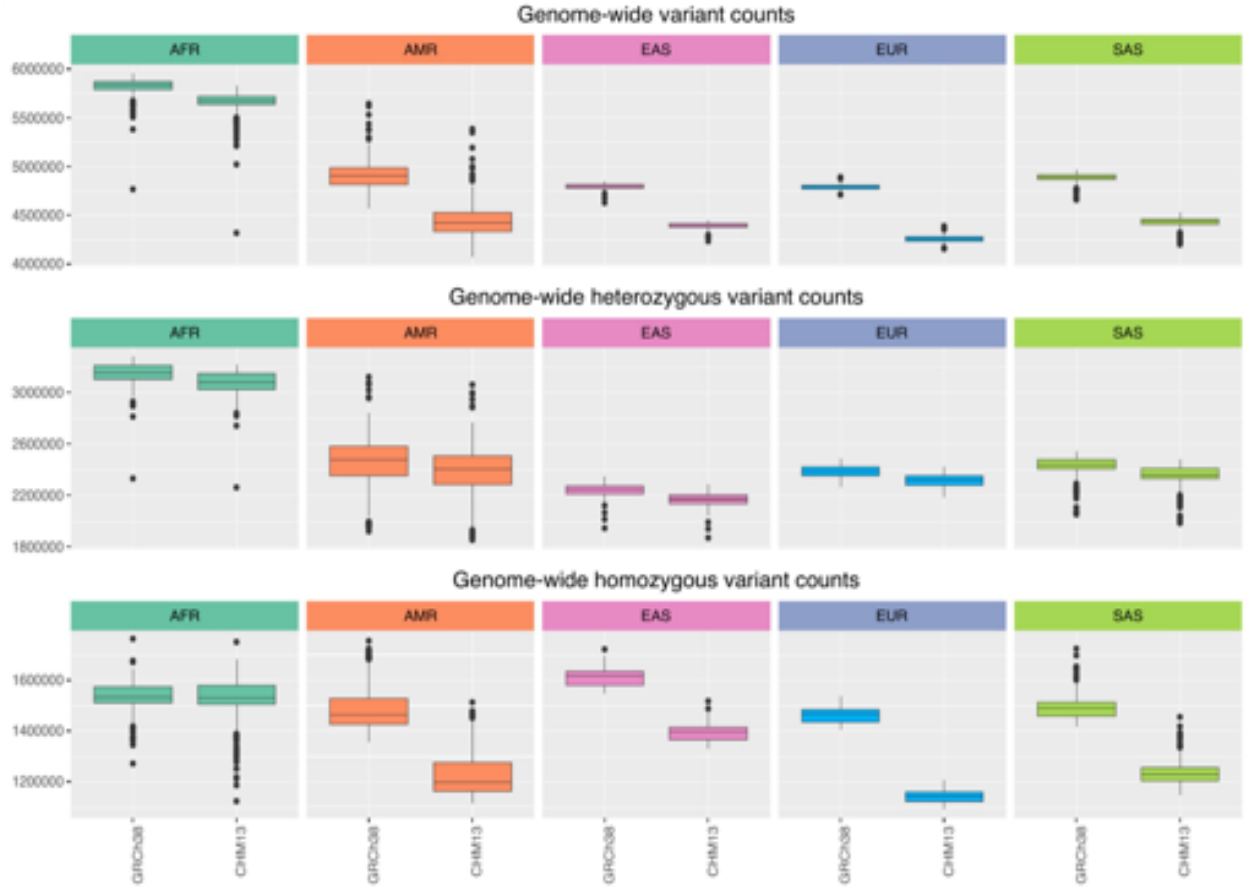
Explaining Decreased Per-Sample Count



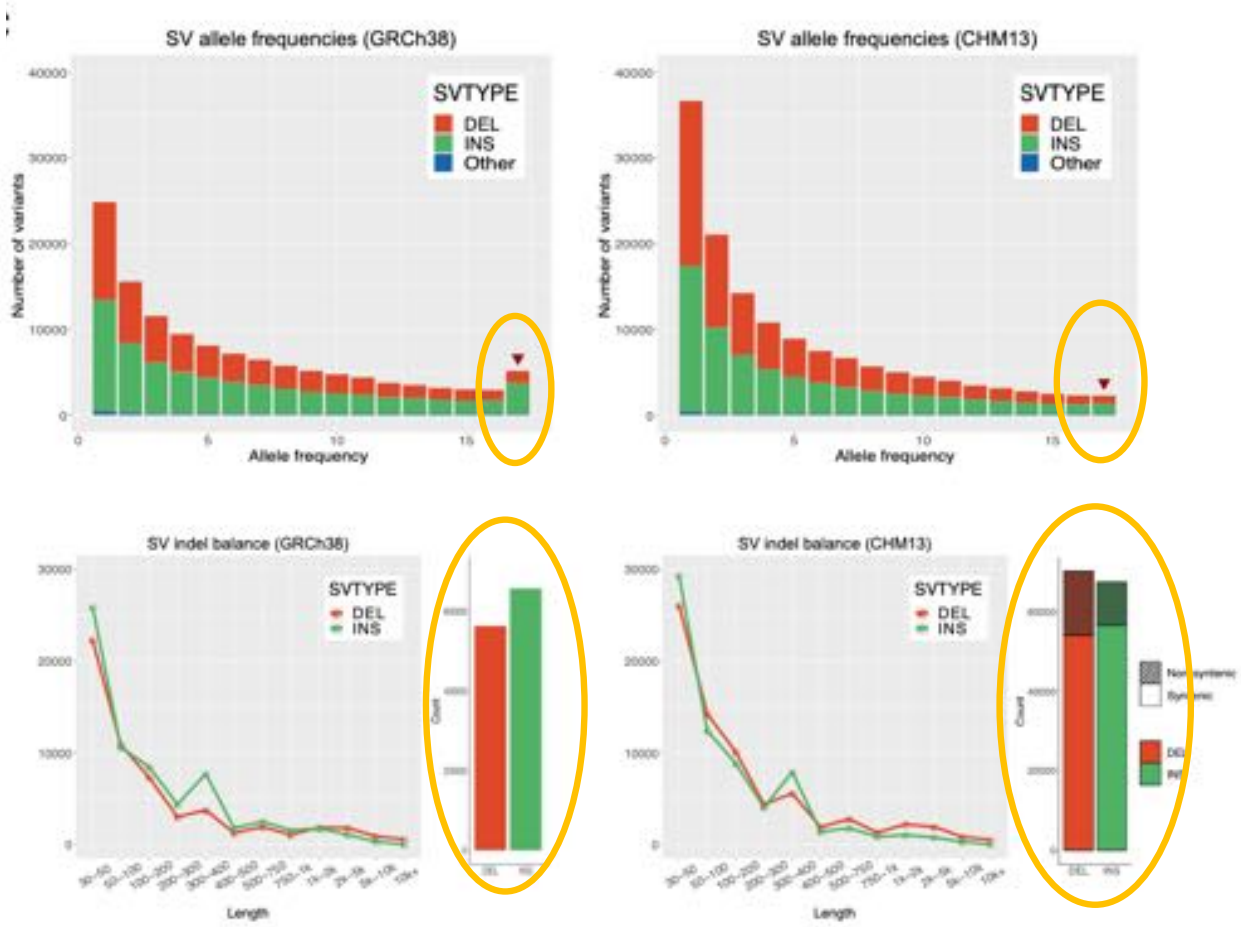
Allele Frequency ≈ 0.5: Collapsed duplication



