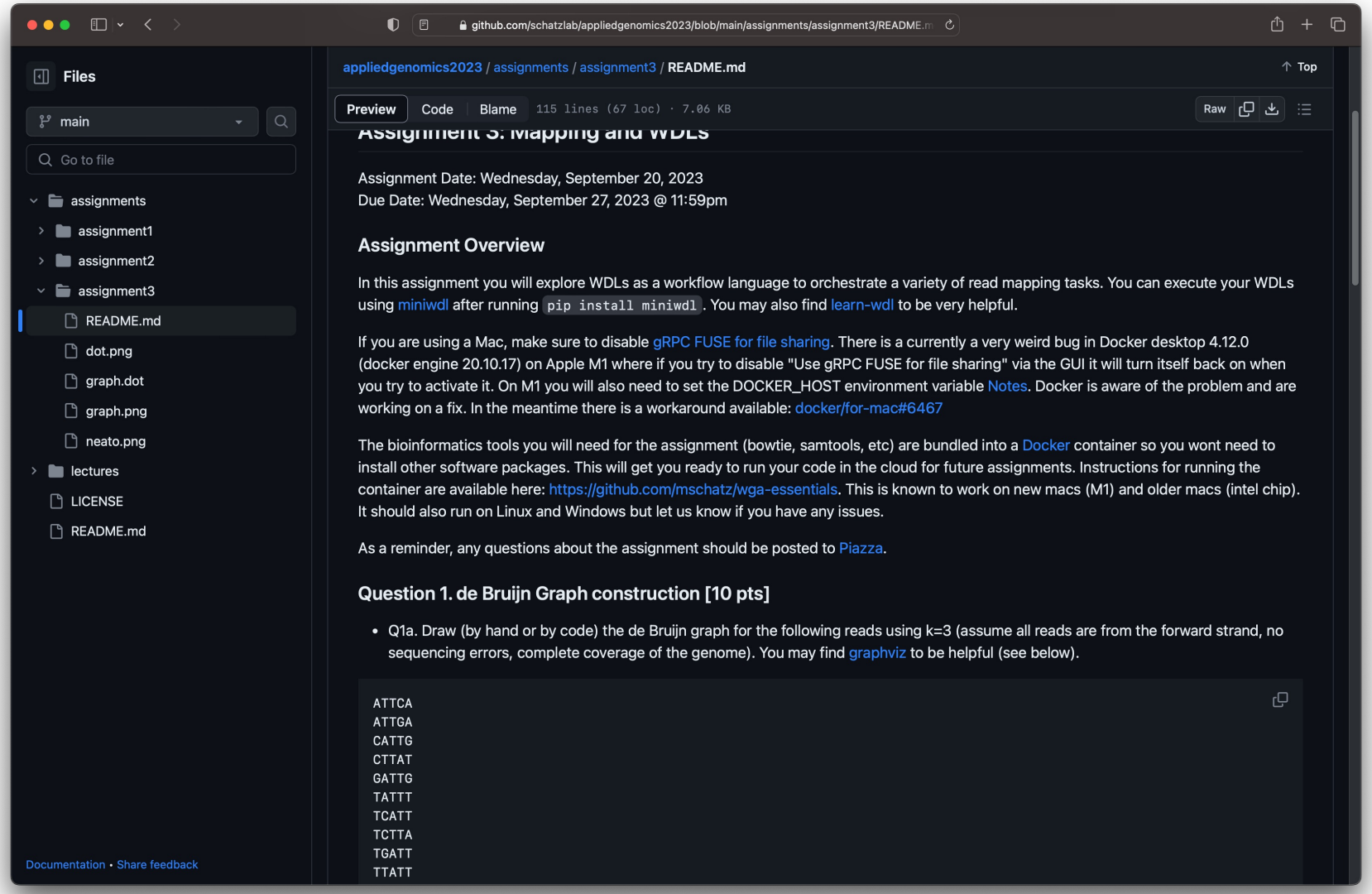
The human genome & cloud computing

Michael Schatz

Sept 19, 2023

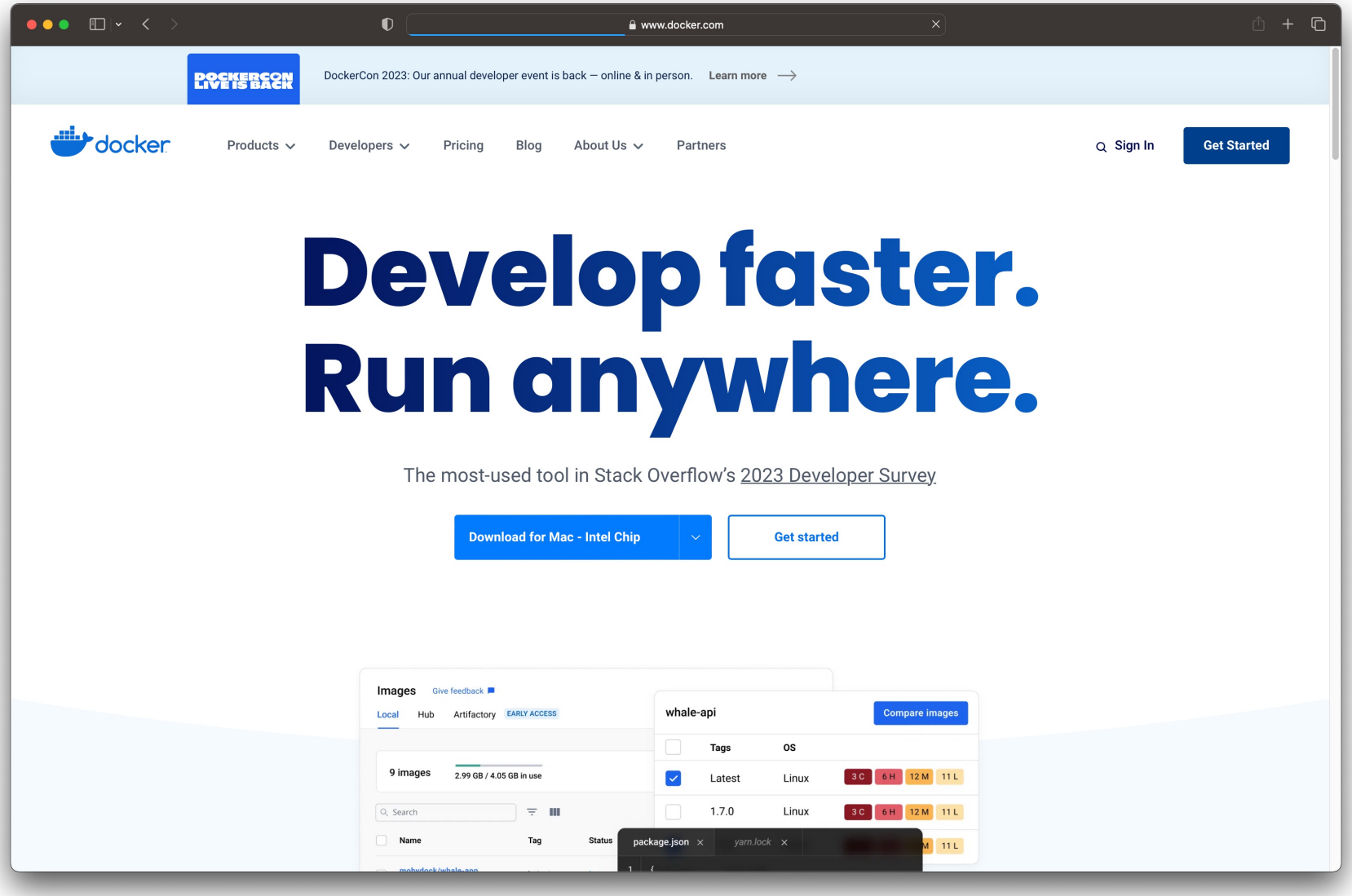
Lecture 7: Applied Comparative Genomics

Assignment 3: Genome Assembly
Due Wednesday Sept 27 by 11:59pm

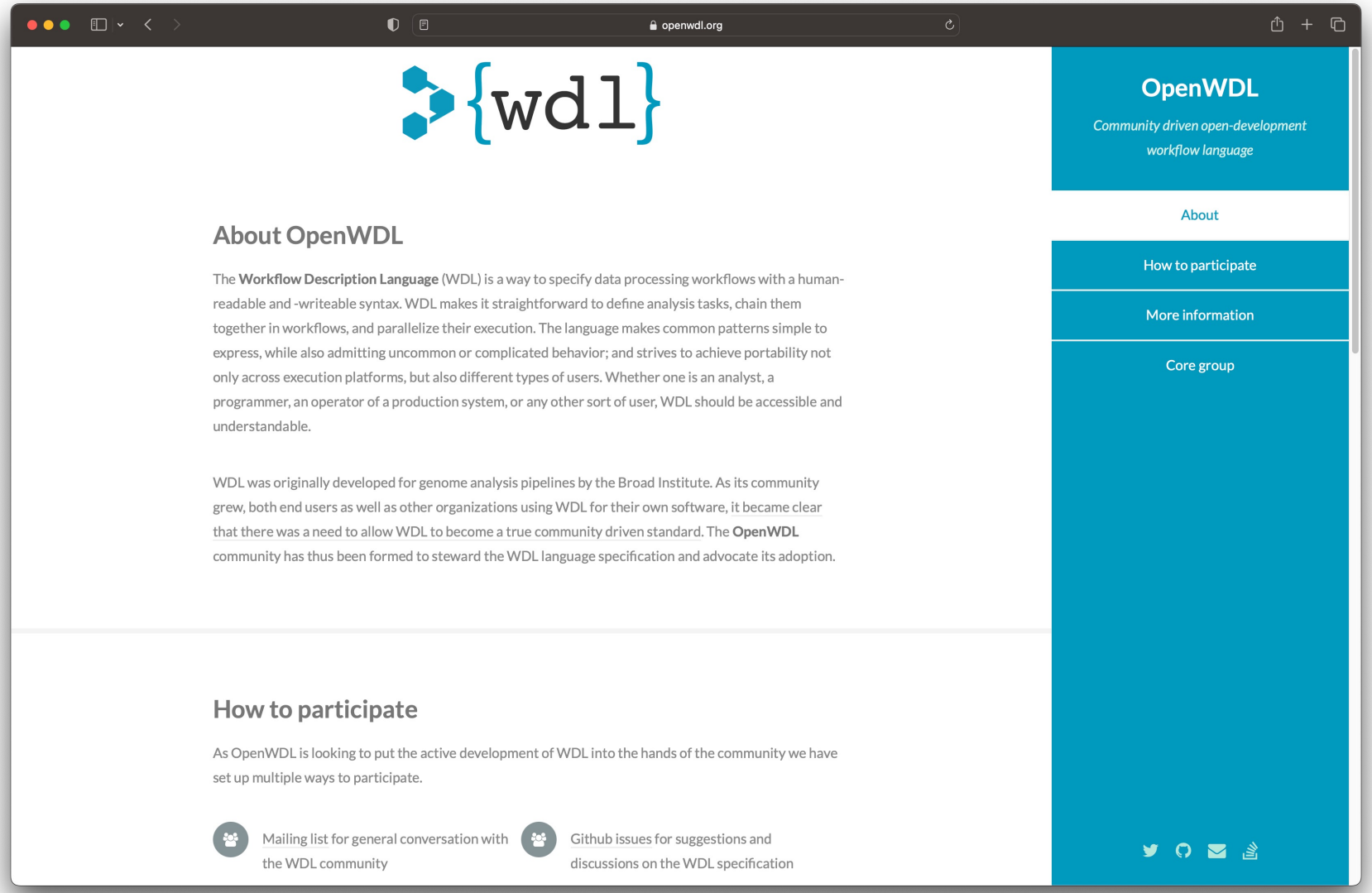


https://github.com/schatzlab/appliedgenomics2023/tree/main/assignments/assignment3

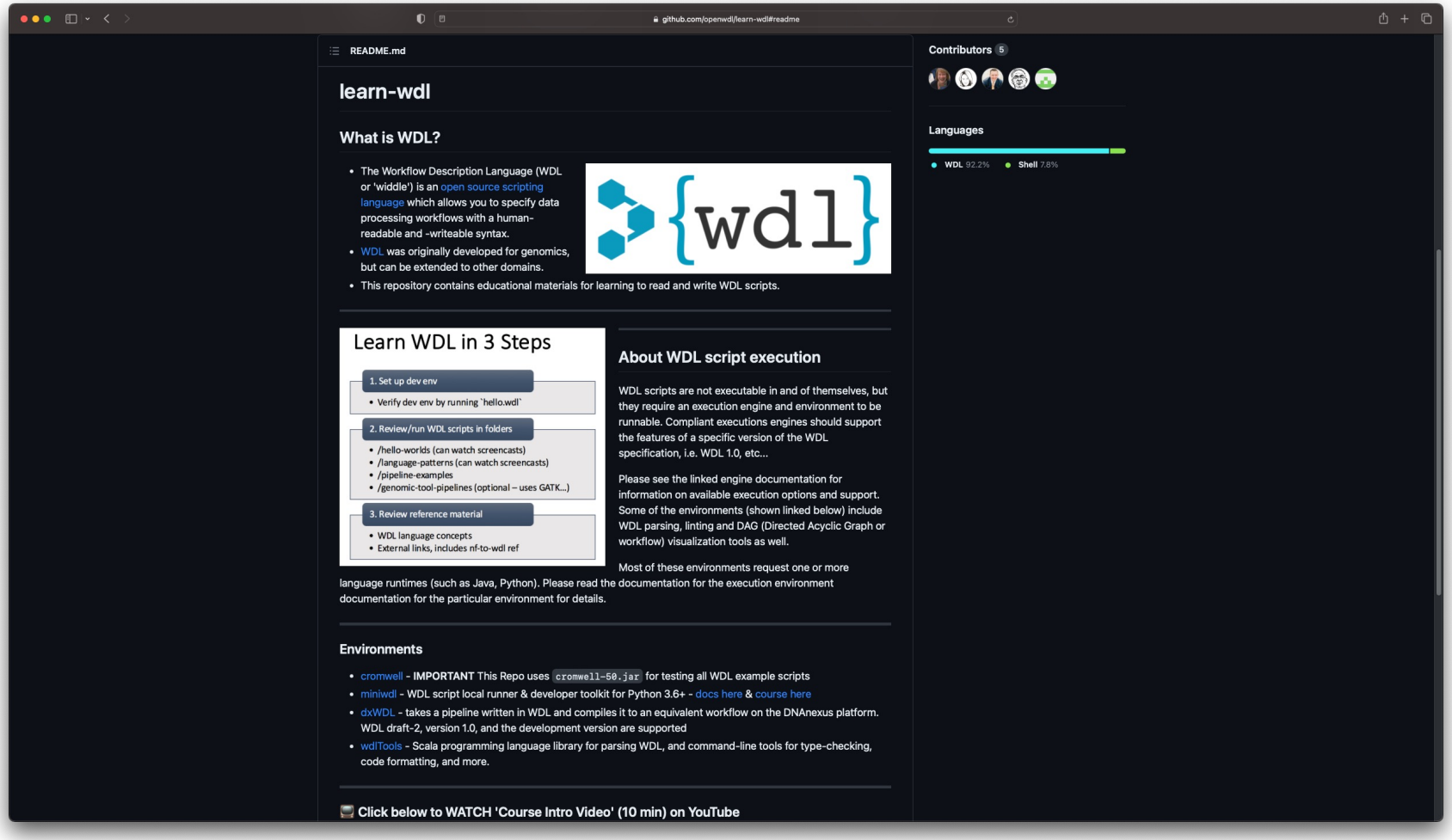
Check Piazza for questions!



