A complete reference genome improves analysis of human genetic variation

Michael Schatz

November 18, 2021
CSHL Advanced Sequencing Course


@mike_schatz

Schatzlab Overview

Human Disease Genetics

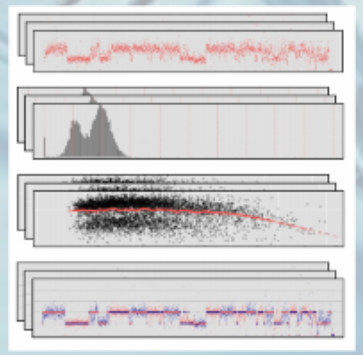
Aganezov *et al.* (2020)
Feigin *et al.* (2017)

Agricultural Genomics

Alonge *et. al.* (2020)
Chen *et al.* (2019)

Algorithmics & Systems Research

Kirsche *et al.* (2020)
Schatz (2009)

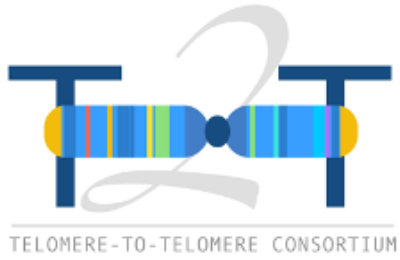


Single Cell & Single Molecule

Kovaka *et al.* (2020)
Sedlazeck *et al.* (2018)

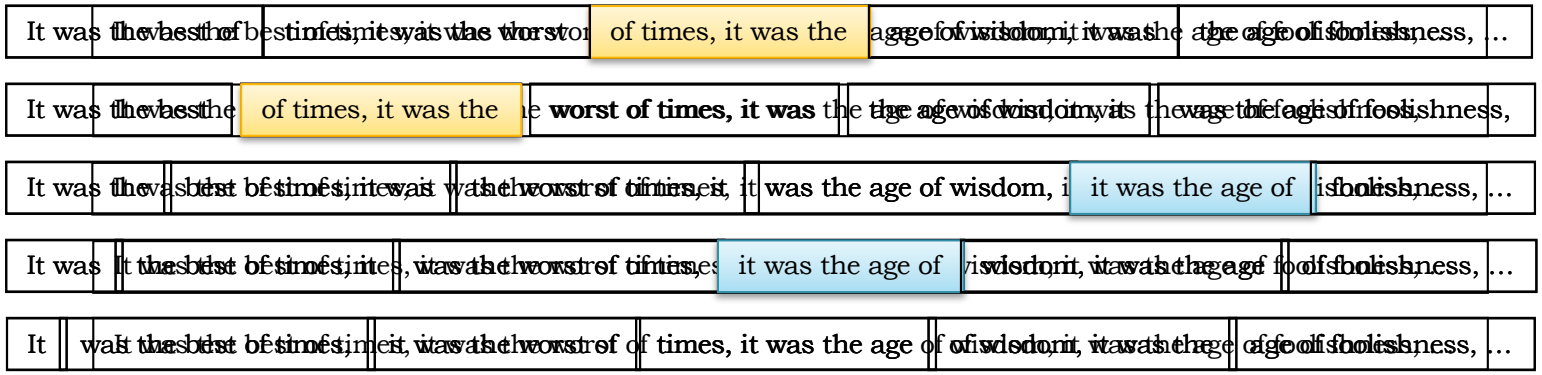
Outline

- Assembly by analogy
- T2T-CHM13 Assembly
  - How the T2T-CHM13 reference improves the analysis of genetic variation
- Beyond T2T-CHM13



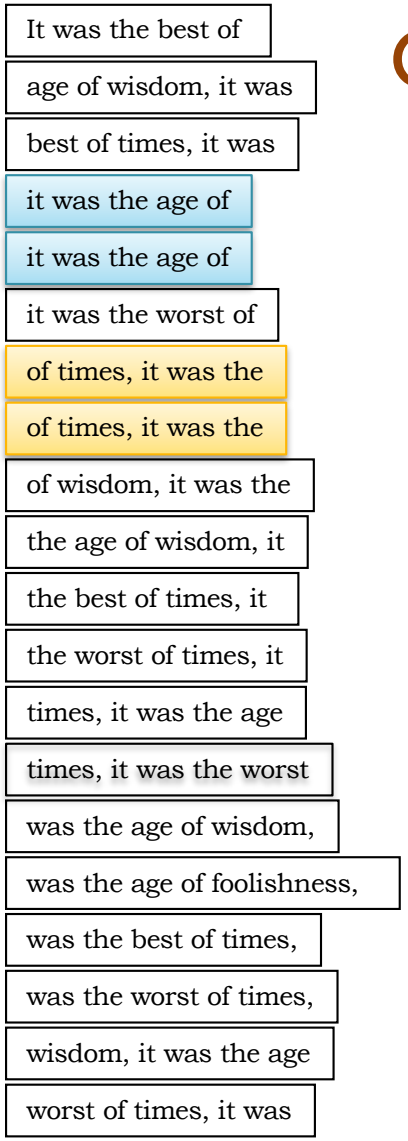
Shredded Book Reconstruction

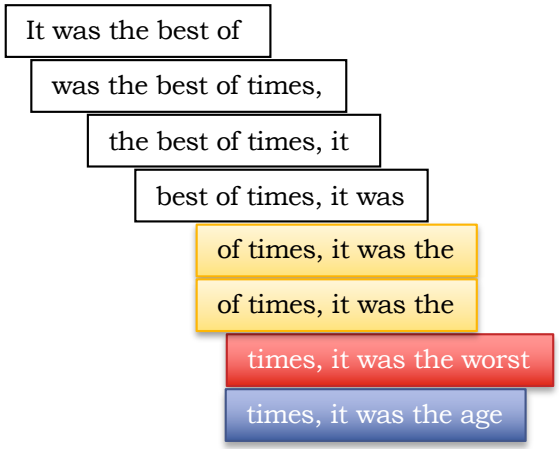
- Dickens accidentally shreds the first printing of A Tale of Two Cities
  - Text printed on 5 long spools



- How can he reconstruct the text?
  - 5 copies x 138,656 words / 5 words per fragment = 138k fragments
  - The short fragments from every copy are mixed together
  - Some fragments are identical

Greedy Reconstruction





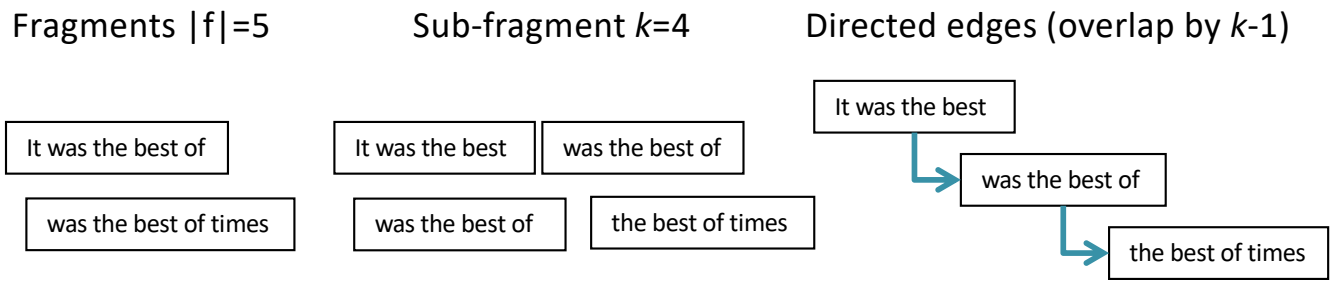
The repeated sequence make the correct reconstruction ambiguous

- It was the best of times, it was the [worst/age]

Model the assembly problem as a graph problem

de Bruijn Graph Construction

- G_k = (V,E)
  - V = Length-*k* sub-fragments
  - E = Directed edges between consecutive sub-fragments
    - Sub-fragments overlap by *k*-1 words

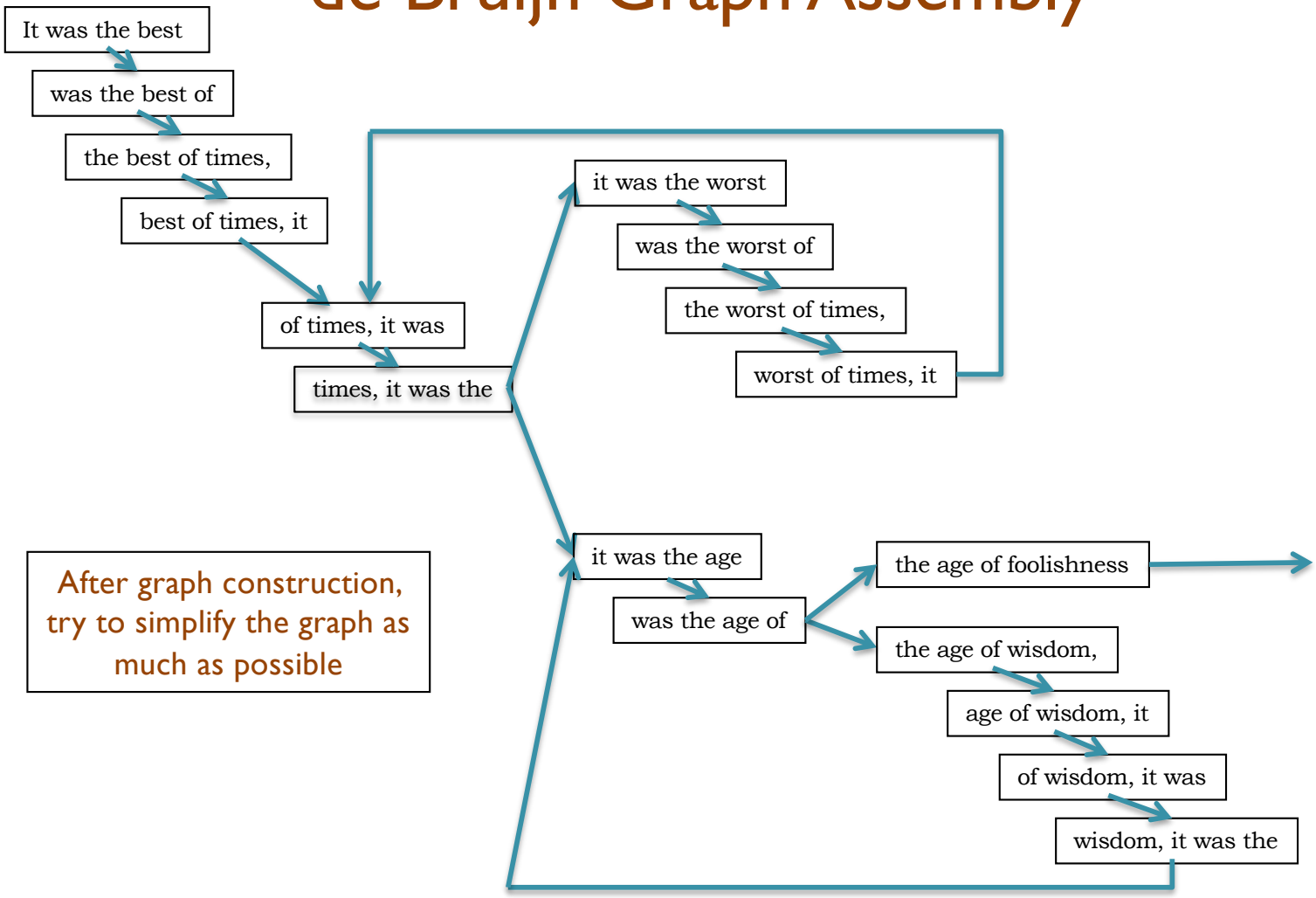


– Overlaps between fragments are implicitly computed

How to pronounce:
https://forvo.com/word/de_bruijn/

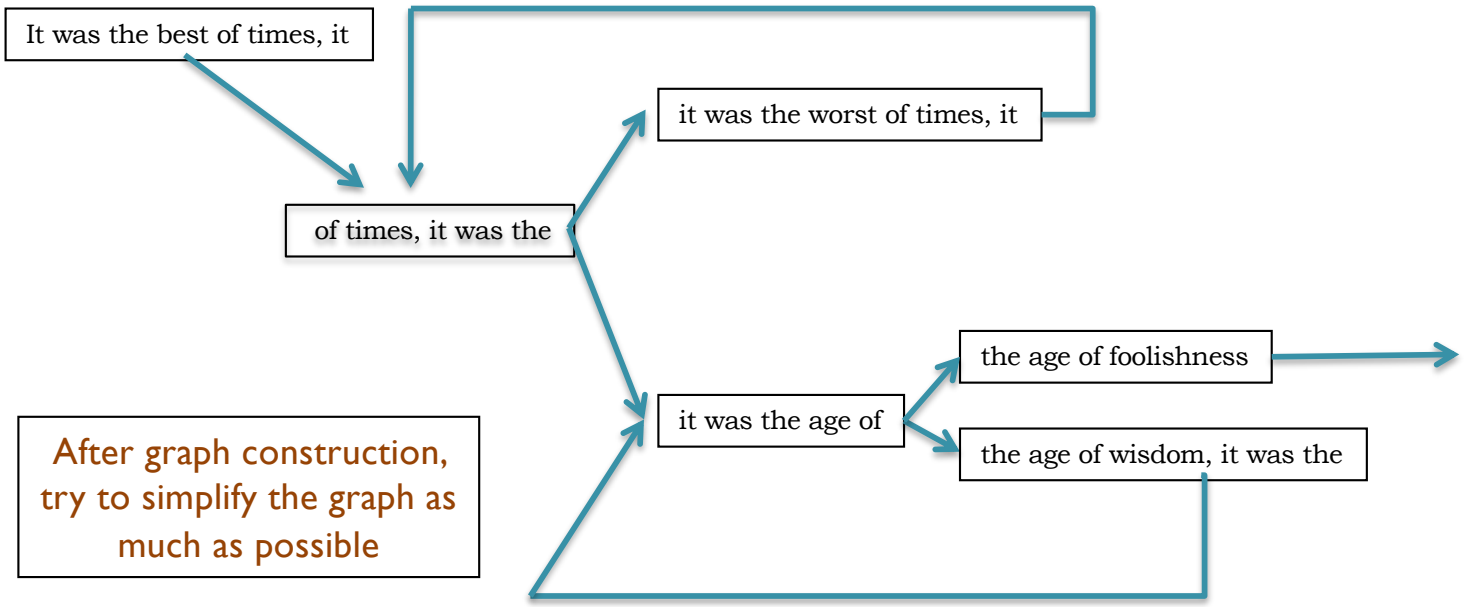
de Bruijn, 1946
Idury et al., 1995
Pevzner et al., 2001