1000G Per-Sample Variant Counts on T2T-CHM13



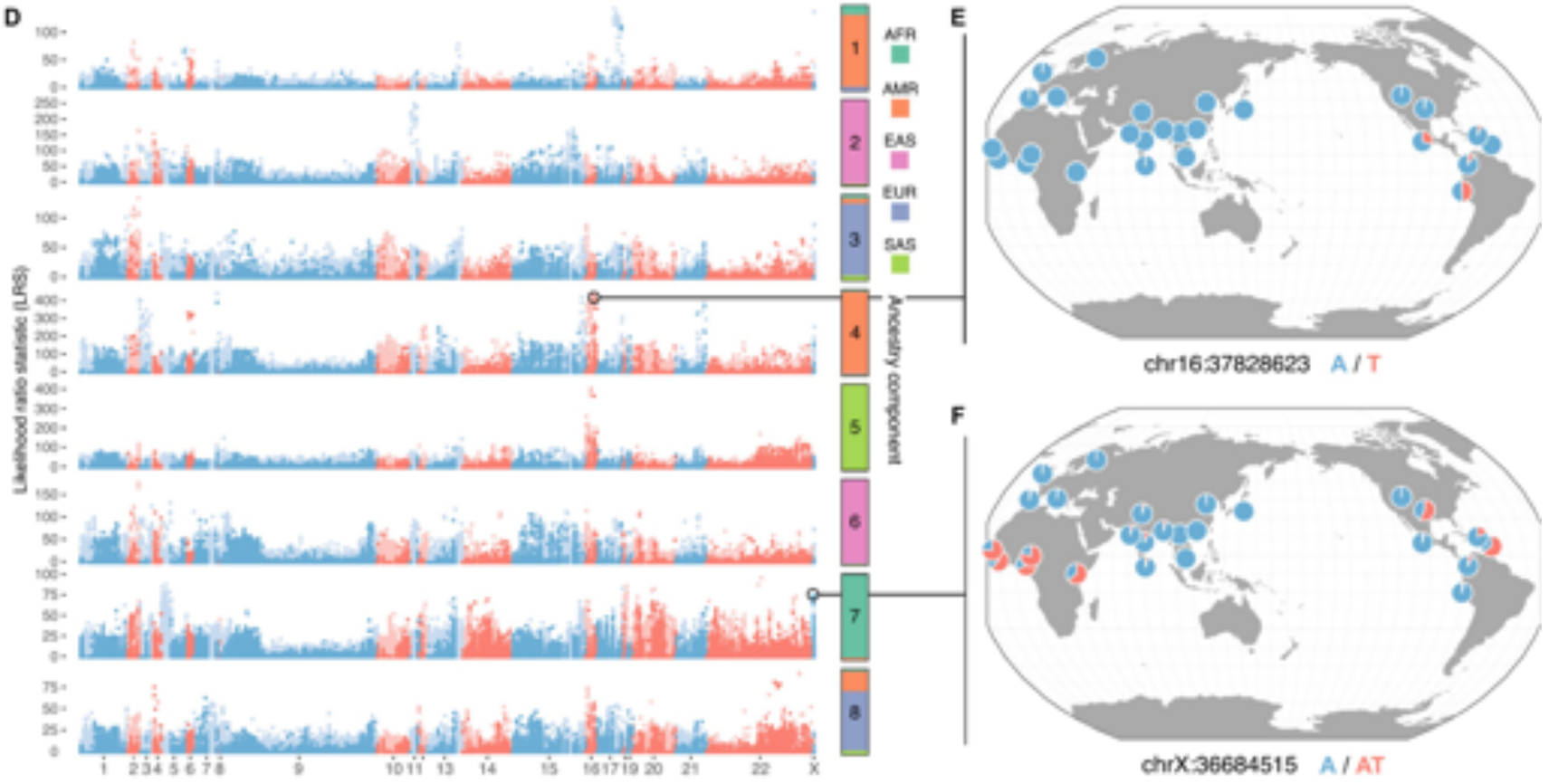
Long-Read Analysis with T2T-CHM13



Melanie Kirsche

Sergey Aganezov

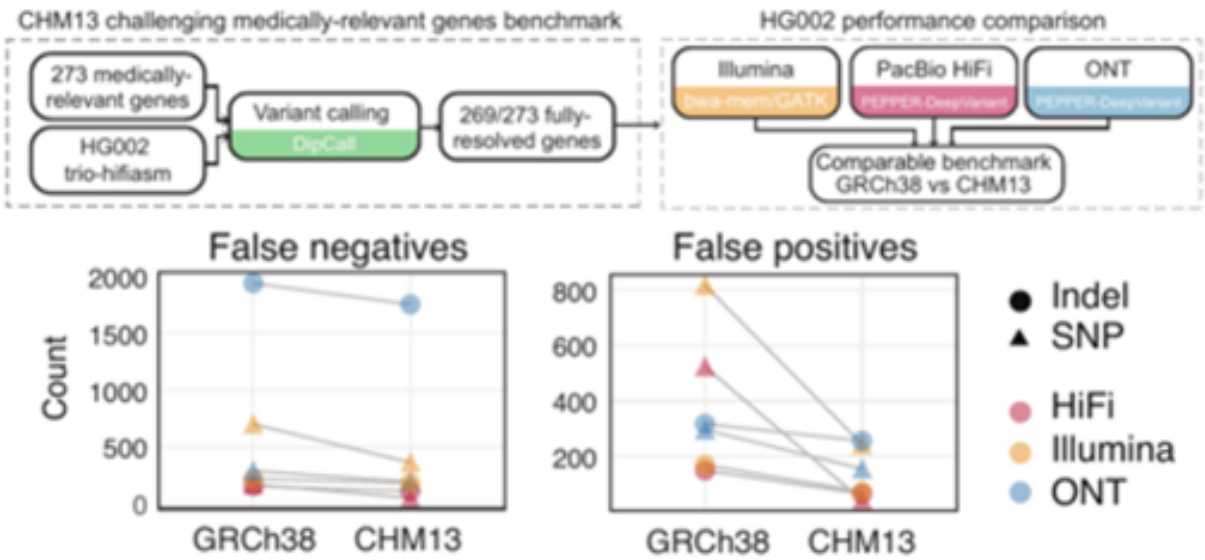
Variants in Newly Resolved Regions



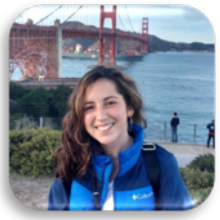
Stephanie Yan

Rajiv McCoy

T2T-CHM13 Improves Clinical Genomics Variant Calling



Danny Miller