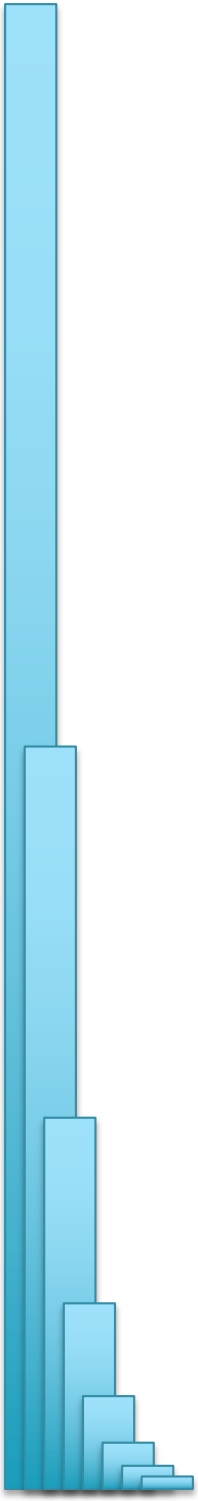
http://docker.com



http://openwdl.org



https://github.com/openwdl/learn-wdl



Part 1: The human genome

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)

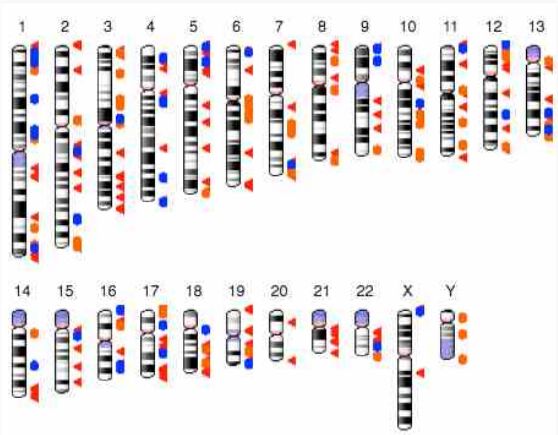
Genome Reference Consortium

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



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

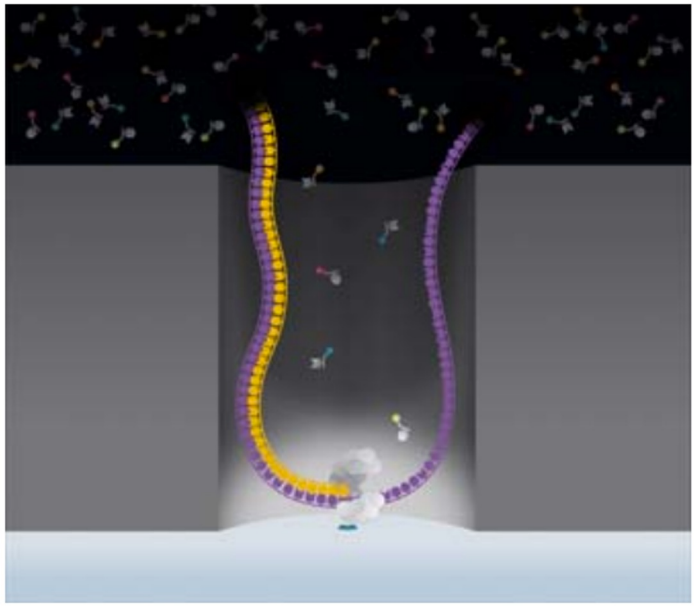
Pseudoautosomal regions

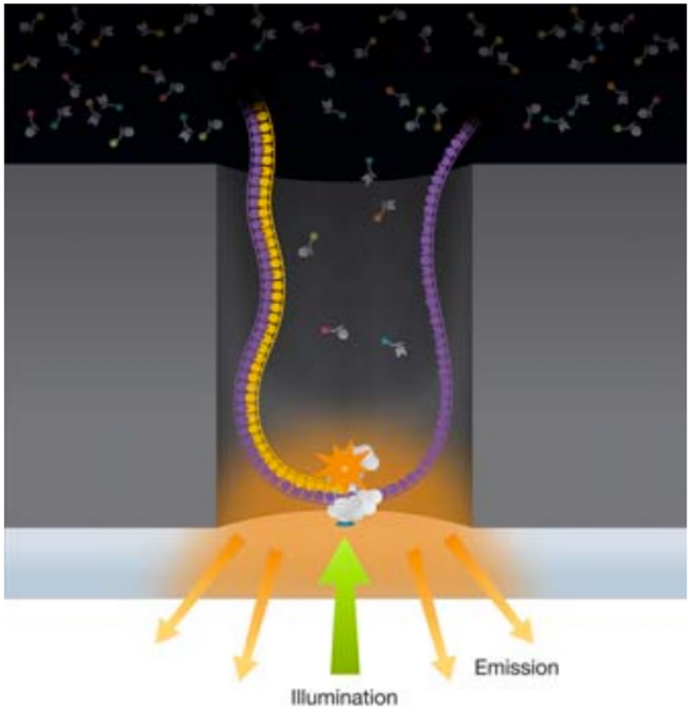
Name	Chr	Start	Stop
PAR#1	X	10,001	2,781,479
PAR#2	X	155,701,383	156,030,895
PAR#1	Y	10,001	2,781,479
PAR#2	Y	56,887,903	57,217,415

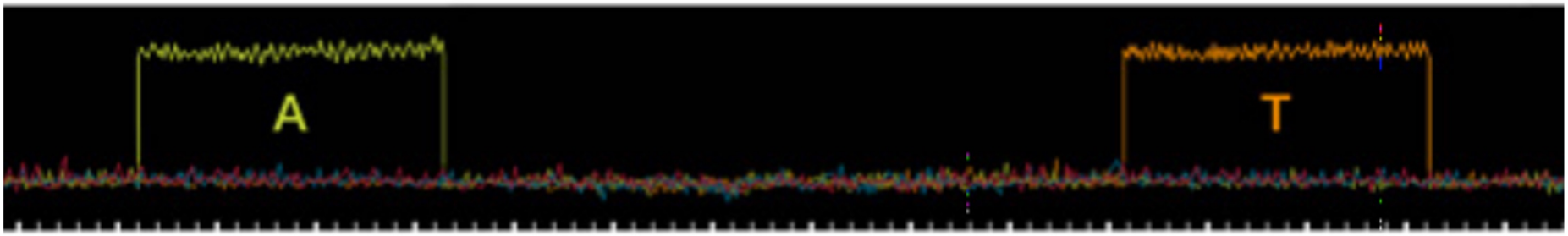
PacBio Single Molecule Real Time Sequencing (SMRT-sequencing)

PacBio: SMRT Sequencing

Imaging of florescent phospholinked labeled nucleotides as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide (ZMW).







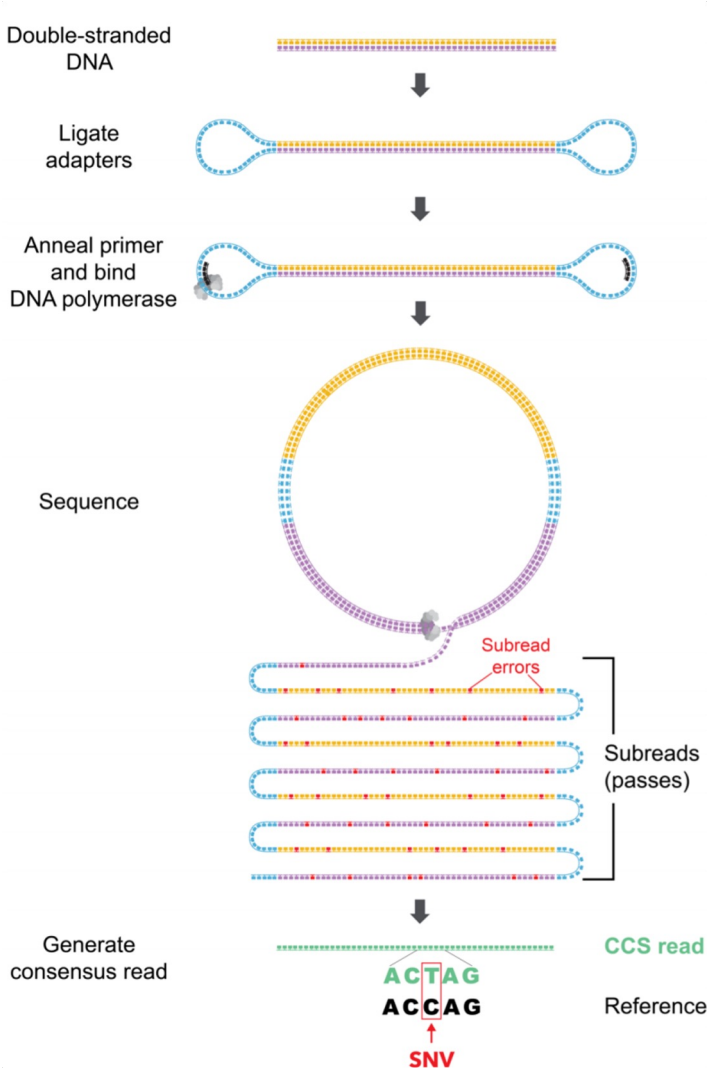
https://www.youtube.com/watch?v=_ID8JyAbwEo

“HiFi” Circular Consensus Reads

High-quality reads produced by sequencing the same molecule multiple times

Higher accuracy for low-coverage sequences like somatic variants or lowly expressed transcripts in RNA-seq, more interpretable alignments, better & faster assembly

Limits read length, used to be very expensive but more manageable now

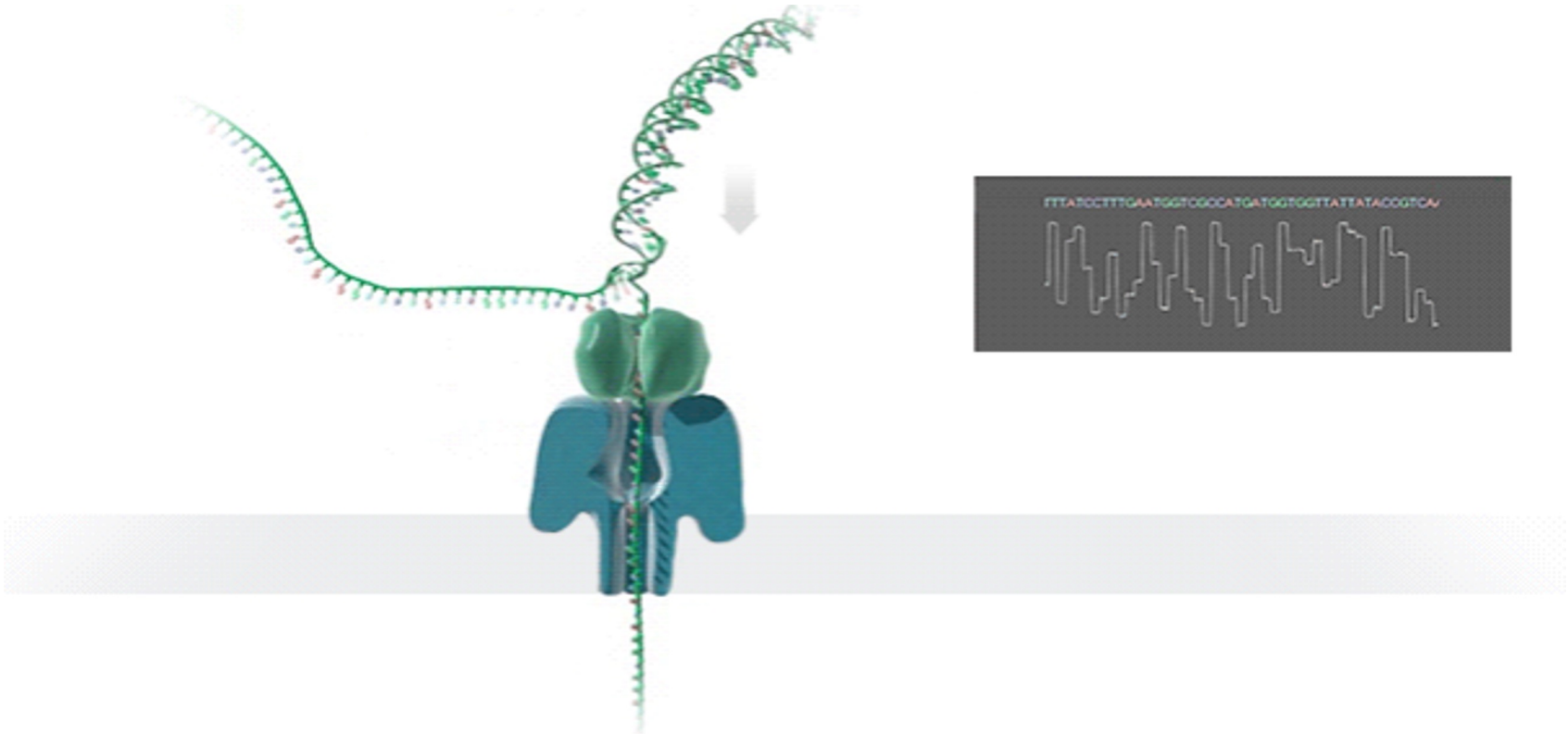


Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome
Wenger et al (2019) Nature Biotechnology doi:10.1038/s41587-019-0217-9

Oxford Nanopore Technologies (ONT)

Nanopore Sequencing

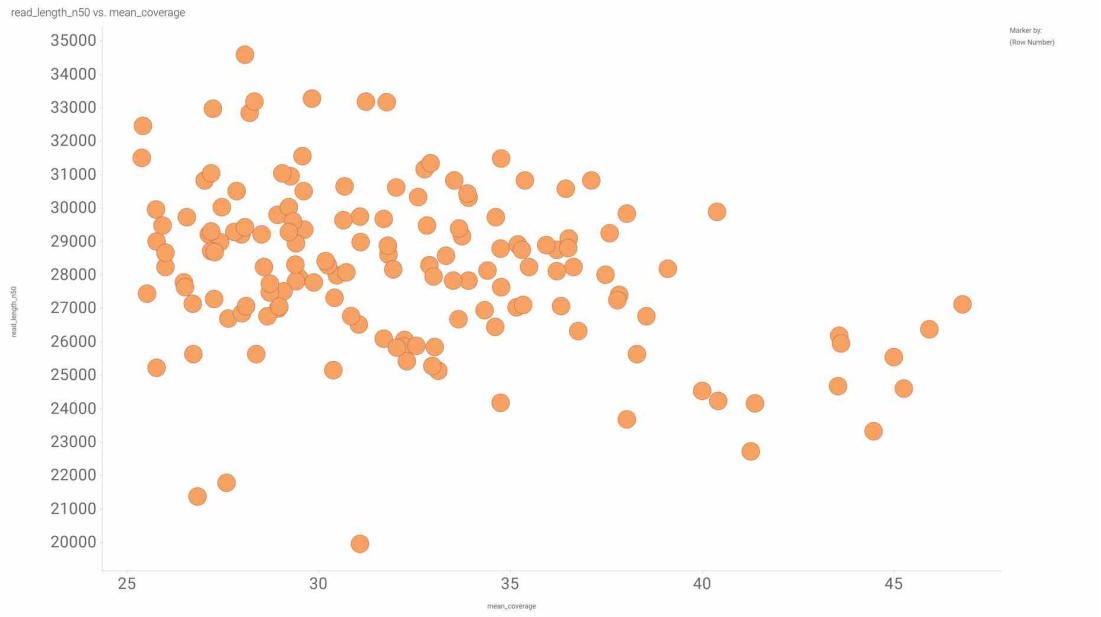
Sequences DNA/RNA by measuring changes in ionic current as nucleotide strand passes through a pore