de Bruijn Graph Assembly



After graph construction, try to simplify the graph as much as possible

de Bruijn Graph Assembly



After graph construction, try to simplify the graph as much as possible

The full tale

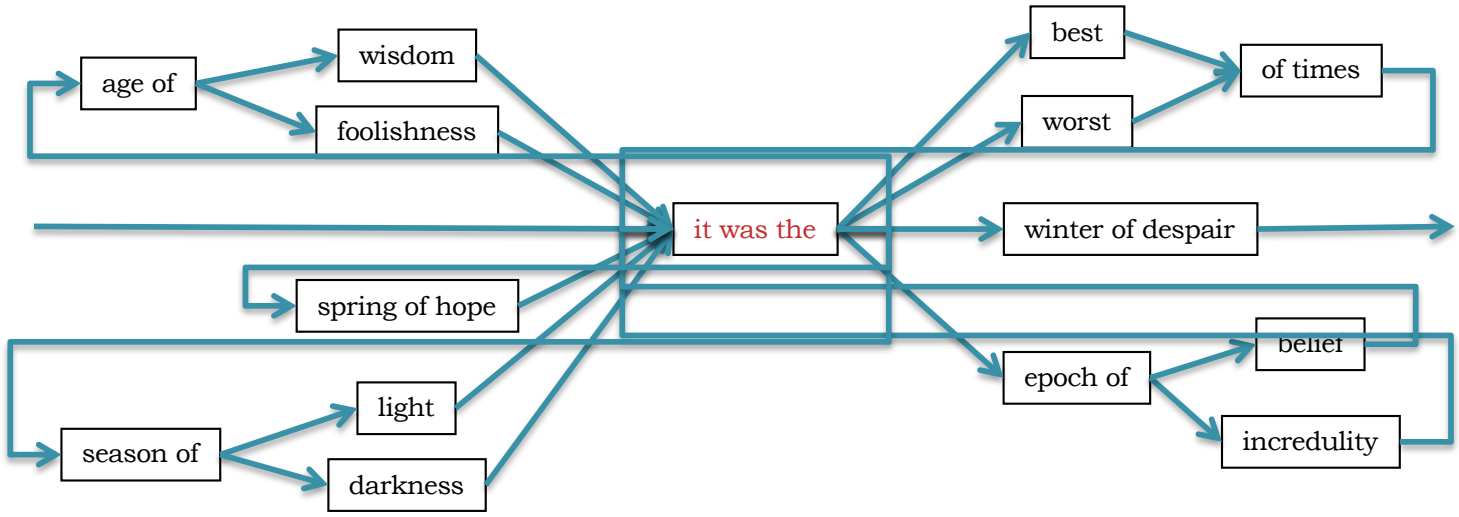
… it was the best of times it was the worst of times …

… it was the age of wisdom it was the age of foolishness …

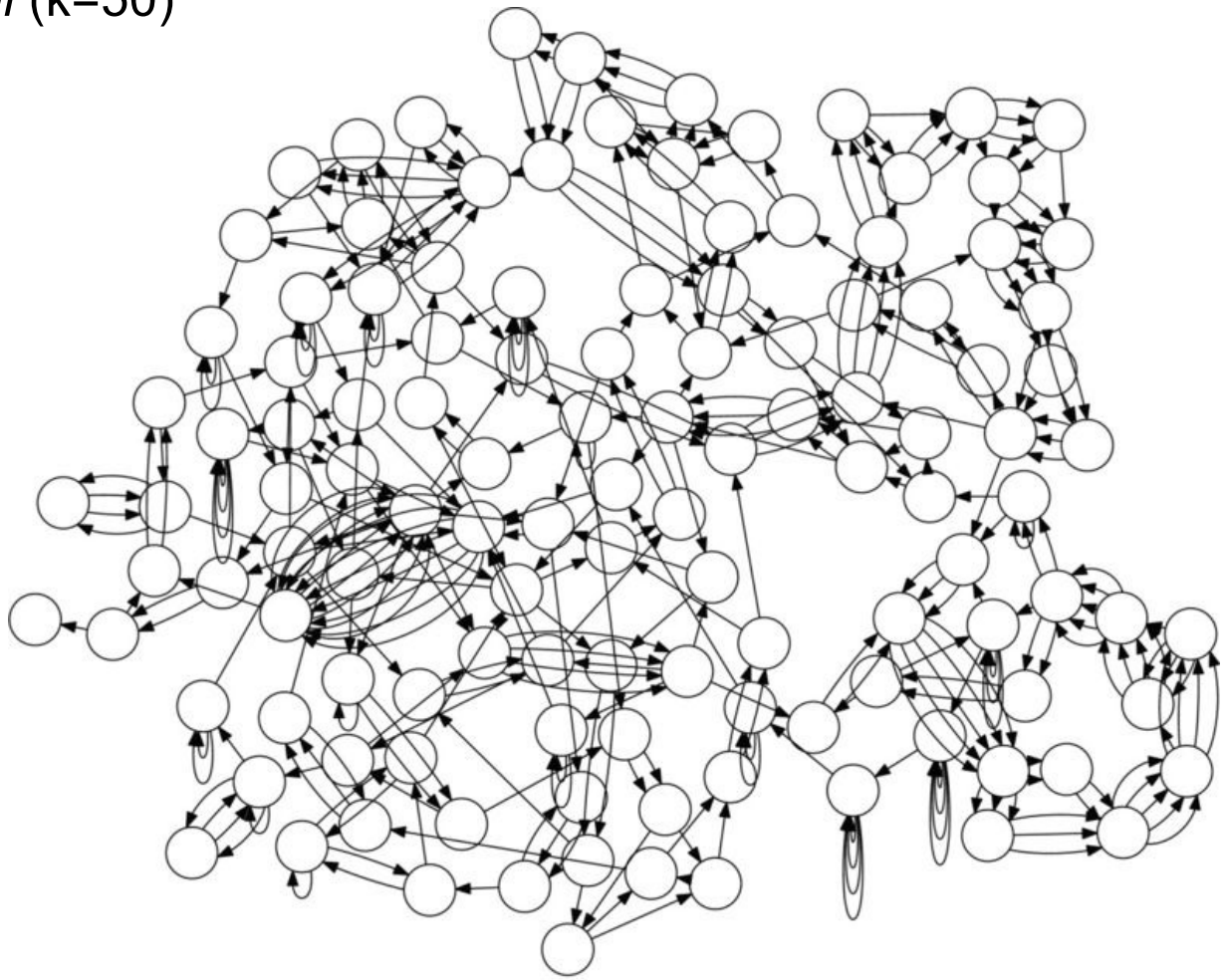
… it was the epoch of belief it was the epoch of incredulity …

… it was the season of light it was the season of darkness …

… it was the spring of hope it was the winder of despair …

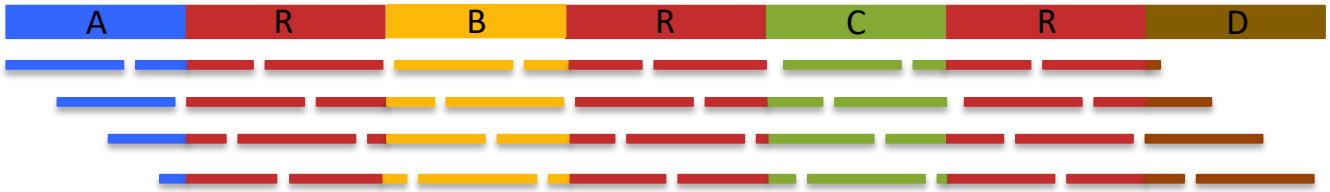


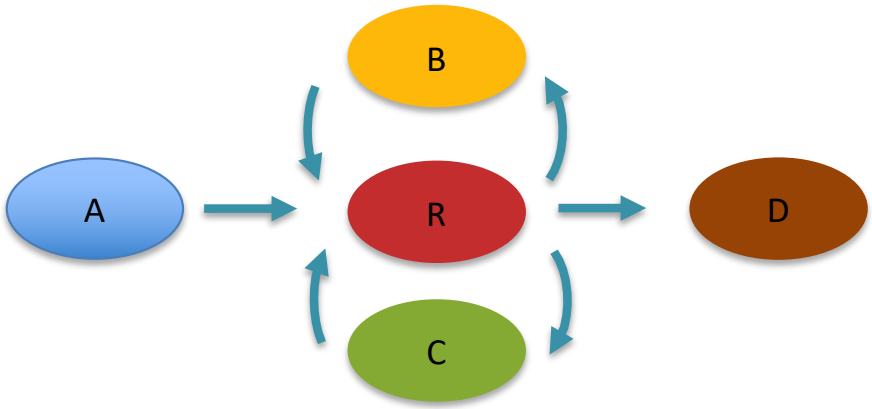
E. coli (k=50)



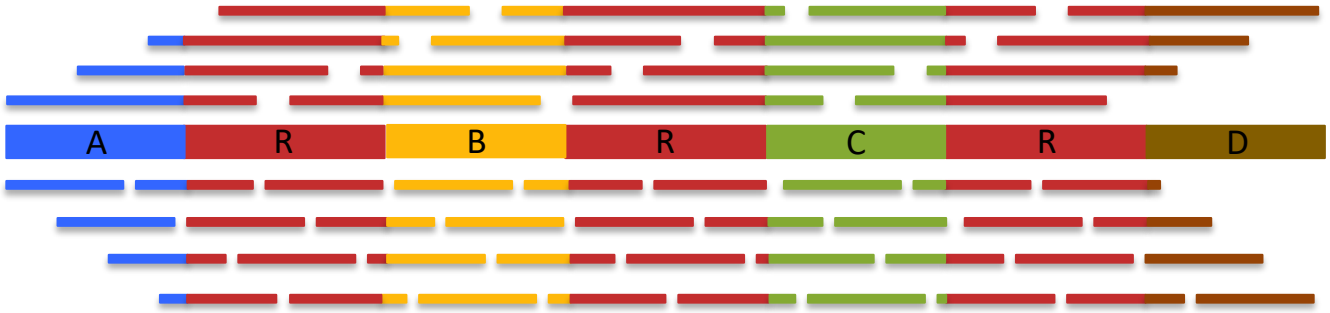
Reducing assembly complexity of microbial genomes with single-molecule sequencing
Koren et al (2013) Genome Biology. **14**:R101 https://doi.org/10.1186/gb-2013-14-9-r101