Daniela Soto

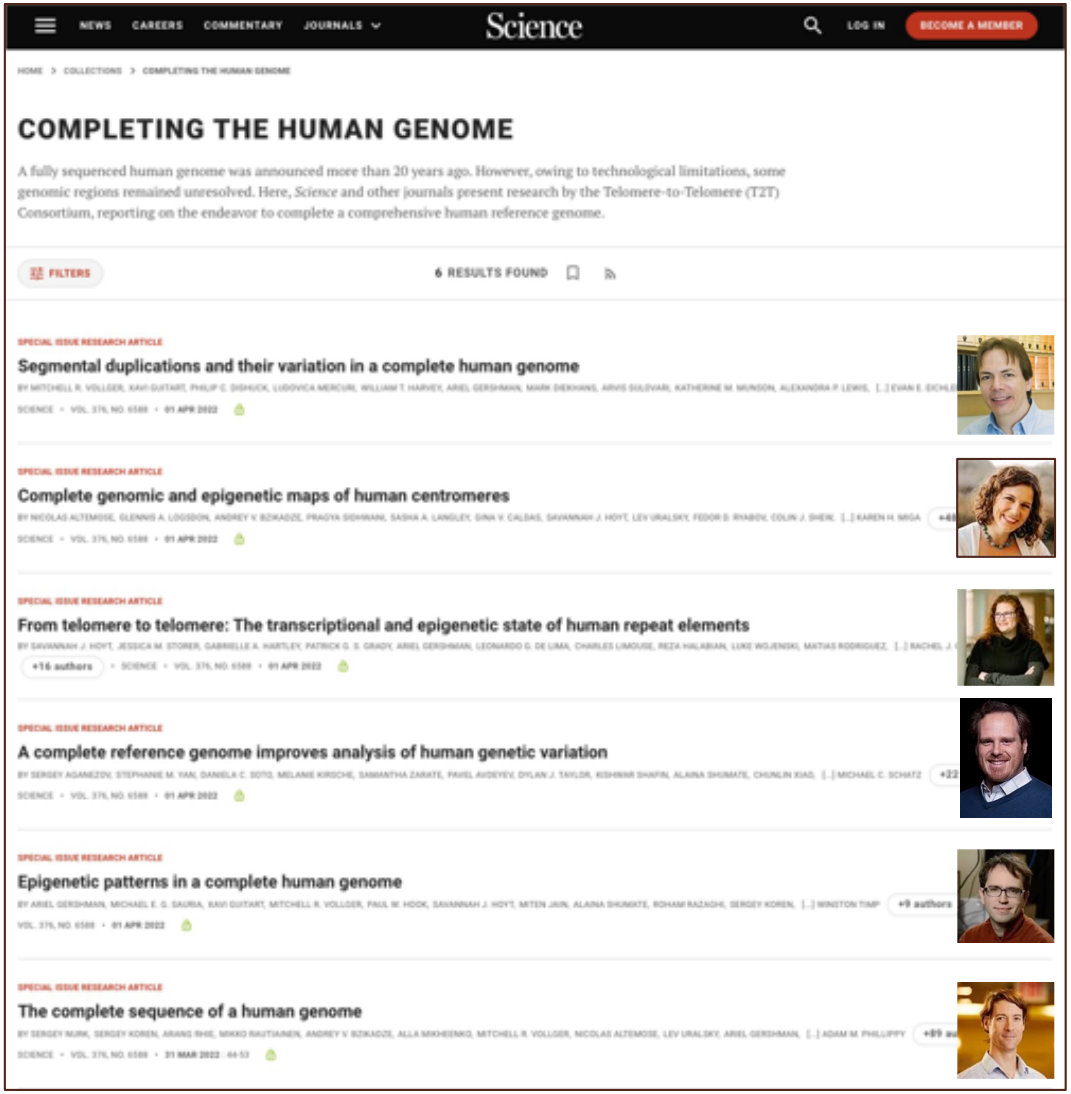
Megan Dennis

Justin Zook

Fritz Sedlazeck

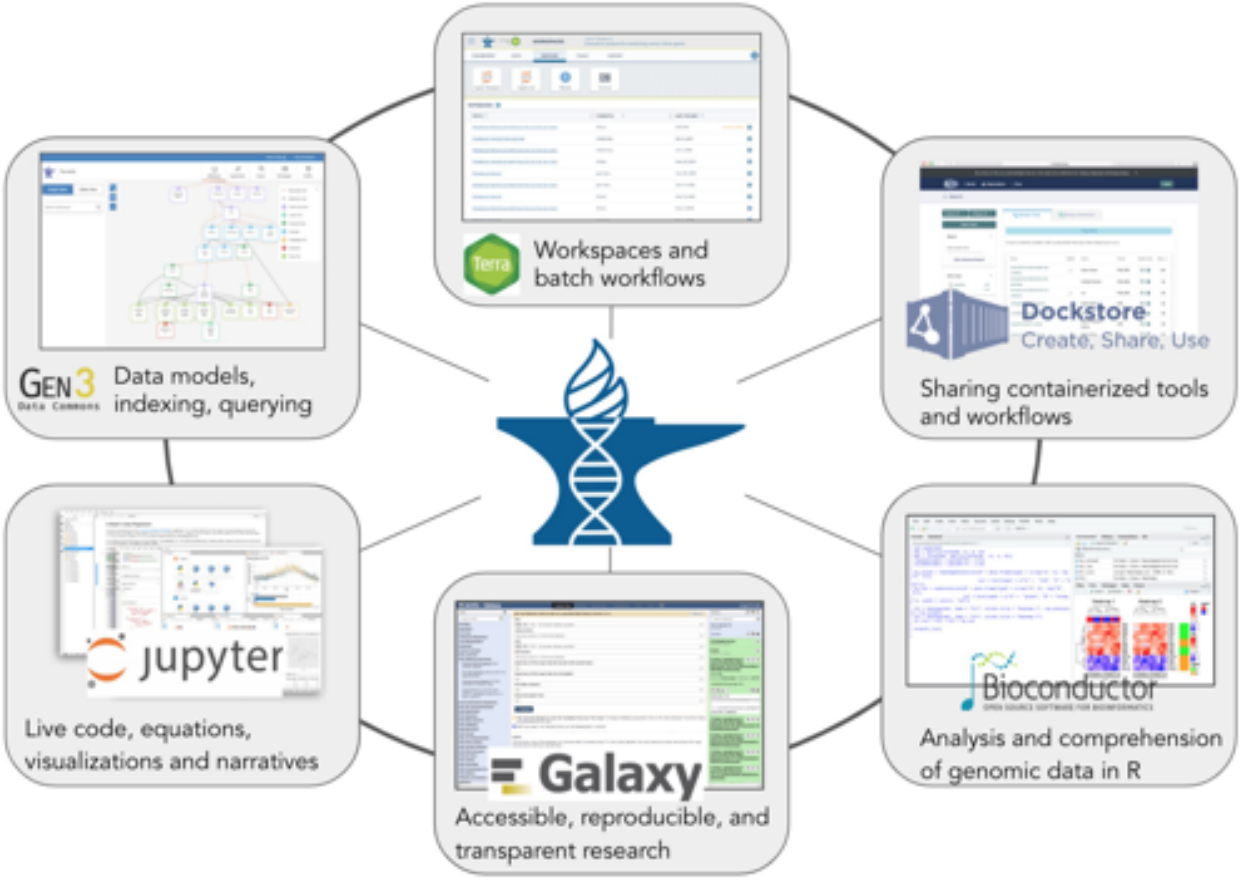
For more information, see: (Wagner, J *et al.*, NBT, 2022)


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


NEWS
CAREERS
COMMENTARY
JOURNALS
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.
FILTERS
6 RESULTS FOUND
