# A complete reference genome improves analysis of human genetic variation

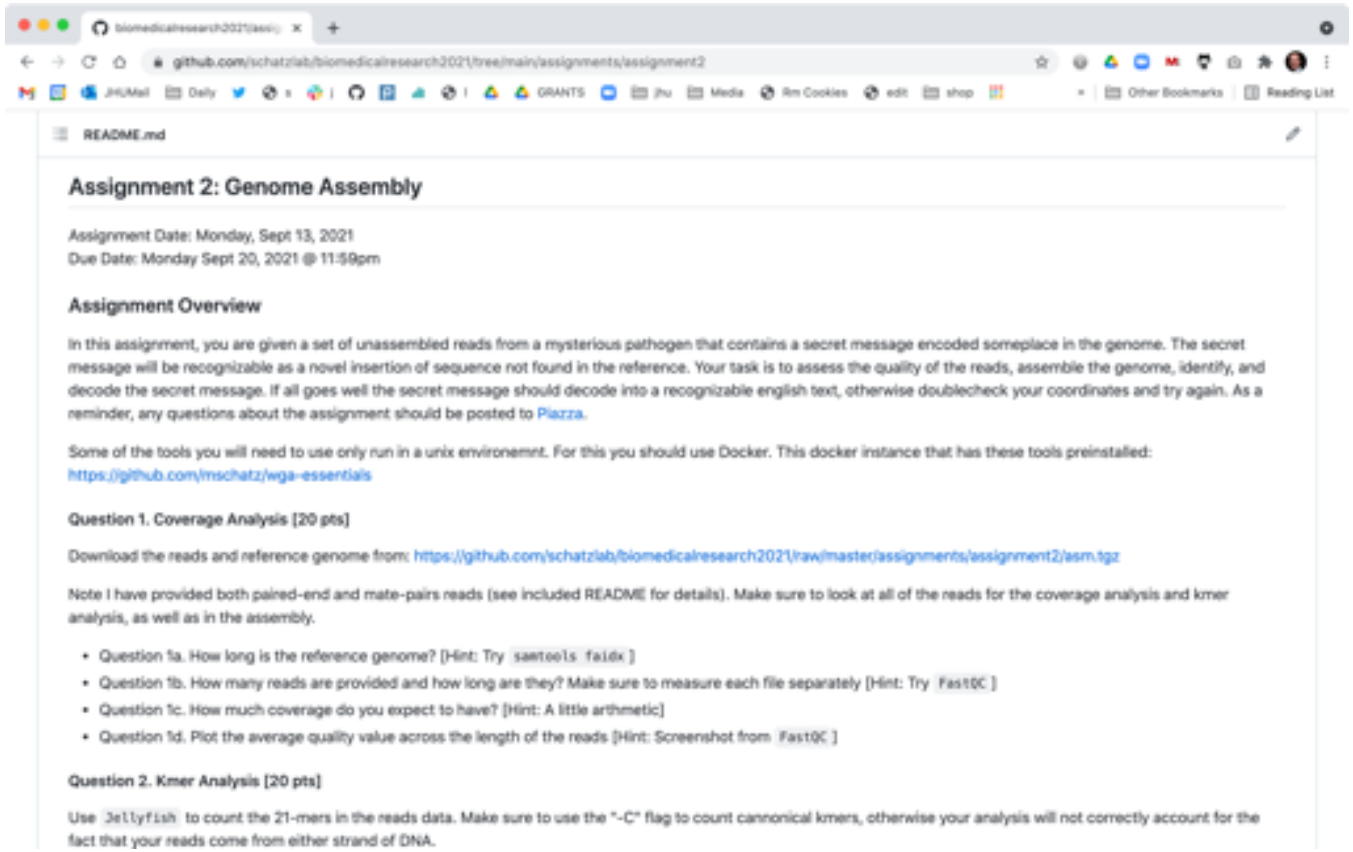
Michael Schatz

Sept 20, 2021

Lecture 6. Computational Biomedical Research

@mike_schatz

# Assignment 2: Genome Assembly
# Due Monday Sept 20 @ 11:59pm



README.md

## Assignment 2: Genome Assembly

Assignment Date: Monday, Sept 13, 2021
Due Date: Monday Sept 20, 2021 @ 11:59pm

### Assignment Overview

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

Some of the tools you will need to use only run in a unix environemnt. For this you should use Docker. This docker instance that has these tools preinstalled: https://github.com/mschatz/wga-essentials

**Question 1. Coverage Analysis [20 pts]**

Download the reads and reference genome from: https://github.com/schatzlab/biomedicalresearch2021/raw/master/assignments/assignment2/asm.tgz

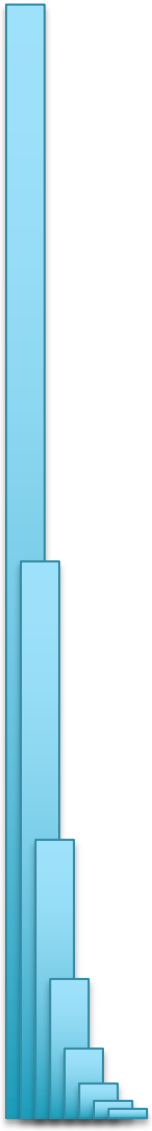
Note I have provided both paired-end and mate-pairs reads (see included README for details). Make sure to look at all of the reads for the coverage analysis and kmer analysis, as well as in the assembly.

- Question 1a. How long is the reference genome? [Hint: Try `samtools faidx`]
- Question 1b. How many reads are provided and how long are they? Make sure to measure each file separately [Hint: Try `FastQC`]
- Question 1c. How much coverage do you expect to have? [Hint: A little arthmetic]
- Question 1d. Plot the average quality value across the length of the reads [Hint: Screenshot from `FastQC`]

**Question 2. Kmer Analysis [20 pts]**

Use `Jellyfish` to count the 21-mers in the reads data. Make sure to use the "-C" flag to count cannonical kmers, otherwise your analysis will not correctly account for the fact that your reads come from either strand of DNA.

https://github.com/schatzlab/biomedicalresearch2021



# Assignment 3: Variant Calling

Postponed for 1 week!

https://github.com/schatzlab/biomedicalresearch2021

# The human genome

# Who is the reference human?

ment abuse, civil disobedience

WANTED

20 Volunteers

to participate in the

Human Genome Project

a very large international scientific research effort.

*Pieter de Jong, RPCI*

WANTED

20 Volunteers

to participate in the

Human Genome Project

a very large international scientific research effort.

The goal is to decode the human hereditary information (*human blueprint*) that determines all individual traits inherited from parents. The outcome of the project will have tremendous impact on future progress of medical science and lead to improved diagnosis and treatment of hereditary diseases.

Volunteers will receive information about the project from the Clinical Genetics Service at Roswell Park, and sign a consent form before participating.

*No personal information will be maintained or transferred.*

Volunteers will provide a one-time donation of a small blood specimen. A small monetary reimbursement will be provided to the participants for their time and effort.

Individuals must be at least 18 years of age.
Persons who have undergone chemotherapy are not eligible.

ROSWELL PARK CANCER INSTITUTE

For more information please contact the
**Clinical Genetics Service**
845-5720 (9:00 am - 3:00 pm)
March 24 - 26, 1997

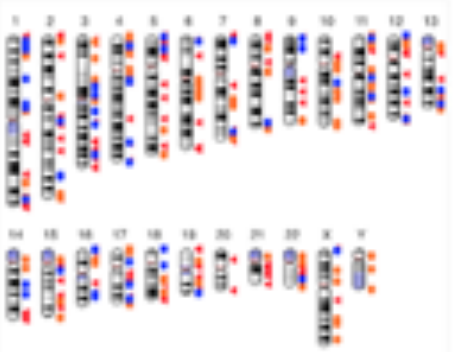
# Genome Reference Consortium

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.p13*

The GRC is working hard to provide the best possible reference assembly for human. We do this by both generating multiple representations (alternate loci) for regions that are too complex to be represented by a single path. Additionally, we are releasing regional fixes known as patches. This allows users who are interested in a specific locus to get an improved representation without affecting users who need chromosome coordinate stability.

**Download data:**

- GRCh38.p13 (latest minor release) FTP
- GRCh38 (latest major release) FTP
- Genomic regions under review FTP
- Current Tiling Path Files (TPFs)

**GRC trackhub:**
http://ngs.sanger.ac.uk/production/grc/track_hub/hub.txt

Slides from GRC presentations in ASHG 2019.

Transitioning to GRCh38? Try the NCBI Remapping Service, which uses the same assembly-assembly alignments used by the GRC.

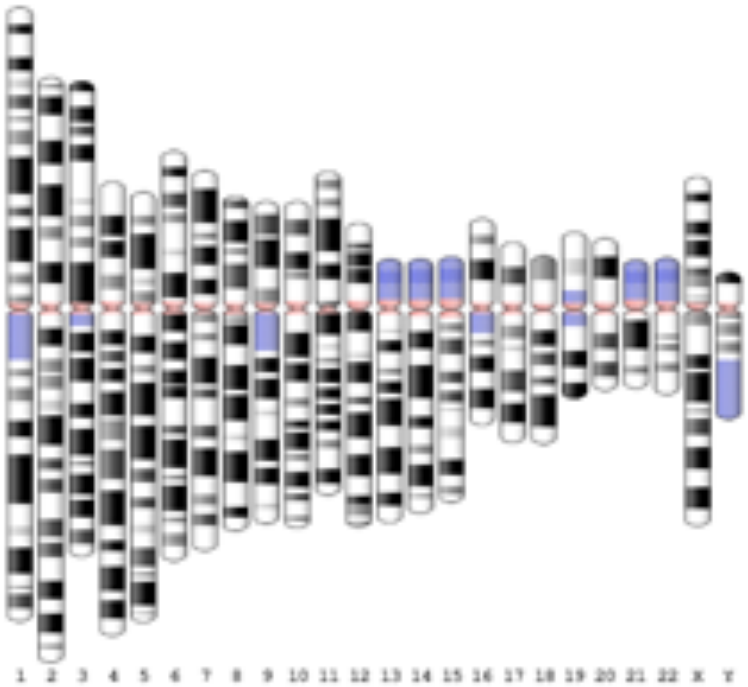
**Next assembly update**

The next assembly update (GRCh38.p14) will be a minor (patch) release planned for release in the second half of 2020. The GRC remains committed to its mission to improve the human reference genome assembly, correcting errors and adding sequence to ensure it provides the best representation of the human genome to meet basic and clinical research needs. We will continue to make these updates publicly available at regular intervals in the form of patch releases, but have decided to indefinitely postpone our next coordinate-changing update (GRCh39) while we evaluate new models and sequence content for the human reference assembly currently in development.