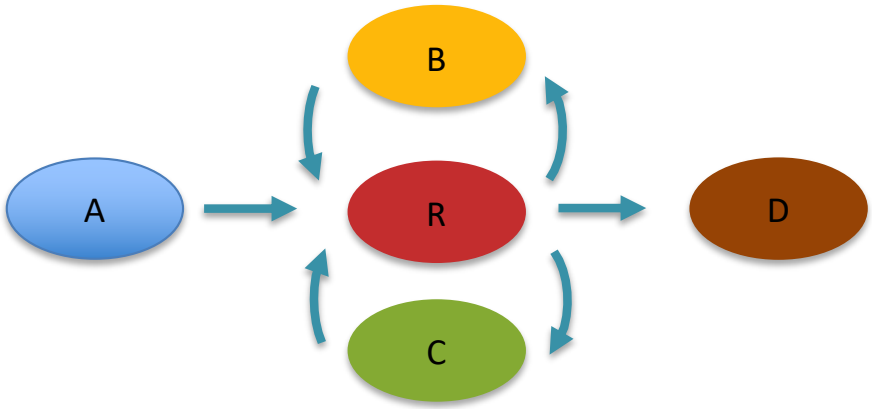
Assembly Complexity



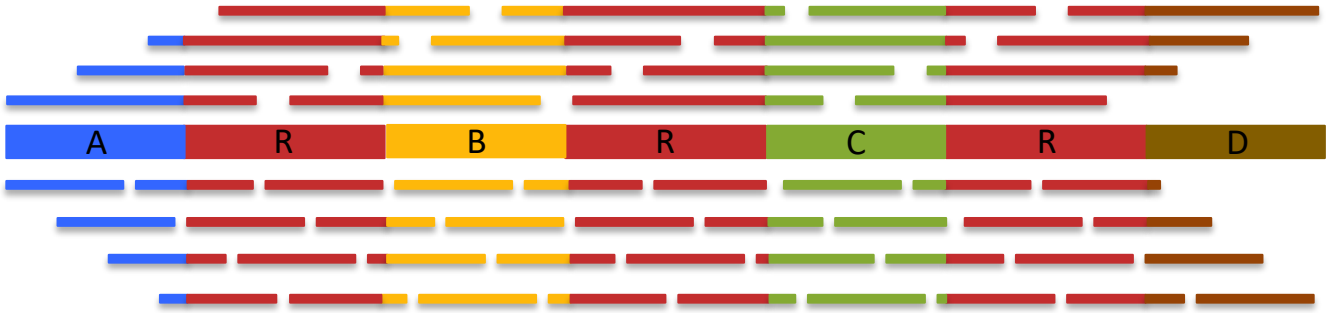


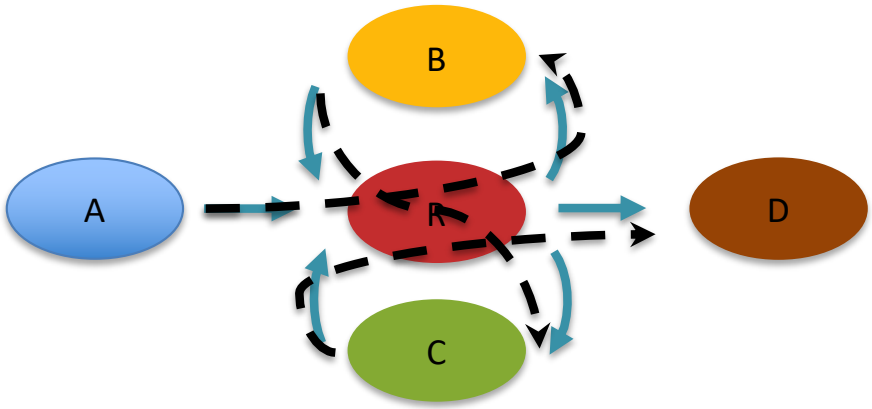
Assembly Complexity



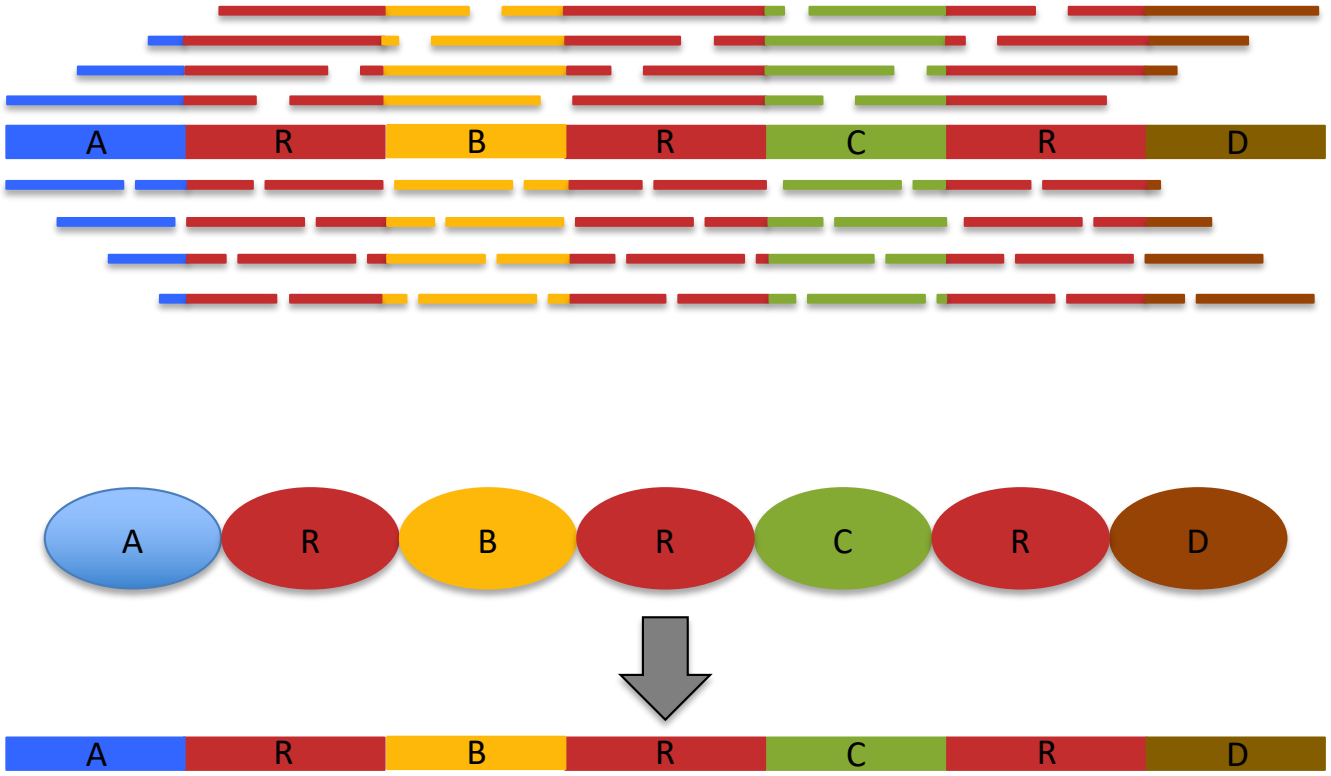


Assembly Complexity





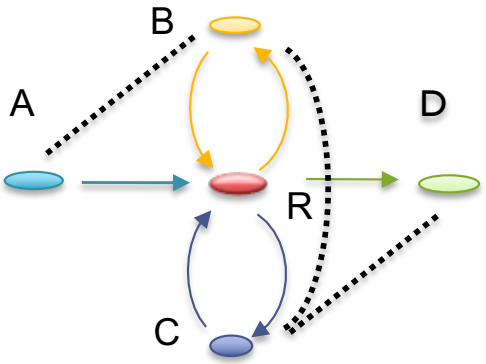
Assembly Complexity



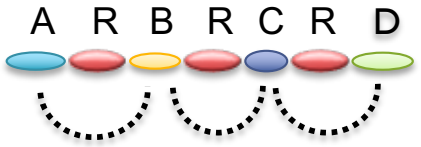
The advantages of SMRT sequencing
Roberts, RJ, Carneiro, MO, Schatz, MC (2013) *Genome Biology.* 14:405

Scaffolding

- Initial contigs (*aka* unipaths, unitigs) terminate at
  - *Coverage gaps*: especially extreme GC
  - *Conflicts*: errors, repeat boundaries, heterozygosity

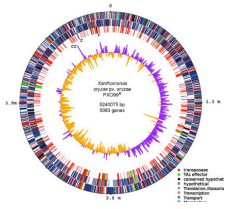


- Use mate-pairs or other mapping data to resolve correct order through assembly graph
  - Place sequence to satisfy the mate constraints
  - Mates through repeat nodes are tangled



- Final scaffold may have internal gaps called "sequencing gaps"
  - We know the order, orientation, and spacing, but just not the bases. Fill with Ns instead

Assembly Summary



Assembly quality depends on

1. ***Coverage***: low coverage is mathematically hopeless
2. ***Repeat composition***: high repeat content is challenging
3. ***Read length***: longer reads help resolve repeats
4. ***Error rate***: errors reduce coverage, obscure true overlaps

- Assembly is a hierarchical, starting from individual reads, build high confidence contigs/unitigs, incorporate the mates to build scaffolds
  - Extensive error correction is the key to getting the best assembly possible from a given data set

- Watch out for collapsed repeats & other misassemblies
  - Globally/Locally reassemble data from scratch with better parameters & stitch the 2 assemblies together

Pop Quiz

Assemble these reads using a de Bruijn graph approach (k=3):

ATTA

GATT

TACA

TTAC

Pop Quiz

Assemble these reads using a de Bruijn graph approach (k=3):

```
ATTA: ATT -> TTA
GATT: GAT -> ATT
TACA: TAC -> ACA
TTAC: TTA -> TAC
```

Pop Quiz

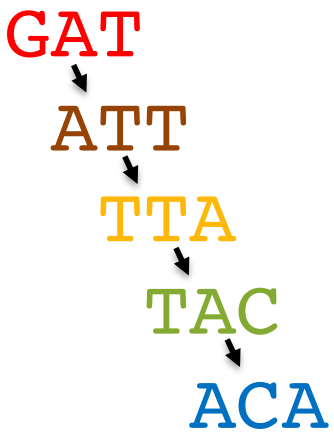
Assemble these reads using a de Bruijn graph approach (k=3):

ATTA: ATT -> TTA

GATT: GAT -> ATT

TACA: TAC -> ACA

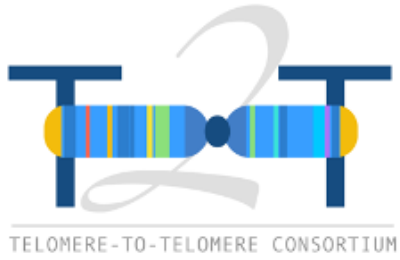
TTAC: TTA -> TAC



GATTACA

Outline

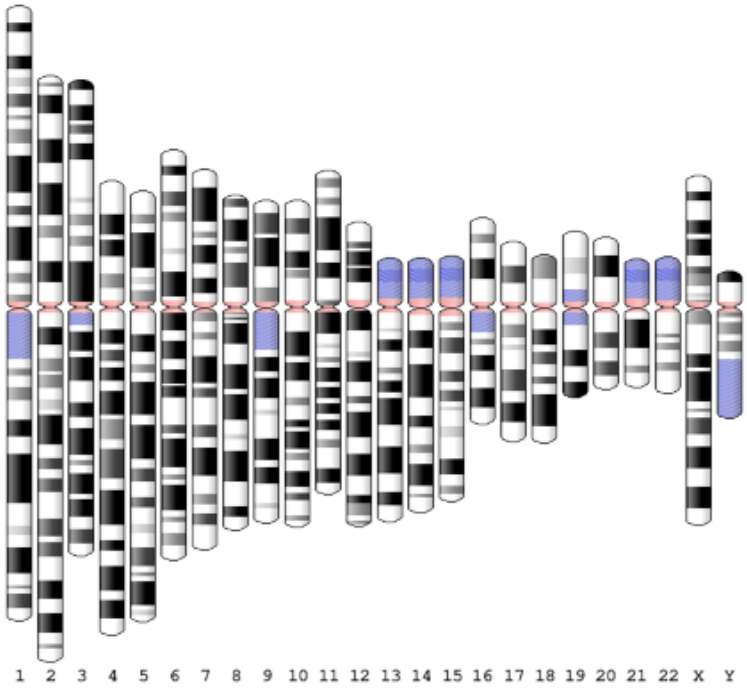
- Assembly by analogy
- T2T-CHM13 Assembly
  - How the T2T-CHM13 reference improves the analysis of genetic variation
- Beyond T2T-CHM13



The Human Genome Today

About 8% is missing or incorrect

- Over 100M "Ns" in the reference
- Centromeres and telomeres
- Segmentally duplicated genes
- Tandem gene arrays (e.g. rDNAs)
- Thousands of haplotype switches
- And an unknown number of errors…



Finishing the human genome

Why does it matter?

Variation in these regions is unexplored

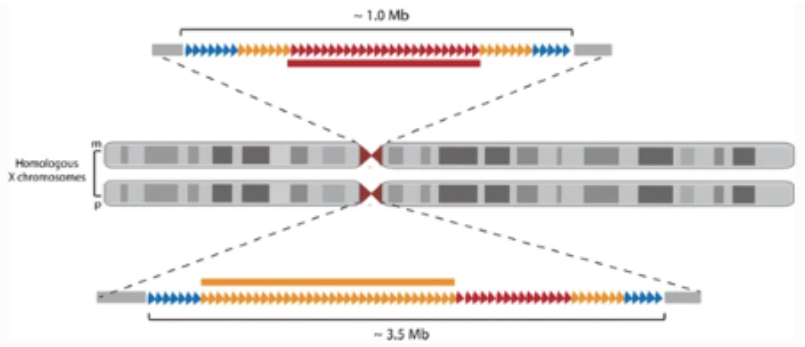
Functional studies need sequence

Reference gaps lead to artifacts

We don't know what we don't know…

Why has it taken so long?

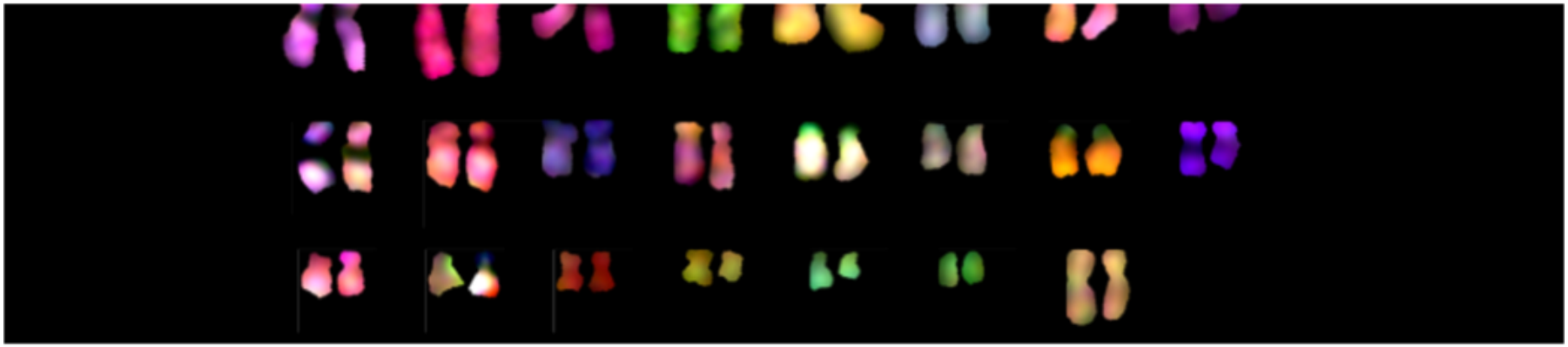
Repeats, repeats, repeats…



Miga
2015

Let's finish a human genome!


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



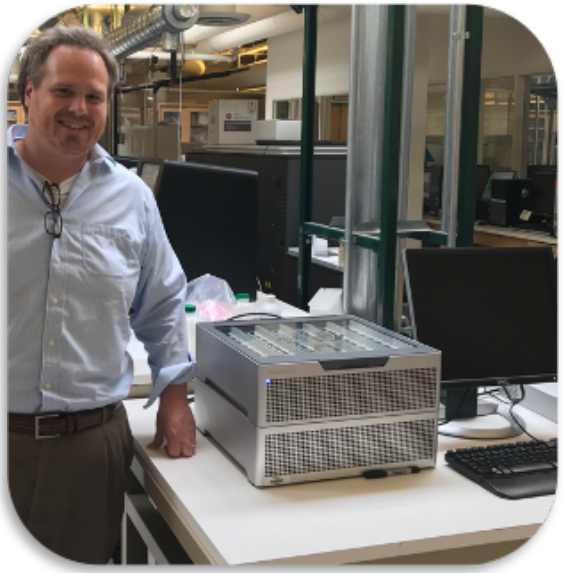
The Telomere-to-Telomere (T2T) consortium is an open, community-based effort to generate the first complete assembly of a human genome.