SPECIAL ISSUE RESEARCH ARTICLE
Segmental duplications and their variation in a complete human genome
SPECIAL ISSUE RESEARCH ARTICLE
Complete genomic and epigenetic maps of human centromeres
SPECIAL ISSUE RESEARCH ARTICLE
From telomere to telomere: The transcriptional and epigenetic state of human repeat elements
+16 authors
SPECIAL ISSUE RESEARCH ARTICLE
A complete reference genome improves analysis of human genetic variation
SPECIAL ISSUE RESEARCH ARTICLE
Epigenetic patterns in a complete human genome
+9 authors
SPECIAL ISSUE RESEARCH ARTICLE
The complete sequence of a human genome

T2T Team @ Santa Cruz, August 2022

T2T Team @ TIME100, June 2022



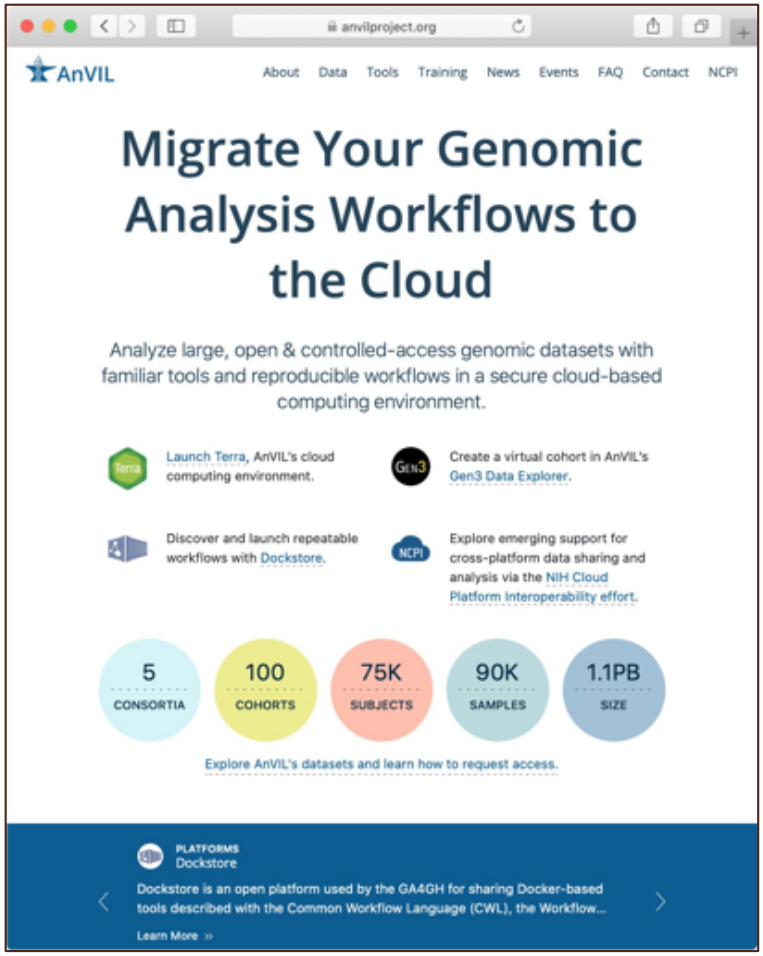
Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL)