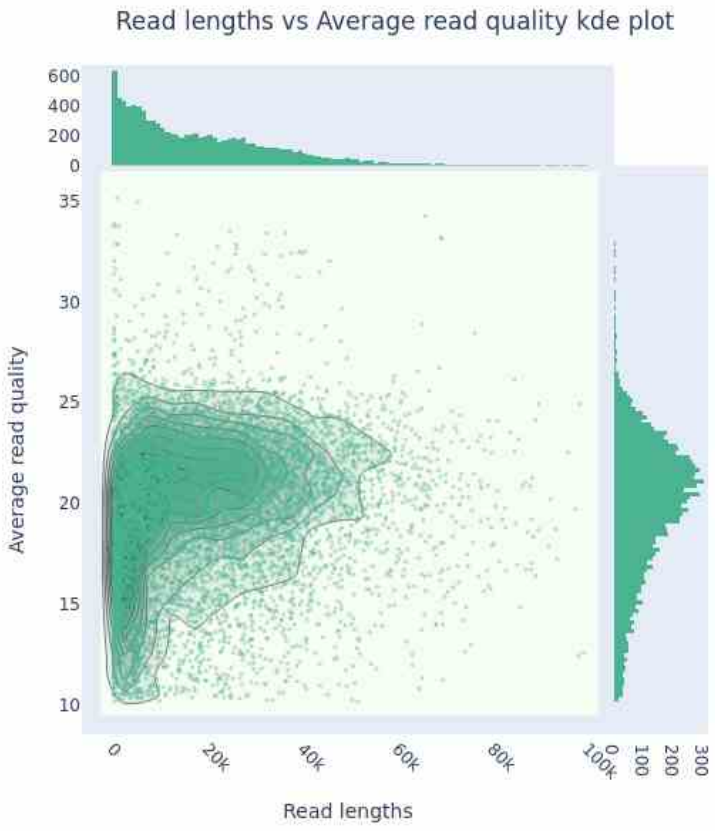


https://www.youtube.com/watch?v=RcP85JHLmnI

nanoporetech.com/applications/dna-nanopore-sequencing

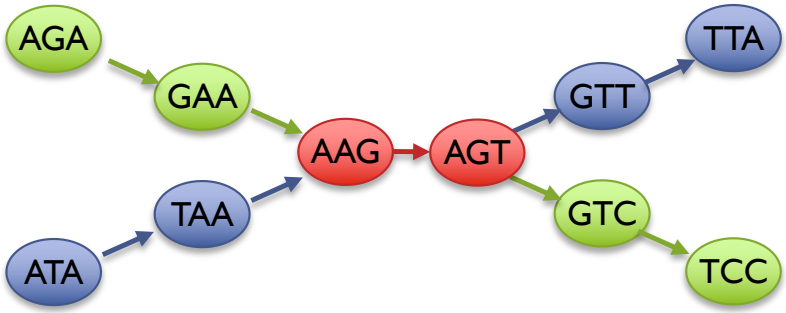
ONT at JHU


read_length_n50 vs. mean_coverage
Marker by:
(Row Number)
35000
34000
33000
32000
31000
30000
29000
28000
27000
26000
25000
24000
23000
22000
21000
20000
read_length_n50
25
30
35
40
45
mean_coverage


Read lengths vs Average read quality kde plot
600
400
200
0
35
30
25
20
15
10
Average read quality
0
20k
40k
60k
80k
100k
0
100
200
300
Read lengths

Two Paradigms for Assembly

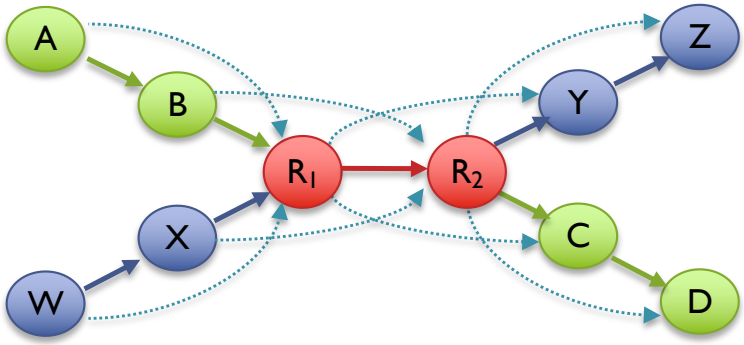
de Bruijn Graph



Short read assemblers

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

Overlap Graph

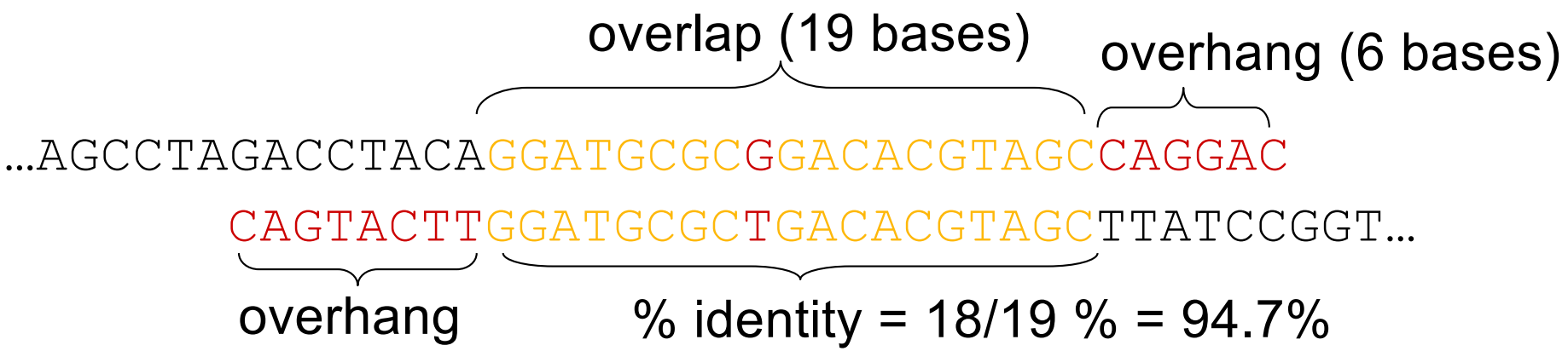


Long read assemblers

- 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.

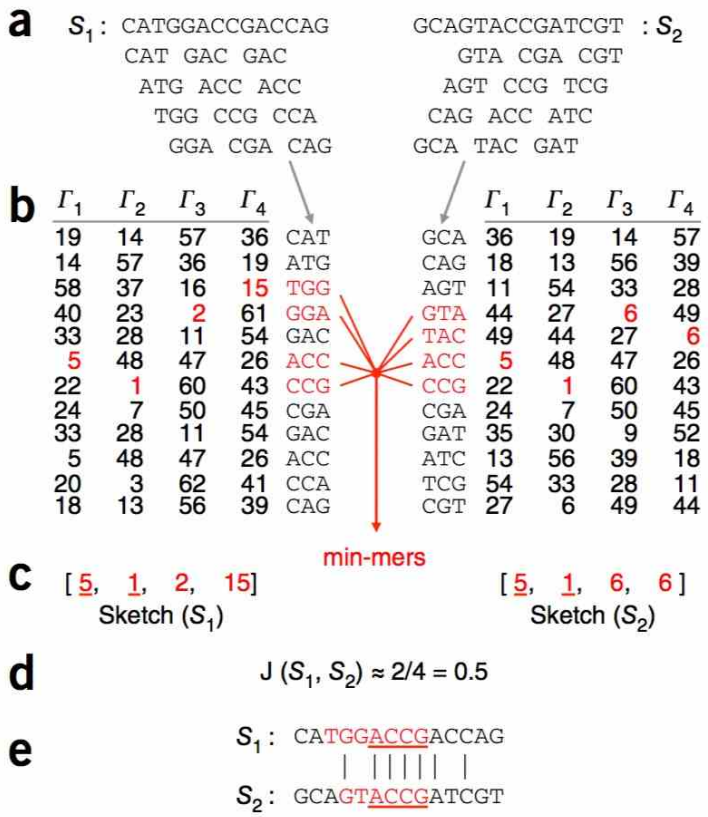
[How do we compute the overlap?]

[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:

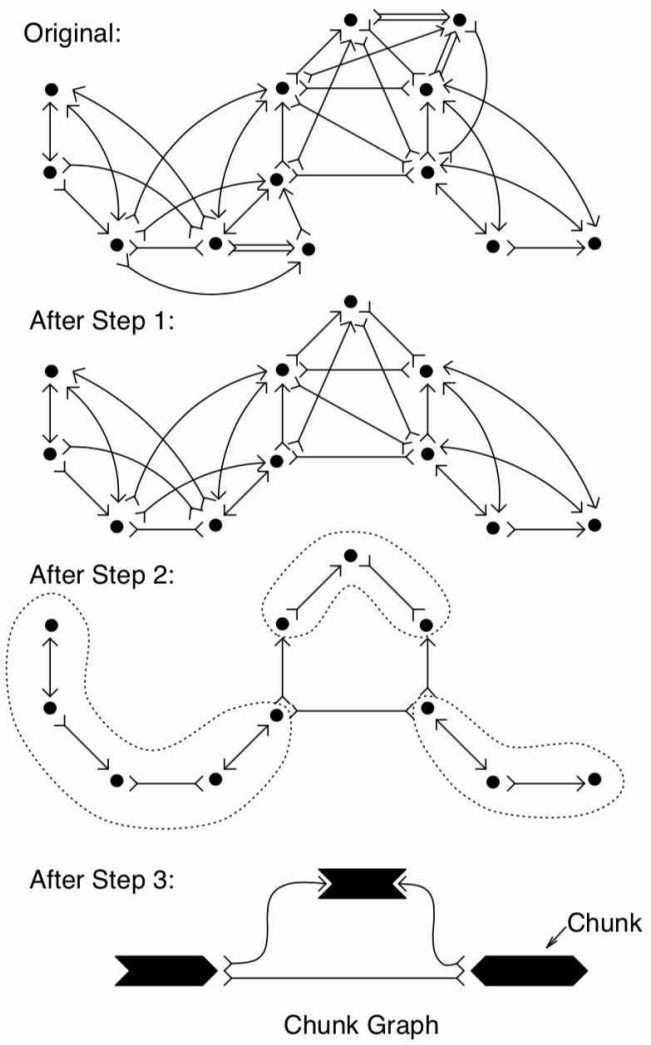
- If the average coverage is D, we should expect D overlaps at the beginning of the read, and D at the end

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

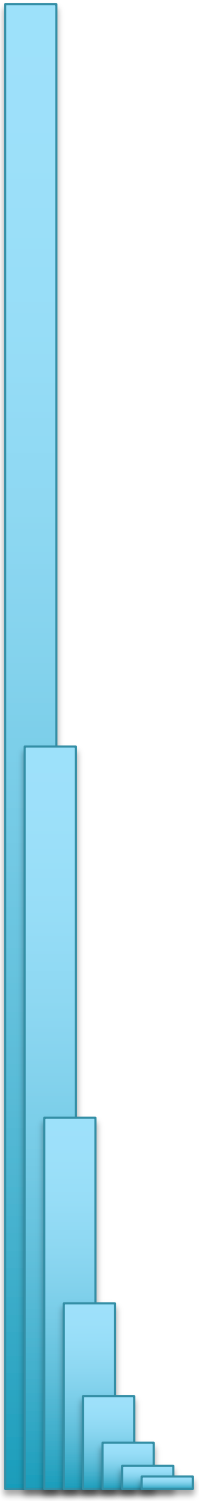
1. ***Contained reads removal:*** Short reads that are substrings of longer reads don't advance the assembly, remove those nodes and all of the edges

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.

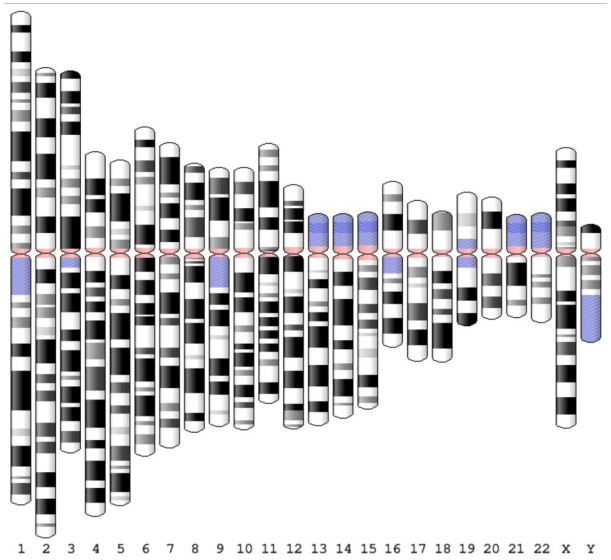


Part 11: The T2T human genome

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

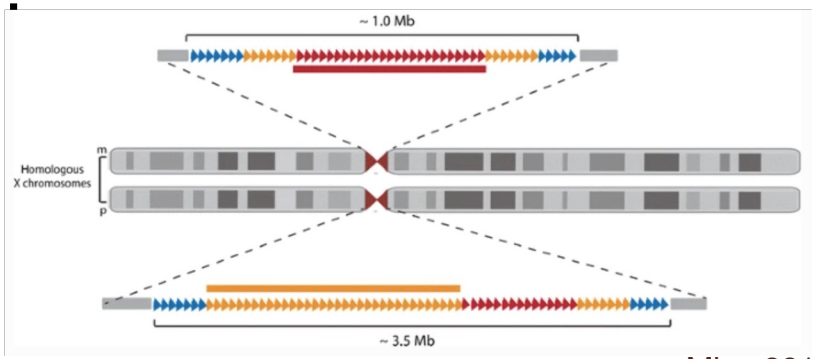
Functional studies need sequence

Reference gaps lead to artifacts

We don't know what we don't

Why has it taken so long?