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

Single Molecule Long Read Sequencing

PacBio
Sequel II

Oxford Nanopore
PromethION

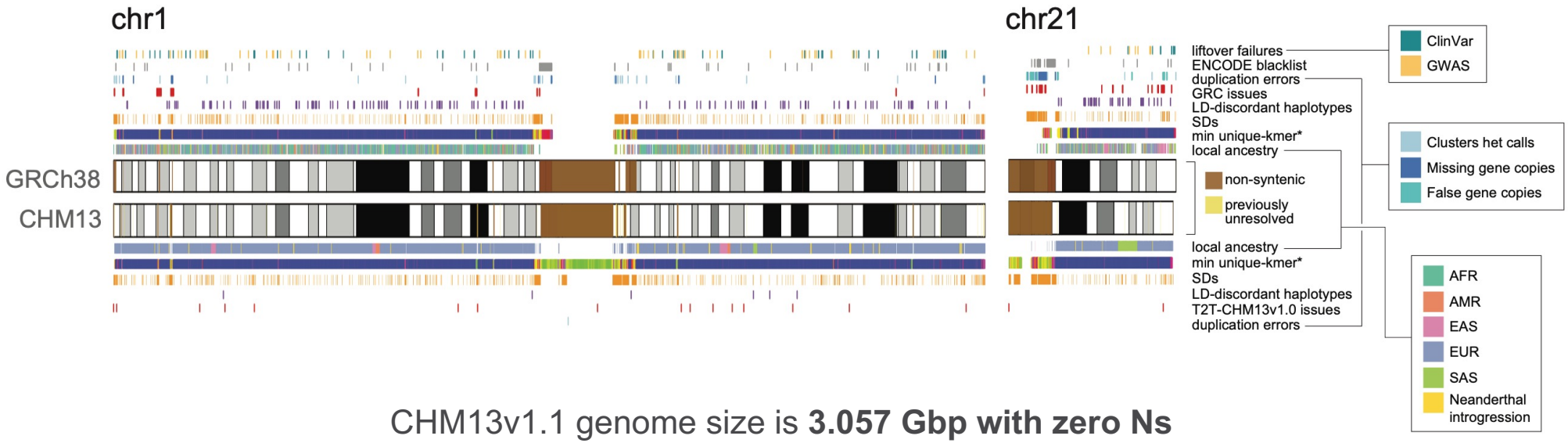


CHM13 HiFi assembly graph

The complete sequence of a human genome

Nurk *et al* (2021) *bioRxiv.* doi: https://doi.org/10.1101/2021.05.26.445798

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 (3–6%) of new sequence vs. GRCh38, fixes thousands of errors

The complete sequence of a human genome
Nurk *et al* (2021) *bioRxiv.* doi: https://doi.org/10.1101/2021.05.26.445798

T2T-Variants Analysis Summary

- Large-scale cloud analysis of short-read data
- Short-read alignment and variant calling
- Long-read alignment and variant calling
- Variants in previously unresolved regions
- Clinical implications of variants


Sergey Aganezov


Stephanie Yan


Daniela Soto


Melanie Kirsche

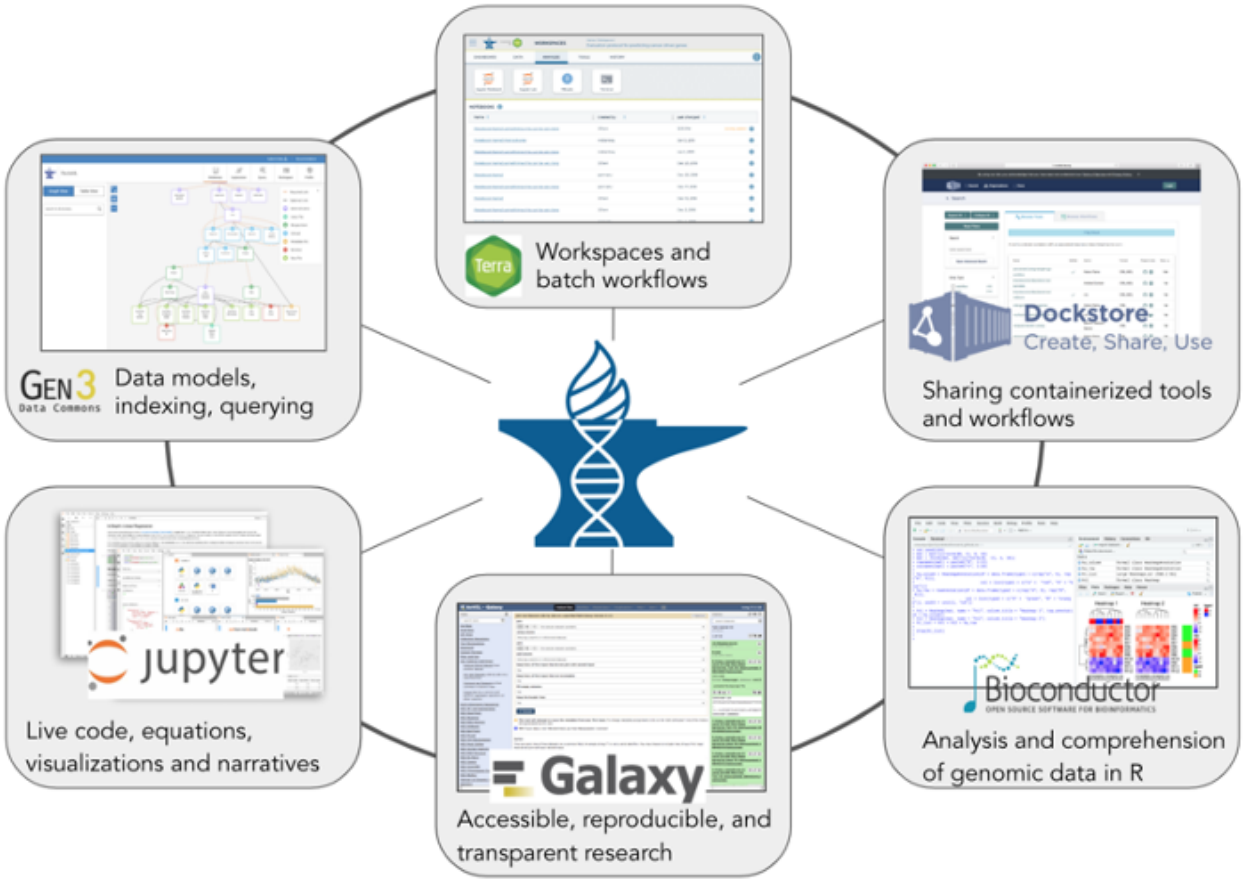

Samantha Zarate

A complete reference genome improves analysis of human genetic variation
Aganezov, S*, Yan, SM*, Soto, DC*, Kirsche, M*, Zarate, S*, *et al.* (2021) *bioRxiv*. doi: https://doi.org/10.1101/2021.07.12.452063

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.* (2021) *bioRxiv* doi: https://doi.org/10.1101/2021.04.22.436044

Samantha Zarate

GIGASCIENCE

TECHNICAL NOTE

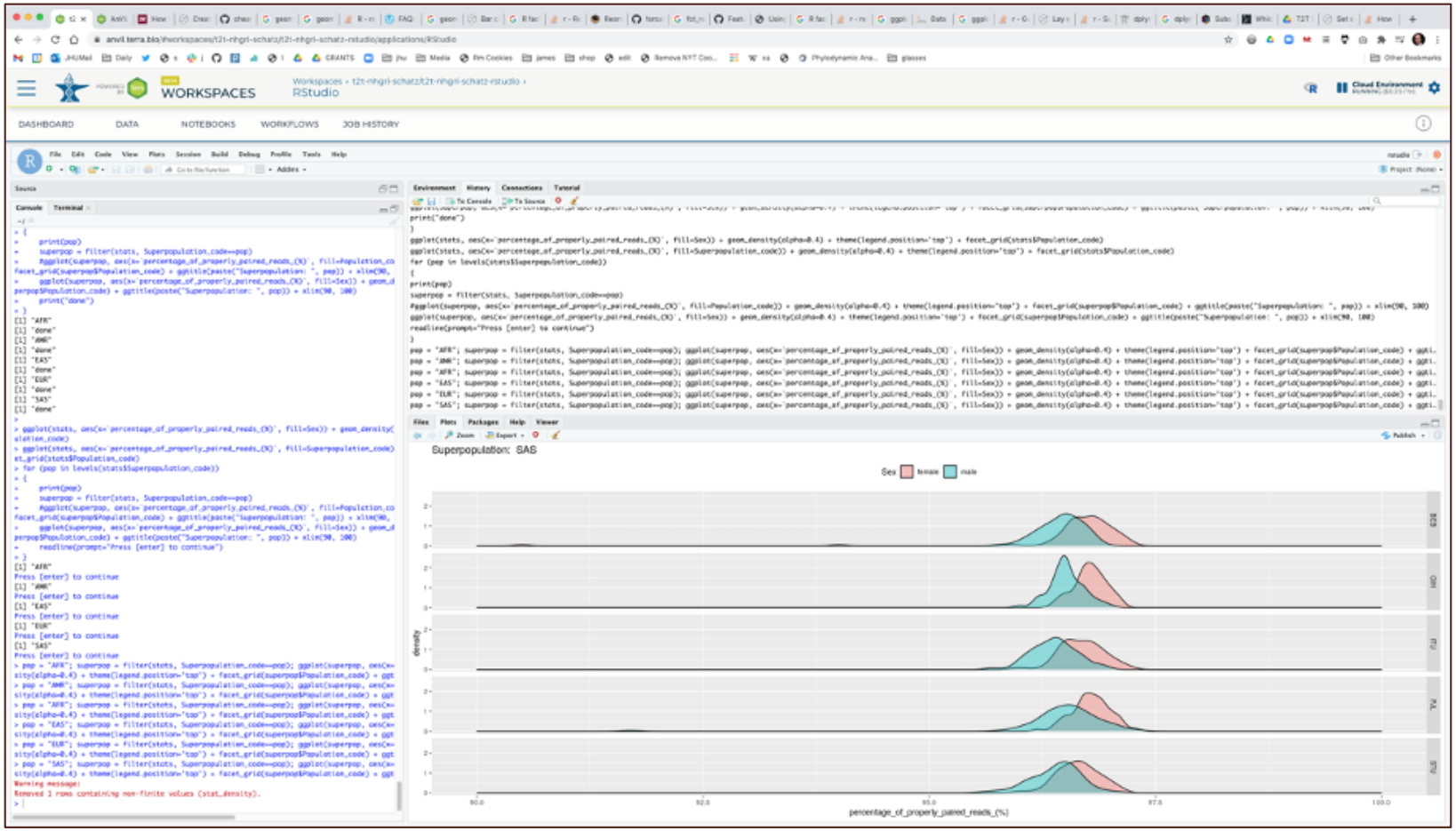
Parliament2: Accurate structural variant calling at scale

Samantha Zarate, Andrew Carroll, Medhat Mahmoud, Olga Krasheninina, Goo Jun, William J. Salerno, Michael C. Schatz, Eric Boerwinkle, Richard A. Gibbs and Fritz J. Sedlazeck

