

A complete reference genome improves analysis of human genetic variation

Michael Schatz

Sept 15, 2021

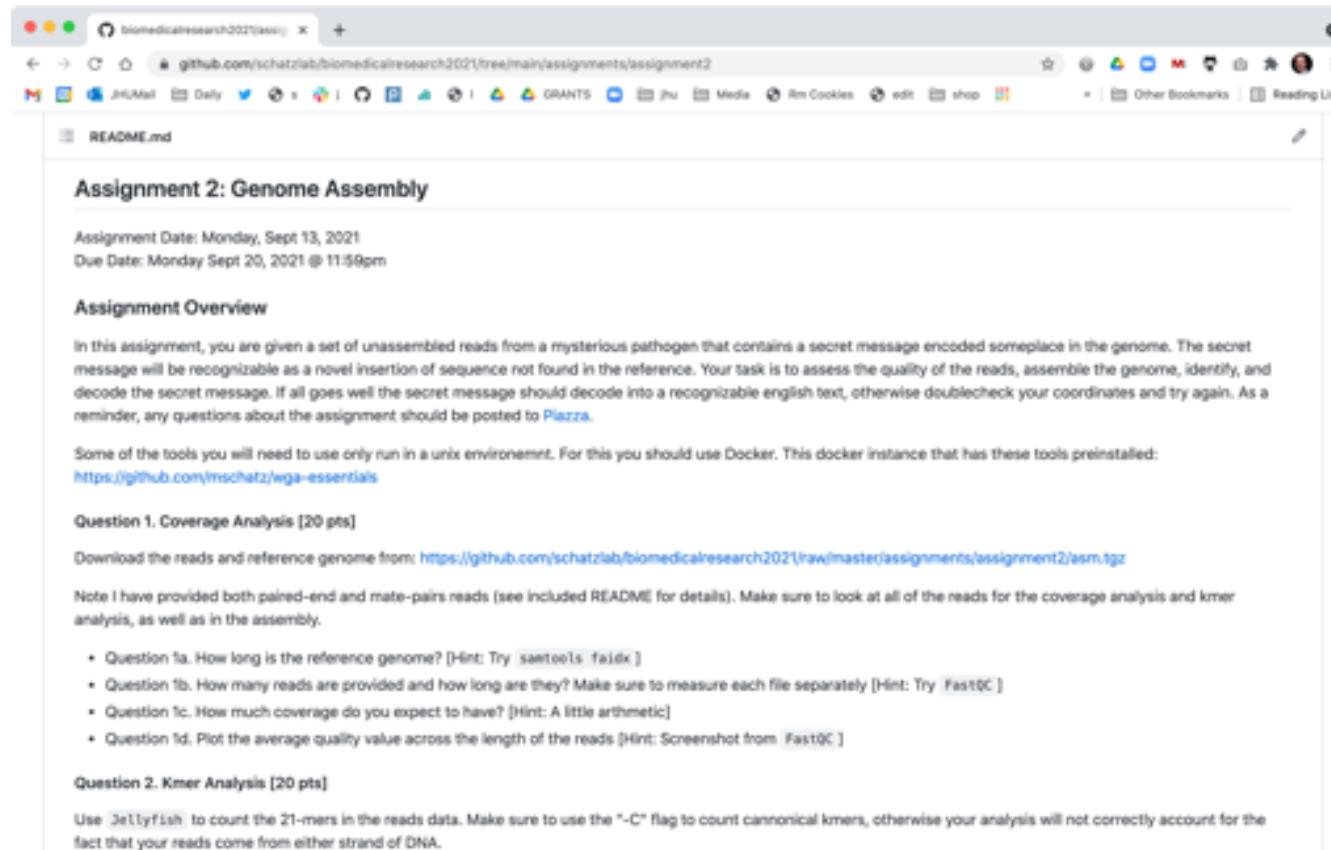
Lecture 5. Computational Biomedical Research



 @mike_schatz

Assignment 2: Genome Assembly

Due Monday Sept 20 @ 11:59pm



The screenshot shows a web browser window with the URL <https://github.com/schatzlab/biomedicalresearch2021/tree/main/assignments/assignment2>. The page displays the contents of the `README.md` file.

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 environment. 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/lsm.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 `saertools 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 arithmetic]
- 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 canonical 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>

Coverage Statistics

$$\text{sequencing_coverage} = \frac{\text{total_bases_sequenced}}{\text{genome_size}}$$

$$\text{genome_size} = \frac{\text{total_bases_sequenced}}{\text{sequencing_coverage}}$$

$$\text{genome_size} = \frac{100\text{Gb}}{50x} = 2\text{Gb}$$

But how can you figure out
the coverage without a genome?

K-mer counting

Kmer-ize

Read 1: GATTACA => GAT, ATT, TTA, TAC, ACA
Read 2: TACAGAG => TAC, ACA, CAG, AGA, GAG
Read 3: TTACAGA => TTA, TAC, ACA, CAG, AGA



GAT	ACA	ACA: 3
ATT	ACA	
TTA	ACA	
TAC	AGA	AGA: 2
ACA	AGA	
TAC	ATT	ATT: 1
ACA	CAG	CAG: 2
CAG	CAG	
AGA	GAG	GAG: 1
GAG	GAT	GAT: 1
TTA	TAC	TAC: 3
TAC	TAC	
ACA	TAC	
CAG	TTA	TTA: 2
AGA	TTA	

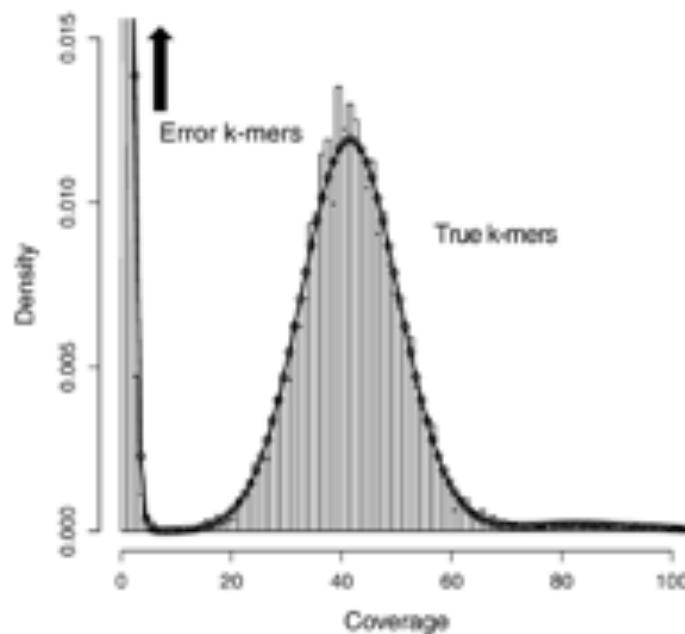
3 kmers occur 1x
3 kmers occur 2x
2 kmers occur 3x



From read k-mers alone, can learn something about how frequently different sequences occur (aka coverage)