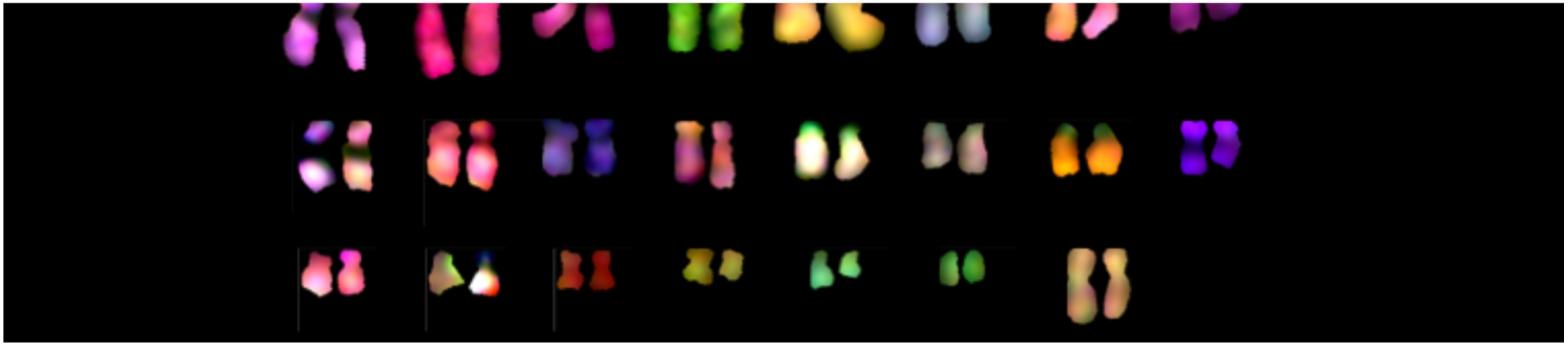
Repeats, repeats, repeats…



Miga 2015

Let's finish a human genome


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



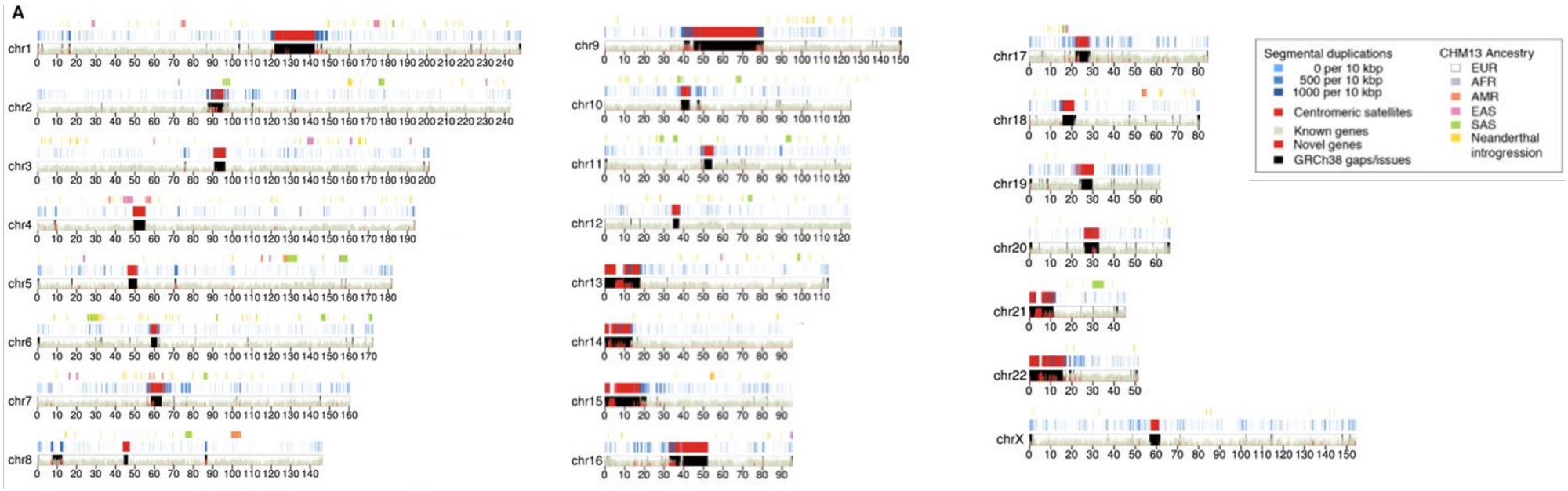
The Telomere-to-Telomere (T2T) consortium is an open, community-based effort to generate the first complete assembly of a human genome.

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

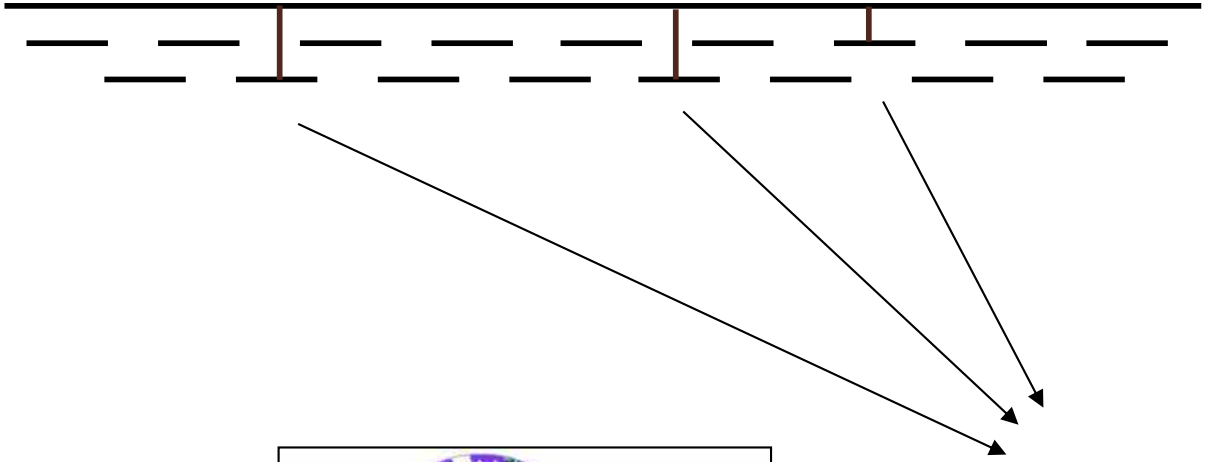


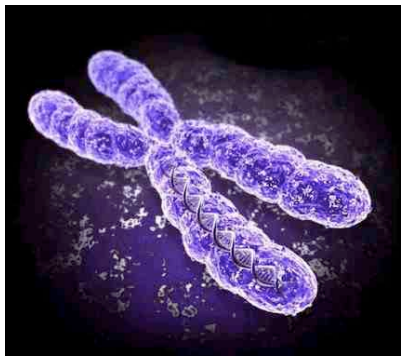
CHM13v1.1 genome size is **3.057 Gbp with zero Ns**
Every chromosome is telomere-to-telomere, quality estimated >Q70
~190 Mbp (~8%) of new sequence vs. GRCh38, fixes thousands of errors

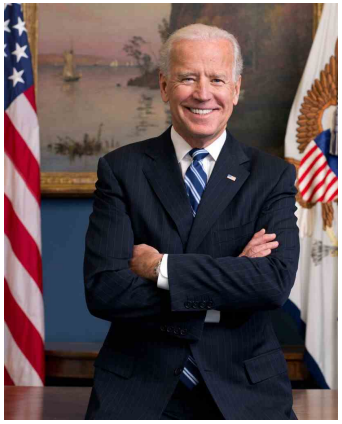
(Nurk et al. Science, 2022)

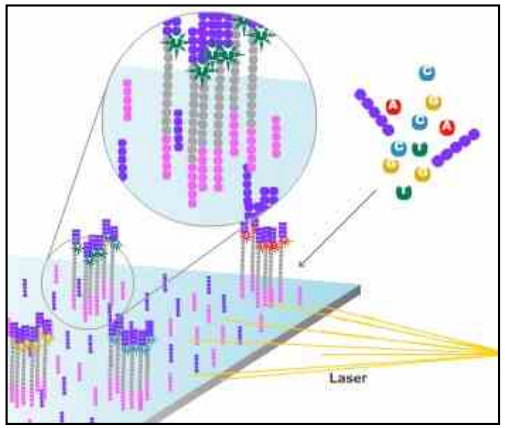
Personal Genomics

How does your genome compare to the reference?







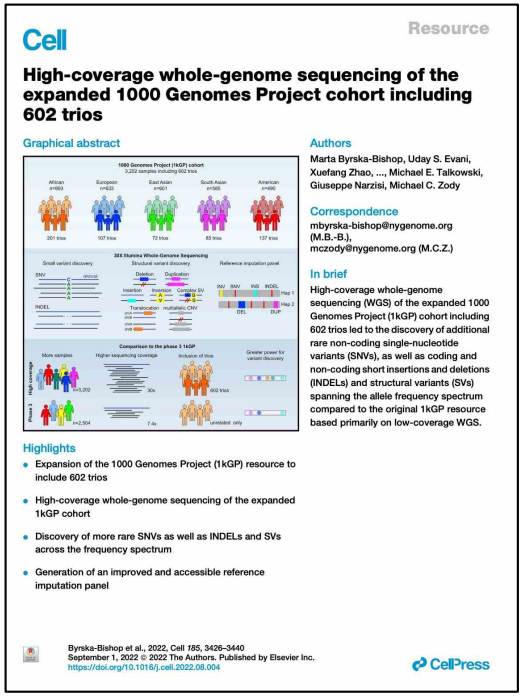


Heart Disease

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

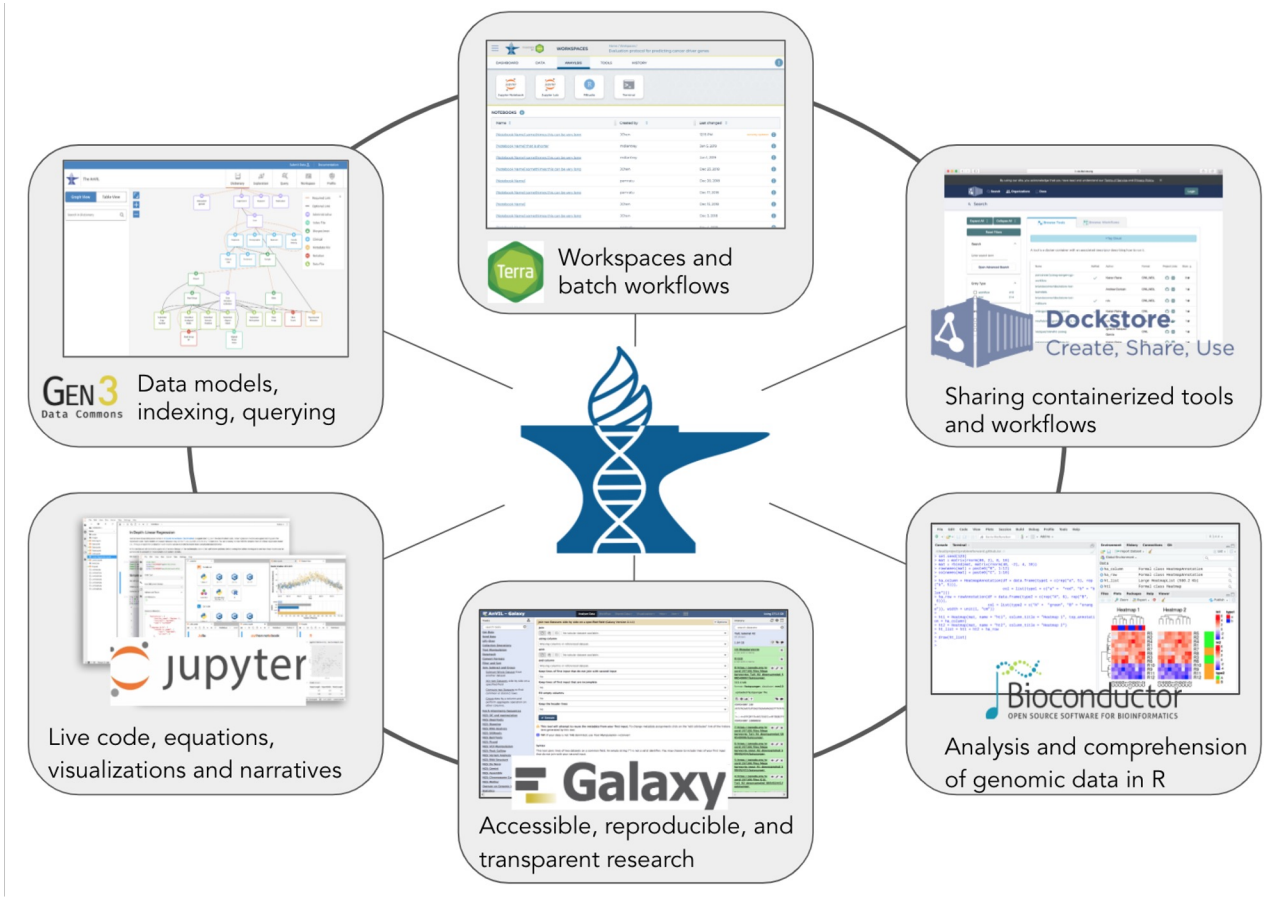
- Expansion of the 1000 Genomes Project (1kGP) resource to include 602 trios
- High-coverage whole-genome sequencing of the expanded 1kGP cohort
- Discovery of more rare SNVs as well as INDELs and SVs across the frequency spectrum
- Generation of an improved and accessible reference imputation panel

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

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

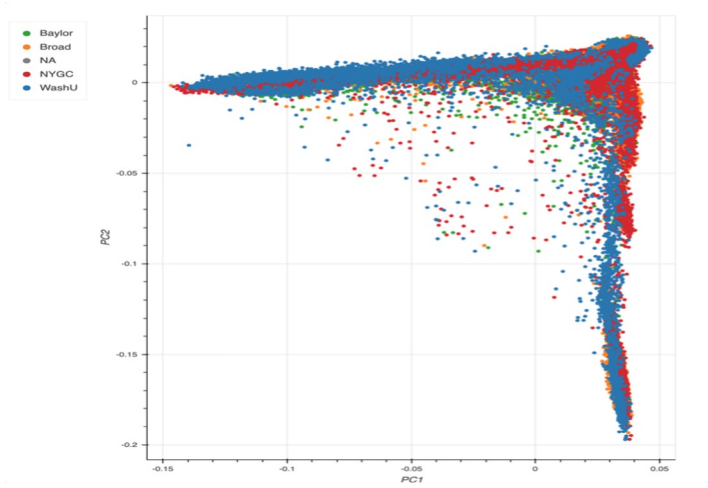


Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL)

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

Powered by AnVIL

Centers for
Common Disease Genomics



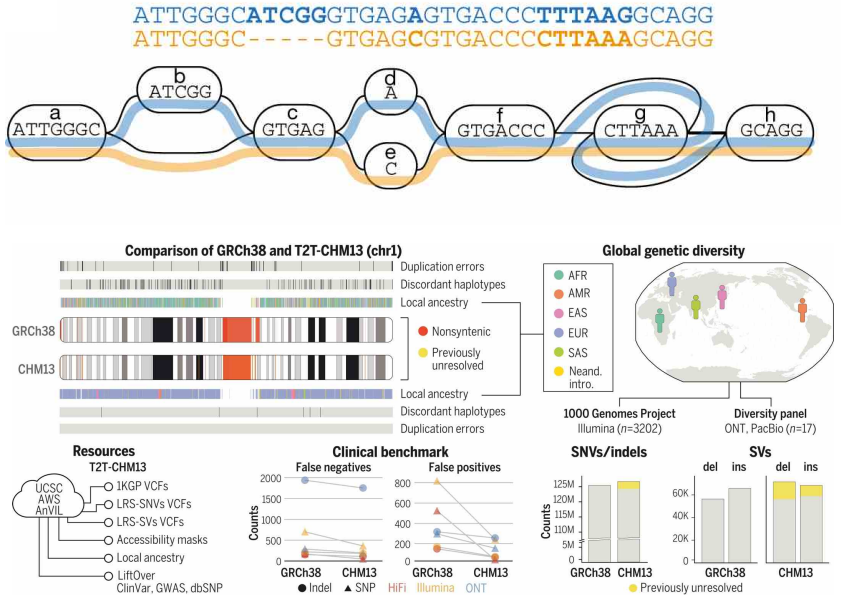
136,959 WGS samples

Broad Institute: 16,828
Baylor College: 35,493
NY Genome Center: 40,975
WUSTL: 43,620

TELOMERE-TO-TELOMERE CONSORTIUM

TOWARDS A COMPLETE REFERENCE OF HUMAN GENOME DIVERSITY

HUMAN PANGENOMI



Nurk et al. (2022) Science
Liao et al. (2023) Nature

WDL 101: Reproducible genomics at scale

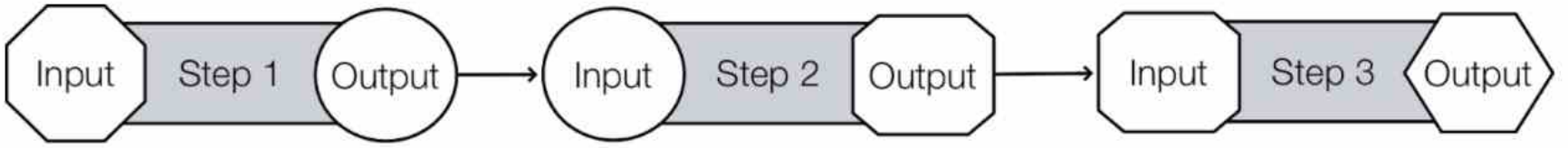
```
task samtoolsStats {
    input {
        File inputCram
        File cramIndex
        File targetRef
        String sampleName
    }

    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"
    }
}
```

What are the inputs?
Automatically copy from buckets to VMs

What should we run?
Essentially any command line tool!