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



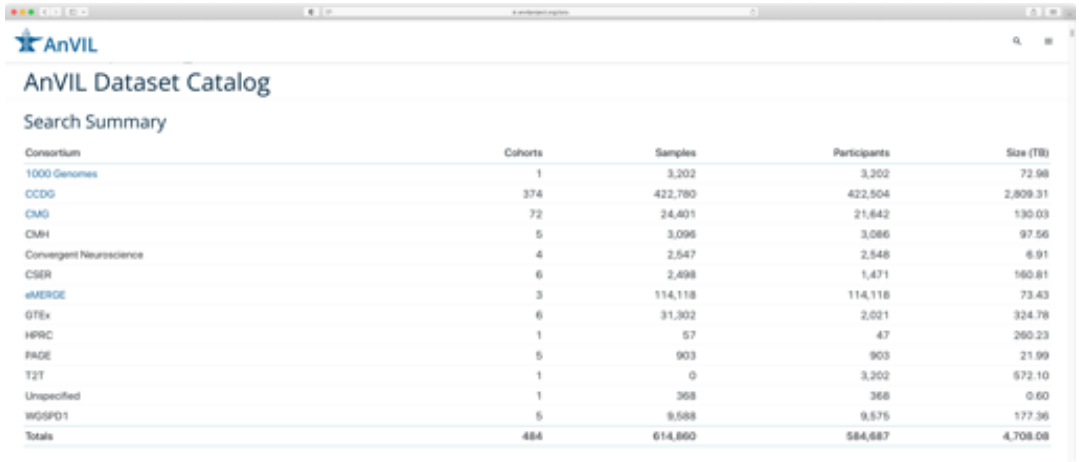
https://anvilproject.org

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

AnVIL Data Dashboard

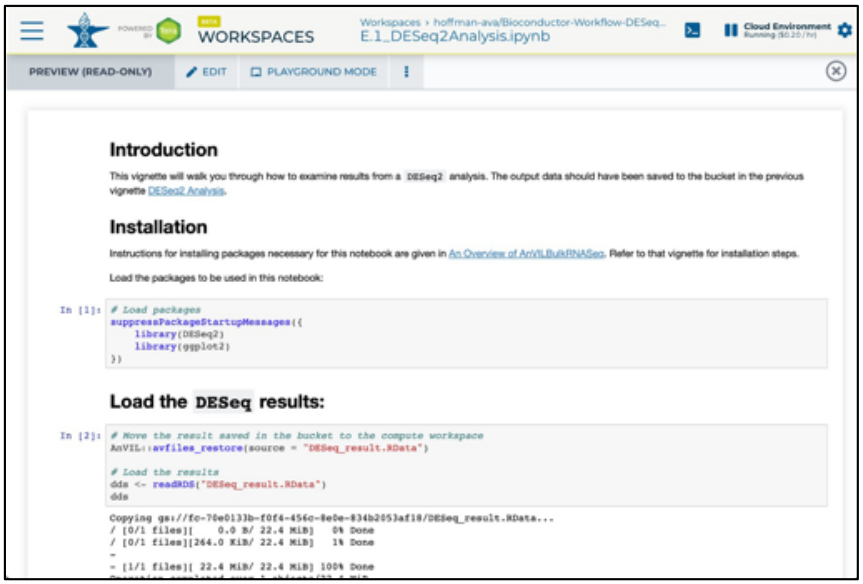
Extensive unrestricted and protected data sets already available within AnVIL

- 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



https://anvilproject.org/data

AnVIL is “renting computers”



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

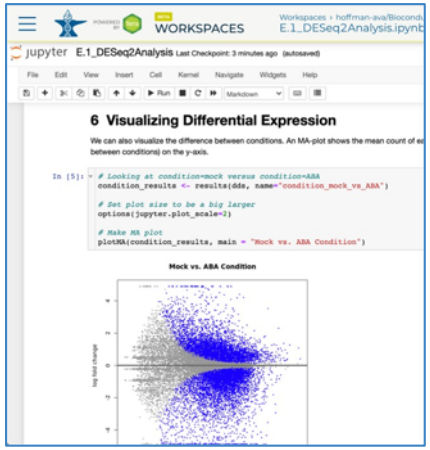
- The cloud is built from several very large clusters of computers
  - Effectively infinite resources
  - High-end servers with many cores, many GB RAM, high speed networking, and exabytes of storage