# Current Status of the Reference Genome

**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…



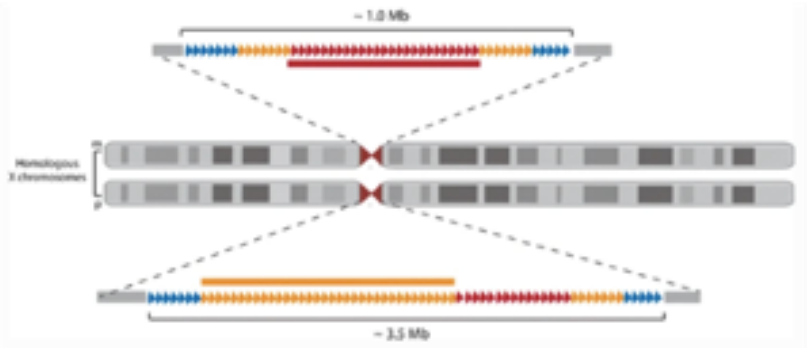
# Finishing the human genome

## 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

# Single Molecule Long Read Sequencing

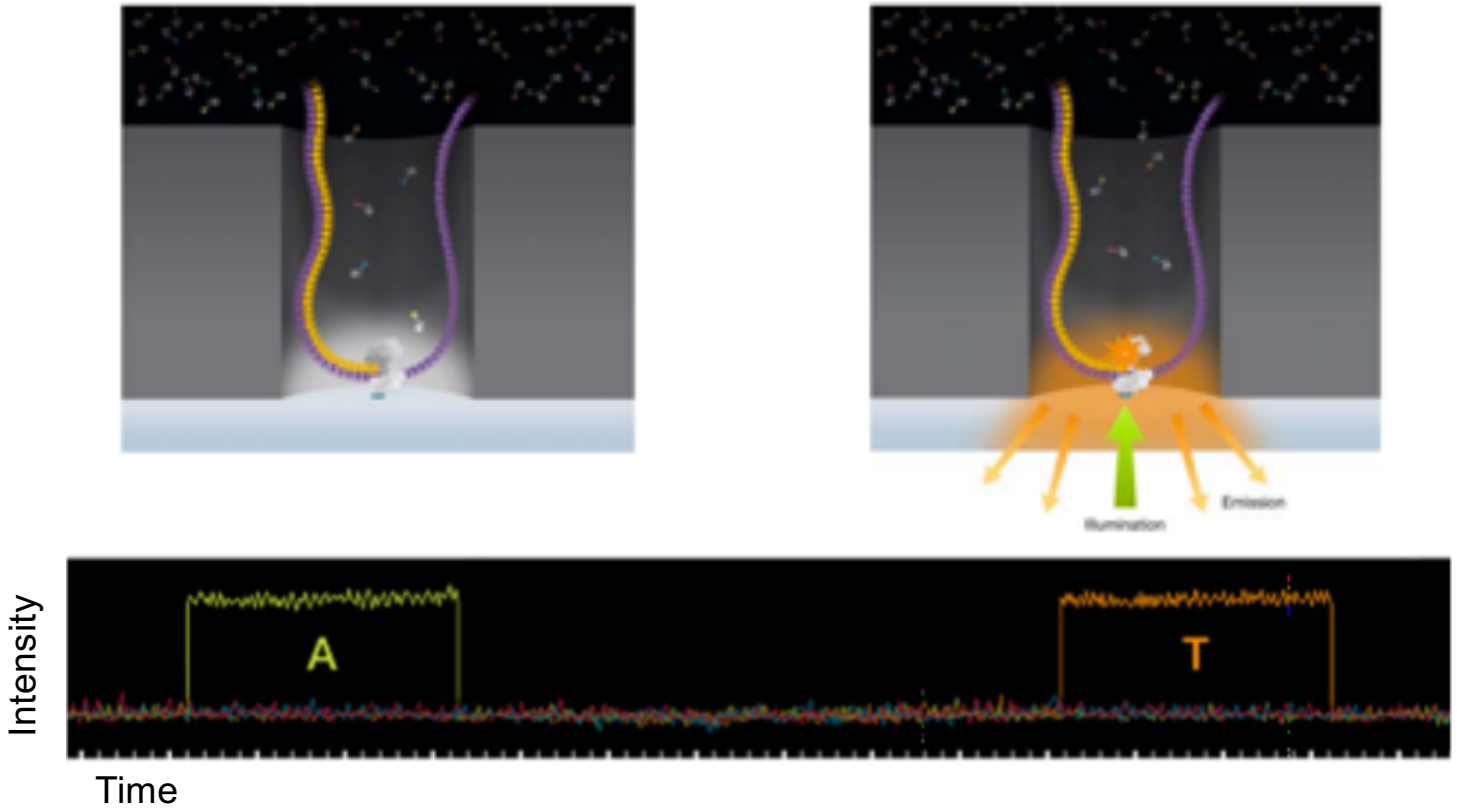
***PacBio***
***Sequel II***

***Oxford Nanopore***
***PromethION***

# 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).

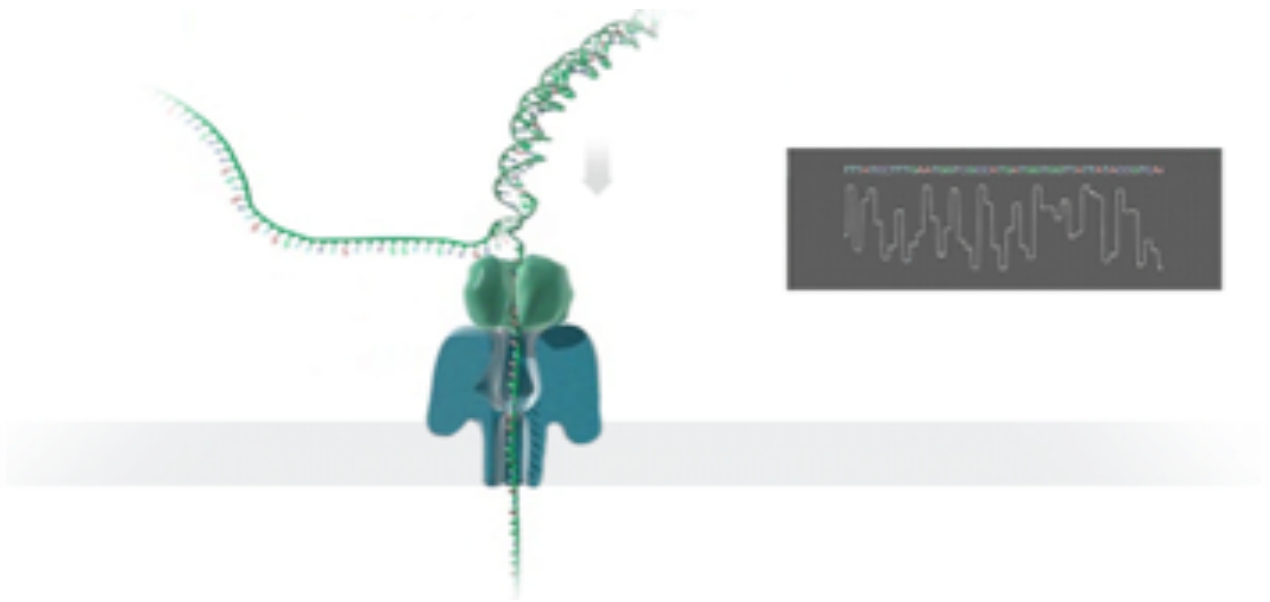



http://www.youtube.com/watch?v=v8p4ph2MAvI

# Oxford Nanopore Technologies (ONT)

# Nanopore Sequencing

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

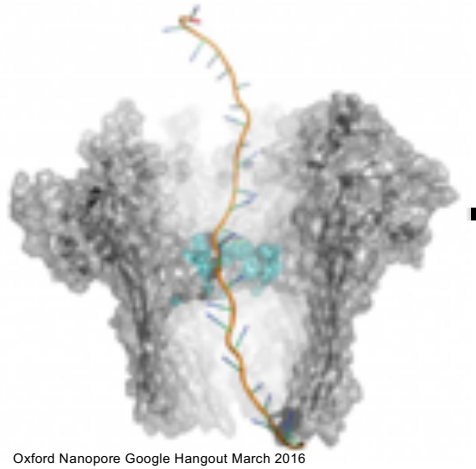


https://www.youtube.com/watch?v=CE4dW64x3Ts

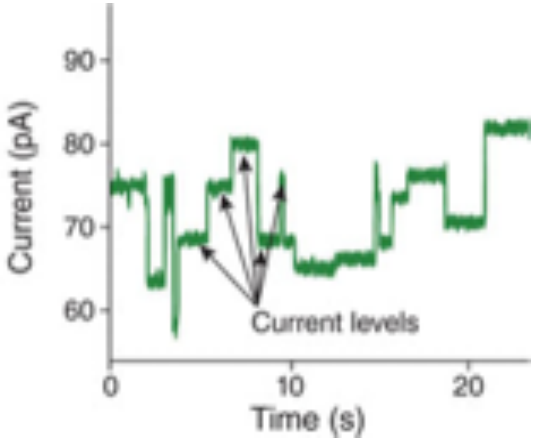
nanoporetech.com/applications/dna-nanopore-sequencing

# Nanopore: Single Molecule Sequencing

- Oxford Nanopore Technologies, CsgG biological pore
- No theoretical upper limit to sequencing read length, practical limit only in delivering DNA to the pore intact
- Palm sized sequencer
- Typical sequencing output 10-20Gb+