Fast to compute even over huge datasets

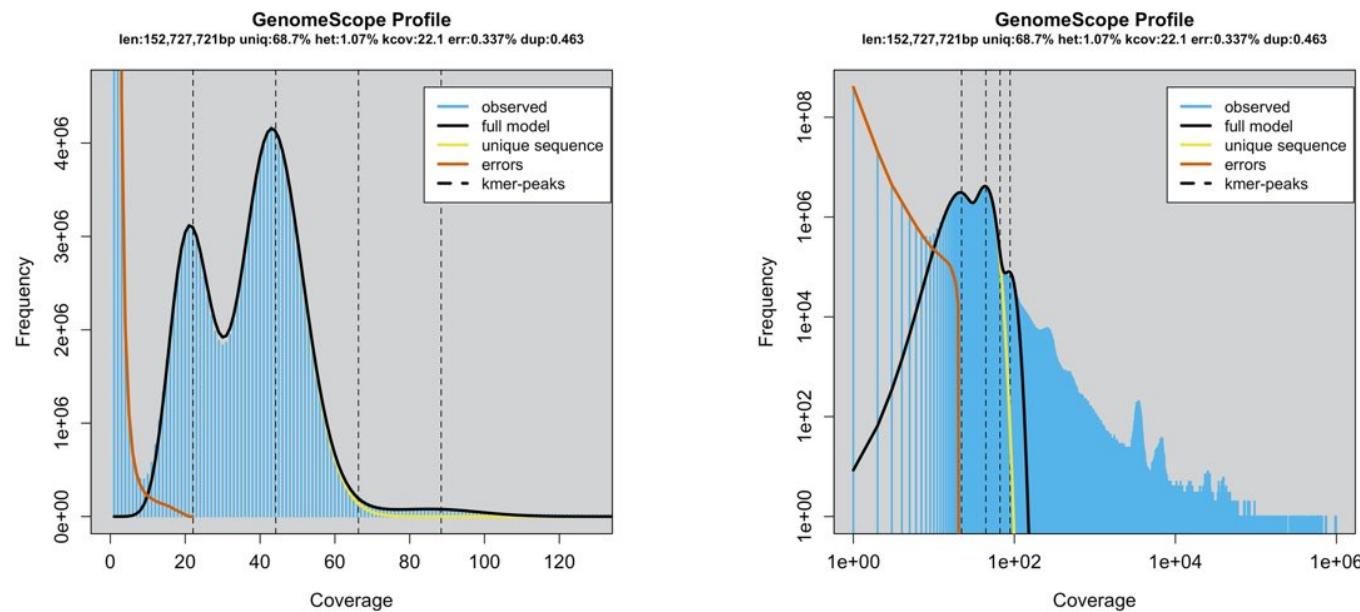
K-mer counting in real genomes



- The tally of k-mer counts in real genomes reveals the coverage distribution.
- Here we sequenced 120Gb of reads from a female human (haploid human genome size is 3Gb), and indeed we see a clear peak centered at 40x coverage
- There are also many kmers that only occur <5 times. These are from errors in the reads
- There are also kmers that occur many times (>>70 times). These are repeats in the genome

GenomeScope: Fast genome analysis from short reads

<http://genomescope.org>



- Theoretical model agrees well with published results:
 - Rate of heterozygosity is higher than reported by other approaches but likely correct.
 - Genome size of plants inflated by organelle sequences (exclude very high freq. kmers)

Vulture, GW*, Sedlazeck FJ*, et al. (2017) *Bioinformatics*
Ranallo-Benavidez, TR. et al. (2020) *Nature Communication*

Pop Quiz I

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

ATTA

GATT

TACA

TTAC

Pop Quiz I

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 I

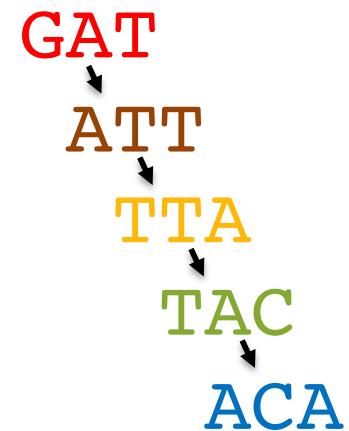
Assemble these reads using a de Bruijn graph approach (k=3):