What type of computers should we use?
Match resources (cores, RAM, disk) to needs

What are the outputs?
Automatically copy to buckets for next stage



github.com/openwdl/learn-wdl/blob/master/1_script_examples/2_language_...

Product Solutions Open Source Pricing Search Sign in Sign up

openwdl / learn-wdl Public

Notifications Fork 32 Star 72

Code Issues 1 Pull requests Actions Projects 1 Security Insights

master learn-wdl / 1_script_examples / 2_language_patterns / 1_linear_two_task / linear-docker.wdl Go to file

lynnlangit rename again Latest commit d643282 on May 22, 2020 History

1 contributor

54 lines (46 sloc) | 774 Bytes Raw Blame

```
version 1.0

workflow SayItTwice {
  input {
    String name
  }
  call WriteGreeting {
    input:
       name = name
  }
  call ReadItBackToMe {
    input:
       written_name = WriteGreeting.output_name
  }

  output {
    File outfile = ReadItBackToMe.repeated_name
  }
}

task WriteGreeting {

  input {
    String name
  }

  command {
    echo "${name}"
  }
  output {
    File output_name = stdout()
  }
  runtime {
          docker: "ubuntu:latest"
  }
}

task ReadItBackToMe {

  input {
    File written_name
    String original_name = read_string(written_name)
  }

  command {
      echo "${original_name} to you too"
  }
  output {
    File repeated_name = stdout()
  }
  runtime {
          docker: "ubuntu:latest"
  }
}
```

© 2022 GitHub, Inc. Terms Privacy Security Status Docs Contact GitHub Pricing API Training Blog About

SayItTwice.input

↓

WriteGreeting

↓

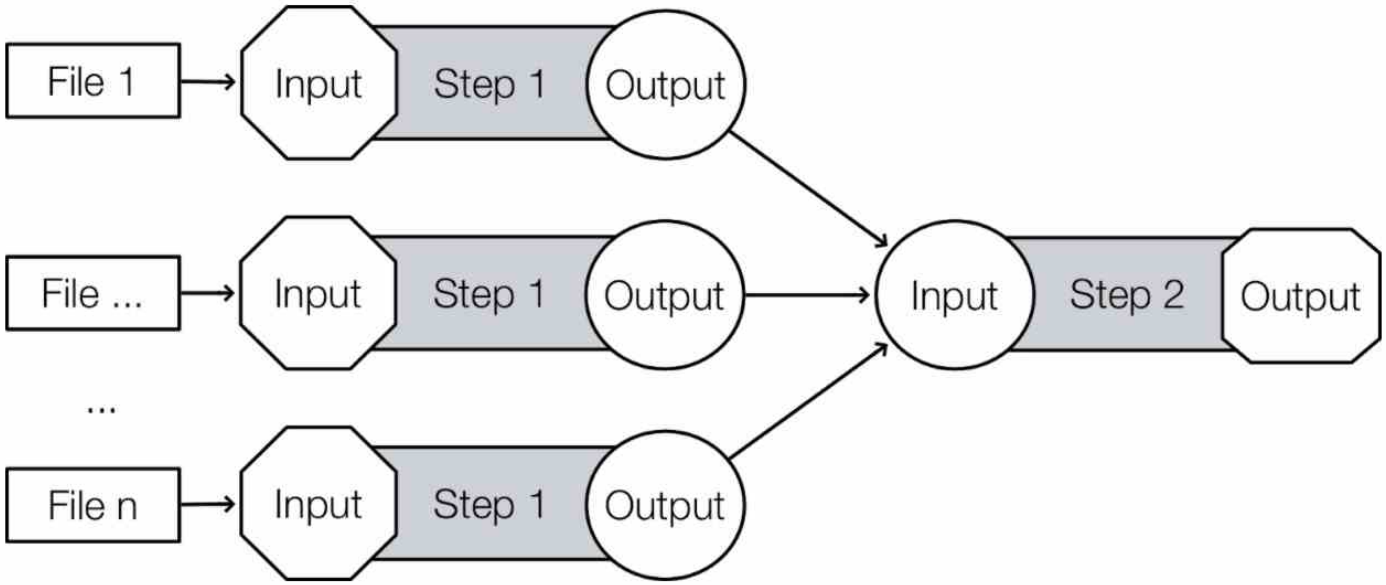
ReadGreeting

↓

SayItTwice.outfile

https://github.com/openwdl/learn-wdl/blob/master/1_script_examples/2_language_patterns/1_linear_two_task/linear-docker.wdl

Scatter-Gather



17 lines (14 sloc) | 220 Bytes

```
version 1.0

workflow HelloWorld {
  scatter(i in range(15)) {
                call WriteGreeting
        }

}

task WriteGreeting {
  command {
     echo "Hello Scattered World"
  }
  output {
     File output_greeting = stdout()
  }
}
```

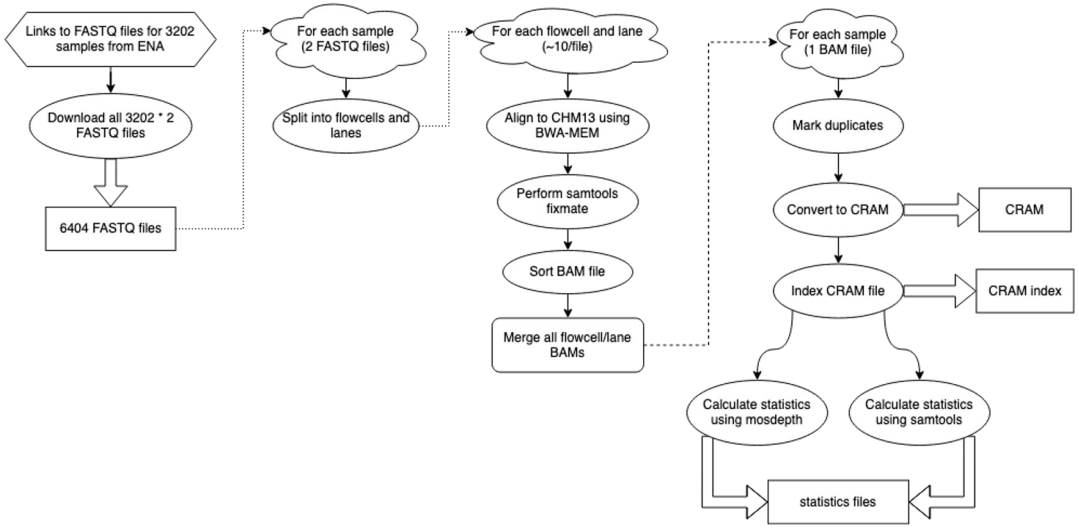
26 lines (22 sloc) | 369 Bytes

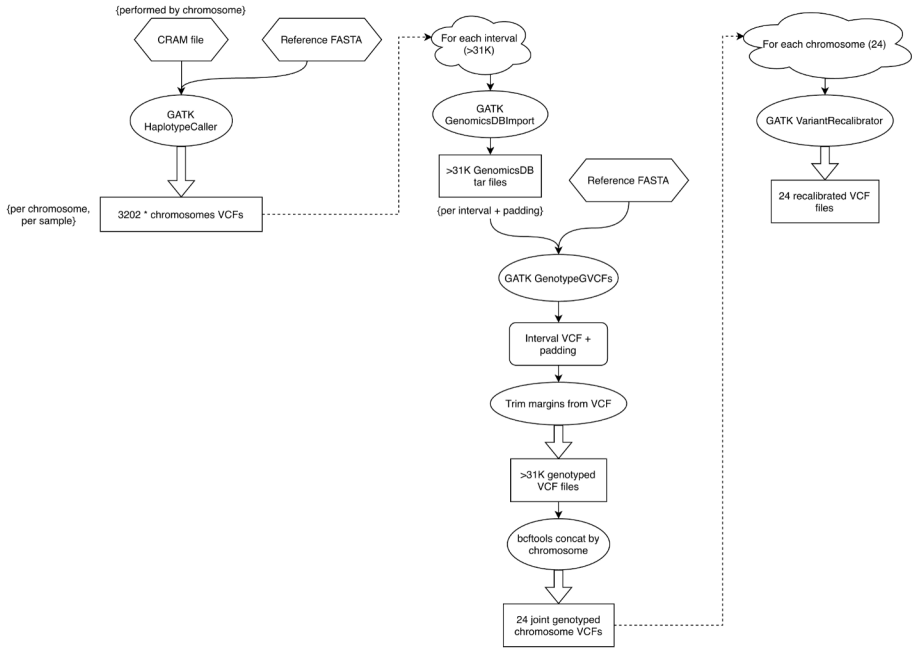
Raw Blame

```
version 1.0

workflow TimingWorkflow {
        input {
                Int sleep_time
        }
        # scatter keyword parallelizes this Sleep task execution
        scatter(i in range(15)) {
                call Sleep { input: sleep_time = i }
        }
}

task Sleep {
        input {
                Int sleep_time
        }
        command {
                echo "I slept for ${sleep_time}"
                Sleep ${sleep_time}
        }
        output {
                String out = read_string(stdout())
        }
}