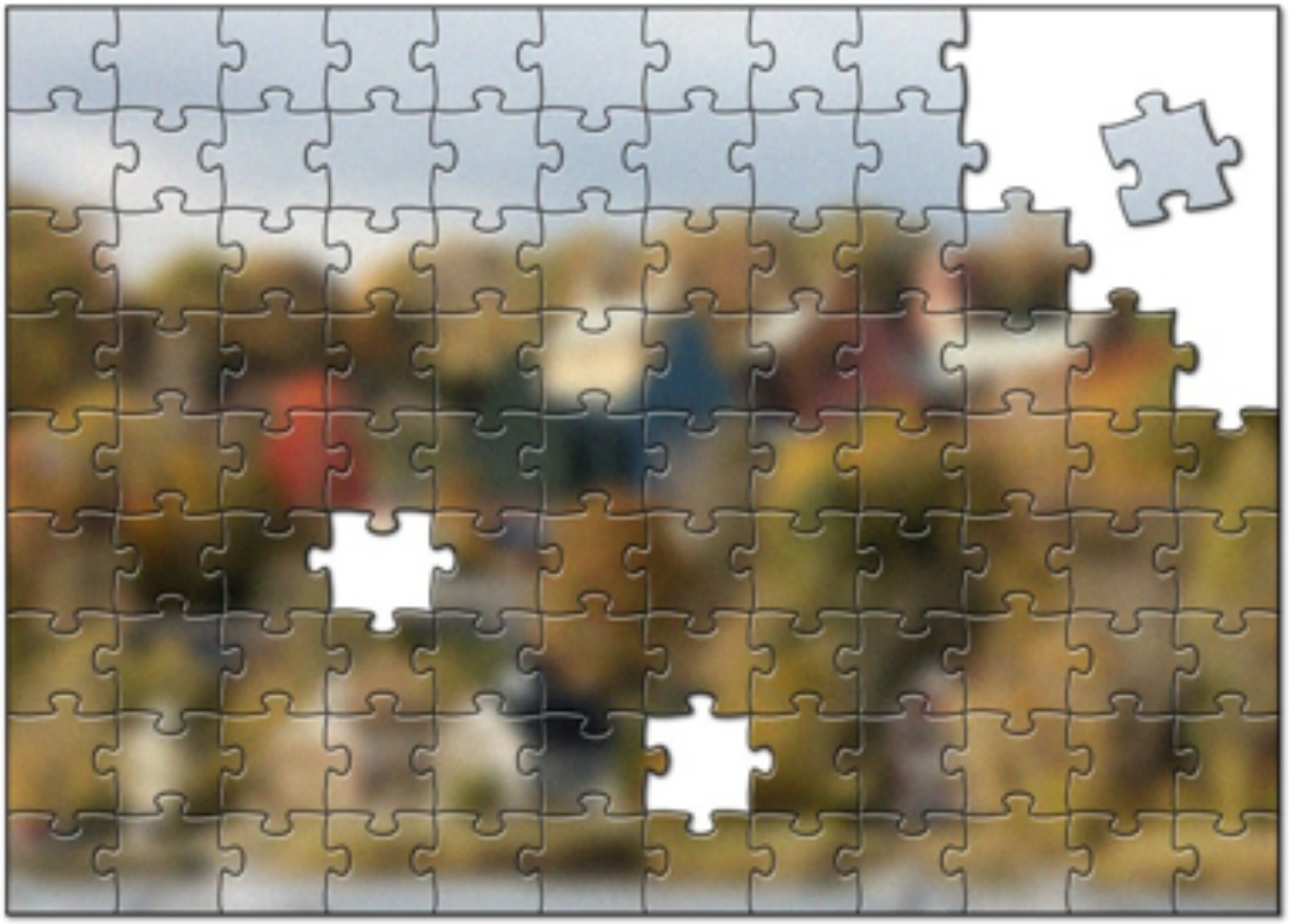
Oxford Nanopore Google Hangout March 2016



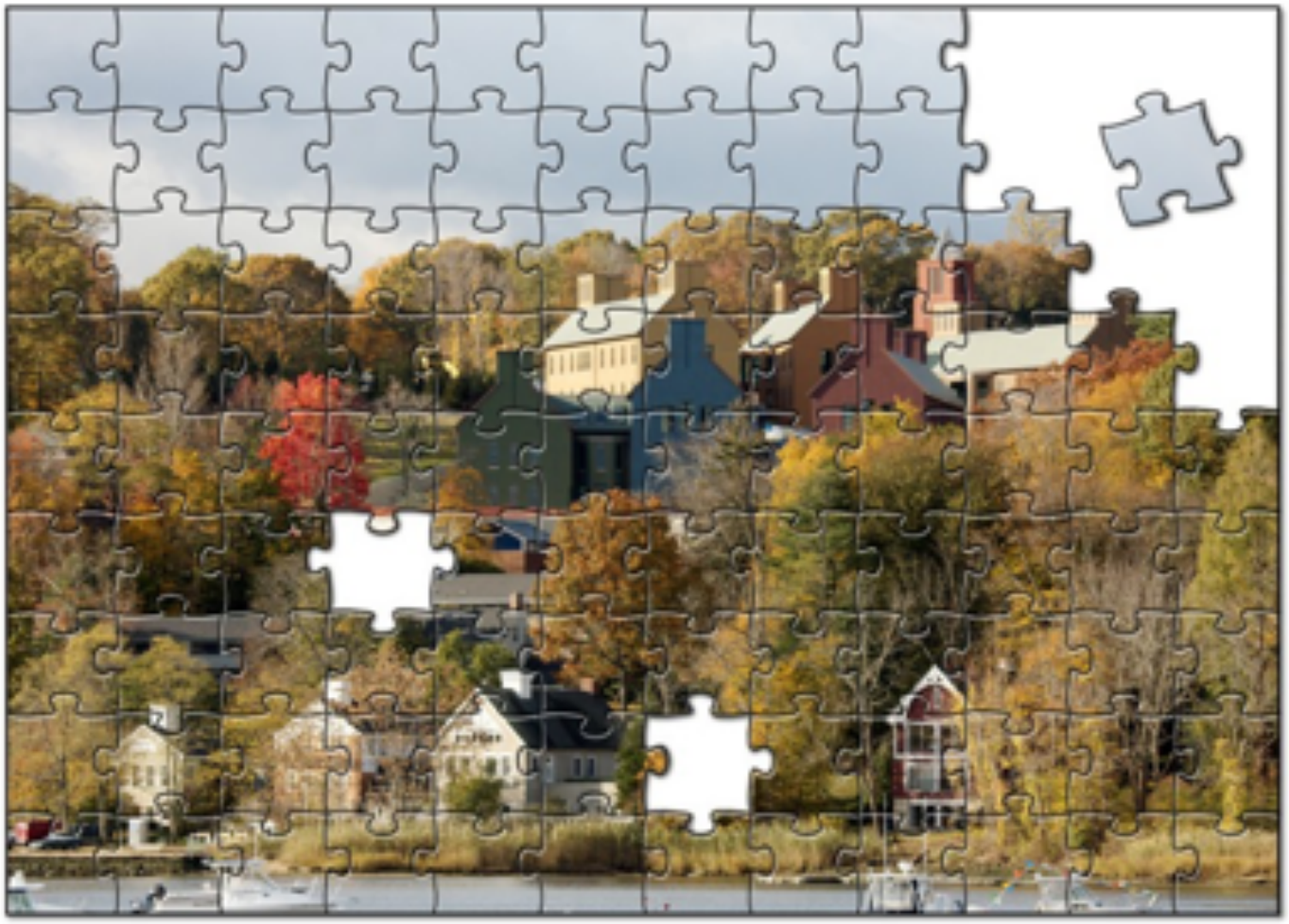

Deamer et al 2016, Nature Biotech

```
ATCGATCGATAGTAT
TAGATACGACTAGCG
ATCAG...
```