ATTA: ATT → TTA

GATT: GAT → ATT

TACA: TAC → ACA

TTAC: TTA → TAC



GATTACA

Pop Quiz 2

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

ACGA

ACGT

ATAC

CGAC

CGTA

GACG

GTAT

TACG

Pop Quiz 2

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

~~ACGA~~

ACGT

ATAC

CGAC

CGTA

GACG

GTAT

TACG

ACG
 ↑
CGA

Pop Quiz 2

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

~~ACGA~~

~~ACGT~~

ATAC

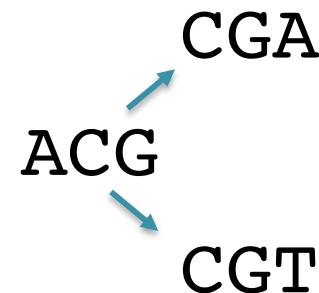
CGAC

CGTA

GACG

GTAT

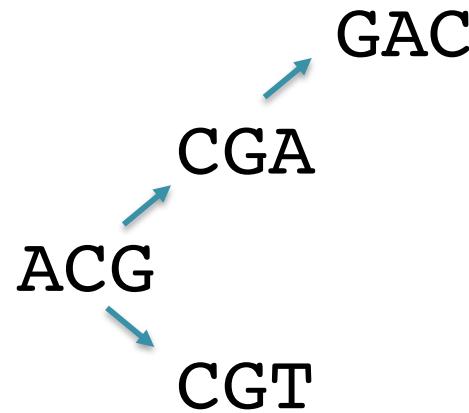
TACG



Pop Quiz 2

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

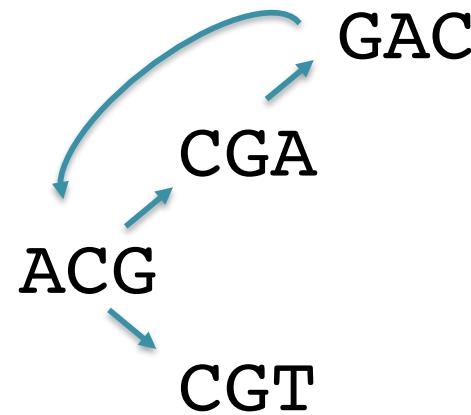
~~ACGA~~
~~ACGT~~
ATAC
~~CGAC~~
CGTA
GACG
GTAT
TACG



Pop Quiz 2

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

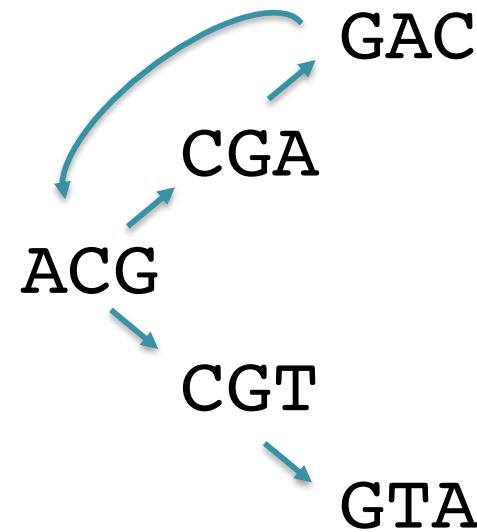
~~ACGA~~
~~ACGT~~
ATAC
~~CGAC~~
CGTA
~~GACG~~
GTAT
TACG



Pop Quiz 2

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

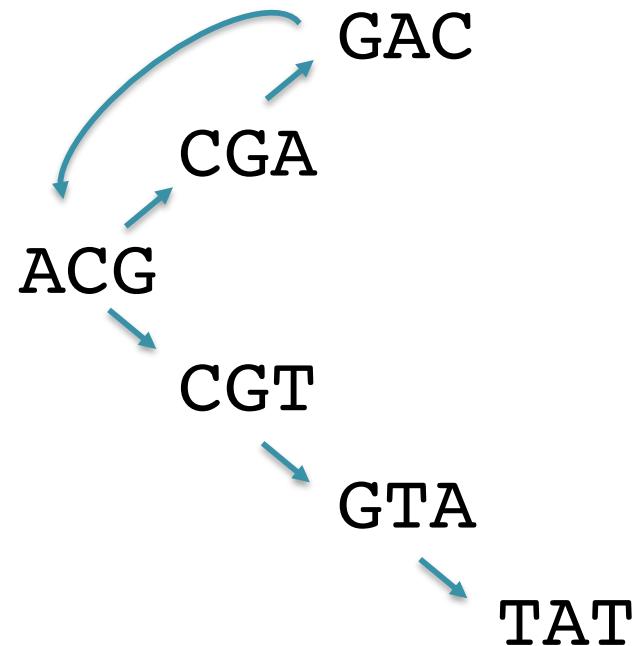
~~-ACGA~~
~~-ACGT~~
ATAC
~~-CGAC~~
~~-CGTA~~
~~-GACG~~
GTAT
TACG



Pop Quiz 2

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

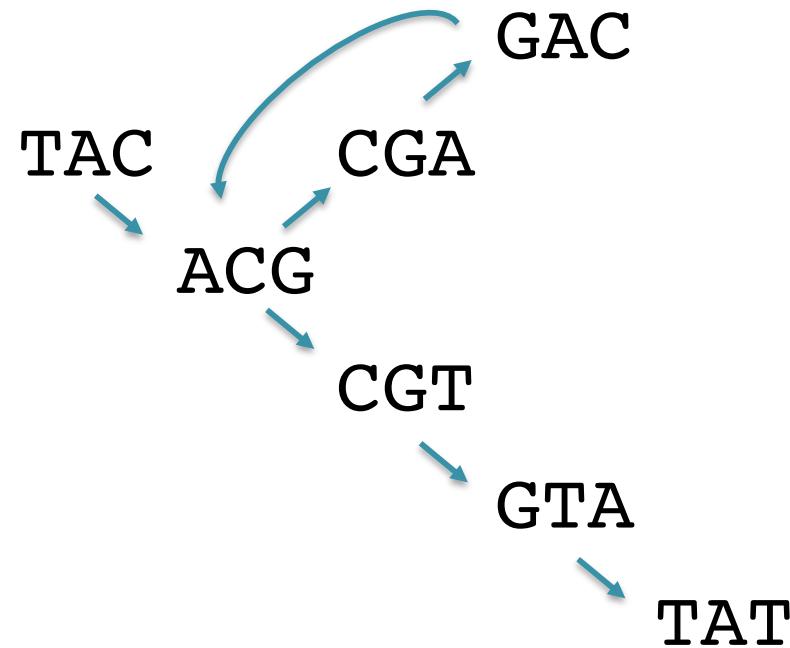
~~-ACGA~~
~~-ACGT~~
ATAC
~~-CGAC~~
~~-CGTA~~
~~-GACG~~
~~-GTAT~~
TACG



Pop Quiz 2

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

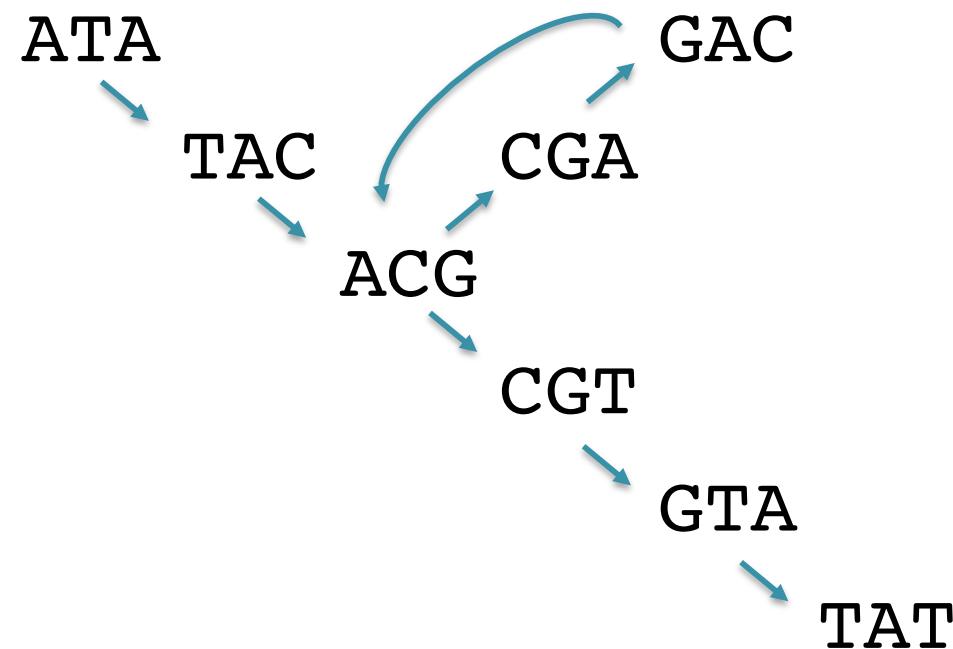
~~ACGA~~
~~ACGT~~
ATAC
~~CGAC~~
~~CGTA~~
~~GACG~~
~~GTAT~~
~~TACG~~



Pop Quiz 2

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

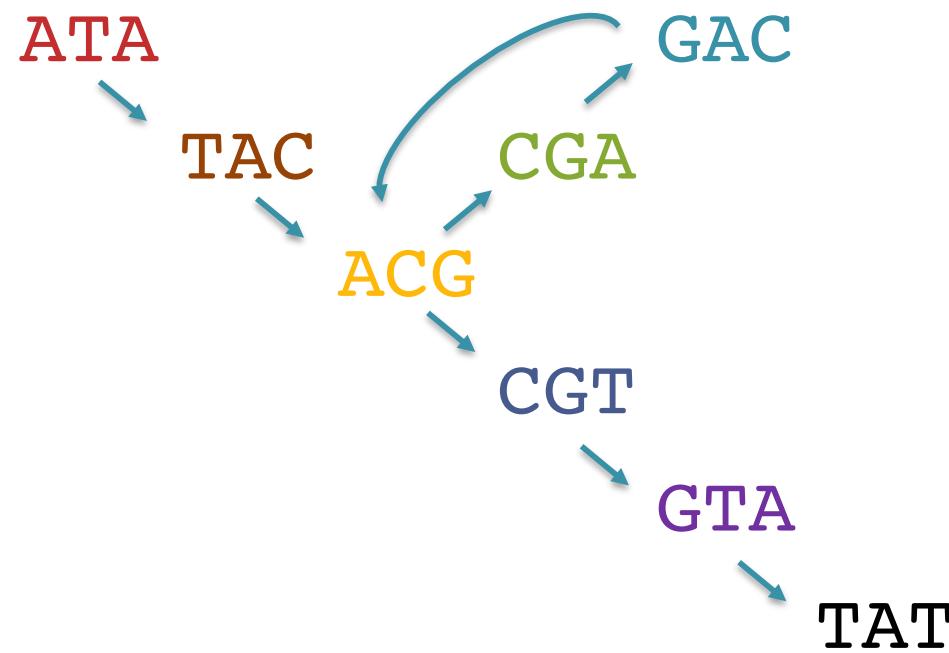
~~ACGA~~
~~ACGT~~
~~ATAC~~
~~CGAC~~
~~CGTA~~
~~GACG~~
~~GTAT~~
~~TACG~~



Pop Quiz 2

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

~~ACGA~~
~~ACGT~~
~~ATAC~~
~~CGAC~~
~~CGTA~~
~~GACG~~
~~GTAT~~
~~TACG~~

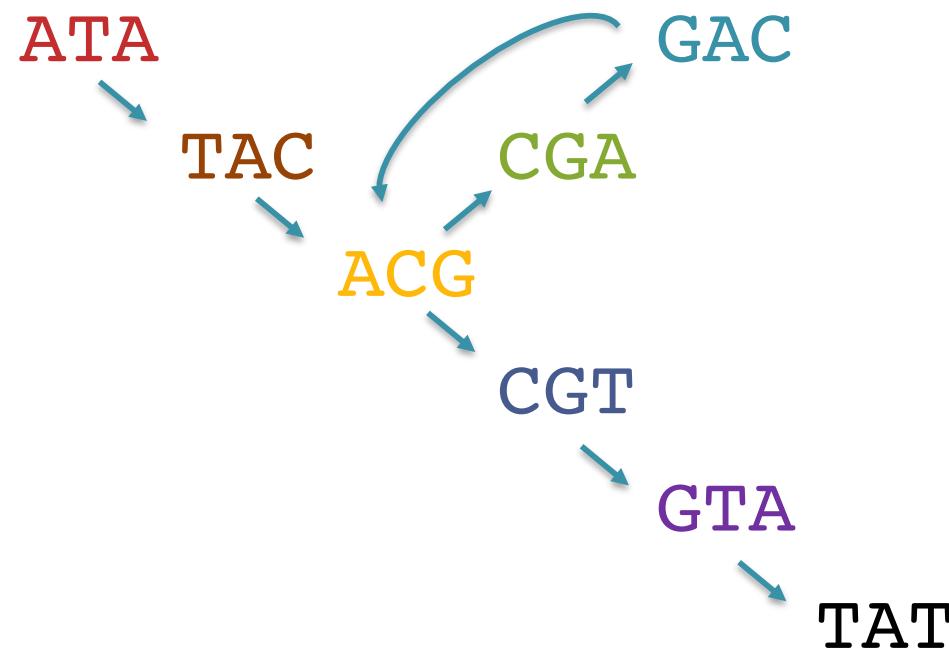


ATACGACGTAT

Pop Quiz 2

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

~~ACGA~~
~~ACGT~~
~~ATAC~~
~~CGAC~~
~~CGTA~~
~~GACG~~
~~GTAT~~
~~TACG~~



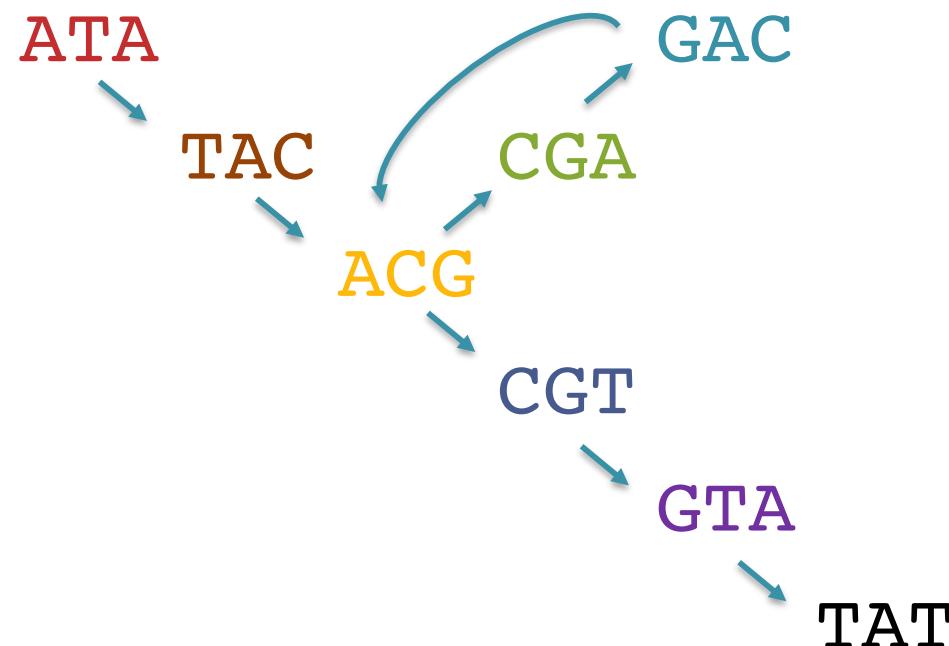
Whats another possible genome?

ATACGACGTAT

Pop Quiz 2

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

~~ACGA~~
~~ACGT~~
~~ATAC~~
~~CGAC~~
~~CGTA~~
~~GACG~~
~~GTAT~~
~~TACG~~



Should we add the edge TAT → ATA?

ATACGACGTAT

The human genome



"Without a doubt, this is the most important, most wondrous map ever produced by humankind."

Bill Clinton
June 26, 2000

Who is the reference human?

The Buffalo News-Sunday, March 29, 1997