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

- 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://en.wikipedia.org/wiki/Virtual_machine

AnVIL Analysis Platforms

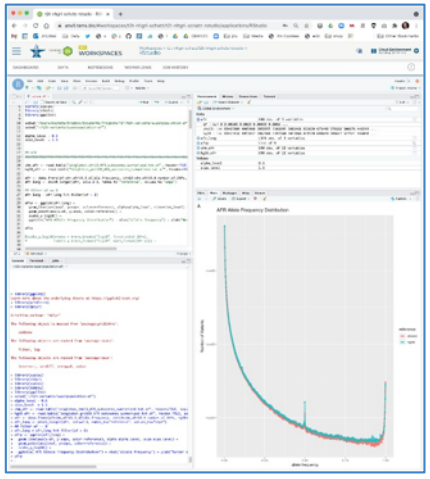


jupyter

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

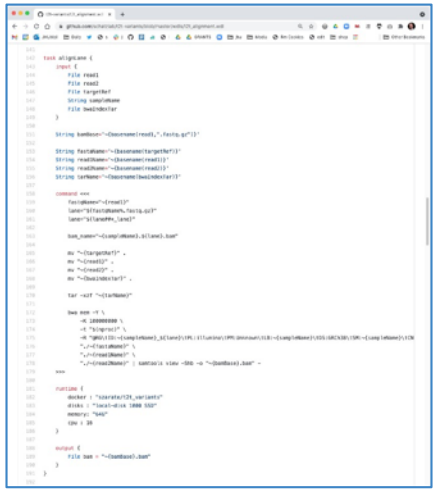
Galaxy

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



RStudio®

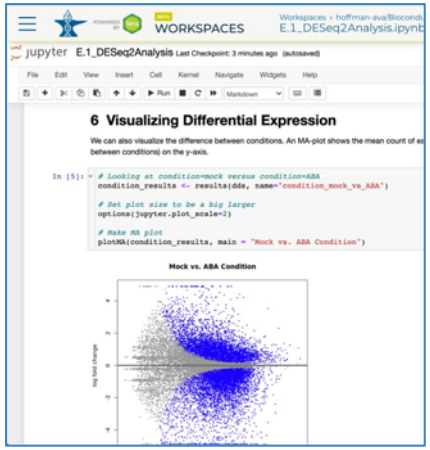
+ Feature rice IDE for programming in R
+ Rich statistics, ML, and visualizations
- Limited support for other programming languages



{wdl}

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

AnVIL Analysis Platforms (coding)

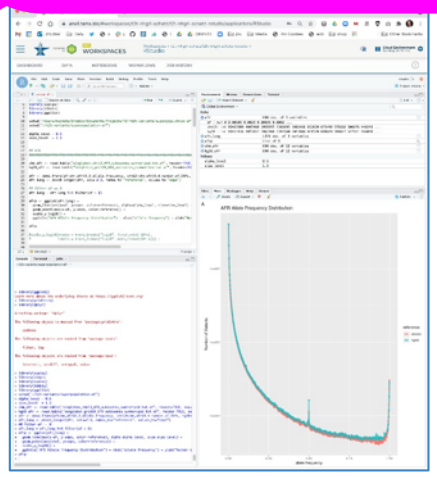


jupyter

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

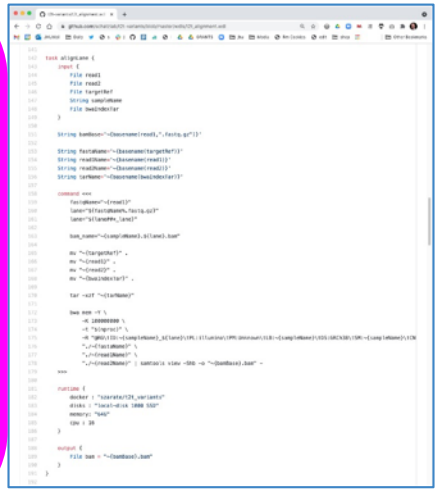
Galaxy

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



RStudio

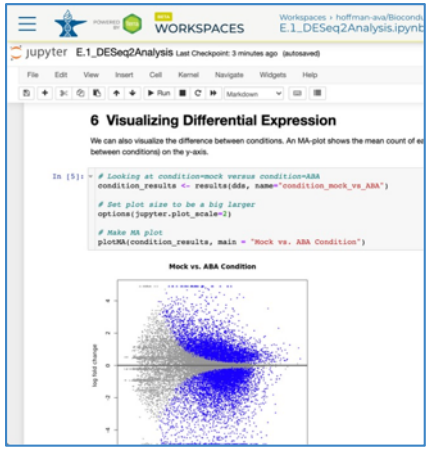
+ Feature rice IDE for programming in R
+ Rich statistics, ML, and visualizations
- Limited support for other programming languages