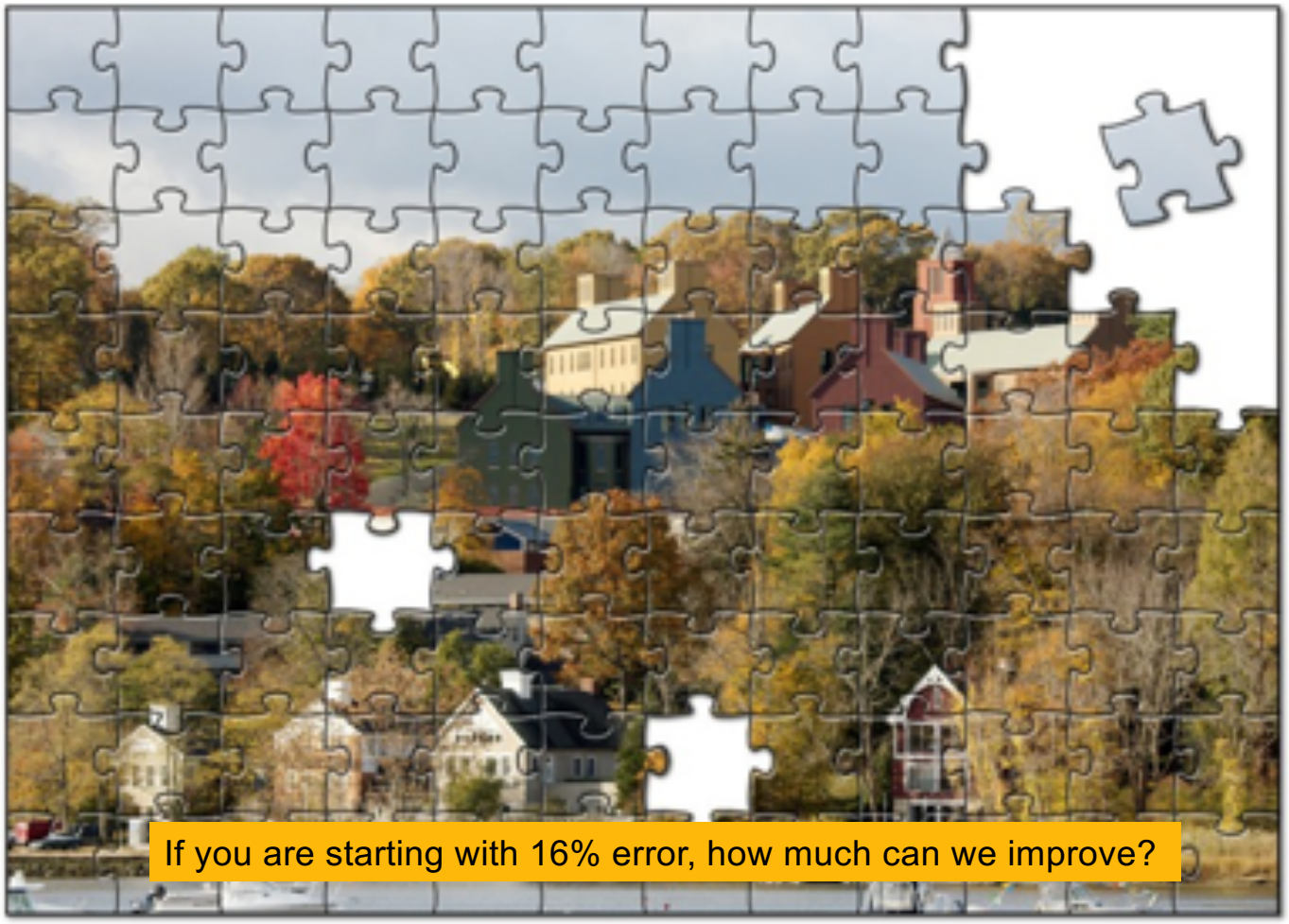
# Single Molecule Sequences



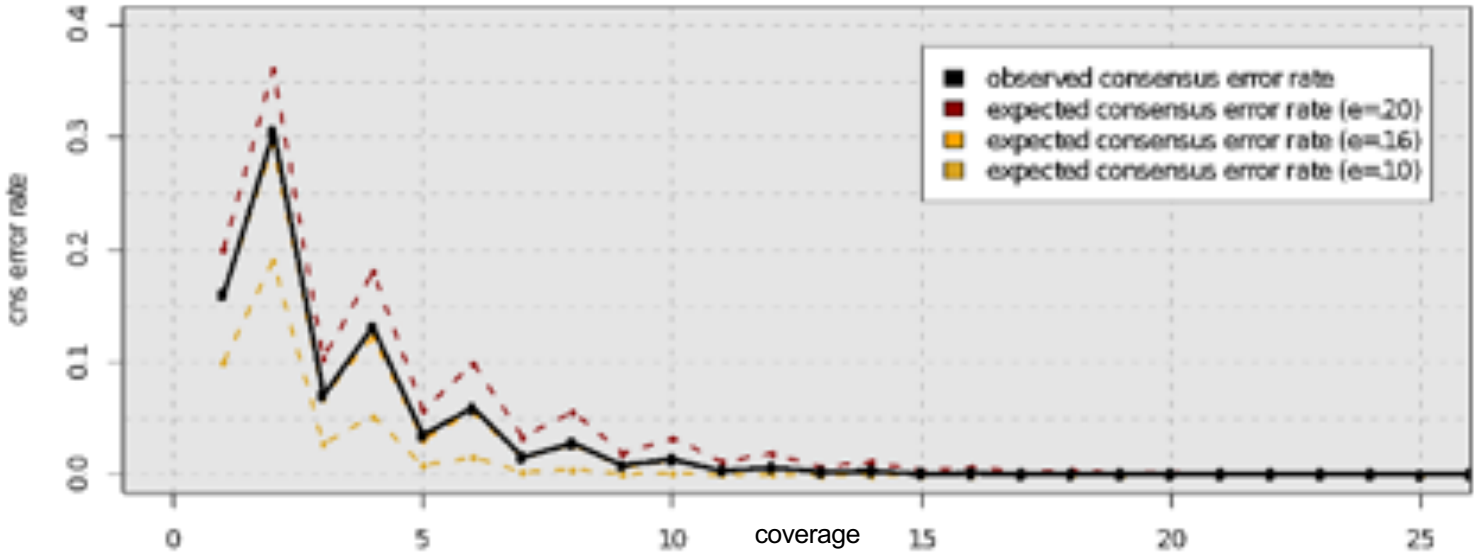
# “Corrective Lens” for Sequencing



# “Corrective Lens” for Sequencing



# Consensus Accuracy and Coverage




## Coverage can overcome random errors

- Dashed: error model from binomial sampling; solid: observed accuracy
- For same reason, CCS is extremely accurate when using 5+ subreads

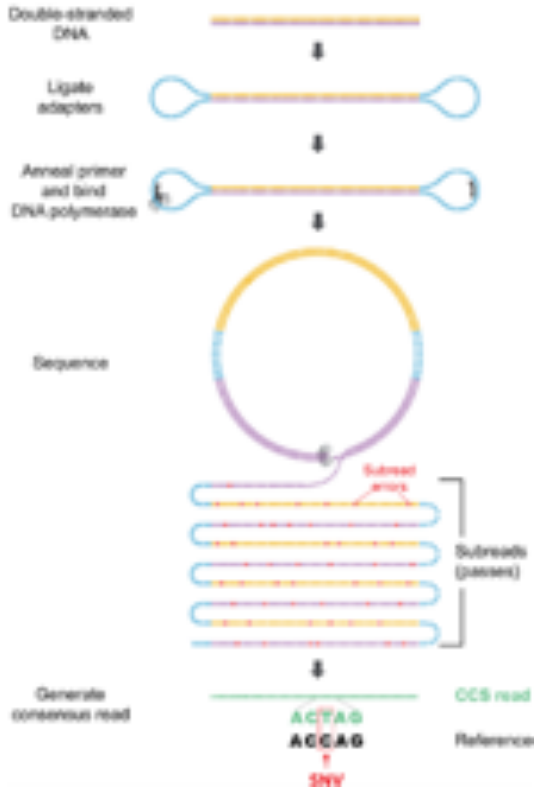
$$CNS\,Error = \sum_{i=\lceil c/2 \rceil}^{c} \binom{c}{i} (e)^{i} (1-e)^{n-i}$$

# “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

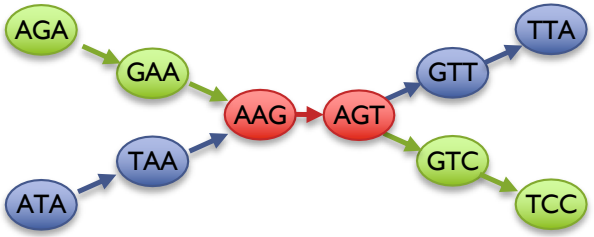
Limits read length, used to be very expensive but more manageable now




**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

# Two Paradigms for Assembly

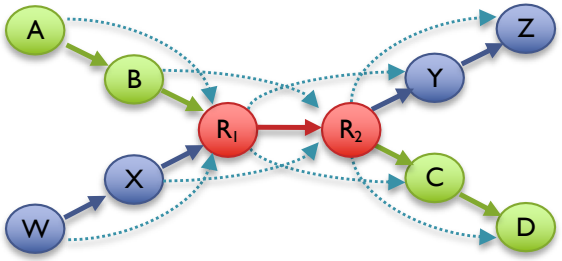
## de Bruijn Graph




Short read assemblers

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

## Overlap Graph

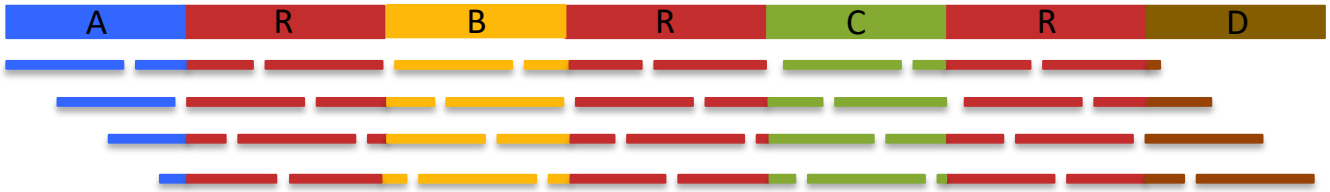


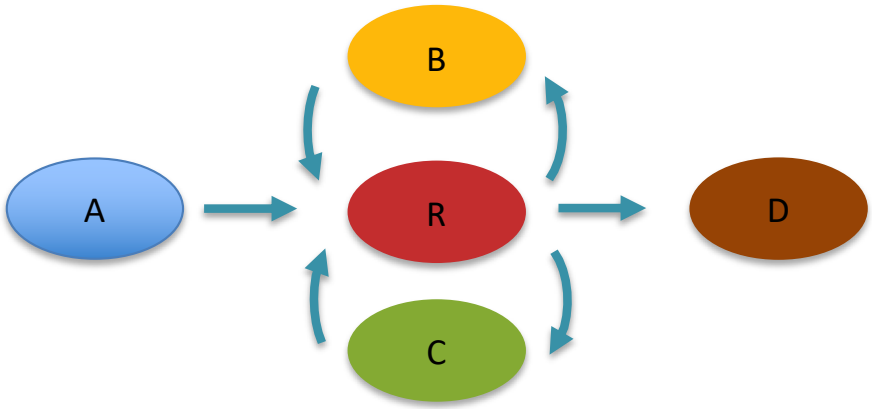

Long read assemblers

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

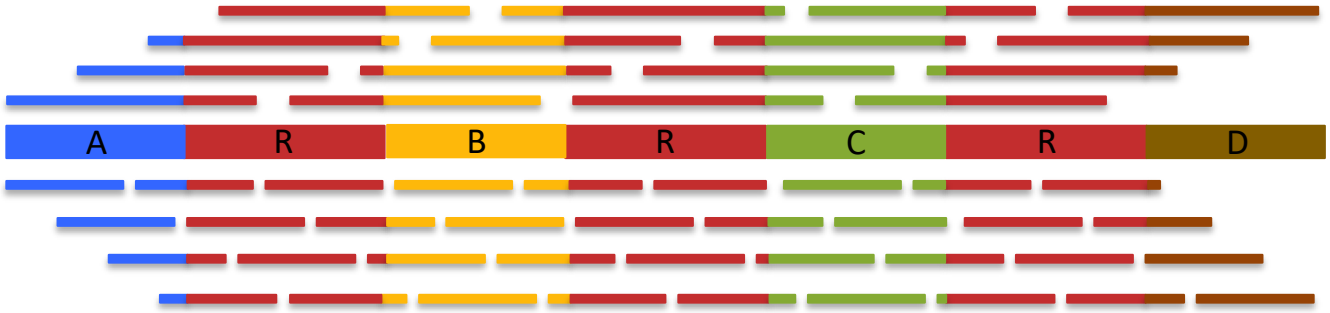
**Assembly of Large Genomes using Second Generation Sequencing**
Schatz MC, Delcher AL, Salzberg SL (2010) *Genome Research.* 20:1165-1173.