ment abuse, civil disobedience

topic. But the very nature of government creates a mind set that inspires increased their authority, above all else of the people," Perkins said.
"The government has interests that aren't of the people," Perkins added, noting more than 90% of the money "goes to and the Lapp states on sliding scale minimum civil disobedience based on being respectful in one's resistance." Barbara Lapp said if we claim to care about our rights, we must protect government institutions.

Violence has to be watched over, and, calling civil disobedience the act of the masses against movements, or violence can serve as an anti-government opposition, he added.
"Law is applied or you're given an ultimatum or legal authority.

Rachel Lapp says she believes government can be good, when it controls the aggression in society. Instead, it too often comes down on the side of the aggressor, who enforces child-protection laws, compulsory education, discipline rules on teachers and seat belt laws.

"We want people to see the correlation between what happened to us and what can happen to anyone when government gets out of hand," Rachel Lapp said.

The Lapp and Perkins will be joined by former Roswell (IL) mayor of public education who was arrested and pleaded guilty to reduced charges following a 1995 disturbance at the City Campus of Elgin Community College.

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.

For more information please contact the
Clinical Genetics Service
645-5720 (9:00 am - 3:00 pm)
March 24 - 26, 1997

ROSWELL
CANCER INSTITUTE

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.

For more information please contact the
Clinical Genetics Service
645-5720 (9:00 am - 3:00 pm)
March 24 - 26, 1997

Pieter de Jong, RPCI

Who is the reference human?



Pieter de Jong, RPCI

Who is the reference human?

The Buffalo News-Sunday, March 29, 1997

ment abuse, civil disobedience

topic. But the very nature of government creates a mind set that inspires increase their authority, above all else of the people," Perkins said. "It's government has forgotten that most of the people." Perkins added, "nothing more than it's the master's to tell the Lapp story on sliding non-violent civil disobedience based on being respectful in one of resistance." Perkins told Lapp that if we claim to care about our rights, we must protect government institutions.