{wdl}

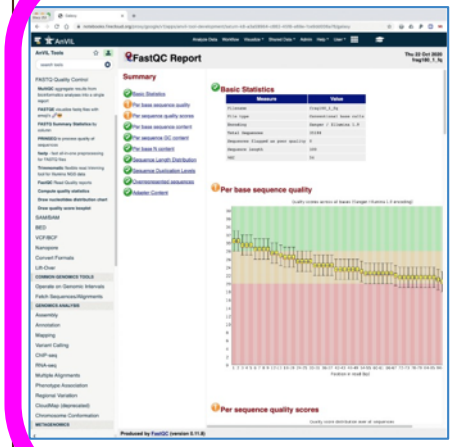
+ Extremely scalable and flexible
- Most technically demanding
- Unpredictable and potentially large costs

AnVIL Analysis Platforms (GUI)



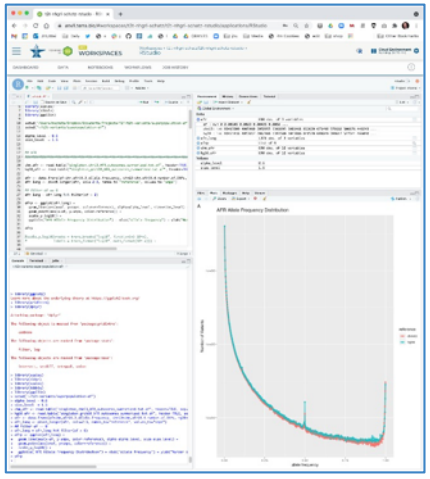
jupyter

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



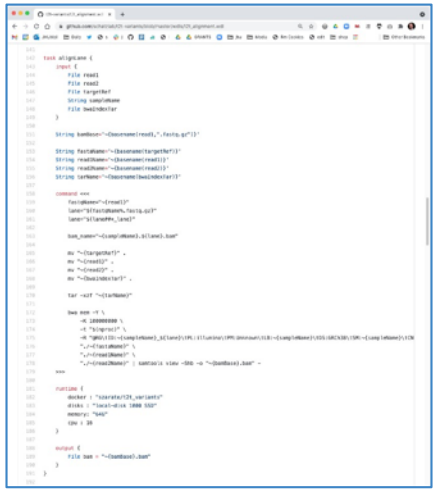
Galaxy

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



RStudio

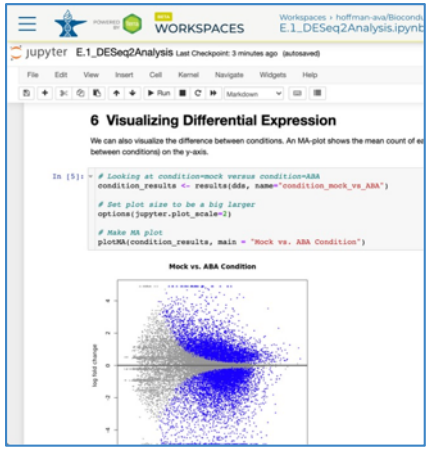
+ Feature rice IDE for programming in R
+ Rich statistics, ML, and visualizations
- Limited support for other programming languages



{wdl}

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

AnVIL Analysis Platforms (workflows)

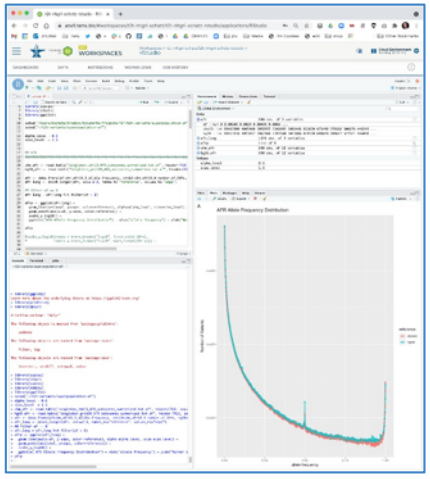


jupyter

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

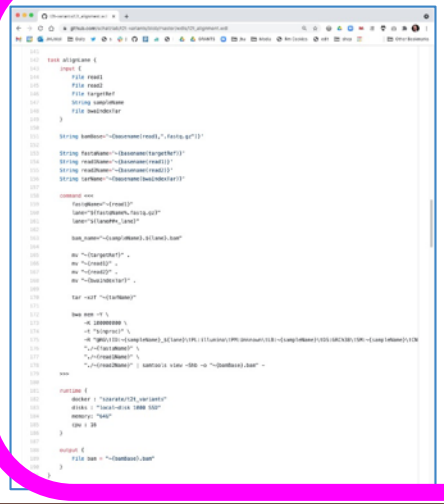
Galaxy

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



RStudio

+ Feature rice 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

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

RStudio®

+ Feature rice 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