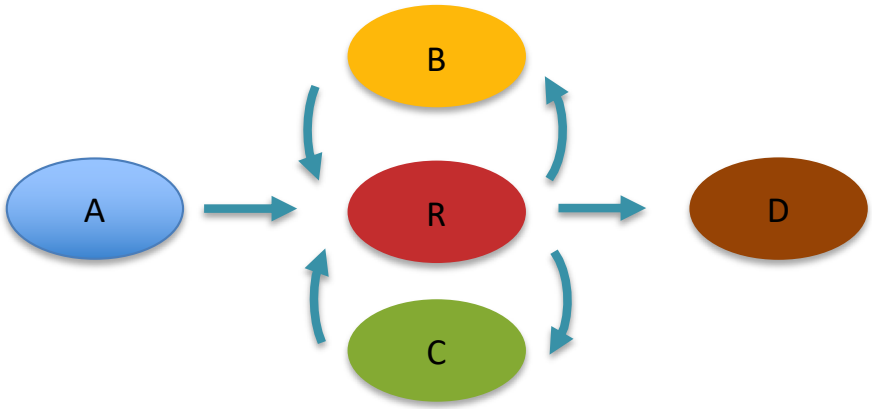
# Assembly Complexity



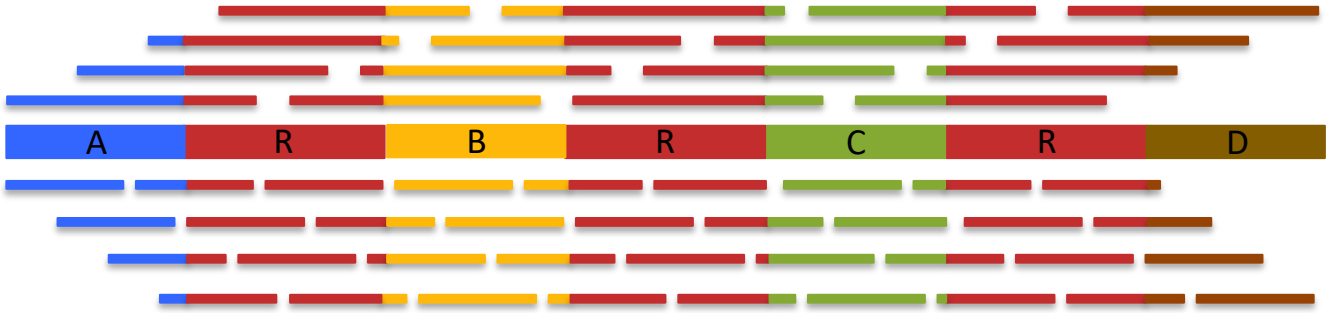



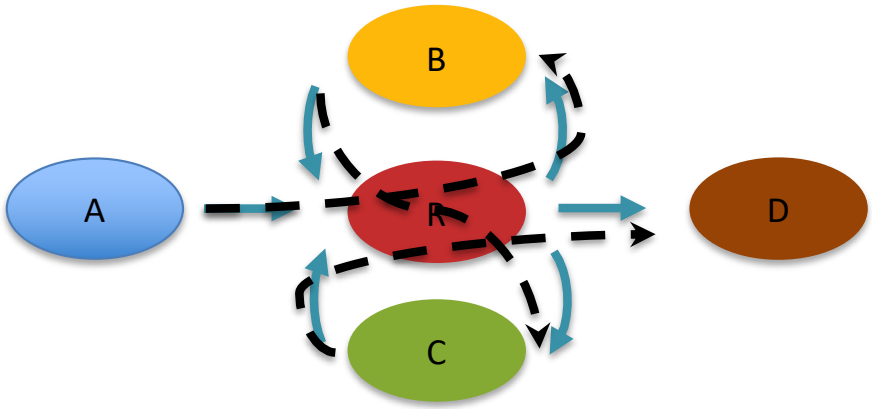
# Assembly Complexity



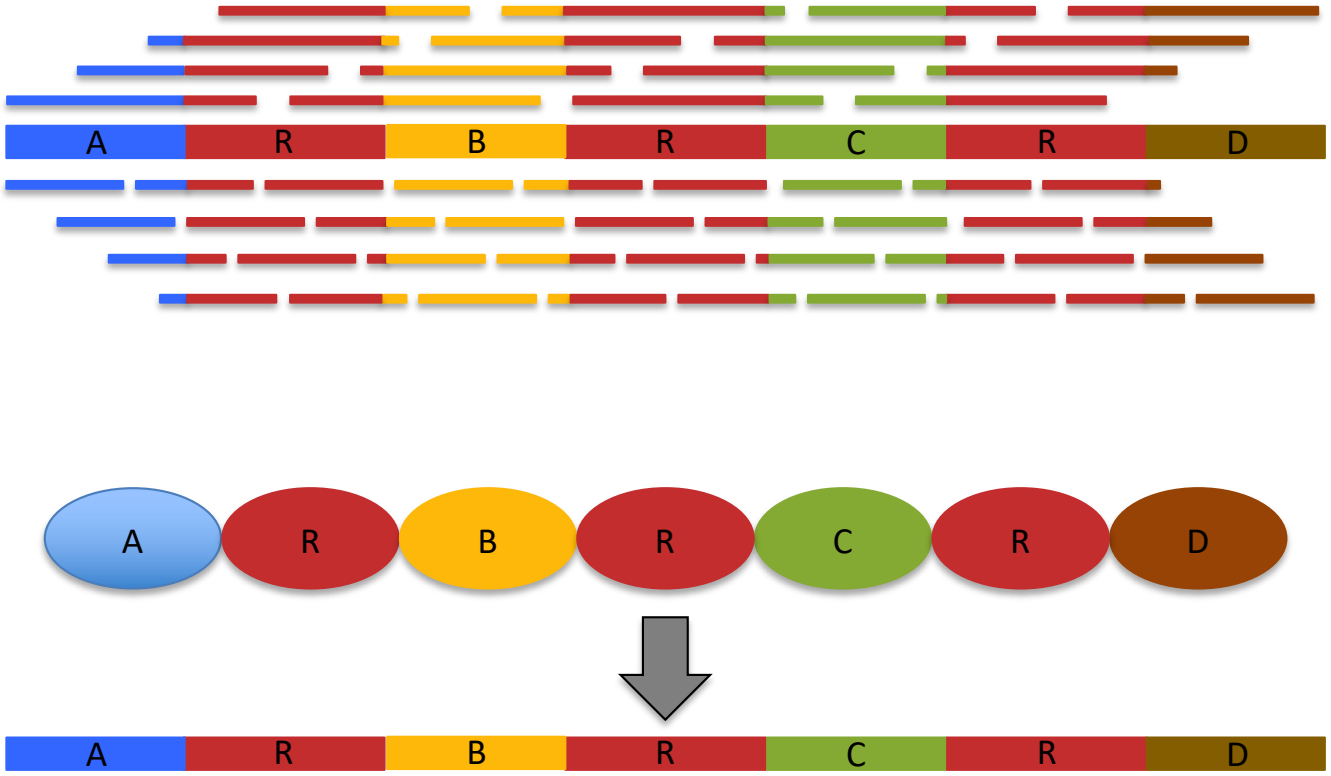



# Assembly Complexity



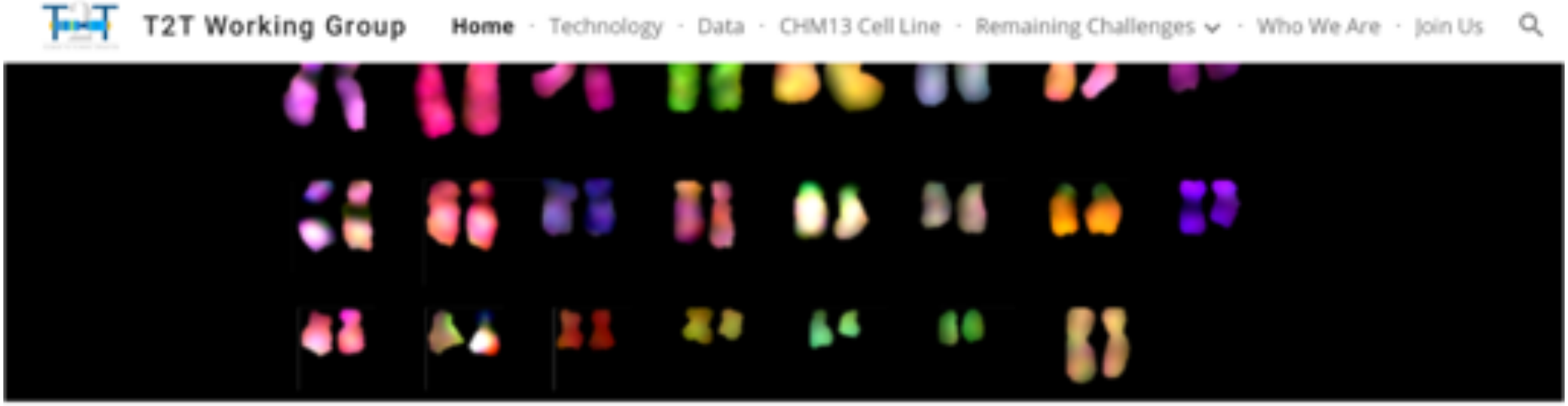



# Assembly Complexity




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

# Let’s finish a human genome

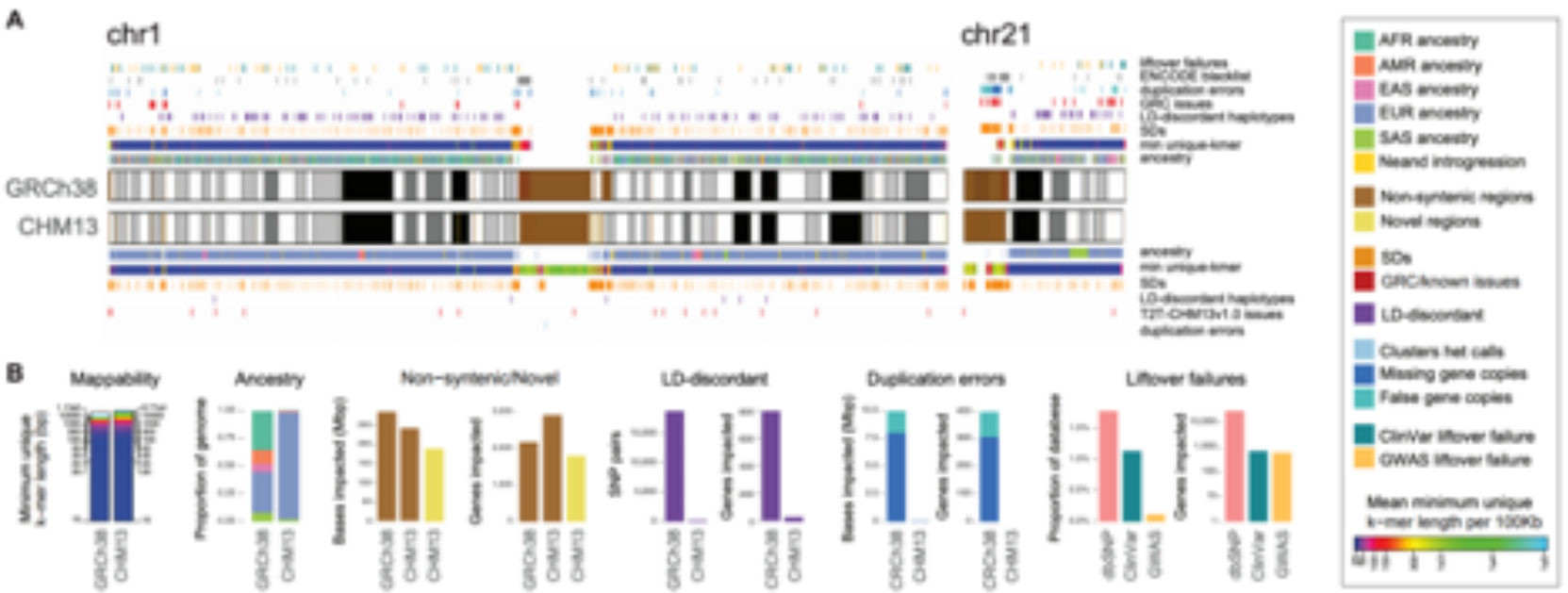

**The Telomere-to-Telomere (T2T) consortium is an open, community-based effort to generate the first complete assembly of a human genome.**