Violence has to be watched over, and, calling civil disobedience the act of the violent nature movement, no violence can serve as an anti-government opposition, he added. "Law is applied or you're given an ultimatum or legal authority,

"you should refuse it," Perkins said. "And, if need be, you have to be brave enough to accept the consequences."

Rachel Lapp says she believes government can be good, when it controls its aggression in society. Instead, it too often comes down on the side of the aggressor, who enforces child-protection laws, compulsory education, discipline rules on law enforcement and war-fight laws.

"We want people to see the correlation between what happened to us and what can happen to anyone when government puts us out of hand," Rachel Lapp said.

The Lapp and Perkins will be joined by former Student DE, a year of public education who was arrested and pleaded guilty to reduced charges following a 1995 disturbance at the City Campus of Erie Community College.

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 genome) that describes all individual traits, interests, from genes. The resources for the project will have enormous impact on future progress of medical science and lead to improved diagnosis and treatment of hereditary diseases.

Individuals will receive information about the project from the CTD (Central Database for the Human Genome Project), and sign a consent form before participating.

An personal information will be maintained as anonymous.

Volunteers will provide a small amount of a small blood sample - a small quantity information will be provided to the participants for their own use.

Individuals must be at least 18 years of age.
Individuals who have undergone chemotherapy are ineligible.

For more information please contact:
Christopher D. Ladd
1000 University Street
Seattle, WA 98101
(206) 543-1367

Dowell R&P

Appendix: Identifying the ancestry of segments of the human genome reference sequence

To compare Neandertal to present-day human haplotypes for the purpose of population genetic analysis, we needed to have long haploid sequences from present-day humans that were of known ancestry. To identify such segments, we took advantage of the fact that the human reference sequence is haploid over scales of tens of kilobases, because it is composed of a tiling-path of Bacterial Artificial Chromosomes (BACs) or other clone types that are of typical size 30-150 kb (38). We do not know of any other substantial source of high quality human haploid sequences of the requisite size.

Determining the ancestry of the libraries in the human genome reference sequence using HAPMIX

It is crucial to know the "ancestry" of a clone to use it in a meaningful population genetic analysis. In what follows, we define "ancestry" as the geographic region in which a clone's ancestor lived 1,000 years ago, inferred based on its genetic proximity to other individuals from that region today. This definition allows us to classify clones from Chinese Americans as "East Asian," from European Americans as "European," and from African Americans as either "West African" or "European".

To identify the ancestries of the libraries comprising most of the human genome reference sequence, we used a list of 26,358 clones tiling the great majority of the genome, most of which we were able to assign to a library of origin. Restricting to the telomeres, we identified 21,156 clones that seemed to fall into 9 libraries based on the same

RPCI-3 (n=456) examination, we from RPCI-11, 8 clones we analyze

1. **RPCI-11 is an African American:** RPCI-11, the individual who contributed most of the human genome reference sequence, is consistent with having African American ancestry, with 42% of the clones of confident West African ancestry and 42% of the clones of confident European ancestry, and the ancestry of the remaining clones less confidently inferred. The finding of likely African American ancestry for RPCI-11 was previously reported in a study of the ancestry of RPCI-11 clones spanning the Duffy blood group locus (59), and here we confirm this finding, and also expand the inference to the whole genome.
2. **CTD is an East Asian:** The majority of clones from CTD, the second largest library in its contribution to the human genome sequence, is likely an East Asian. In a HAPMIX analysis with CEU (European) – CHB+JPT (East Asian) as the proposed ancestral populations, the majority of clones are of confident East Asian origin, and there is no secondary mode of confident European ancestry, as might be expected from a Latino or South Asian individual.
3. **The remaining 7 libraries are European:** The remaining libraries (CTA, CTB, CTC, RPCI-1, RPCI-3, RPCI-4 and RPCI-5) are inferred to be of European ancestry, since they all have consistent distributions of inferred clone ancestries, with the majority of clones of confident European ancestry in both our HAPMIX analyses and no secondary modes.

A Draft Sequence of the Neandertal Genome
Green et al (2010) Science. DOI: 10.1126/science.1188021
Supplemental Note 16 (pg 145-146)

Pieter de Jong, RPCI

Genome Reference Consortium

GRC Home Beta 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



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)

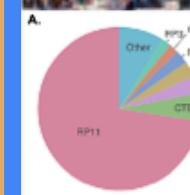
GRC trackhub:
http://img.sanger.ac.uk/production/gtctrack_hubhub.txt

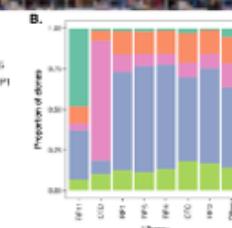
Slides from GRC presentations in ASHG 2019

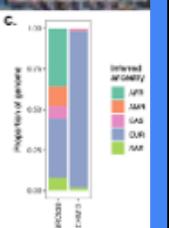
Transitioning to GRCh38! Try the NCBI Remapping Service, which uses the same assembly–assembly alignments used by the GRC.

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 genome currently in development.

A. 

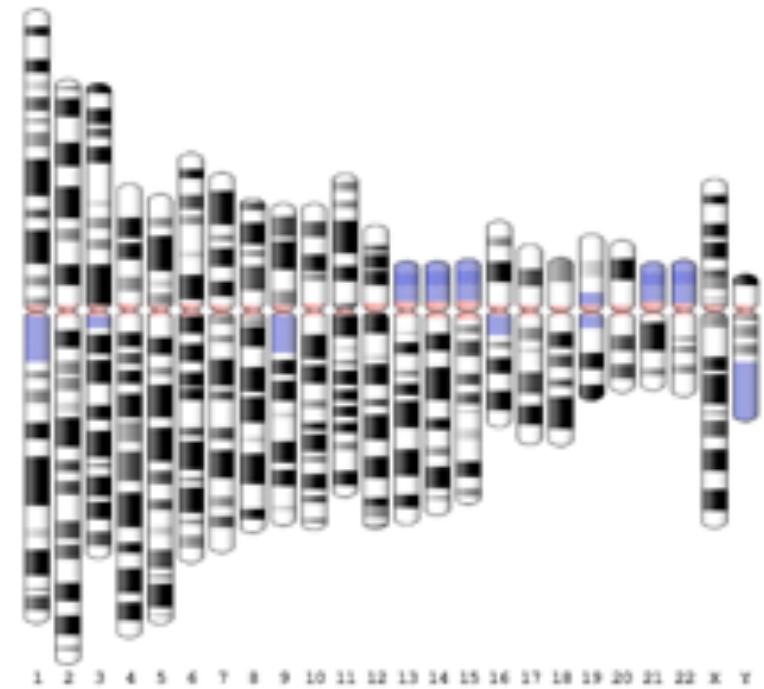
B. 

C. 