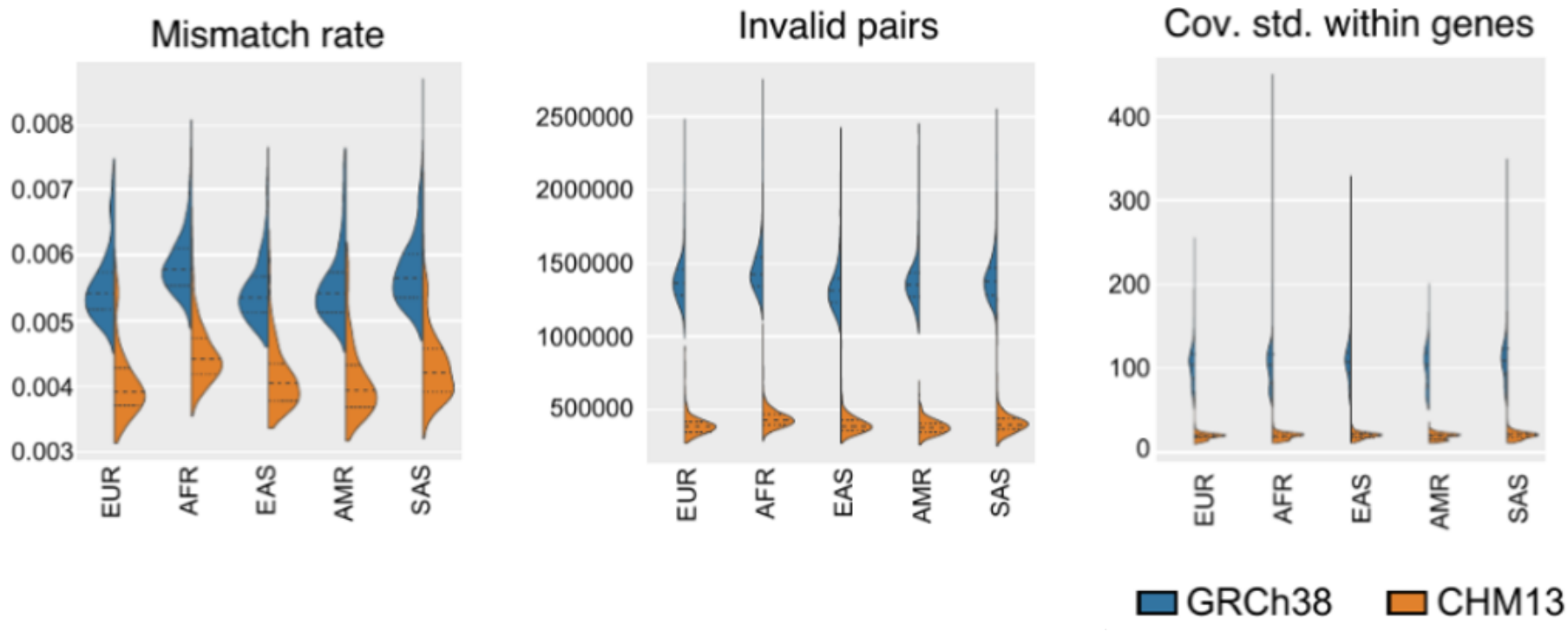
Core usage over 24 hours

Samantha Zarate

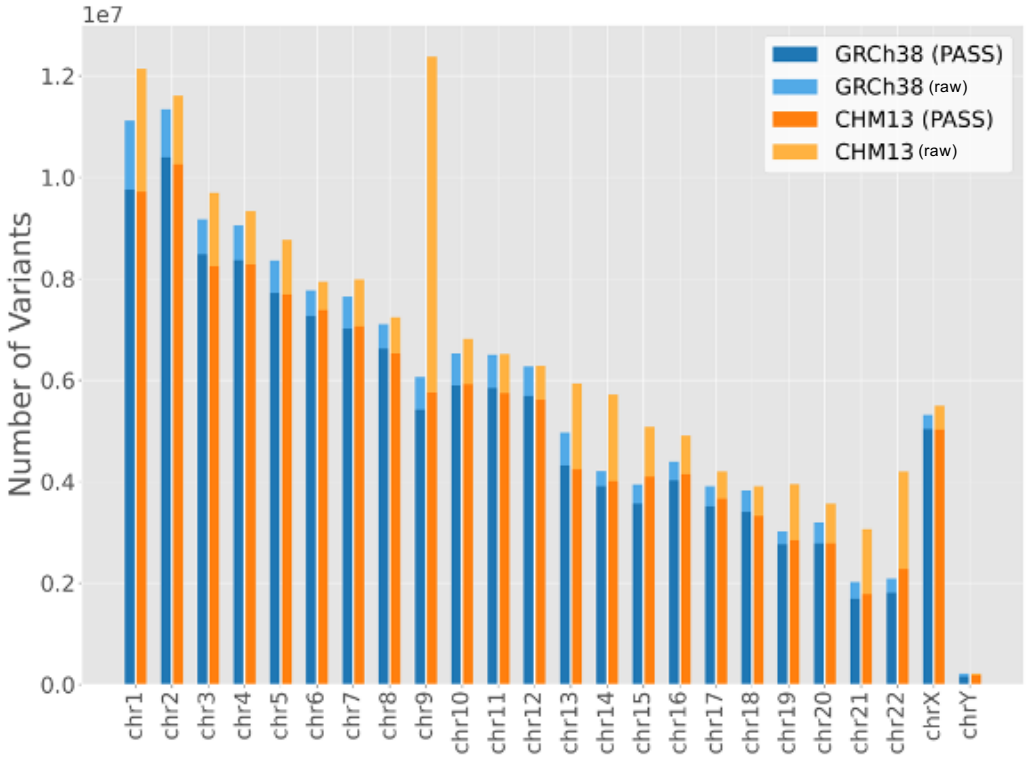
AnVIL: RStudio in the Cloud



1000G Mapping on T2T-CHM13

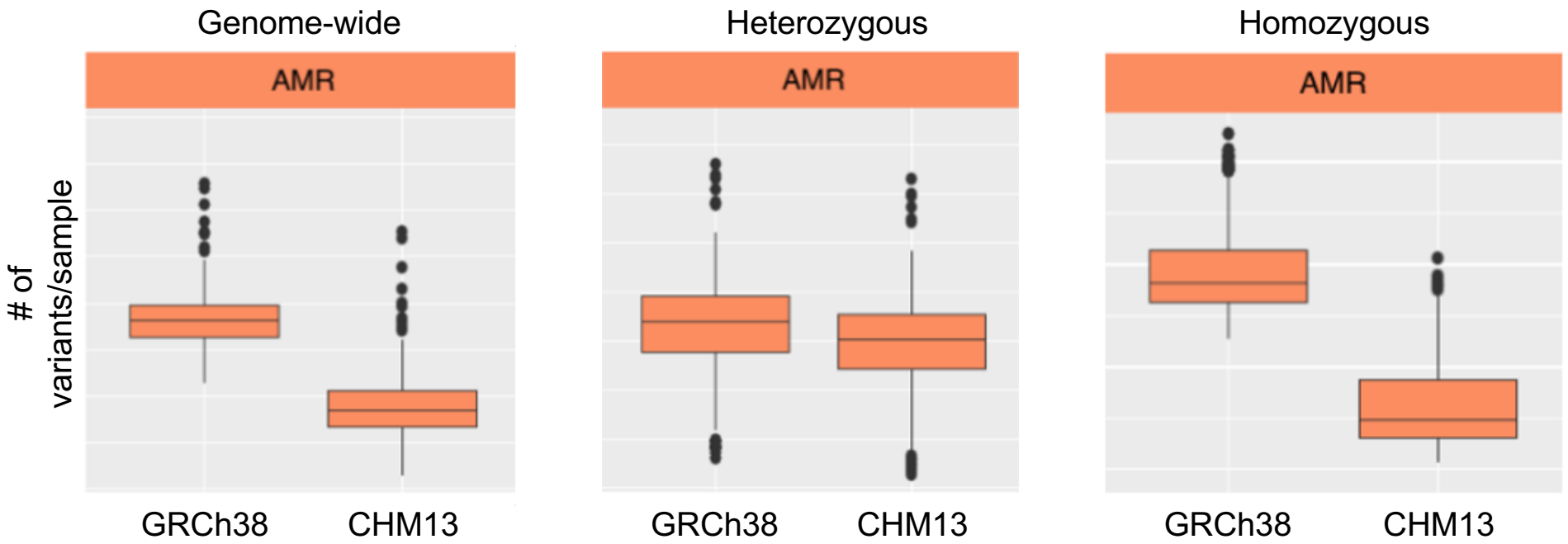


Across 1000G, Many More Variants Found Using CHM13

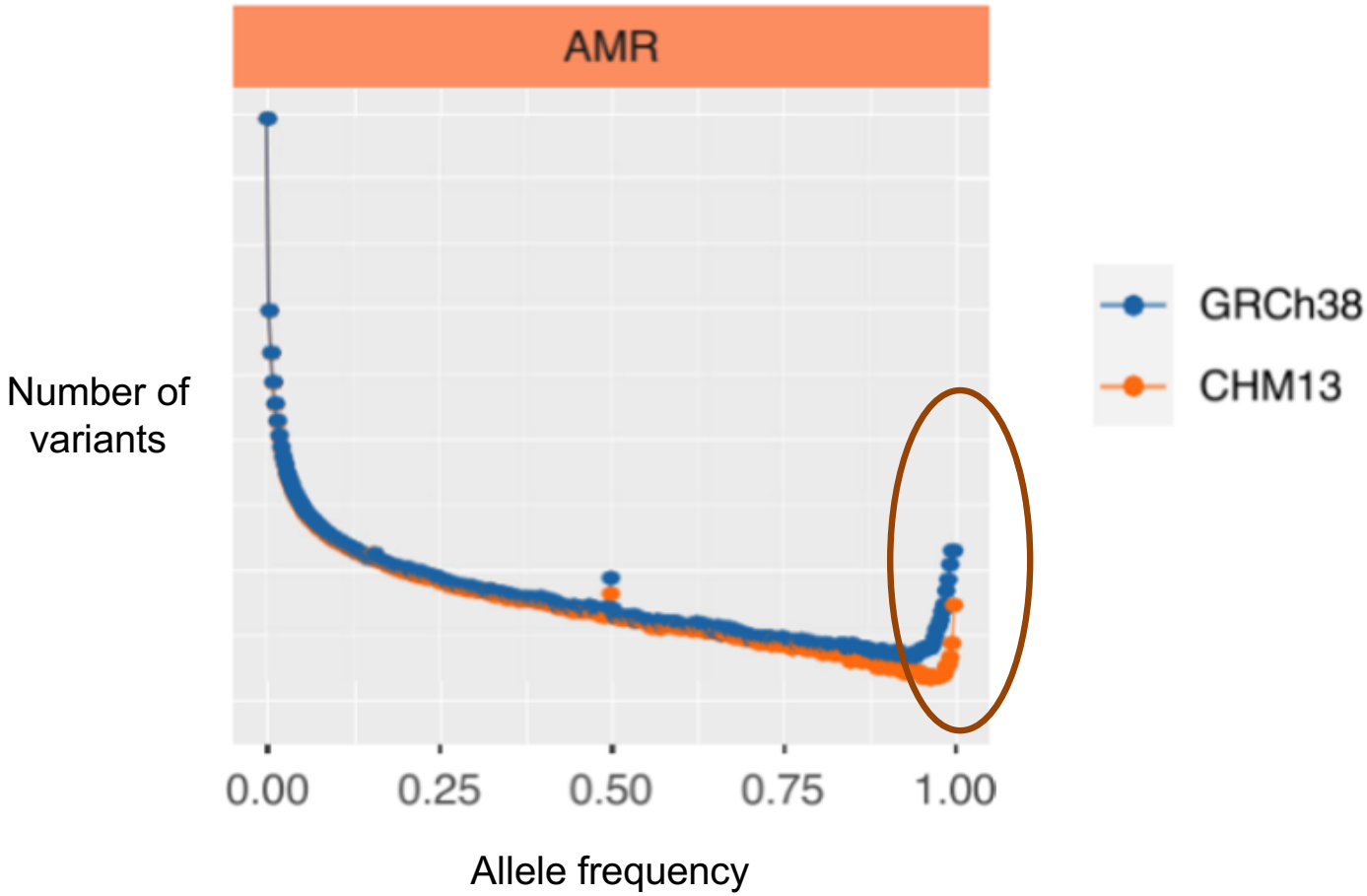


	GRCh38	CHM13
# PASS variants	125,484,020	126,591,489

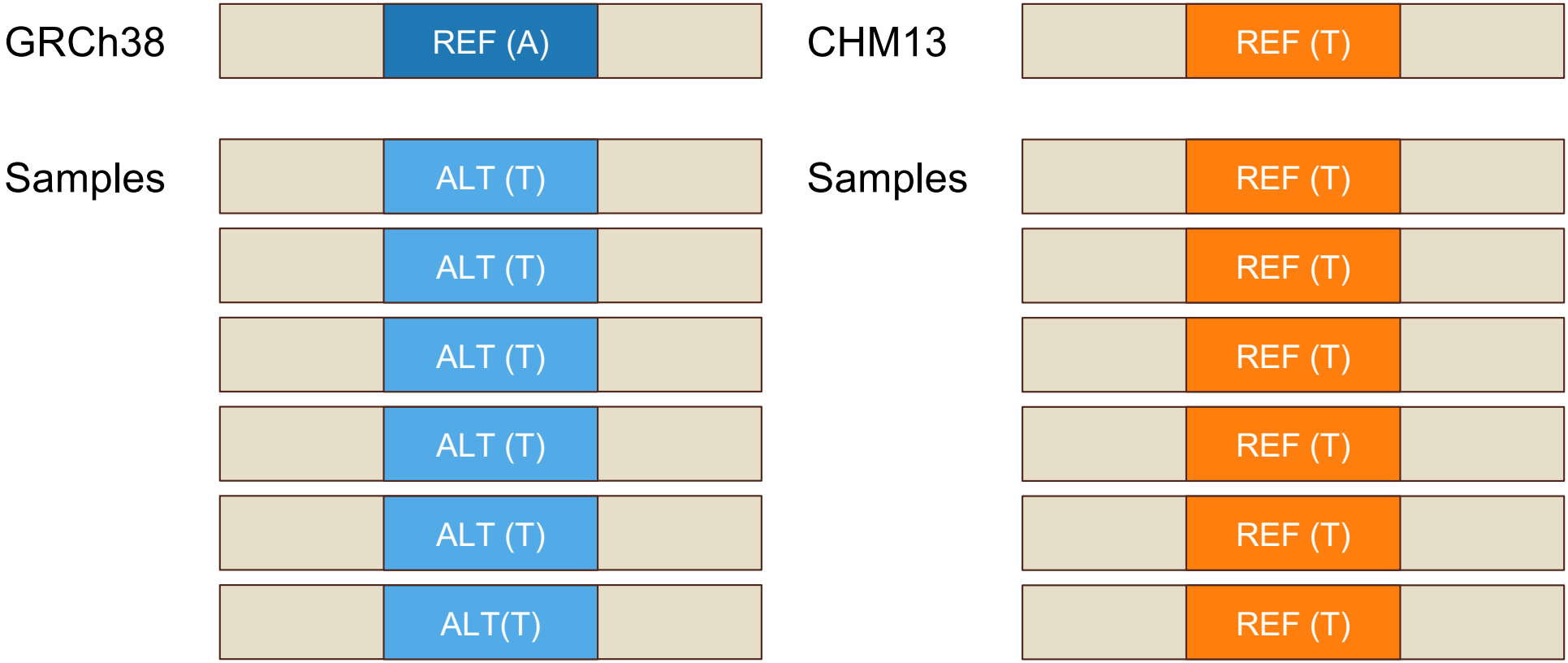
1000G Per-Sample Variant Counts on T2T-CHM13