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

# CHM13 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




Daniela Soto

Megan Dennis

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

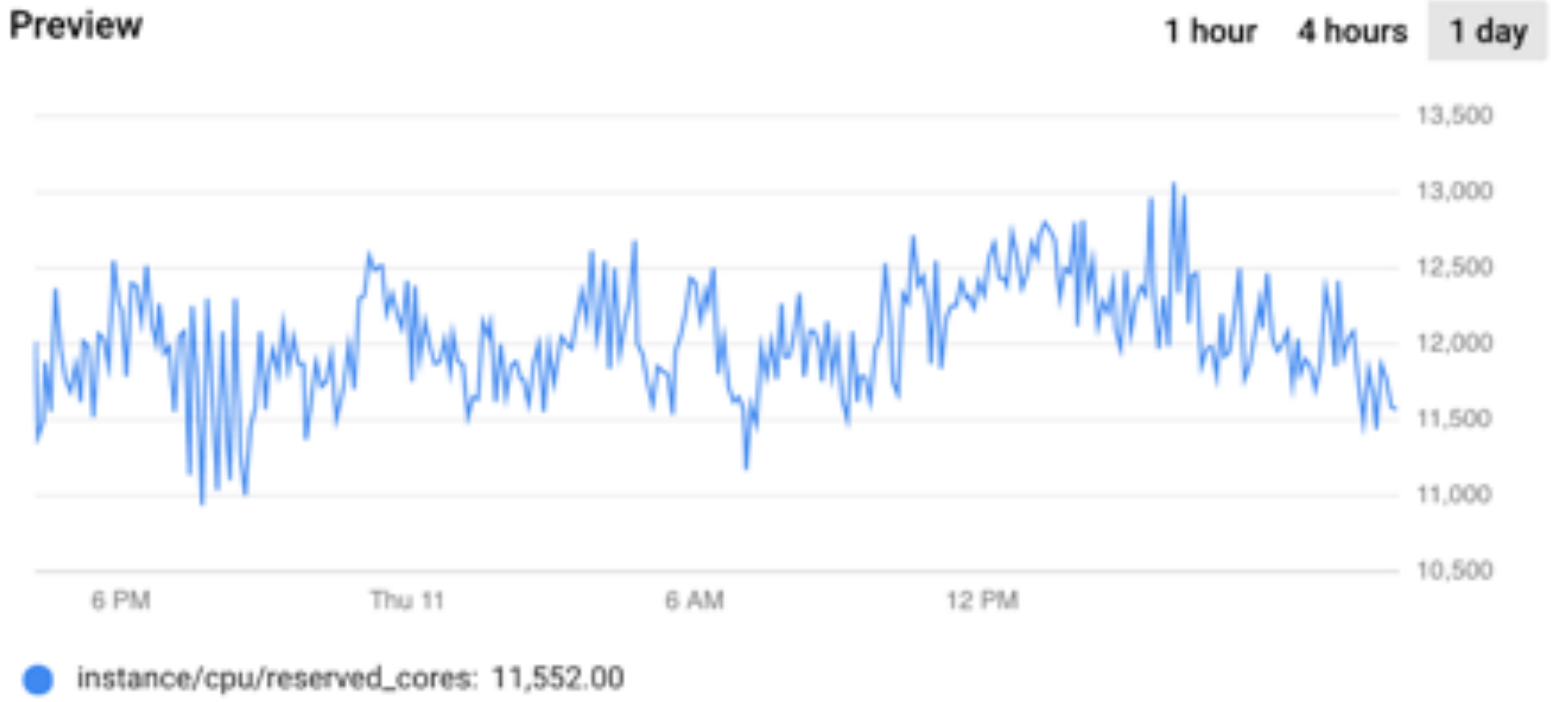
**A complete reference genome improves analysis of human genetic variation**
Aganezov, S*, Yan, SM*, Soto, DC*, Kirsche, M*, Zarate, S*, *et al.* (2021) bioRxiv. doi: https://doi.org/10.1101/2021.07.12.452063