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

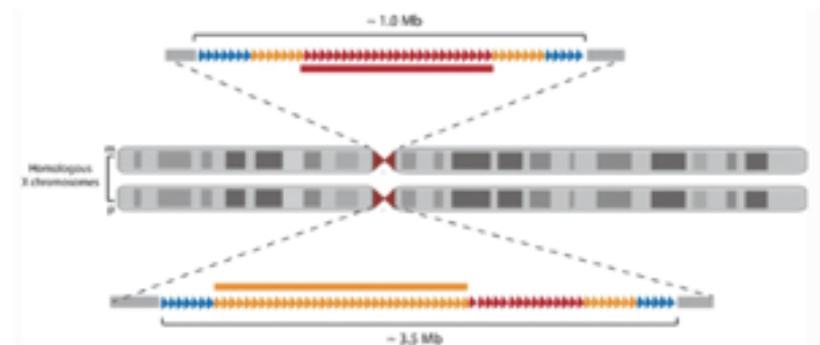
Functional studies need sequence

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**



Let's finish a human genome



A karyotype image showing human chromosomes arranged in three rows, each row containing pairs of chromosomes. The chromosomes are stained with fluorescent dyes, showing various colors including pink, green, yellow, and purple. The background is black.

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



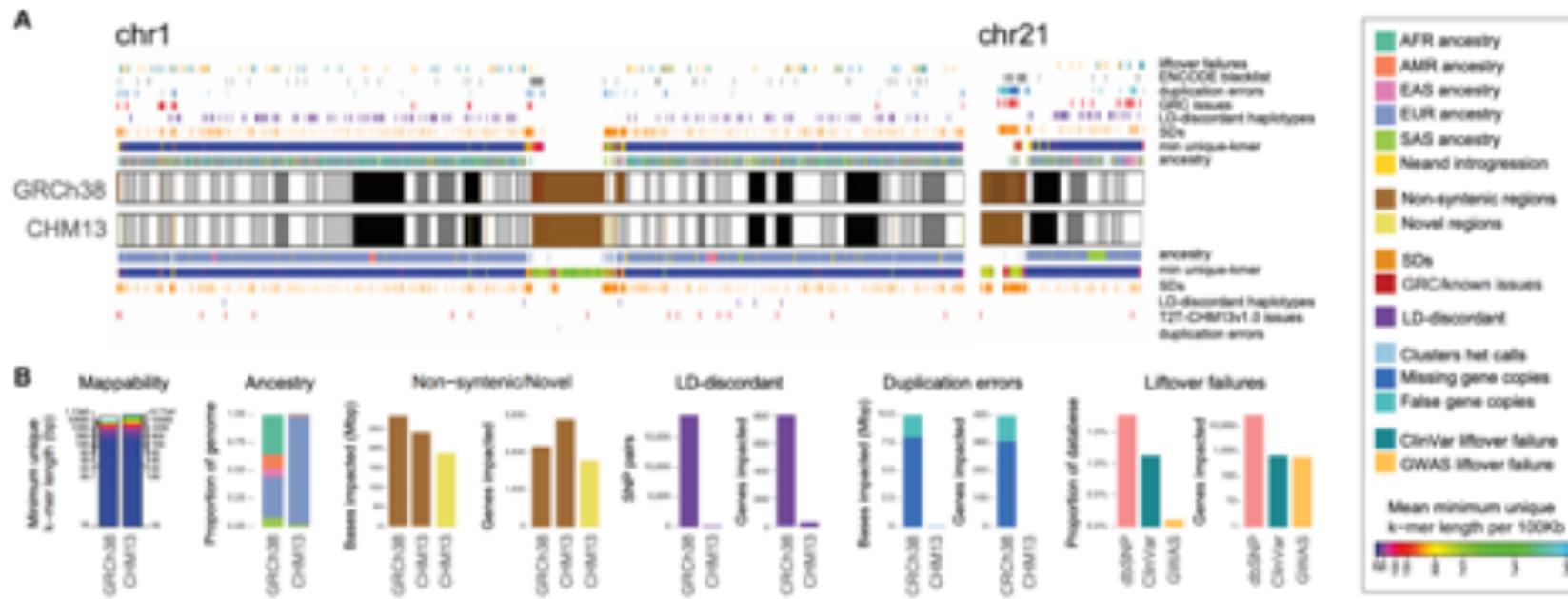
CHM13 HiFi assembly graph



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



A photograph of a young woman with dark hair, smiling at the camera. She is wearing a blue zip-up jacket. In the background, the Golden Gate Bridge is visible across a body of water. Other people can be seen walking along a path near the bridge.

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



Preview

1 hour 4 hours 1 day

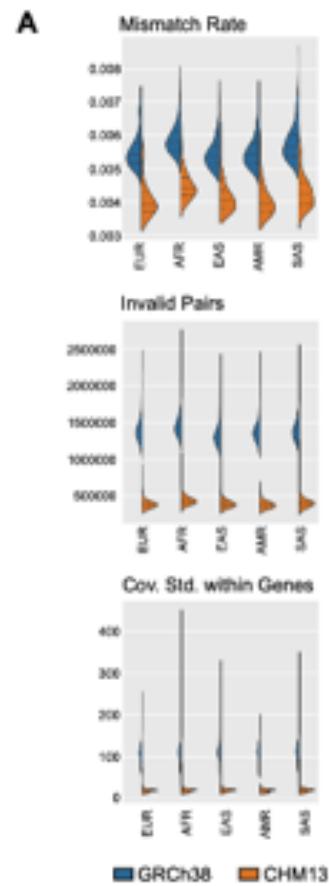
Samantha Zarate



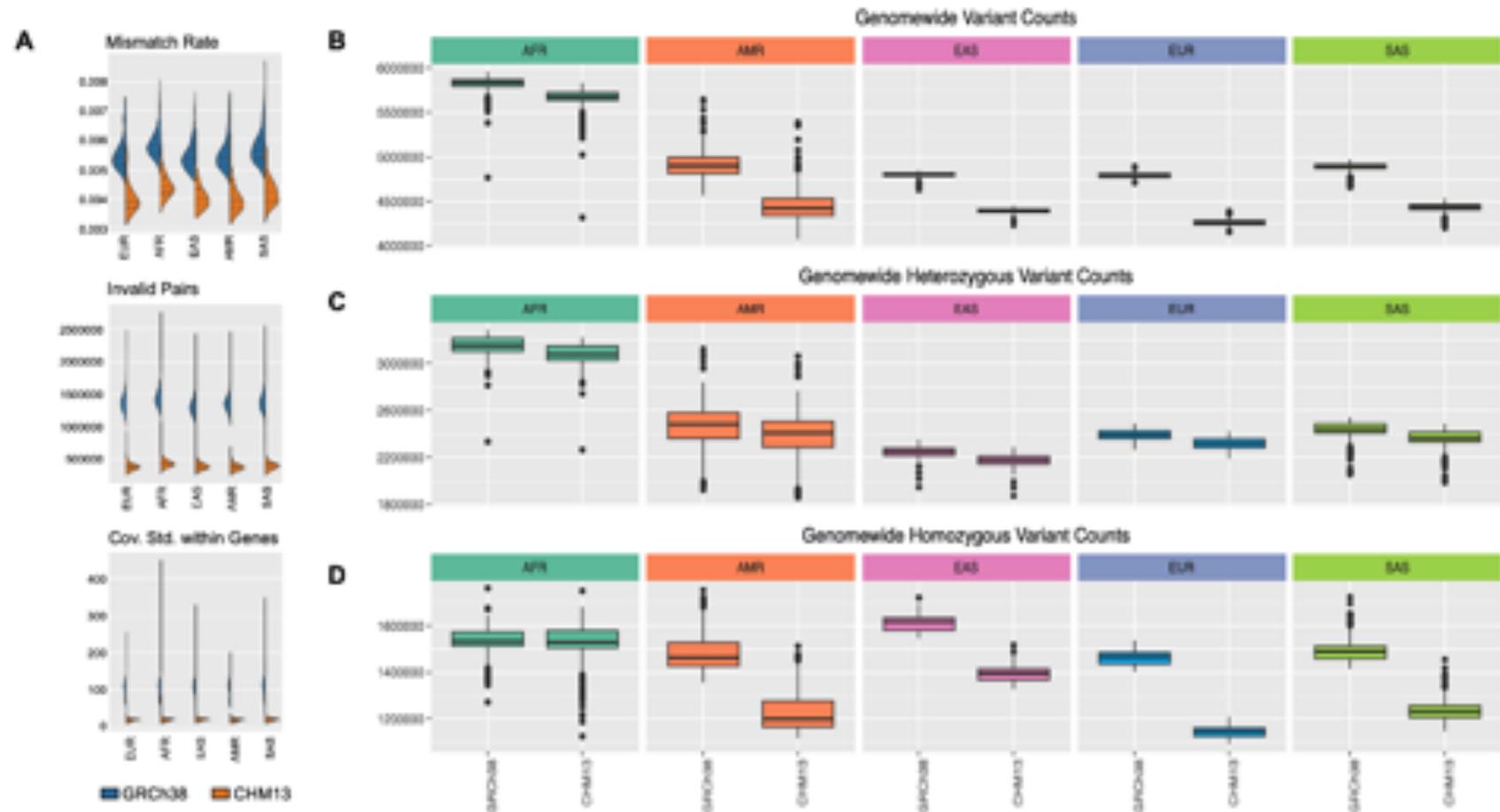
● instance/cpu/reserved_cores: 11,552.00