```

3202 Genomes to go…

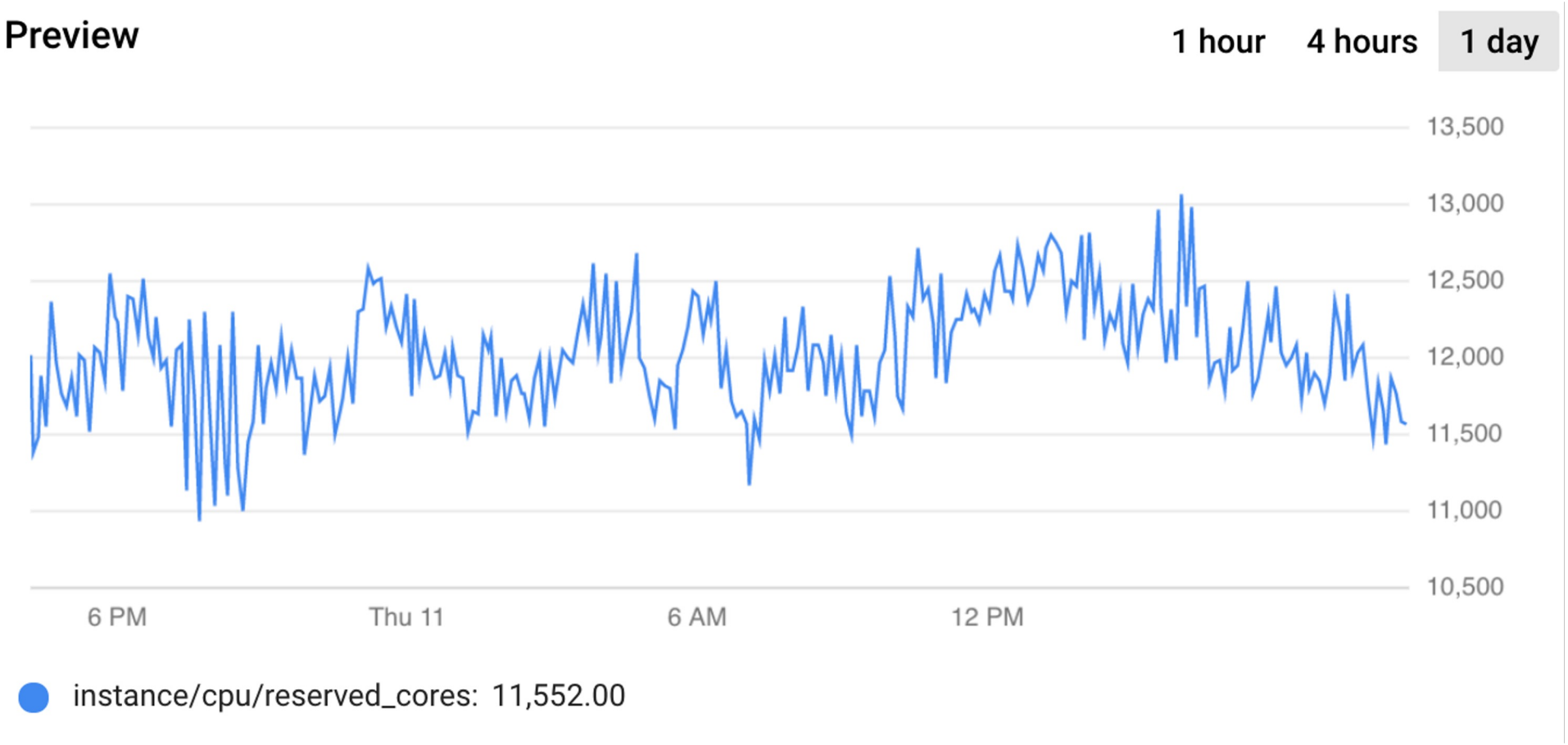
Samantha Zarate



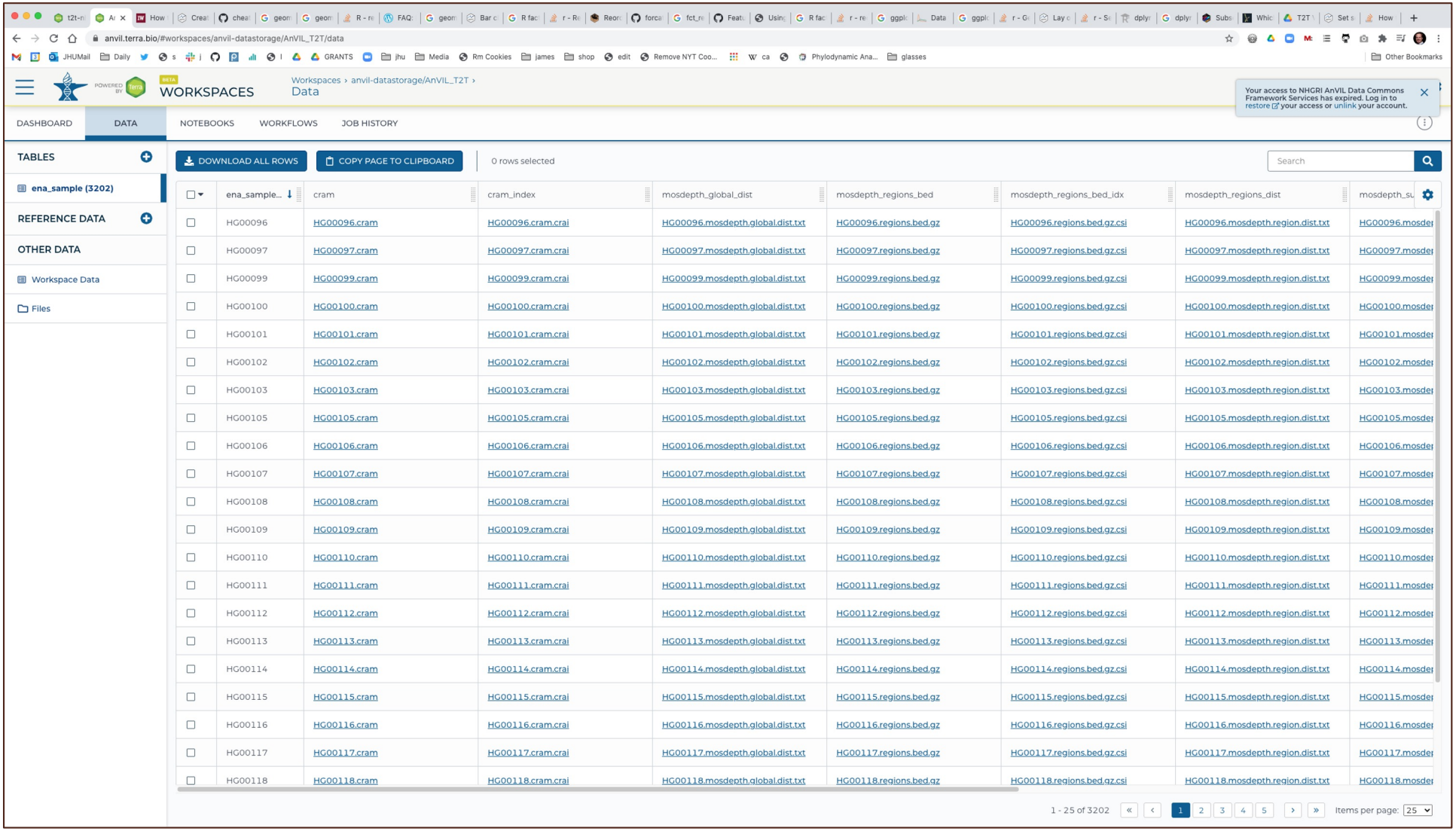



```
task samtoolsStats {
    input {
        File inputCram
        File cramIndex
        File targetRef
        String sampleName
    }

    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"
    }
}
```