# T2T Variants: 1000 Genomes Project



https://www.internationalgenome.org/faq/which-populations-are-part-your-study/

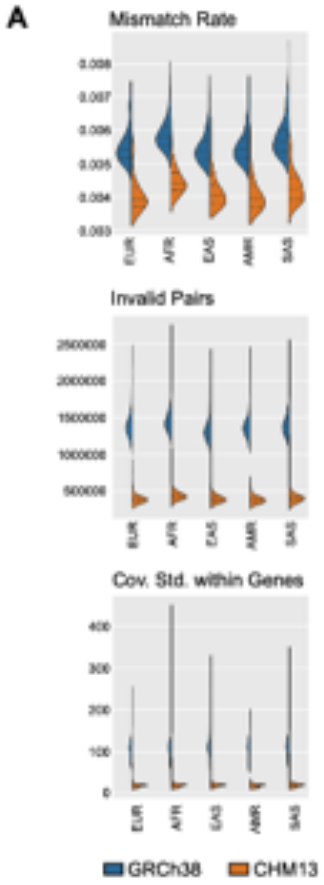
# Core usage over 24 hours



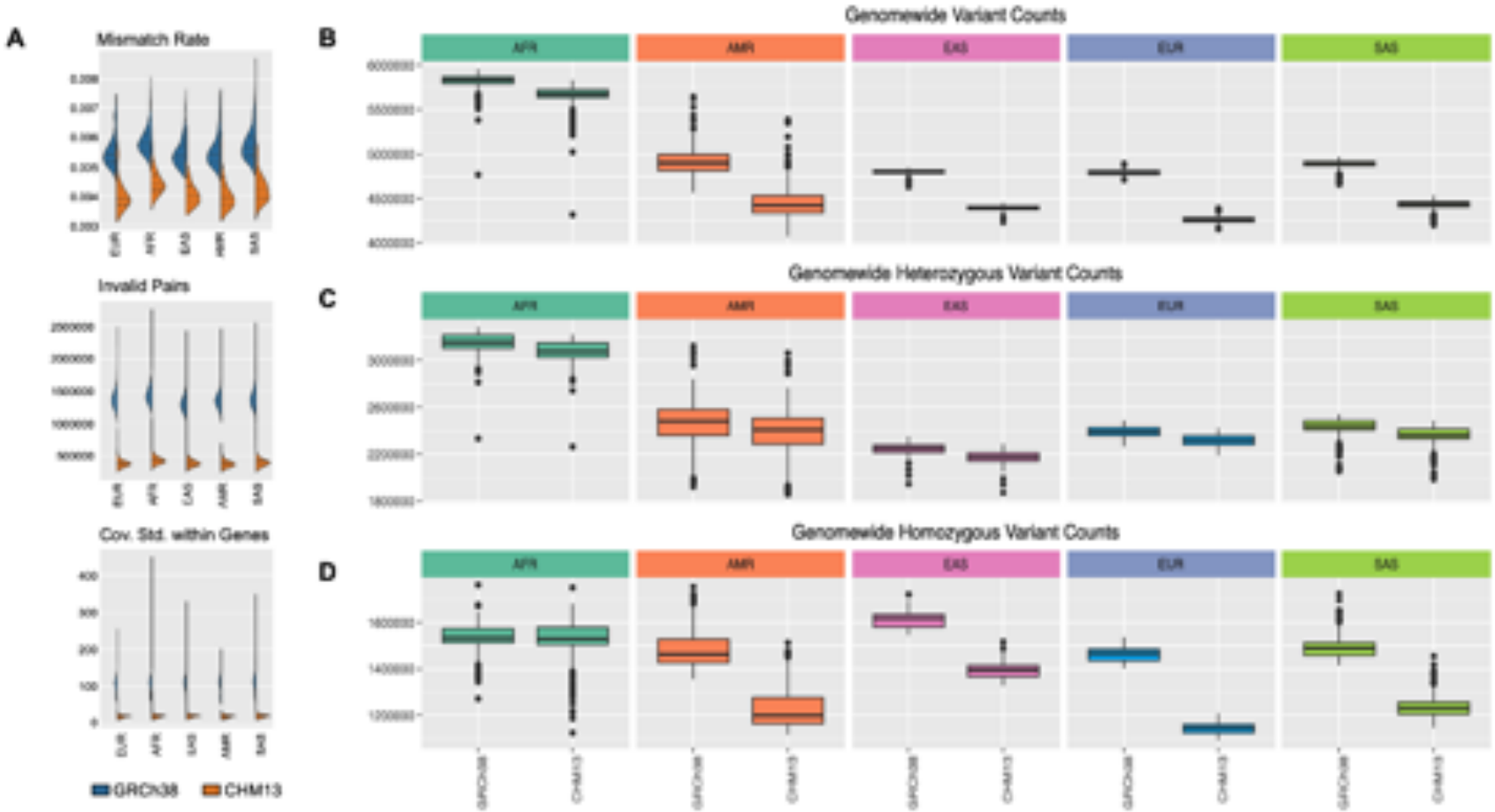

Samantha Zarate

**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

# 1000G Mapping & Variants on T2T-CHM13 (Figure 2 - top)



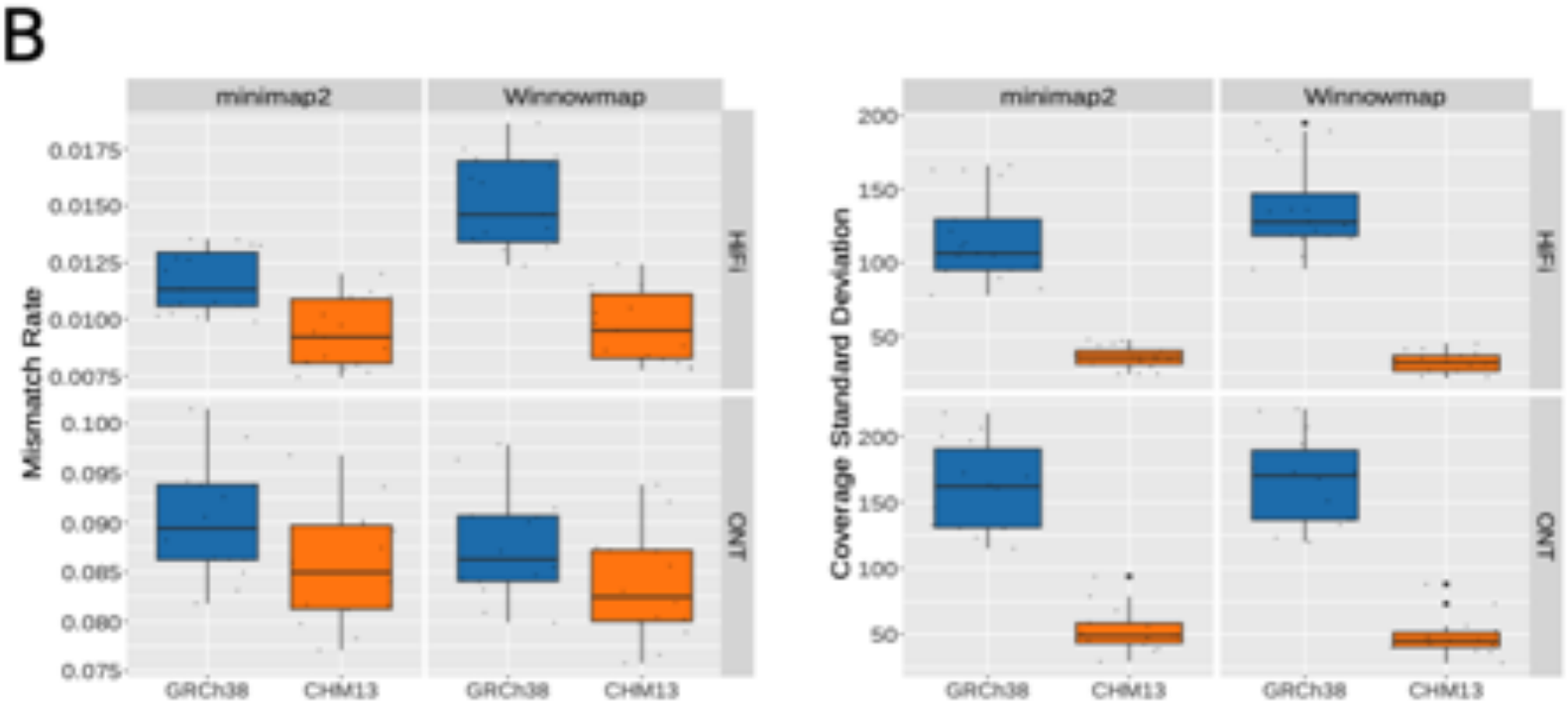
# 1000G Mapping & Variants on T2T-CHM13 (Figure 2 - top)



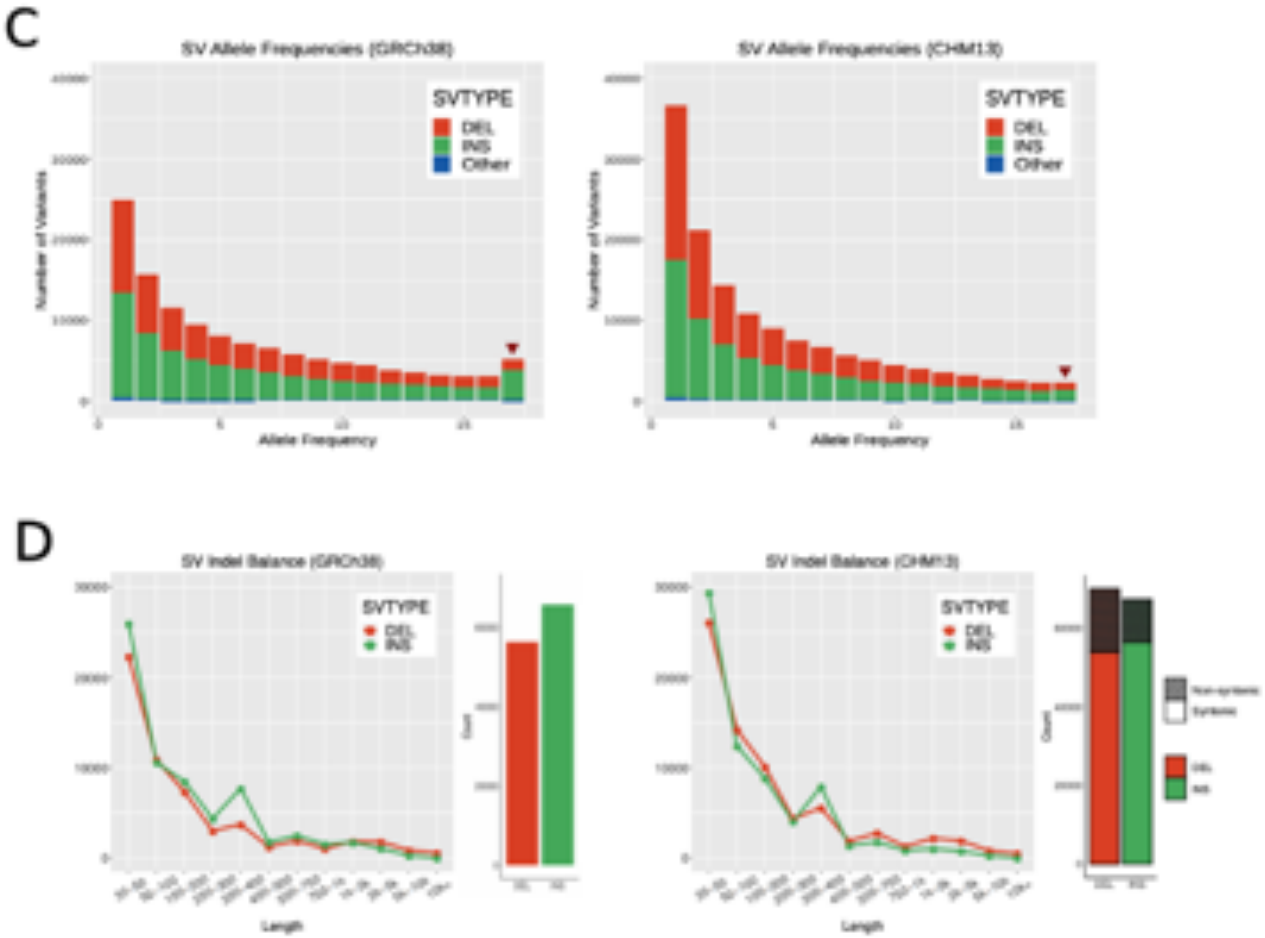
# Long Read Analysis with T2T-CHM13 (Fig 3)

Melanie Kirsche

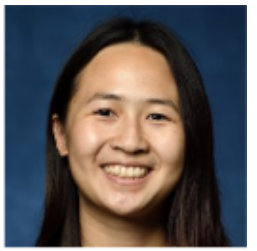
Sergey Aganezov



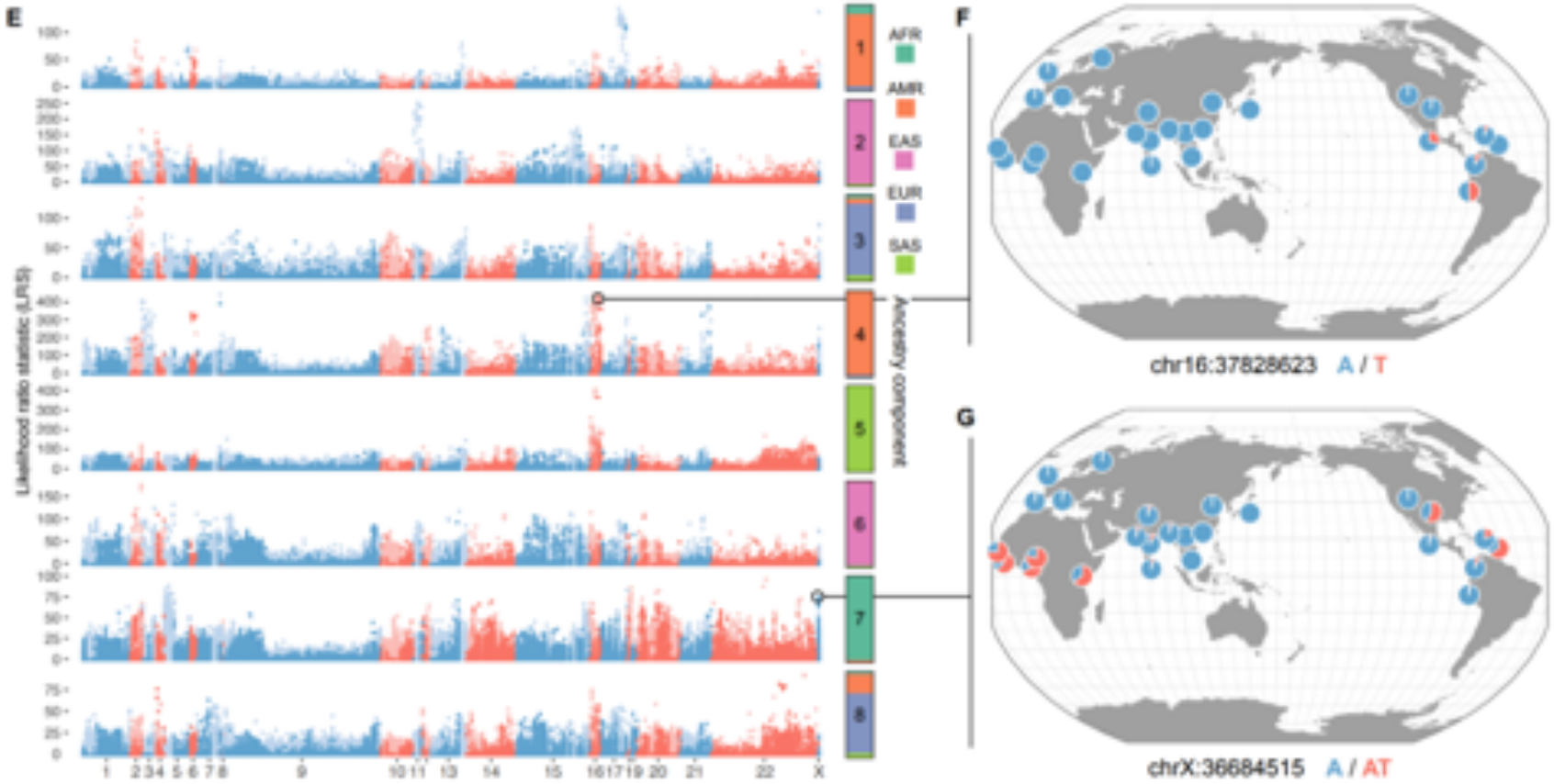
# Long Read Analysis with T2T-CHM13 (Fig 3)