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)

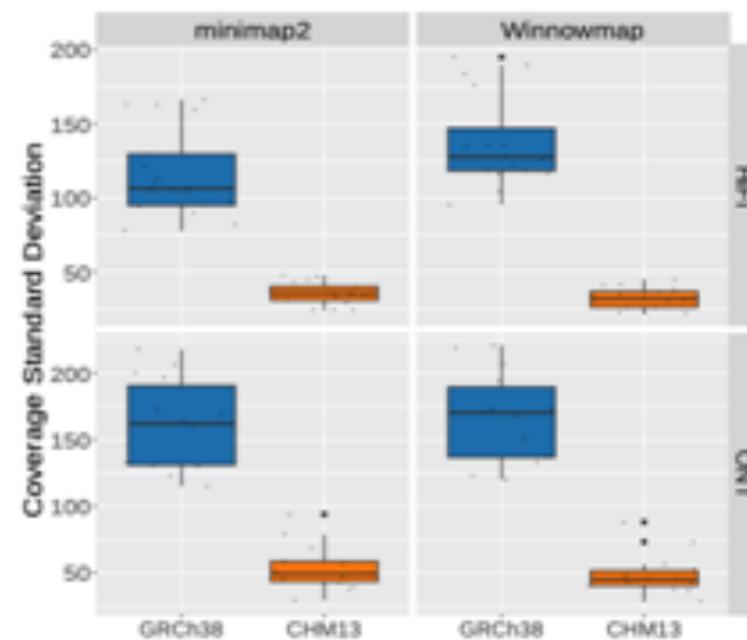
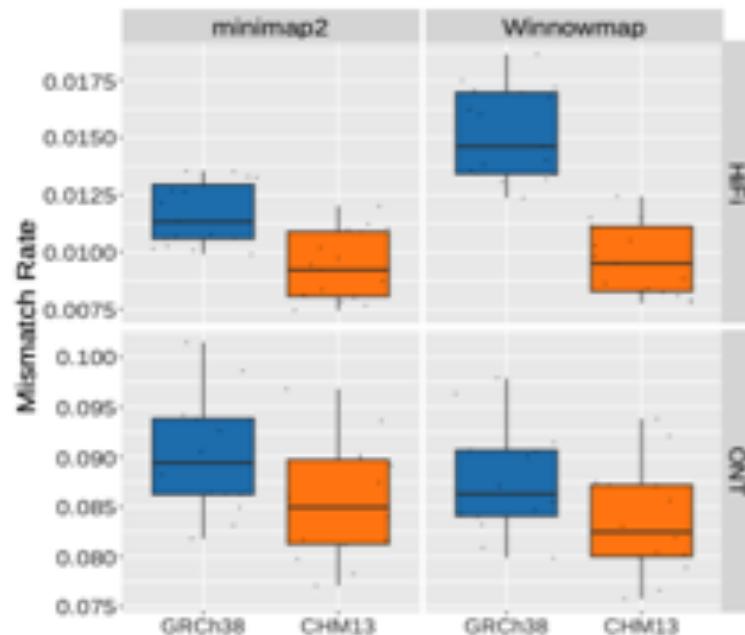


Melanie Kirsche



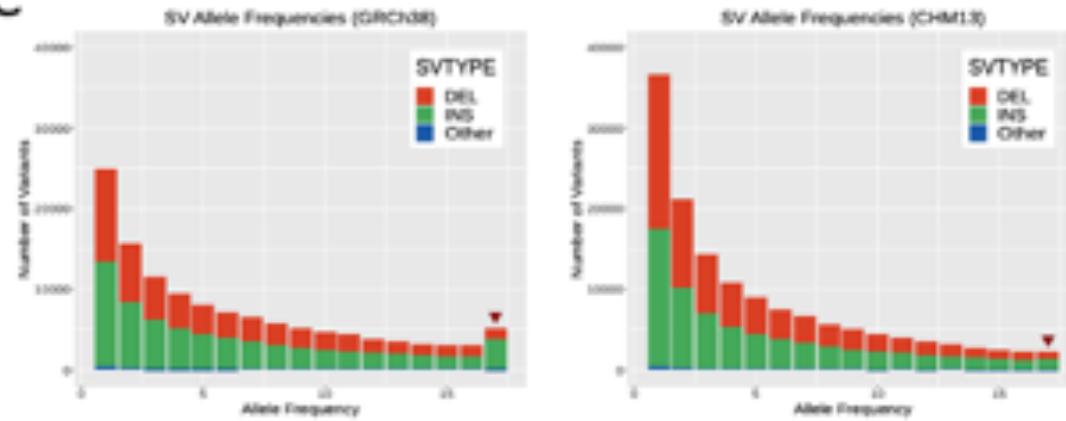
Sergey Aganezov

B

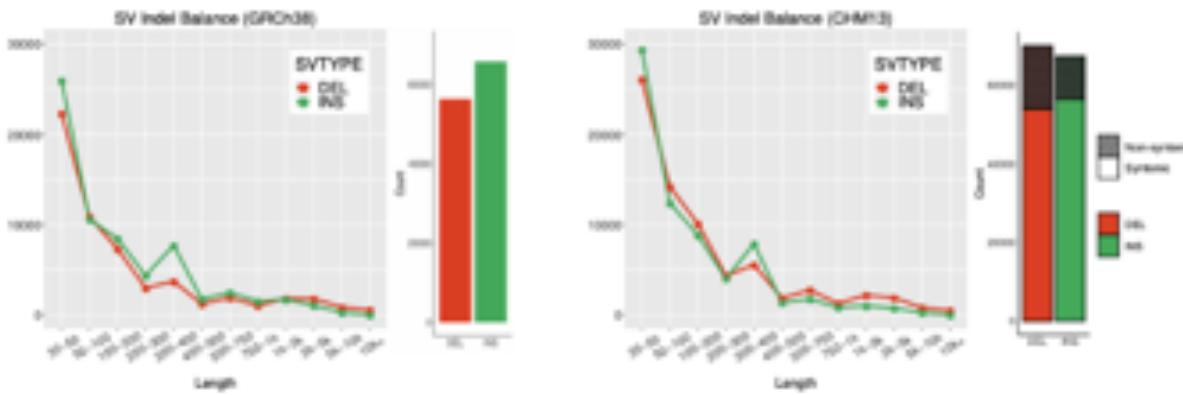


Long Read Analysis with T2T-CHM13 (Fig 3)