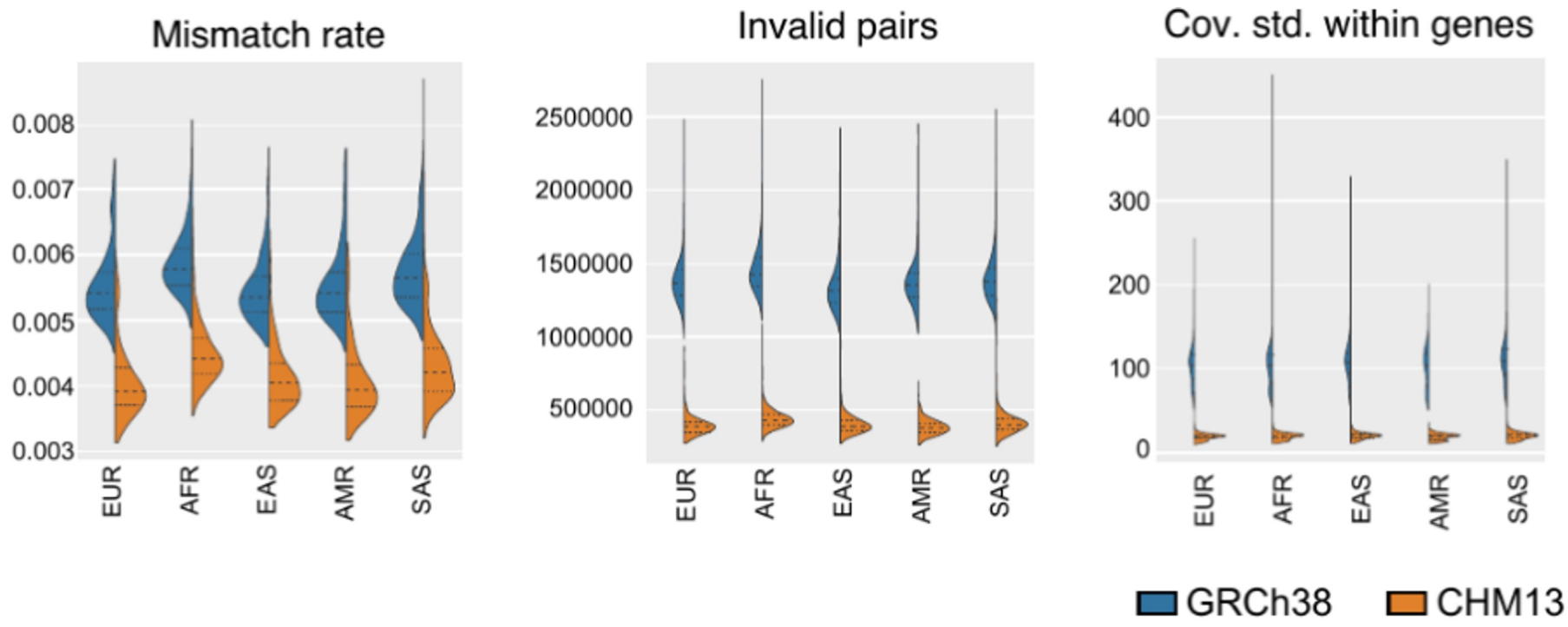
Core usage over 24 hours



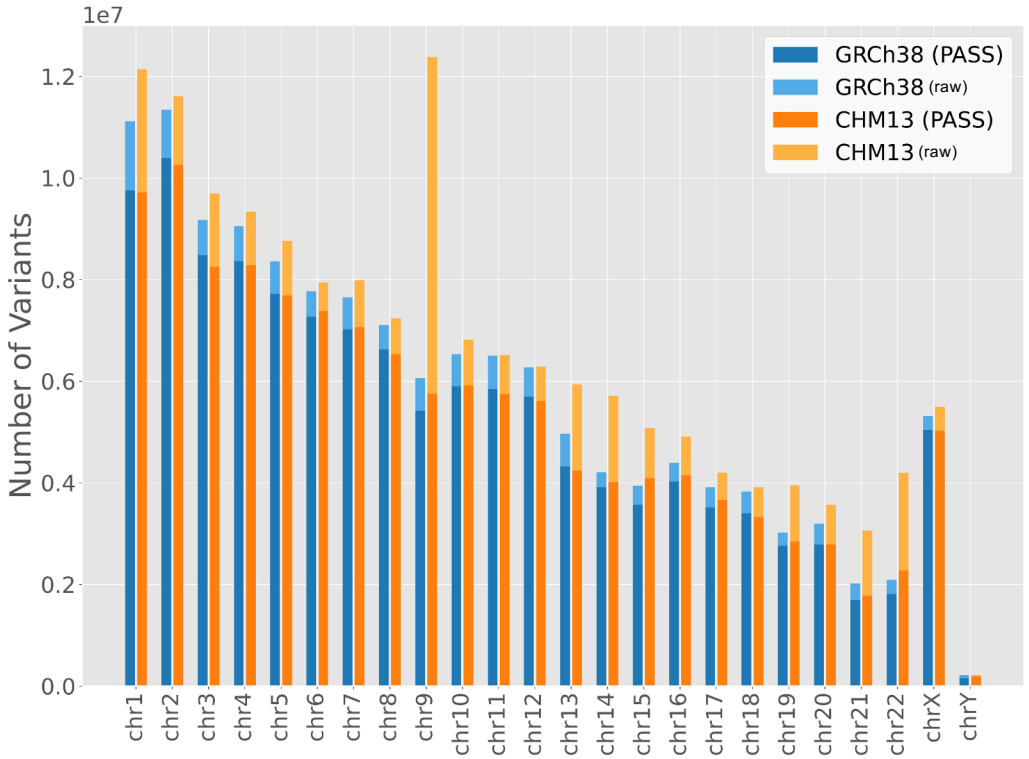
AnVIL Data Table



1000G Mapping on T2T-CHM13

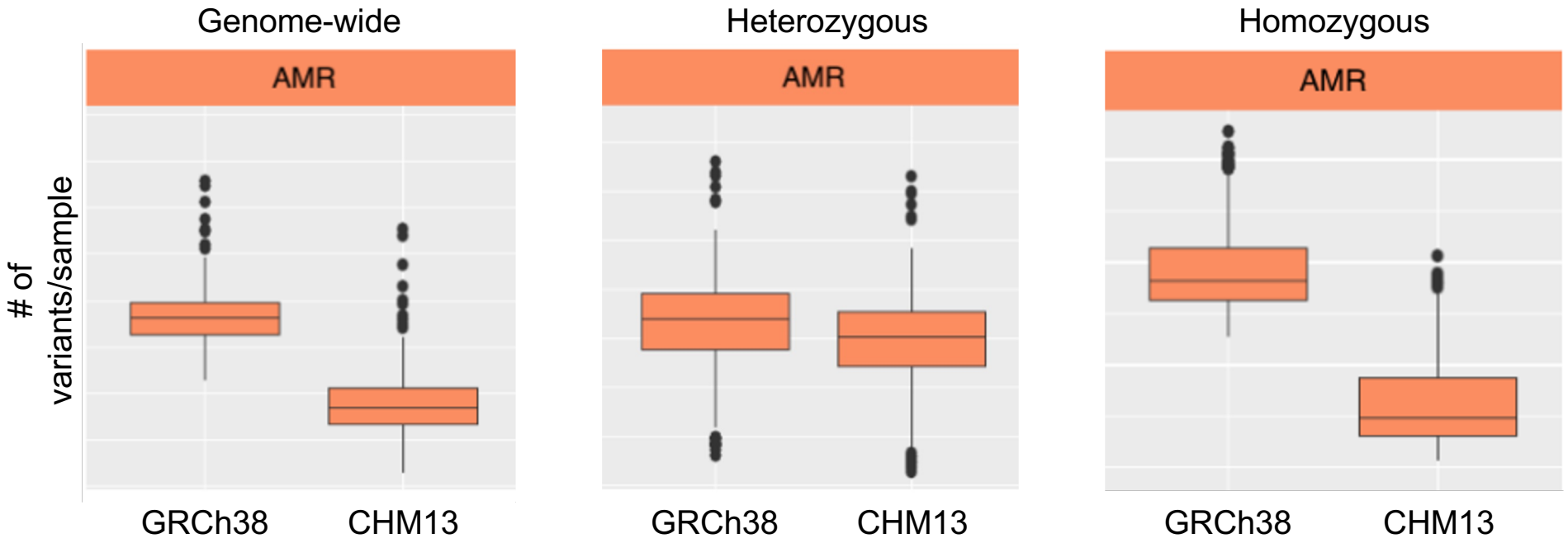


Across 1000G, Many More Variants Found Using CHM13

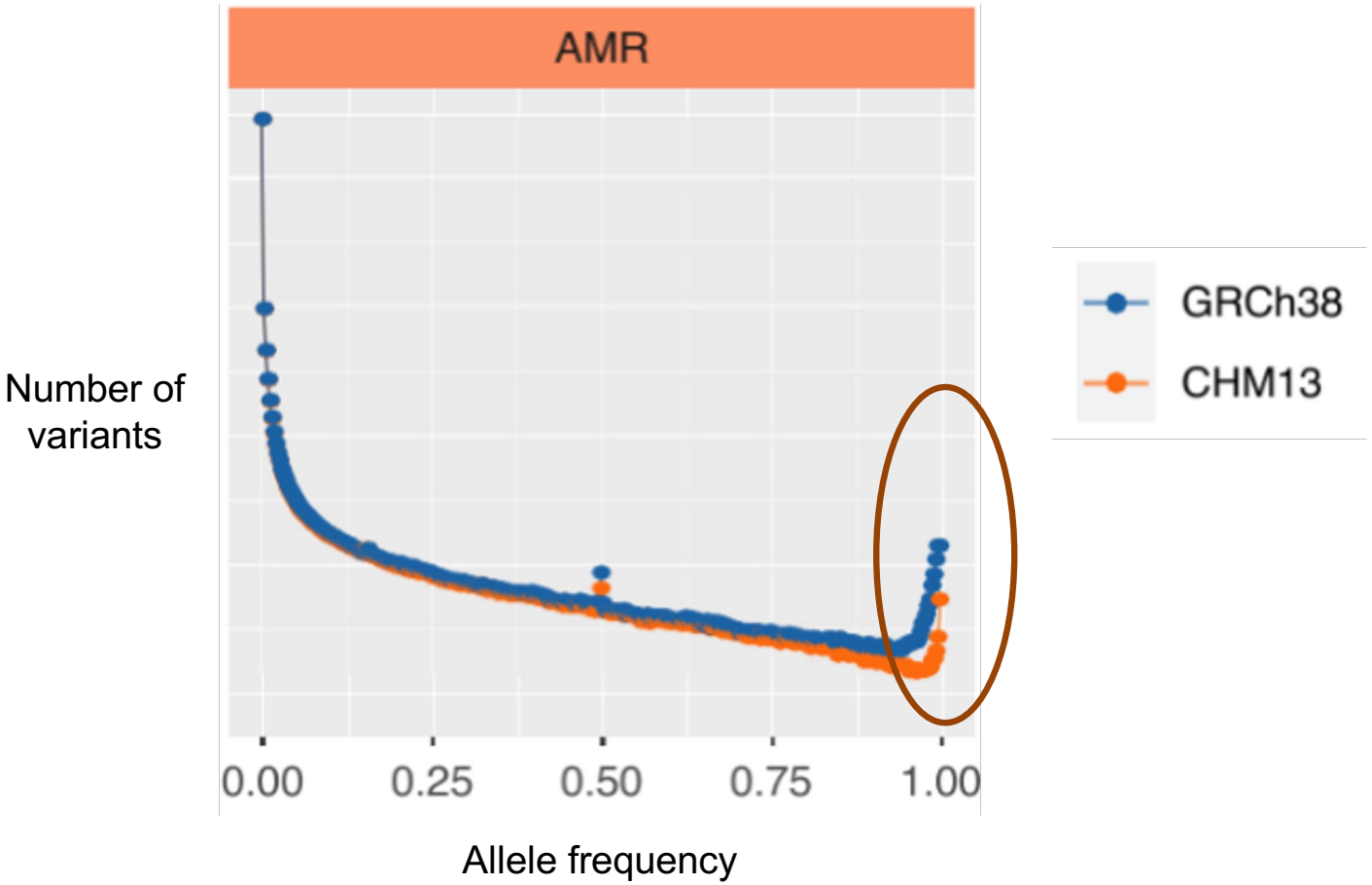


	GRCh38	CHM13
# PASS variants	125,484,020	126,591,489

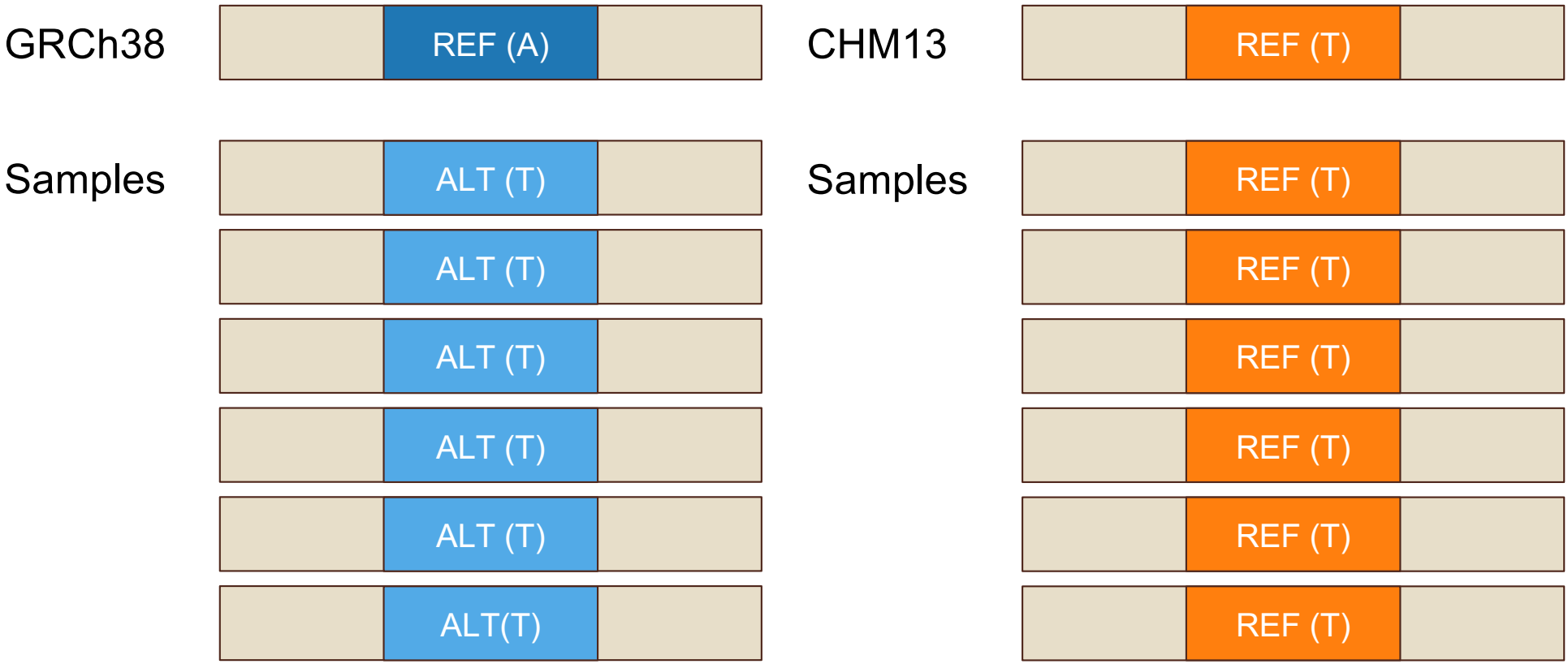
1000G Per-Sample Variant Counts on T2T-CHM13



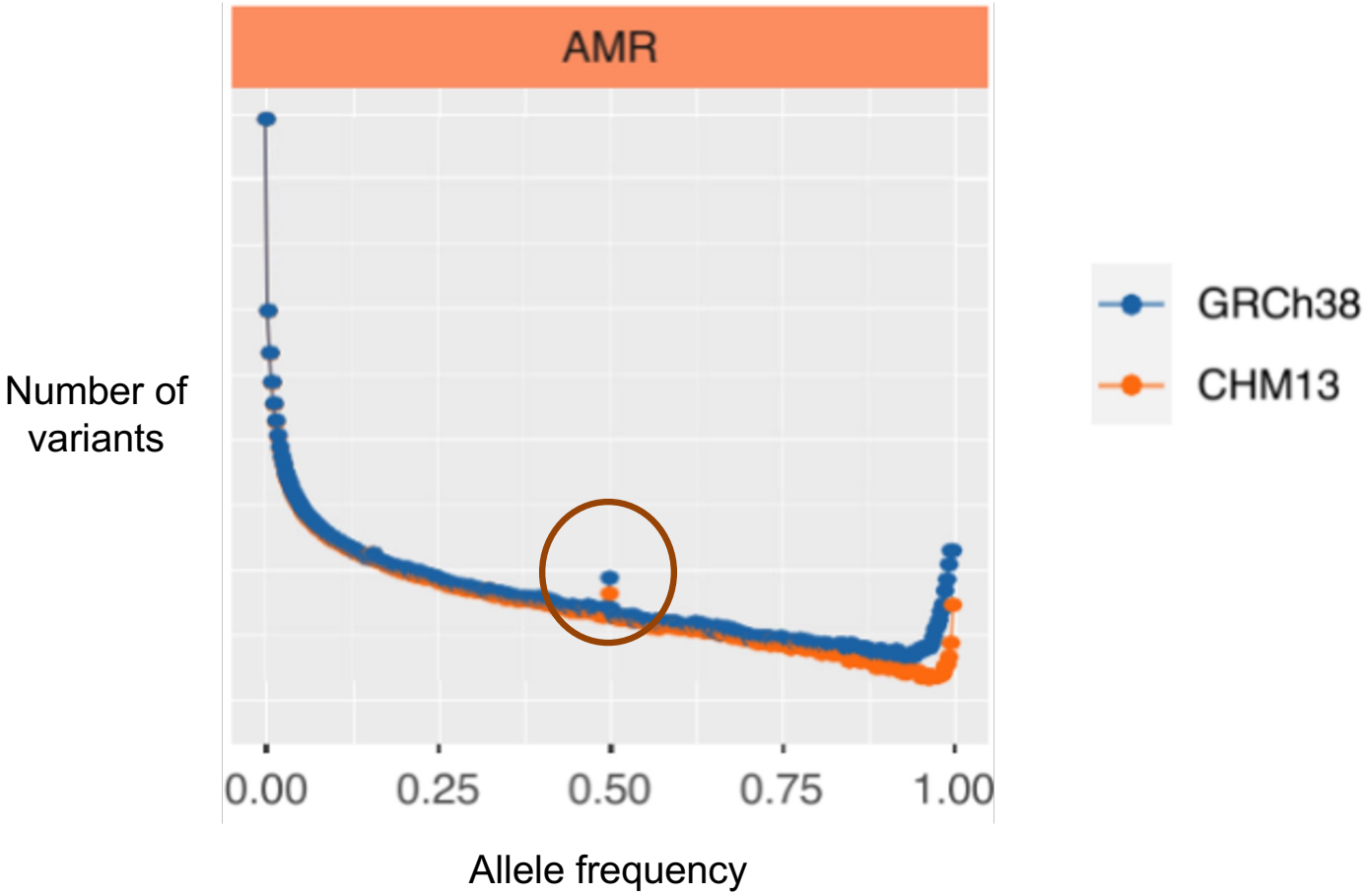
Explaining Decreased Per-Sample Count



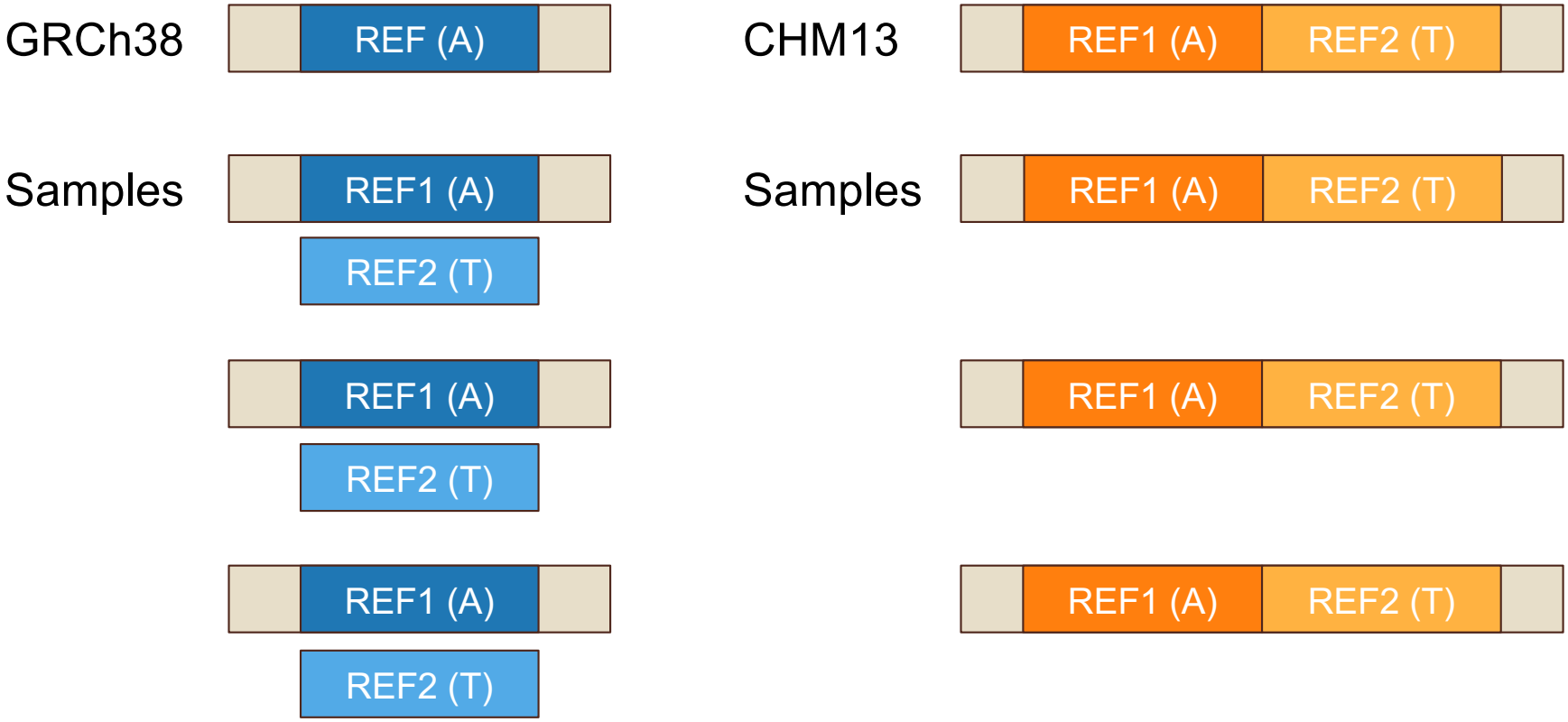
Allele Frequency = 1: Reference error / private variant