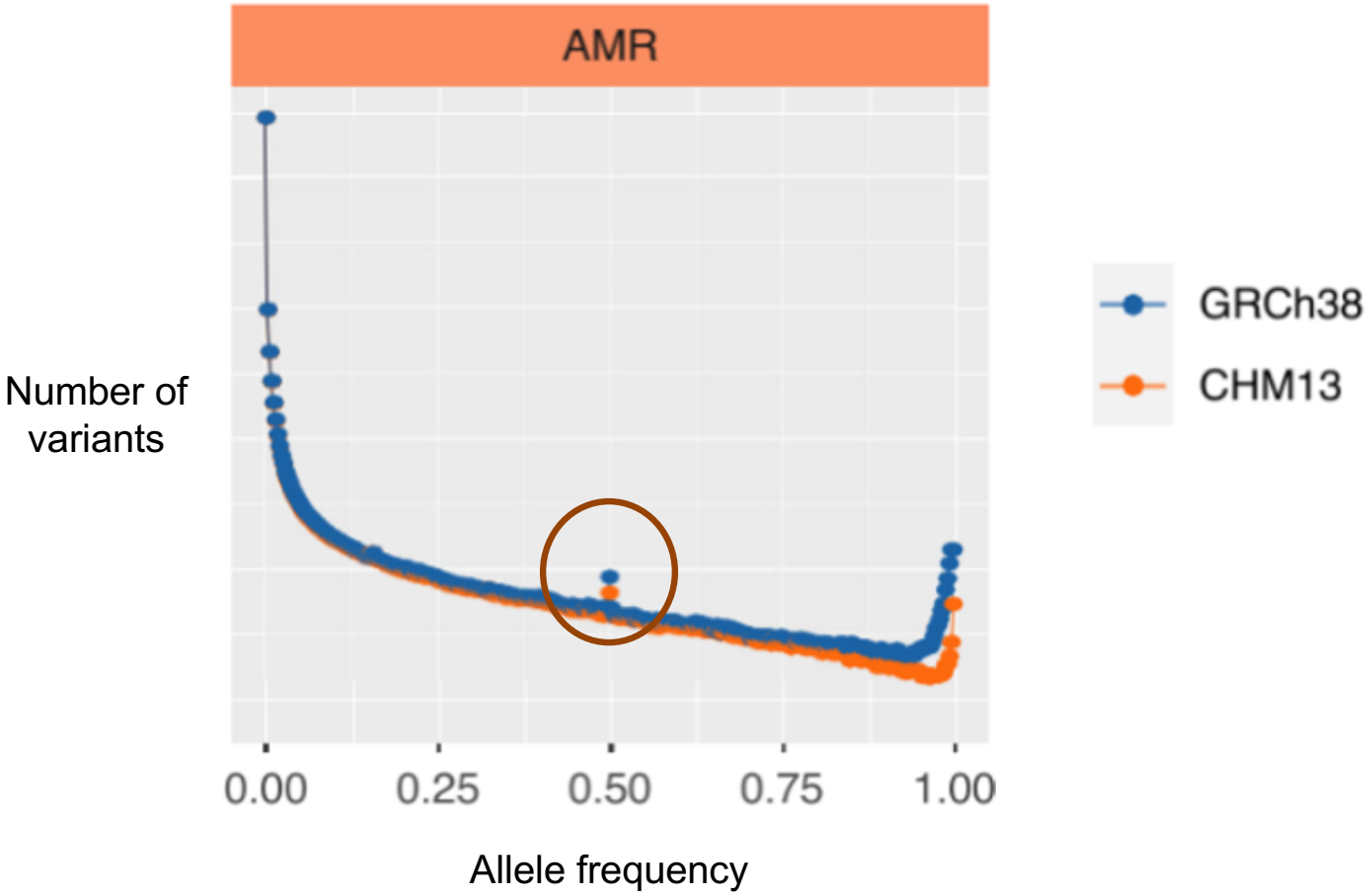
Explaining Decreased Per-Sample Count



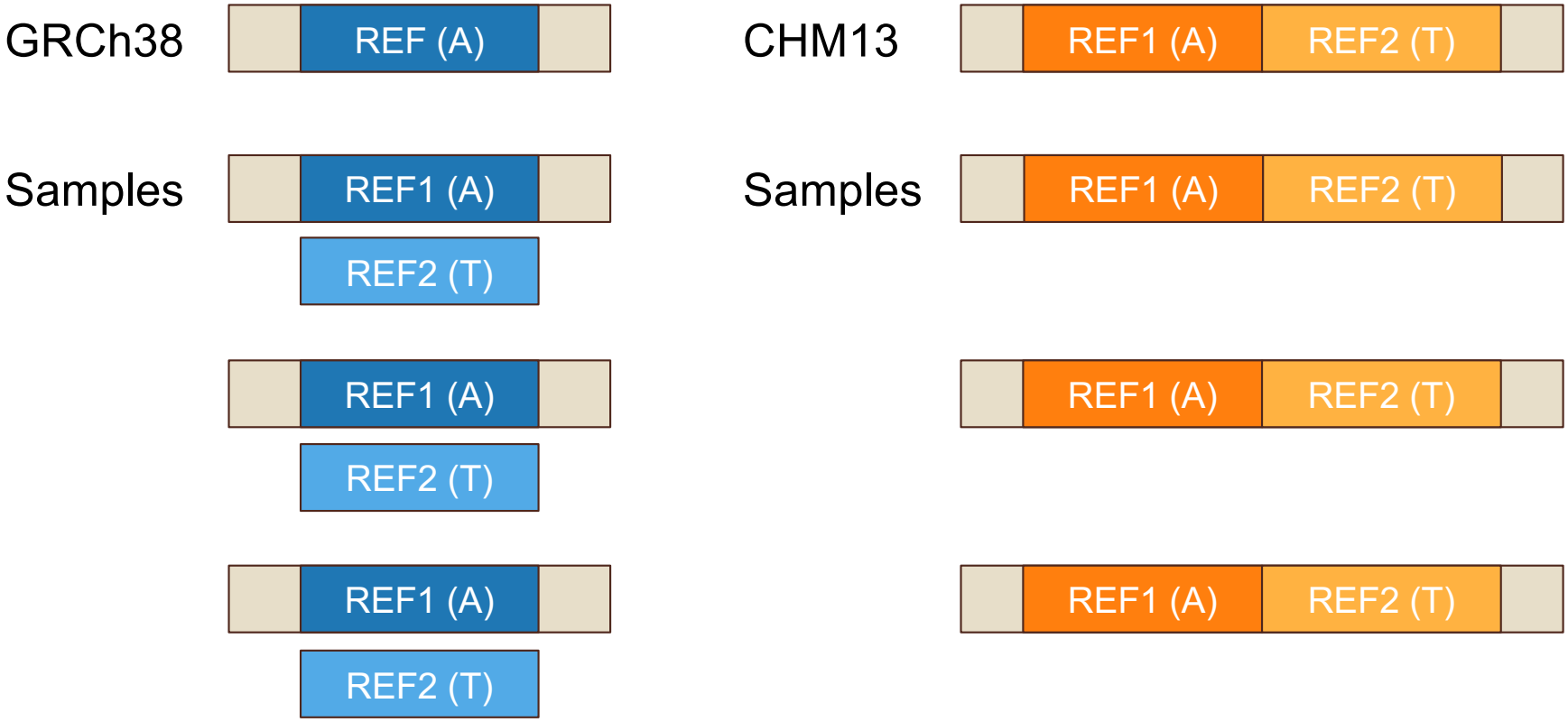
Allele Frequency = 1: Reference error / private variant



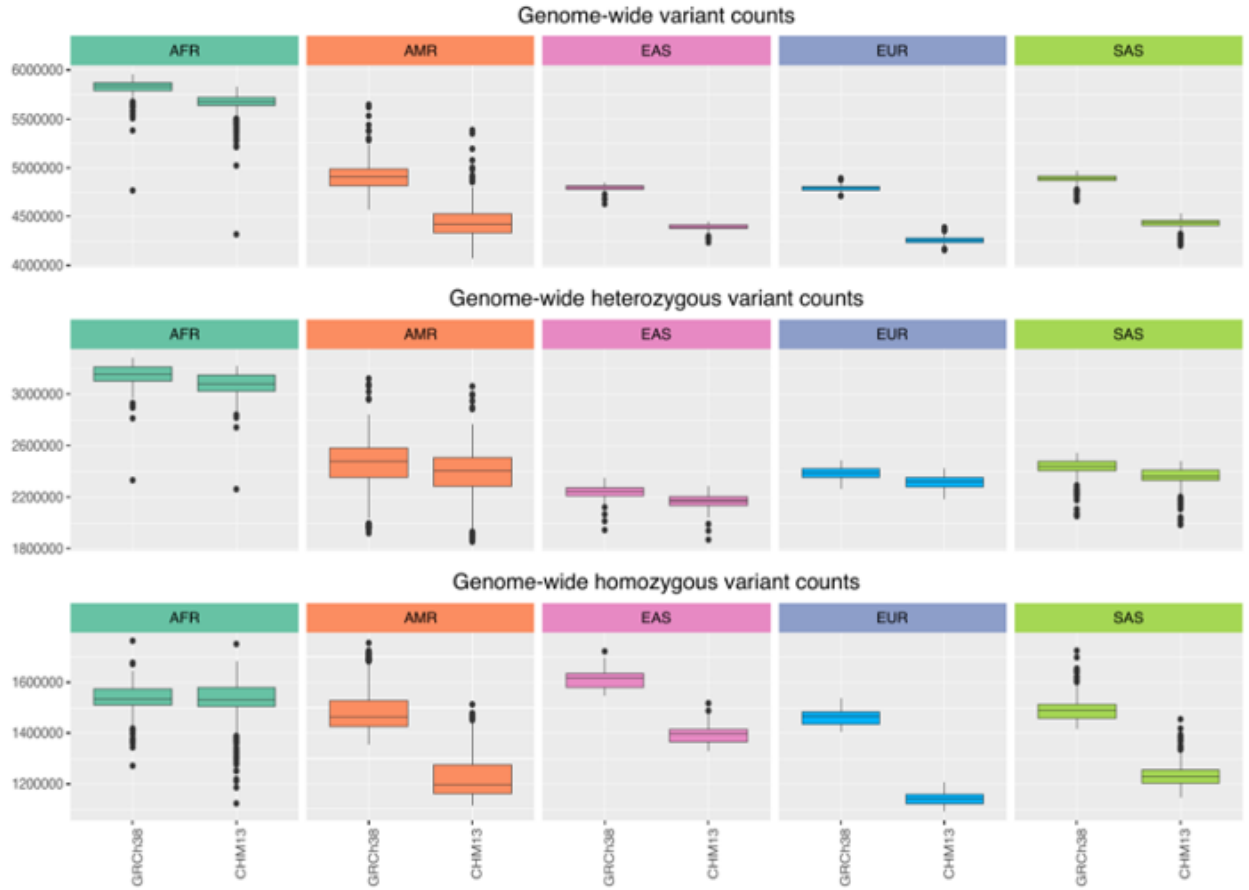
Explaining Decreased Per-Sample Count



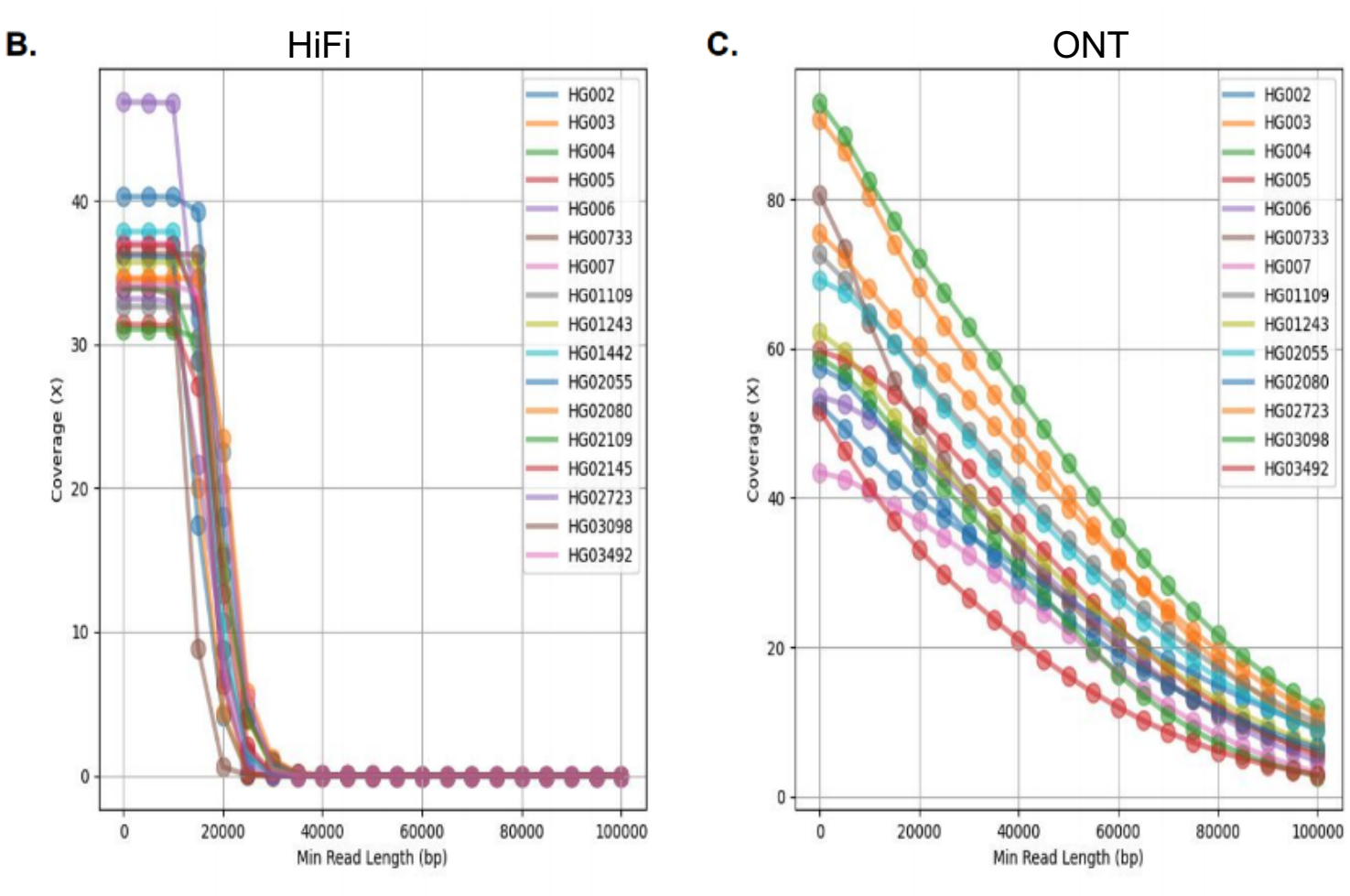
Allele Frequency ≈ 0.5: Collapsed duplication