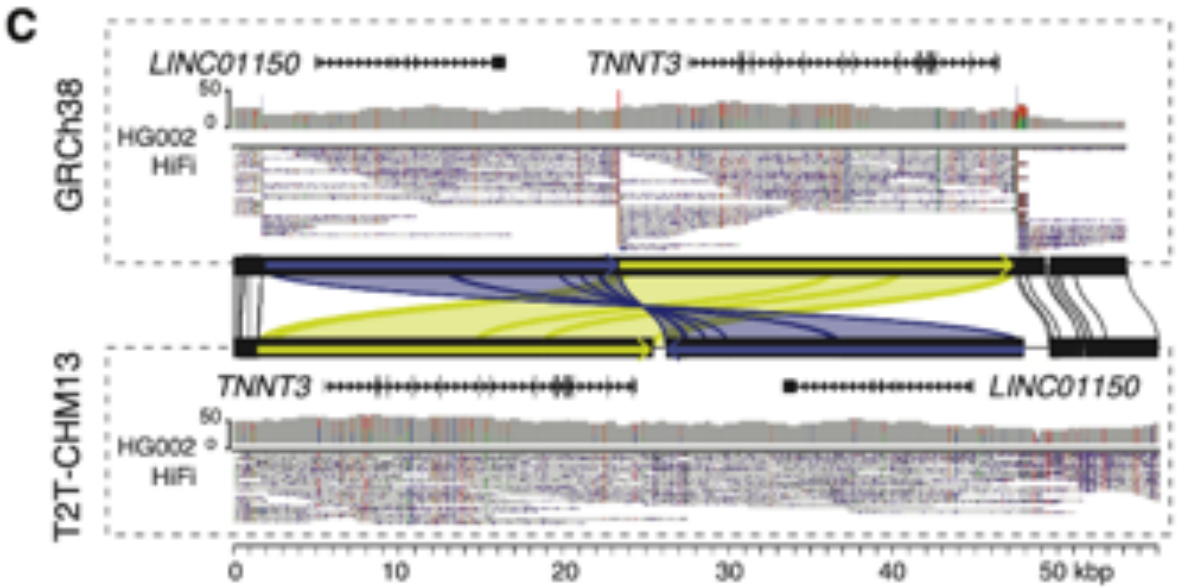
# Novel Variants from 1000G on T2T-CHM13 (Fig 4)



Stephanie Yan



# T2T-CHM13 Improves Clinical Genomics Variant Calling (Fig 5)




Danny Miller

Daniela Soto

Megan Dennis

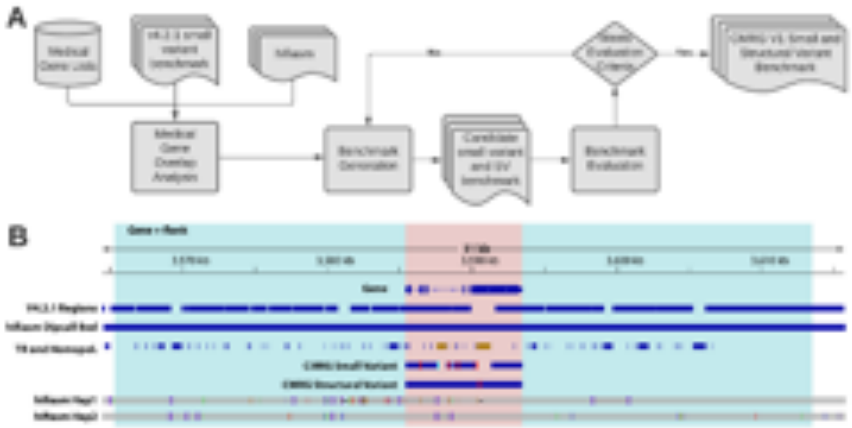
Justin Zook

Fritz Sedlazeck

# T2T-CHM13 Improves Clinical Genomics Variant Calling (Fig 5)

## Towards a Comprehensive Variation Benchmark for Challenging Medically-Relevant Autosomal Genes

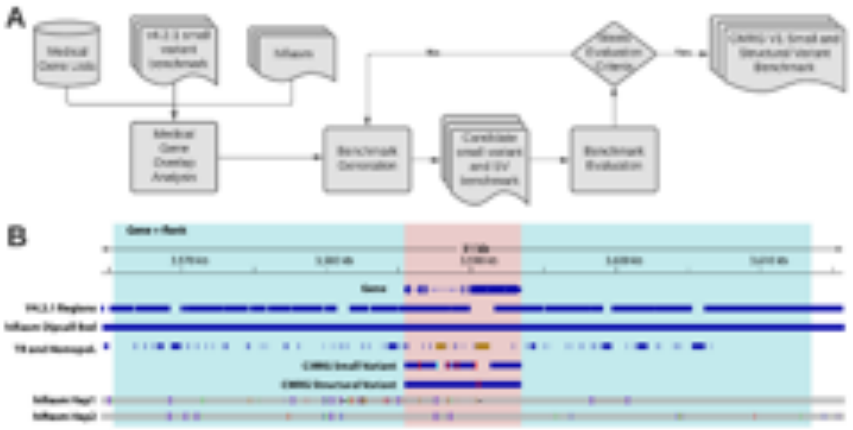
Justin Wagner,[1] Nathan D Olson,[1] Lindsay Harris,[1] Jennifer McDaniel,[1] Haoyu Cheng,[2] Arkarachai Fungtammasan,[3] Yih-Chii Hwang,[3] Richa Gupta,[3] Aaron M Wenger,[4] William J Rowell,[4] Ziad M Khan,[5] Jesse Farek,[6] Yiming Zhu,[6] Aishwarya Pisupati,[6] Medhat Mahmoud,[6] Chunlin Xiao,[6] Byunggil Yoo,[7] Sayed Mohammad Ebrahim Sahraeian,[8] Danny E. Miller,[9] David Jáspez,[10] José M. Lorenzo-Salazar,[10] Adrián Muñoz-Barrera,[10] Luis A. Rubio-Rodríguez,[10] Carlos Flores,[10,11] Giuseppe Narzisi,[12] Uday Shanker Evani,[12] Wayne E. Clarke,[12] Joyce Lee,[13] Christopher E. Mason[14], Stephen E. Lincoln[15], Karen H. Miga[16], Mark T. W. Ebbert,[17] Alaina Shumate,[18] Heng Li,[2] Chen-Shan Chin*[,3], Justin M Zook*[,1], Fritz J Sedlazeck*[,5]

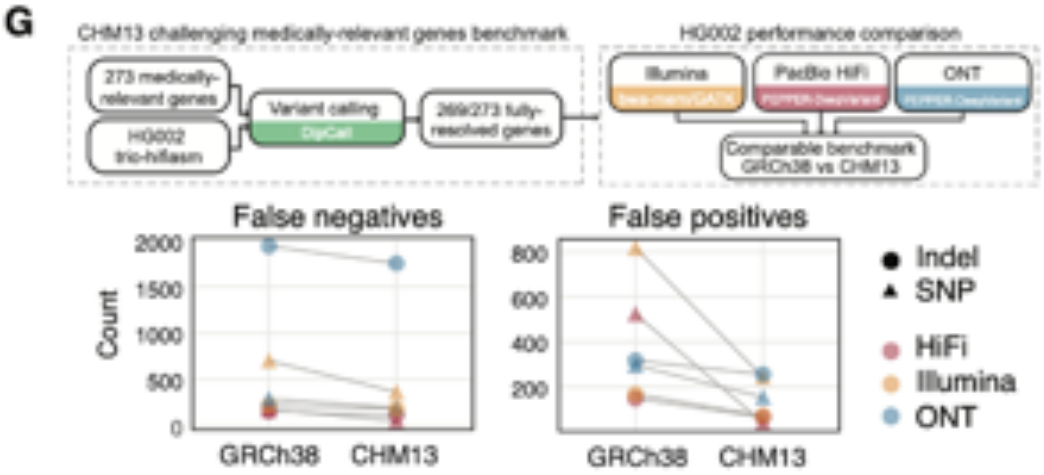


# T2T-CHM13 Improves Clinical Genomics Variant Calling (Fig 5)

## Towards a Comprehensive Variation Benchmark for Challenging Medically-Relevant Autosomal Genes

Justin Wagner,[1] Nathan D Olson,[1] Lindsay Harris,[1] Jennifer McDaniel,[1] Haoyu Cheng,[2] Arkarachai Fungtammasan,[3] Yih-Chii Hwang,[3] Richa Gupta,[3] Aaron M Wenger,[4] William J Rowell,[4] Ziad M Khan,[5] Jesse Farek,[6] Yiming Zhu,[6] Aishwarya Pisupati,[6] Medhat Mahmoud,[6] Chunlin Xiao,[6] Byunggil Yoo,[7] Sayed Mohammad Ebrahim Sahraeian,[8] Danny E. Miller,[9] David Jáspez,[10] José M. Lorenzo-Salazar,[10] Adrián Muñoz-Barrera,[10] Luis A. Rubio-Rodríguez,[10] Carlos Flores,[10,11] Giuseppe Narzisi,[12] Uday Shanker Evani,[12] Wayne E. Clarke,[12] Joyce Lee,[13] Christopher E. Mason[14], Stephen E. Lincoln[15], Karen H. Miga[16], Mark T. W. Ebbert,[17] Alaina Shumate,[18] Heng Li,[2] Chen-Shan Chin*[,3], Justin M Zook*[,1], Fritz J Sedlazeck*[,5]



## 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

# Open Research Questions

- What about other genomes?
  - Simons Genome Diversity Project (279 genomes)
  - GTEx (~1000 samples)
  - Clinical Samples
- What about other assays
  - Short read structural variants
  - Phased analysis
  - Exome-capture
  - RNAseq
  - eQTL Analysis