Explaining Decreased Per-Sample Count

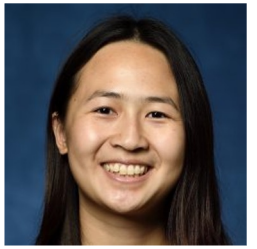


Allele Frequency ≈ 0.5: Collapsed duplication

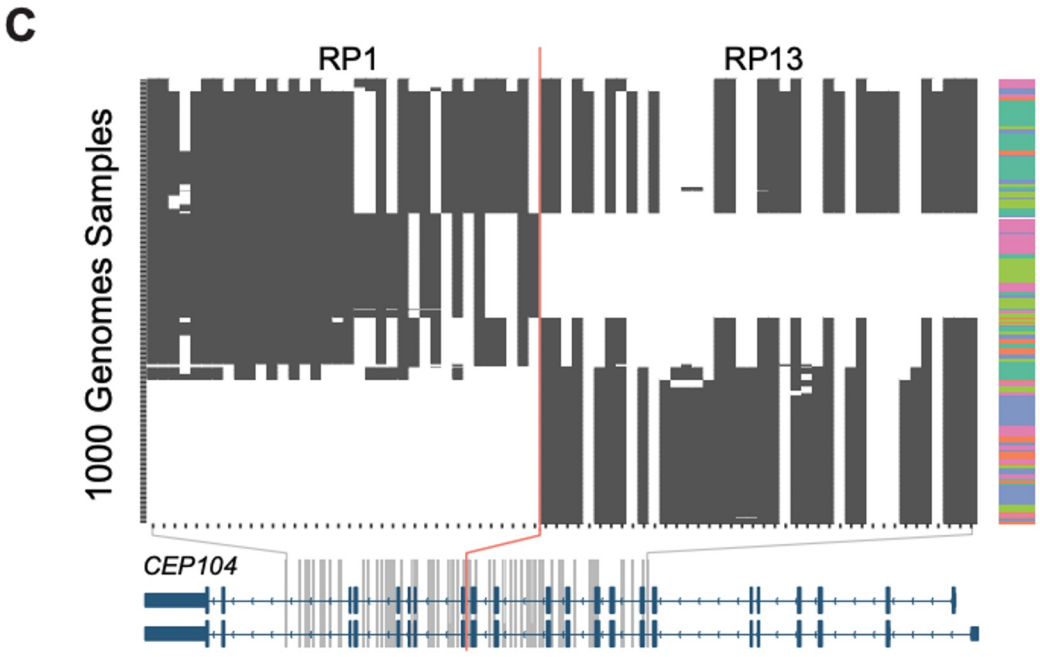


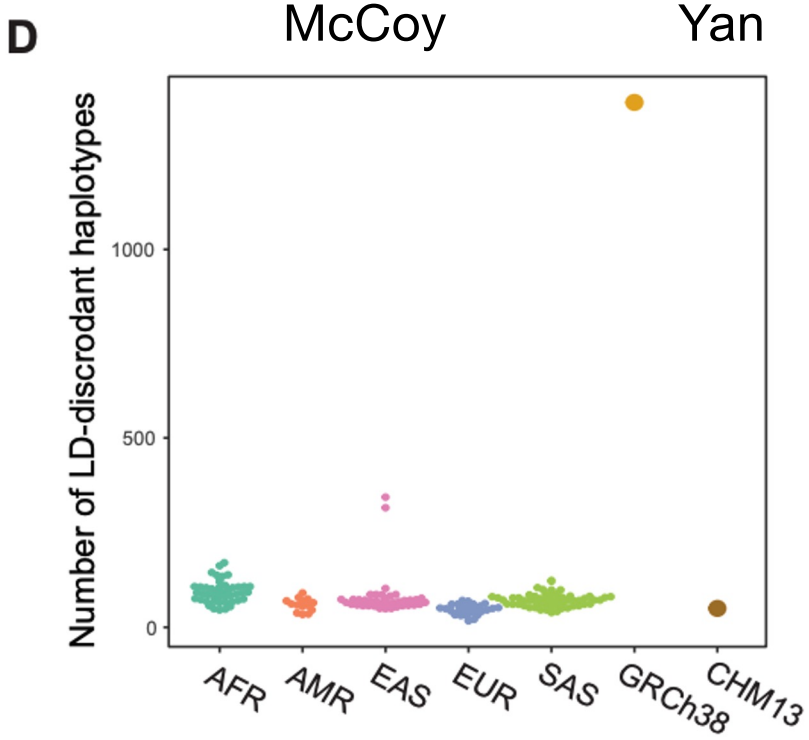
Local Ancestry Comparisons

Rajiv McCoy

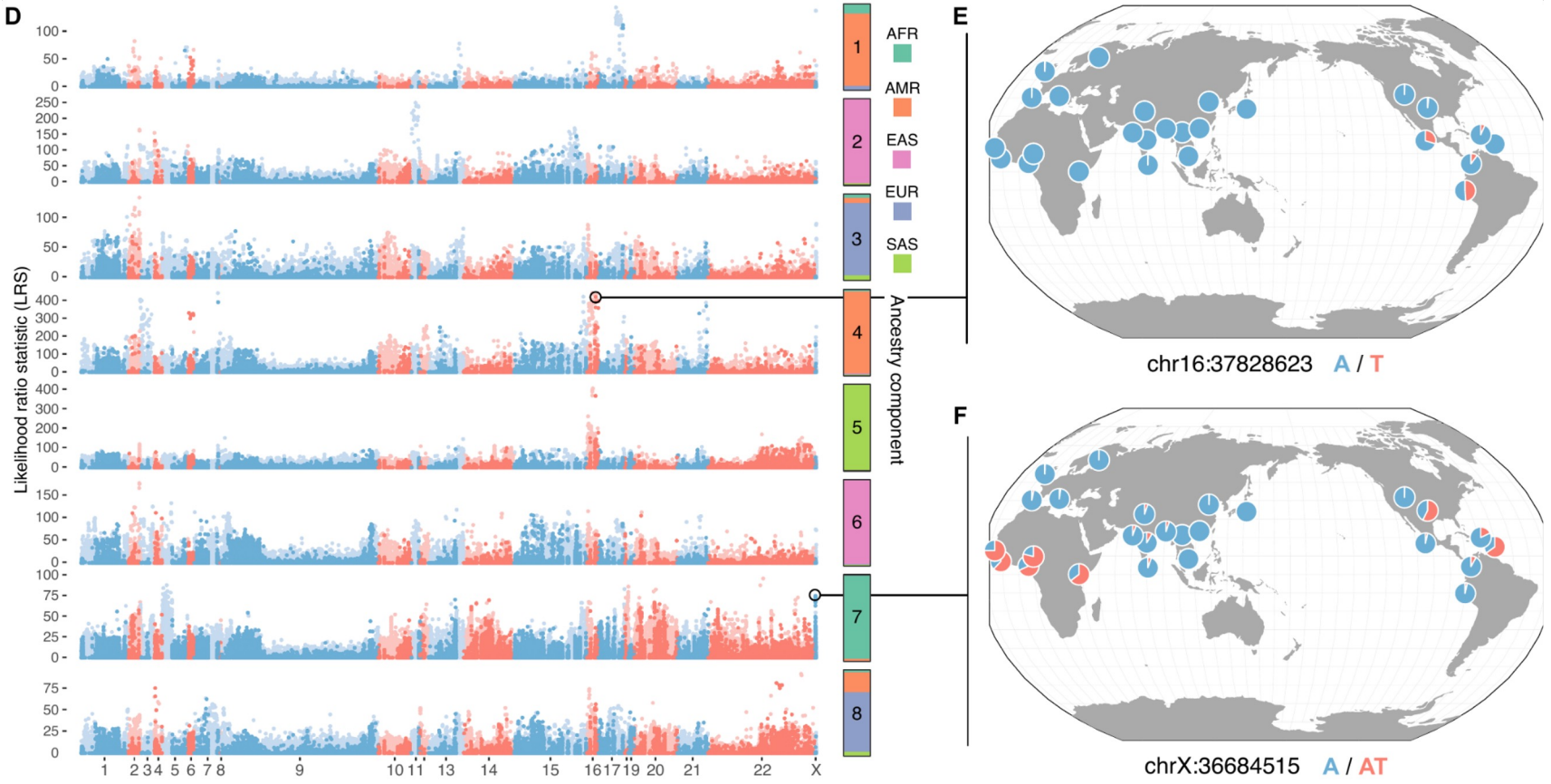


Stephanie Yan





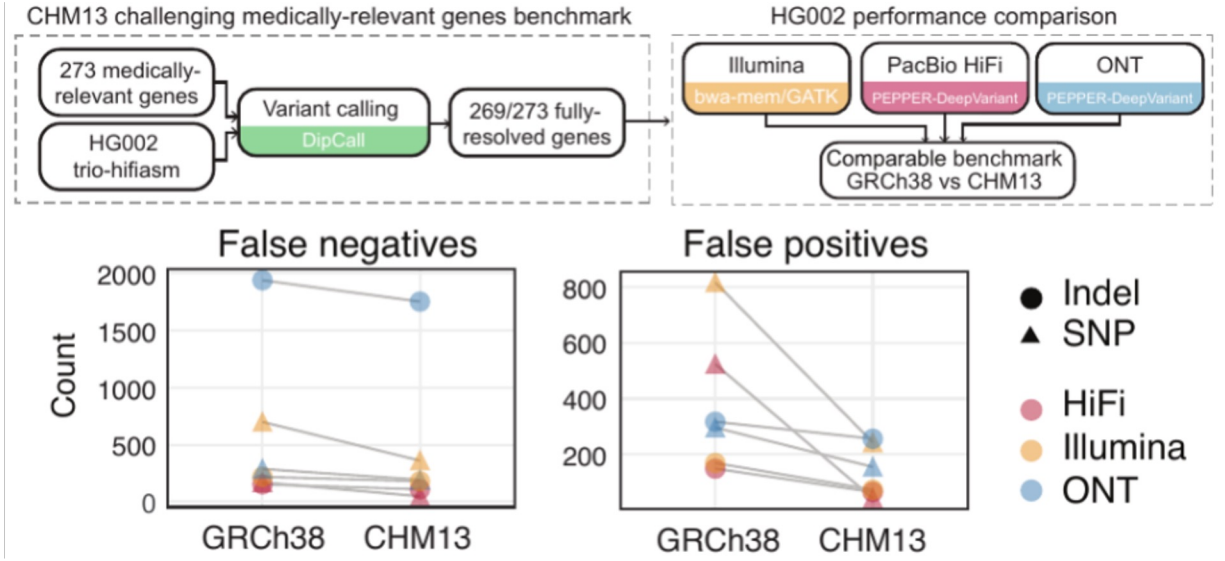
Variants in Newly Resolved Regions



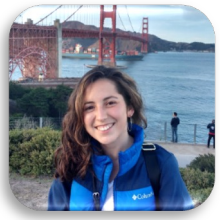
Stephanie Yan

Rajiv McCoy

T2T-CHM13 Improves Clinical Genomics Variant Calling



Danny Miller



Daniela Soto

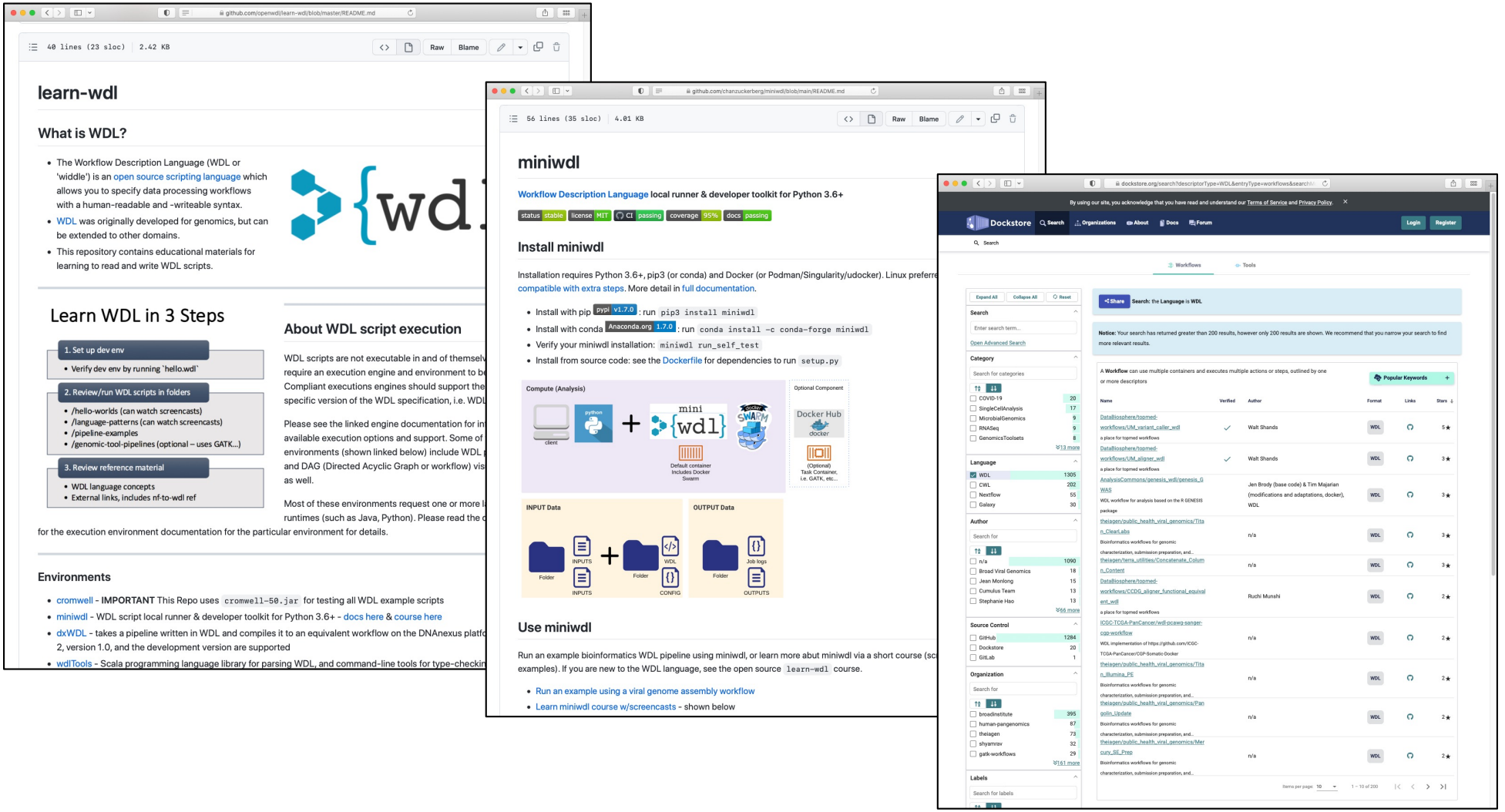
Megan Dennis

Justin Zook

Fritz Sedlazeck

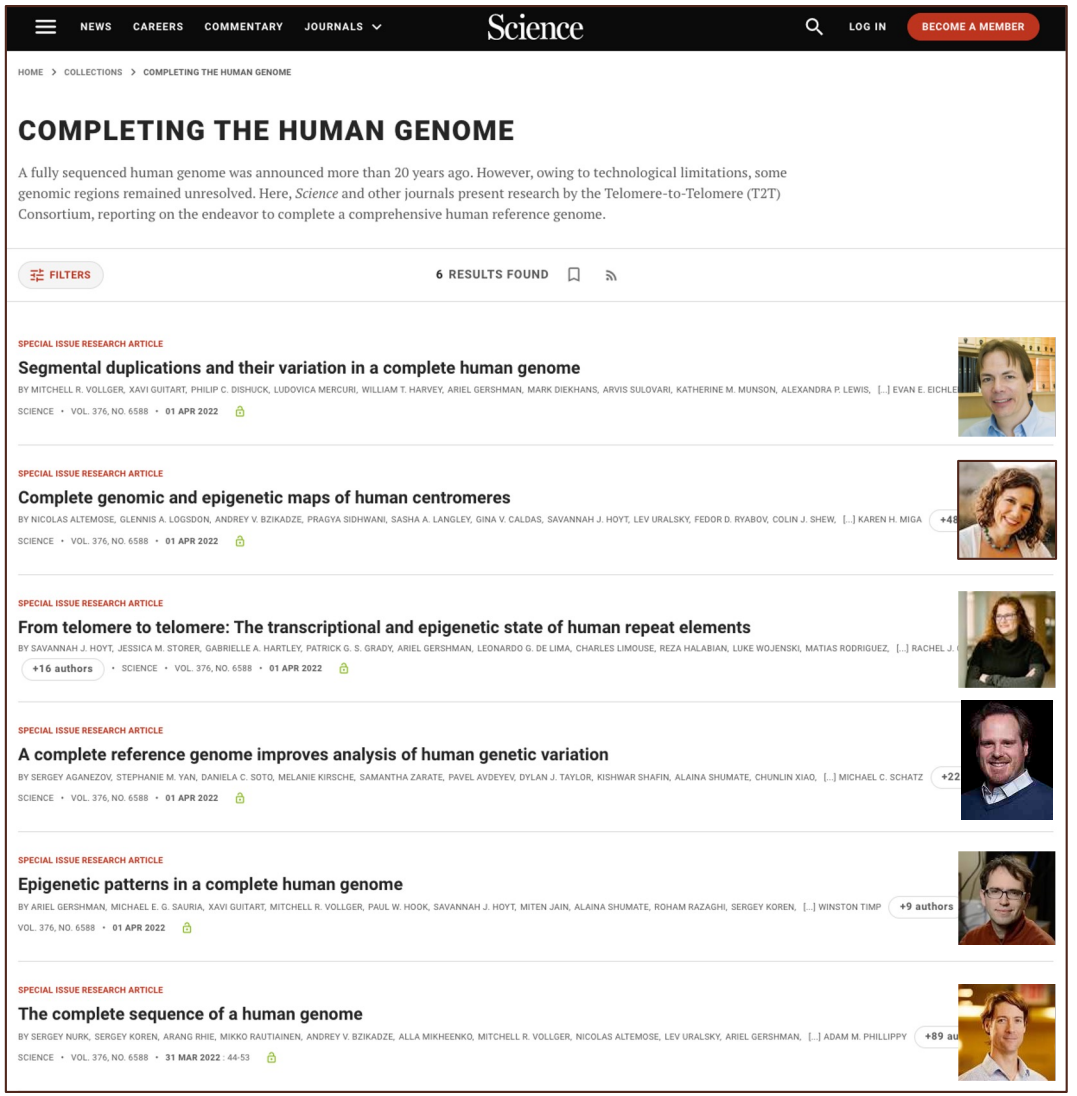
For more information, see: (Wagner, J *et al.*, NBT, 2022)

Learning More about WDL



https://github.com/openwdl/learn-wdl/
https://github.com/chanzuckerberg/miniwdl/
https://dockstore.org/


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
SCIENCE • VOL. 376, NO. 6588 • 01 APR 2022
SPECIAL ISSUE RESEARCH ARTICLE
Complete genomic and epigenetic maps of human centromeres
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
+16 authors • SCIENCE • VOL. 376, NO. 6588 • 01 APR 2022
SPECIAL ISSUE RESEARCH ARTICLE
A complete reference genome improves analysis of human genetic variation
SCIENCE • VOL. 376, NO. 6588 • 01 APR 2022
SPECIAL ISSUE RESEARCH ARTICLE
Epigenetic patterns in a complete human genome
+9 authors
VOL. 376, NO. 6588 • 01 APR 2022
SPECIAL ISSUE RESEARCH ARTICLE
The complete sequence of a human genome
SCIENCE • VOL. 376, NO. 6588 • 31 MAR 2022 : 44-53

T2T Team @ Santa Cruz, August 2022