1000G Per-Sample Variant Counts on T2T-CHM13



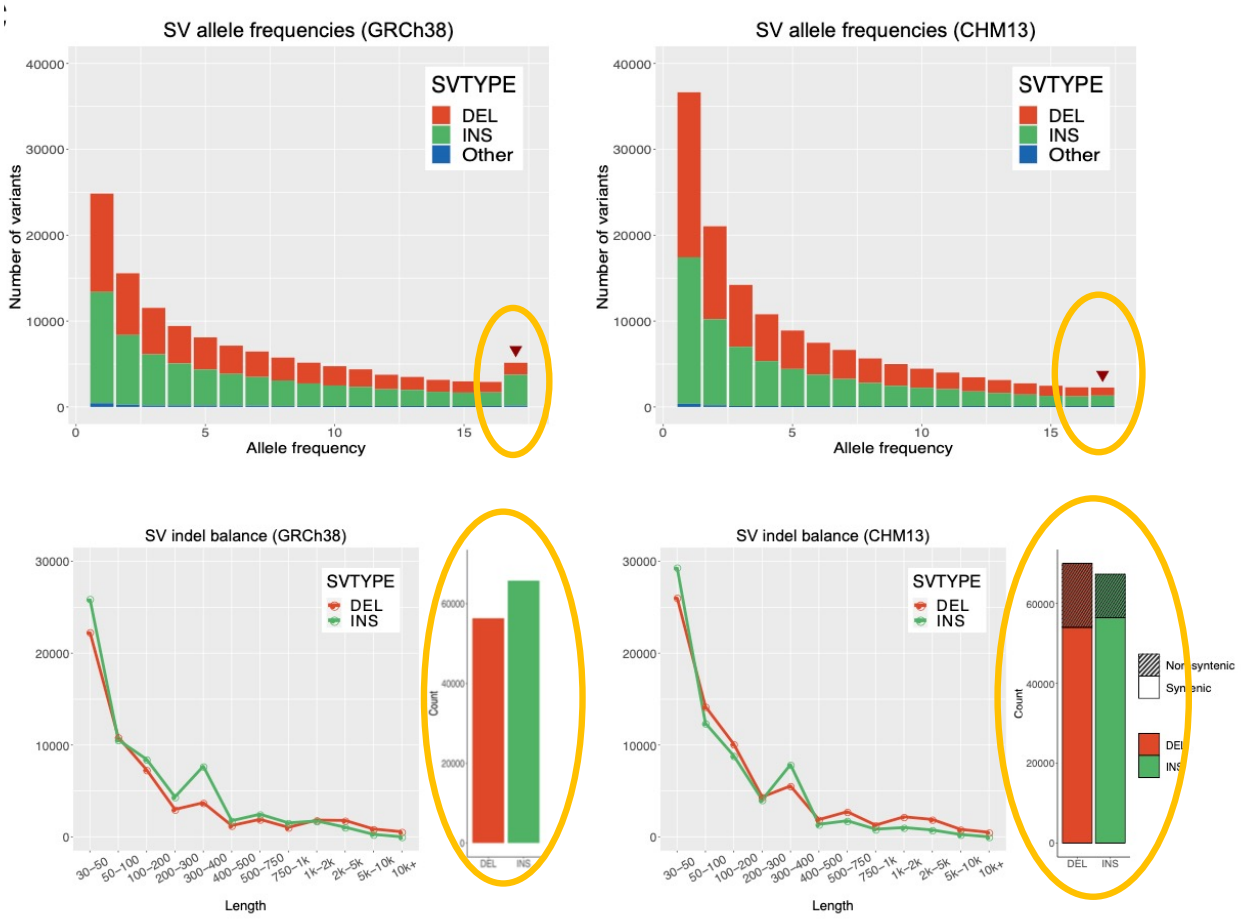
Long Read Analysis



Melanie Kirsche

Sergey Aganezov

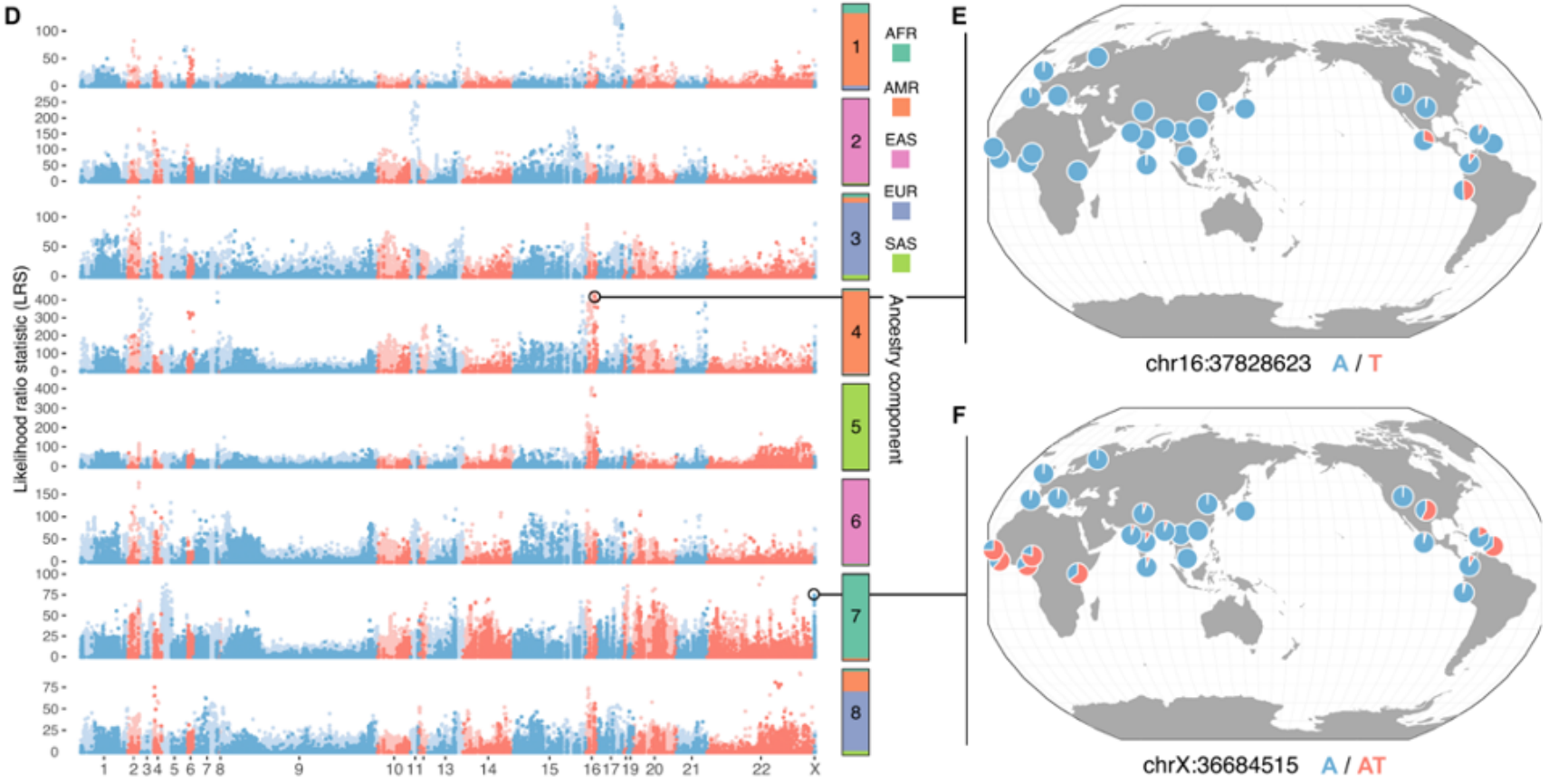
Long-Read Analysis with T2T-CHM13



Melanie Kirsche

Sergey Aganezov

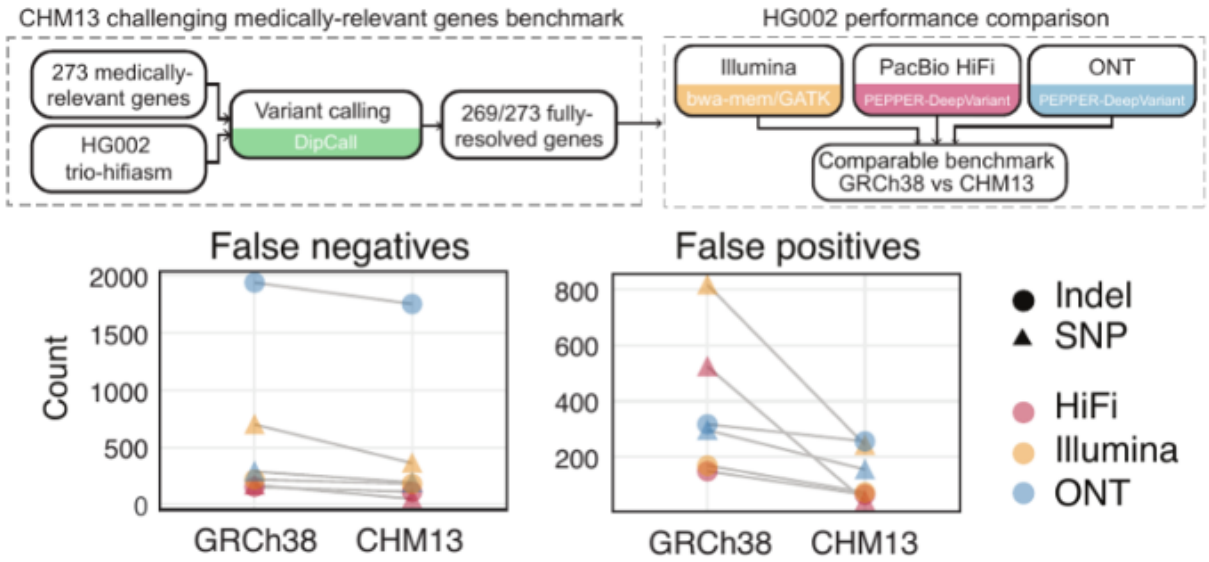
Variants in Newly Resolved Regions



Stephanie Yan

Rajiv McCoy

T2T-CHM13 Improves Clinical Genomics Variant Calling



Danny Miller

Daniela Soto

Megan Dennis

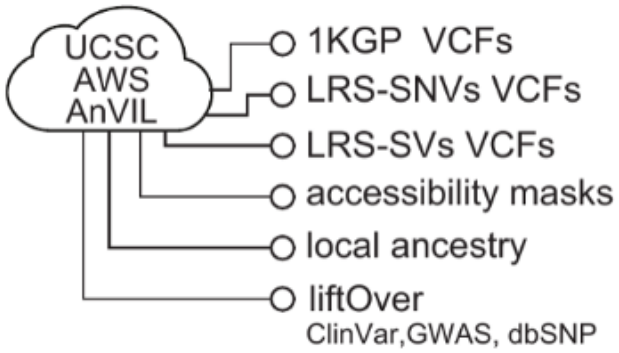
Justin Zook

Fritz Sedlazeck

For more information, see: (Wagner, J *et al.*, bioRxiv, 2021)

Summary of Results

- Large-scale cloud analysis of short-read data
- Short-read alignment and variant calling
- Long-read alignment and variant calling
- Variants in previously unresolved regions
- Clinical implications of variants



Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

The complete sequence of a human genome

Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J. Hoyt, Mark Diekhans, Glennis A. Logsdon, Michael Alonge, Stylianos E. Antonarakis, Matthew Borchers, Gerard G. Bouffard, Shelise Y. Brooks, Gina V. Caldas, Haoyu Cheng, Chen-Shan Chin, William Chow, Leonardo G. de Lima, Philip C. Dishuck, Richard Durbin, Tatiana Dvorkina, Ian T. Fiddes, Giulio Formenti, Robert S. Fulton, Arkarachai Fungtammasan, Erik Garrison, Patrick G.S. Grady, Tina A. Graves-Lindsay, Ira M. Hall, Nancy F. Hansen, Gabrielle A. Hartley, Marina Haukness, Kerstin Howe, Michael W. Hunkapiller, Chirag Jain, Miten Jain, Erich D. Jarvis, Peter Kerpedjiev, Melanie Kirsche, Mikhail Kolmogorov, Jonas Korlach, Milinn Kremitzki, Heng Li, Valerie V. Maduro, Tobias Marschall, Ann M. McCartney, Jennifer McDaniel, Danny E. Miller, James C. Mullikin, Eugene W. Myers, Nathan D. Olson, Benedict Paten, Paul Peluso, Pavel A. Pevzner, David Porubsky, Tamara Potapova, Evgeny I. Rogaev, Jeffrey A. Rosenfeld, Steven L. Salzberg, Valerie A. Schneider, Fritz J. Sedlazeck, Kishwar Shafin, Colin J. Shew, Alaina Shumate, Yumi Sims, Arian F.A. Smit, Daniela C. Soto, Ivan Sović, Jessica M. Storer, Aaron Streets, Beth A. Sullivan, Françoise Thibaud-Nissen, James Torrance, Justin Wagner, Brian P. Walenz, Aaron Wenger, Jonathan M. D. Wood, Chunlin Xiao, Stephanie M. Yan, Alice C. Young, Samantha Zarate, Urvashi Surti, Rajiv C. McCoy, Megan Y. Dennis, Ivan A. Alexandrov, Jennifer L. Gerton, Rachel J. O'Neill, Winston Timp, Justin M. Zook, Michael C. Schatz, Evan E. Eichler, Karen H. Miga, Adam M. Phillippy

doi: https://doi.org/10.1101/2021.05.26.445798

A complete reference genome improves analysis of human genetic variation

Sergey Aganezov, Stephanie M. Yan, Daniela C. Soto, Melanie Kirsche, Samantha Zarate, Pavel Avdeyev, Dylan J. Taylor, Kishwar Shafin, Alaina Shumate, Chunlin Xiao, Justin Wagner, Jennifer McDaniel, Nathan D. Olson, Michael E.G. Sauria, Mitchell R. Vollger, Arang Rhie, Melissa Meredith, Skylar Martin, Joyce Lee, Sergey Koren, Jeffrey A. Rosenfeld, Benedict Paten, Ryan Layer, Chen-Shan Chin, Fritz J. Sedlazeck, Nancy F. Hansen, Danny E. Miller, Adam M. Phillippy, Karen H. Miga, Rajiv C. McCoy, Megan Y. Dennis, Justin M. Zook, Michael C. Schatz

doi: https://doi.org/10.1101/2021.07.12.452063

Complete genomic and epigenetic maps of human centromeres

Nicolas Altemose, Glennis A. Logsdon, Andrey V. Bzikadze, Pragya Sidhwani, Sasha A. Langley, Gina V. Caldas, Savannah J. Hoyt, Lev Uralsky, Fedor D. Ryabov, Colin J. Shew, Michael E.G. Sauria, Matthew Borchers, Ariel Gershman, Alla Mikheenko, Valery A. Shepelev, Tatiana Dvorkina, Olga Kunyavskaya, Mitchell R. Vollger, Arang Rhie, Ann M. McCartney, Mobin Asri, Ryan Lorig-Roach, Kishwar Shafin, Sergey Aganezov, Daniel Olson, Leonardo Gomes de Lima, Tamara Potapova, Gabrielle A. Hartley, Marina Haukness, Peter Kerpedjiev, Fedor Gusev, Kristof Tigyi, Shelise Brooks, Alice Young, Sergey Nurk, Sergey Koren, Sofie R. Salama, Benedict Paten, Evgeny I. Rogaev, Aaron Streets, Gary H. Karpen, Abby F. Dernburg, Beth A. Sullivan, Aaron F. Straight, Travis J. Wheeler, Jennifer L. Gerton, Evan E. Eichler, Adam M. Phillippy, Winston Timp, Megan Y. Dennis, Rachel J. O'Neill, Justin M. Zook, Michael C. Schatz, Pavel A. Pevzner, Mark Diekhans, Charles H. Langley, Ivan A. Alexandrov, Karen H. Miga

doi: https://doi.org/10.1101/2021.07.12.452052

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

Savannah J. Hoyt, Jessica M. Storer, Gabrielle A. Hartley, Patrick G. S. Grady, Ariel Gershman, Leonardo G. de Lima, Charles Limouse, Reza Halabian, Luke Wojenski, Matias Rodriguez, Nicolas Altemose, Leighton J. Core, Jennifer L. Gerton, Wojciech Makalowski, Daniel Olson, Jeb Rosen, Arian F.A. Smit, Aaron F. Straight, Mitchell R. Vollger, Travis J. Wheeler, Michael C. Schatz, Evan E. Eichler, Adam M. Phillippy, Winston Timp, Karen H. Miga, Rachel J. O'Neill

doi: https://doi.org/10.1101/2021.07.12.451456

Epigenetic Patterns in a Complete Human Genome

Ariel Gershman, Michael E.G. Sauria, Paul W. Hook, Savannah J. Hoyt, Roham Razaghi, Sergey Koren, Nicolas Altemose, Gina V. Caldas, Mitchell R. Vollger, Glennis A. Logsdon, Arang Rhie, Evan E. Eichler, Michael C. Schatz, Rachel J. O'Neill, Adam M. Phillippy, Karen H. Miga, Winston Timp

doi: https://doi.org/10.1101/2021.05.26.443420

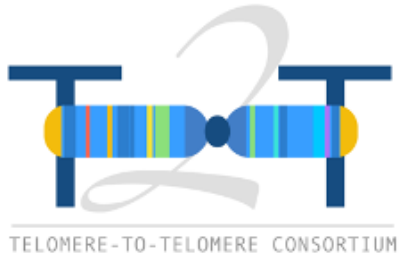
Segmental duplications and their variation in a complete human genome

Mitchell R. Vollger, Xavi Guitart, Philip C. Dishuck, Ludovica Mercuri, William T. Harvey, Ariel Gershman, Mark Diekhans, Arvis Sulovari, Katherine M. Munson, Alexandra M. Lewis, Kendra Hoekzema, David Porubsky, Ruiyang Li, Sergey Nurk, Sergey Koren, Karen H. Miga, Adam M. Phillippy, Winston Timp, Mario Ventura, Evan E. Eichler

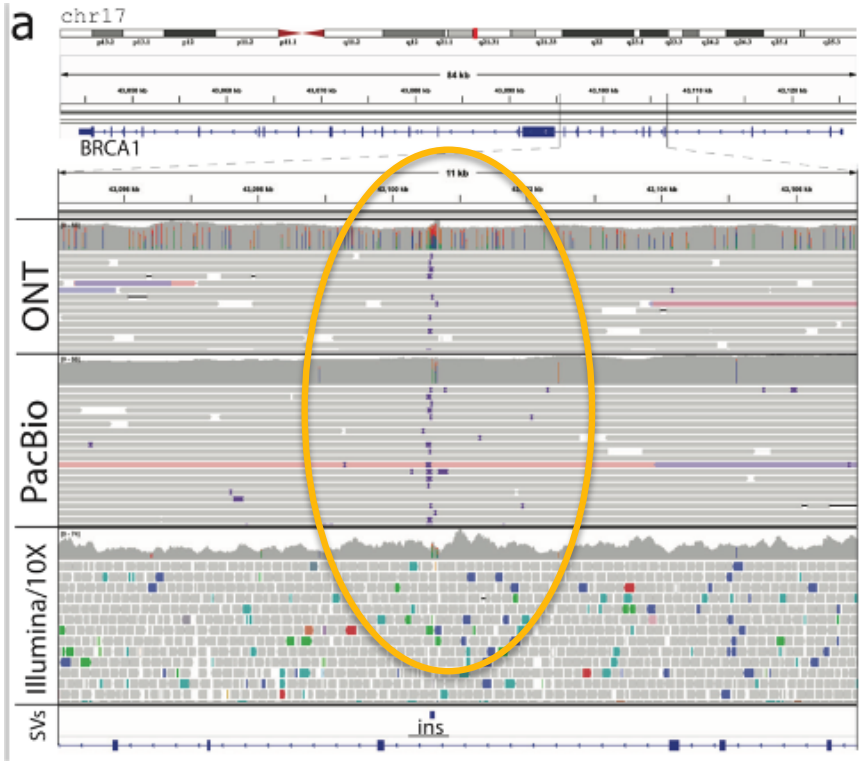
doi: https://doi.org/10.1101/2021.05.26.445678

Outline

- Assembly by analogy
- T2T-CHM13 Assembly
  - How the T2T-CHM13 reference improves the analysis of genetic variation
- Beyond T2T-CHM13



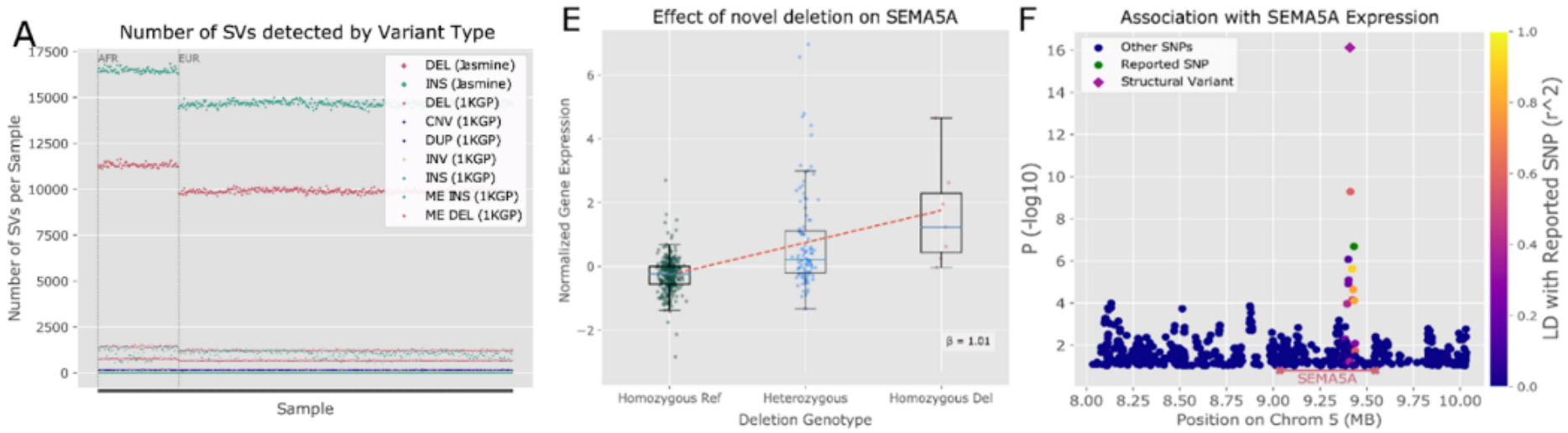
Hidden Variants in Breast Cancer Genes



62bp repeat expansion in BRCA1 detected in normal tissue that is undetectable using a cancer panel or short read sequencing

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

Novel association of 37bp deletion with SEMA5A expression



- Genotyped ~200k SVs from our long-read cohort in 444 samples from the 1000 Genomes Project using ParaGRAPH (Chen et al, 2019) which also had gene expression data available
- The deletion is in strong LD with the reported SNP-eQTL, but has a much higher effect size on SEMA5A

Kirsche, et al (2021) bioRxiv doi: https://doi.org/10.1101/2021.05.27.445886

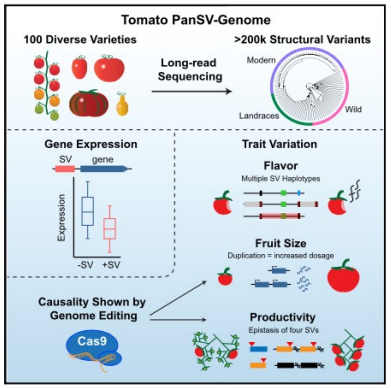
T2T Genomes and Epigenomes Across the Tree of Life

Cell

Article

Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato

Graphical Abstract



Authors

Michael Alonge, Xingang Wang, Matthias Benoit, ..., Esther van der Knaap, Michael C. Schatz, Zachary B. Lippman

Correspondence

mschatz@cs.jhu.edu (M.C.S.), lippman@cshl.edu (Z.B.L.)

In Brief

Comprehensive structural variant identification in tomato genomes allows insight into the evolution and domestication of tomato and serves as a resource for phenotype-directed breeding.

Highlights

- Long-read sequencing of 100 tomato genomes uncovered 238,490 structural variants
- Transposons underlie many SVs, and SV hotspots revealed large introgressions
- SVs associated with genes are predictive of population-scale changes in expression
- New genome assemblies resolved complex breeding QTLs caused by SVs

Alonge et al., 2020, Cell *182*, 145–161
July 9, 2020 © 2020 Elsevier Inc.
https://doi.org/10.1016/j.cell.2020.05.021

CellPress

RESEARCH ARTICLE SUMMARY

PLANT SCIENCE

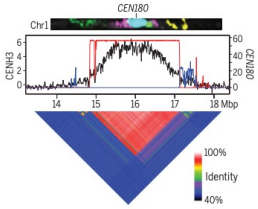
The genetic and epigenetic landscape of the *Arabidopsis* centromeres

Matthew Naish†, Michael Alonge†, Piotr Wlodzimierz†, Andrew J. Tock, Bradley W. Abramson, Anna Schmücker, Terezie Mandáková, Bhagyshree Jamge, Christophe Lambing, Pallas Kuo, Natasha Yelina, Nolan Hartwick, Kelly Colt, Lisa M. Smith, Jurriaan Ton, Tetsuji Kakutani, Robert A. Martienssen, Korbinian Schneeberger, Martin A. Lysak, Frédéric Berger, Alexandros Bousios, Todd P. Michael, Michael C. Schatz*, Ian R. Henderson*

INTRODUCTION: The centromeres of eukaryotic chromosomes assemble the multiprotein kinetochore complex and thereby position attachment to the spindle microtubules, allowing chromosome segregation during cell division. The key function of the centromere is to load nucleosomes containing the CENTROMERE SPECIFIC HISTONE H3 (CENH3) histone variant [also known as centromere protein A (CENPA)], which directs kinetochore formation. Despite their conserved function during chromosome segregation, centromeres show radically diverse organization between species at the sequence level, ranging from single nucleosomes to megabase-scale satellite repeat arrays, which is termed the centromere paradox. Centromeric satellite repeats are variable in sequence composition and length when compared between species and show a high capacity for evolutionary change, both at the levels of primary sequence and array position along the chromosome. However, the genetic and epigenetic features that contribute to centromere function and evolution are incompletely understood, in part because of the challenges of centromere sequence assembly and functional genomics of highly repetitive sequences. New long-read DNA sequencing technologies can now resolve these complex repeat arrays, revealing insights into centromere architecture and chromatin organization.

RATIONALE: *Arabidopsis thaliana* is a model plant species; its genome was first sequenced in 2000, yet the centromeres, telomeres, and ribosomal DNA repeats have remained unassembled, owing to their high repetition and similarity. Genomic repeats are difficult to assemble from fragmented sequencing reads, with longer, high-identity repeats being the most challenging to correctly assemble. As sequencing reads have become longer and more accurate, eukaryotic de novo genome assemblies have captured an increasingly complete picture of the repetitive component of the genome, including the centromeres. For example, Oxford Nanopore Technologies (ONT) reads have become longer and more accurate, now reaching >100 kilo–base pairs (kbp) in length with 95 to 99% modal accuracy. PacBio high-fidelity (HiFi) reads, although shorter (~15 kbp), are >99% accurate. Using ONT and HiFi reads, it is possible to bridge across interspersed unique marker sequences and accurately assemble centromere sequences. In this study, we used long-read DNA sequencing to generate a genome assembly of the *A. thaliana* accession Columbia (Col-CEN) that resolves all five centromeres. We use the Col-CEN assembly to derive insights into the chromatin and recombination landscapes within the *Arabidopsis* centromeres and how these regions evolve.

RESULTS: The Col-CEN assembly reveals that the *Arabidopsis* centromeres consist of megabase-scale tandemly repeated satellite arrays, which support high CENH3 (the centromere-specific histone variant that recruits kinetochores) occupancy and are densely DNA methylated. We show patterns of higher-order repetition within centromeres and that many satellite variants are private to each chromosome, which has implications for the recombination pathways acting in the centromeres. CENH3 preferentially occupies the satellites with the least amount of divergence and that show higher-order repetition. The *Arabidopsis* centromeres are mainly composed of satellite repeats that are ~178 bp in length, termed the *CEN180* satellites. *Arabidopsis* centromeres have also been invaded by *ATHILA* long terminal repeat–class retrotransposons, which disrupt the genetic and epigenetic organization of the centromeres. Using chromatin immunoprecipitation sequencing (ChIP-seq) and immunofluorescence, we demonstrate that the centromeres show a hybrid chromatin state that is distinct from euchromatin and heterochromatin. We show that crossover recombination is suppressed within the centromeres, yet low levels of meiotic double-strand breaks occur, which are regulated by DNA methylation. Together, our Col-CEN assembly reveals the genetic and epigenetic landscapes within the *Arabidopsis* centromeres.



Assembly of the *Arabidopsis* centromeres. The structure of *Arabidopsis* centromere 1 is shown by fluorescence in situ hybridization (top) [upper-arm bacterial artificial chromosomes (BACs) (green), *ATHILA* (purple), *CEN180* (blue), the telomeric repeat (green), and bottom-arm BACs (yellow)] and a long-read genome assembly (middle). The density of centromeric histone CENH3 binding measured by ChIP-seq is shown (black), alongside the frequency of *CEN180* centromere satellite repeats. Red and blue represent forward- and reverse-strand satellites, respectively. The heatmap (bottom) shows patterns of sequence identity across the centromere between nonoverlapping 5-kbp windows. Chr, chromosome 1.

CONCLUSION: Our Col-CEN assembly and functional genomics analysis have implications for understanding centromere sequence evolution in eukaryotes. We propose that a recombination-based homogenization process, occurring between allelic or nonallelic locations on the same chromosome, maintains the *CEN180* library close to the consensus optimal for CENH3 recruitment. The advantage conferred to *ATHILA* retrotransposons by integration within the centromeres is presently unclear. They may be engaged in centromere drive, supporting the hypothesis that centromere satellite homogenization acts as a mechanism to purge driving elements. Each *Arabidopsis* centromere appears to represent different stages in cycles of satellite homogenization and *ATHILA*-driven diversification. These opposing forces provide both a capacity for homeostasis and a capacity for change during centromere evolution. In the future, assembly of centromeres from multiple *Arabidopsis* accessions and closely related species may further clarify how centromeres form and the evolutionary dynamics of *CEN180* and *ATHILA* repeats. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: mschatz@cs.jhu.edu (M.C.S.); irh25@cam.ac.uk (I.R.H.)
†These authors contributed equally to this work.
Cite this article as M. Naish *et al.*, *Science* 374, eabi7489 (2021). DOI: 10.1126/science.abi7489

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abi7489

Our genomics future

We are on the cusp of a biotechnology revolution: "tricorders for all"

- We all carry professional digital cameras in our pocket with quality unimaginable 20 years ago
- Endless applications to learn more about ourselves, improve our health, and study the world around us

Genomics 2041

- Interpreting genomics results requires very deep knowledge that is currently out of reach for many
- We must immediately take action to ensure genomics equality for all
- In 1000+ years, we will still be learning new meaning in the genome

Acknowledgements

Schatz Lab

Enis Afghan
Mike Alonge
Arun Das
Dannon Baker
David Clements
Sergey Golitsynskiy
Sam Guerler
Katie Jenike
Melanie Kirsche
Sam Kovaka
Natalie Kucher
Alex Mahmoud
Stephen Mosher
Bohan Ni
Alex Ostrovsky
Cokie Parker
Gautam Prabhu
Srividya Ramakrishnan
Michael Sauria
Vikram Shivakumar
Margaret Starostik
Jenn Vessio
Samantha Zarate

AnVIL, Galaxy, & T2T Teams

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

JHU

Battle Lab
Langmead Lab
Leek Lab
McCoy Lab
Salzberg Lab
Taylor Lab
Timp Lab

CSHL

McCombie Lab
Lippman Lab

Baylor

Sedlazeck Lab

U. Cambridge

Henderson Lab

National Human Genome Research Institute

NIH NATIONAL CANCER INSTITUTE

The Mark Foundation® for Cancer Research

NSF

Bloomberg Professors

Thank you!

@mike_schatz
http://schatz-lab.org