C



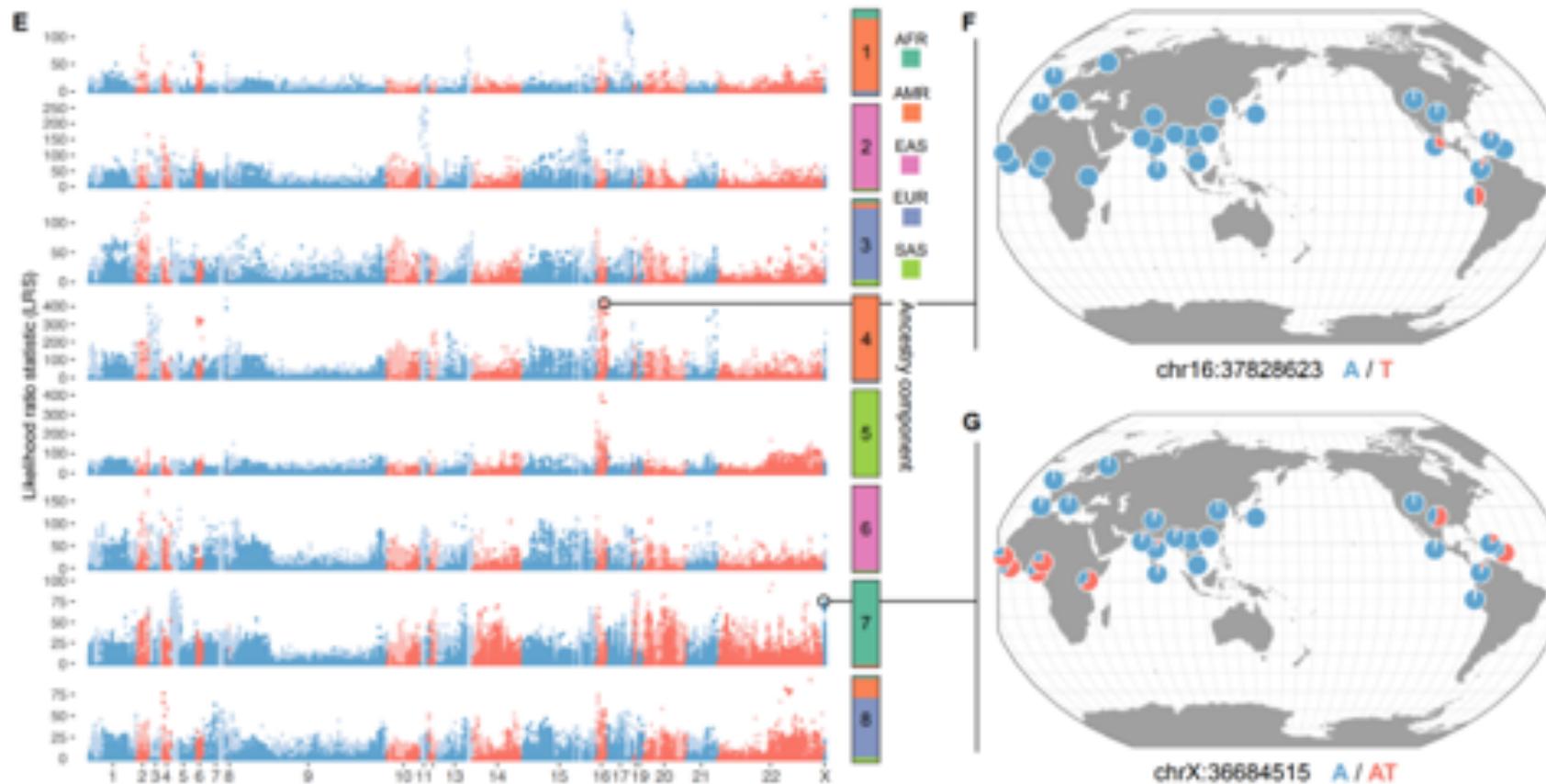
D



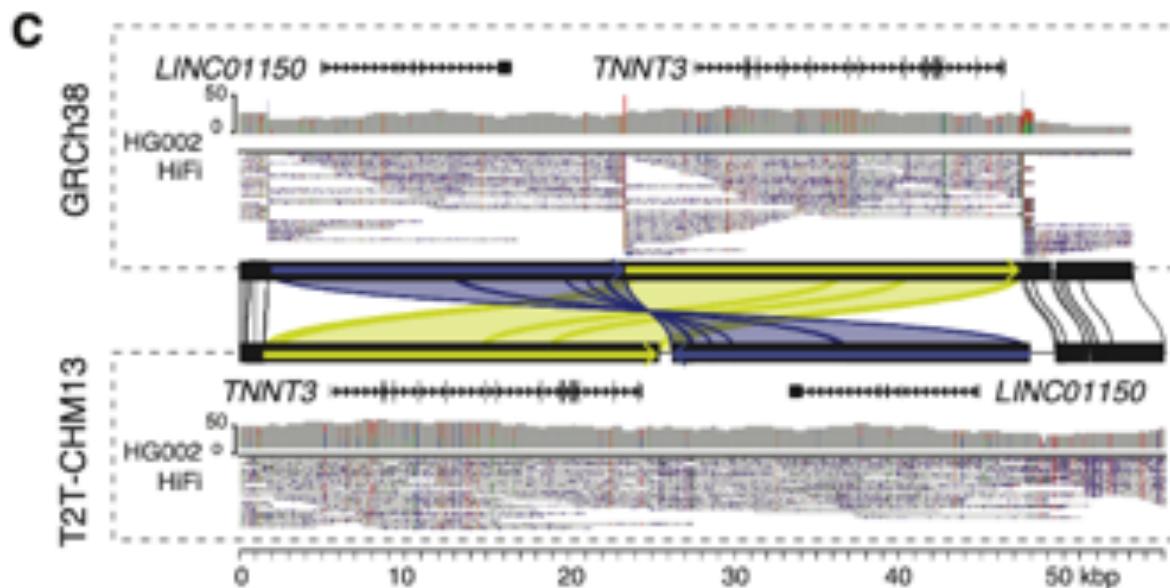
Novel Variants from 1000G on T2T-CHM13 (Fig 4)



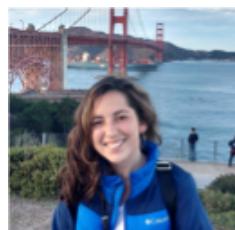
Stephanie
Yan



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



Danny Miller



Daniela Soto



Megan Dennis



Justin Zook

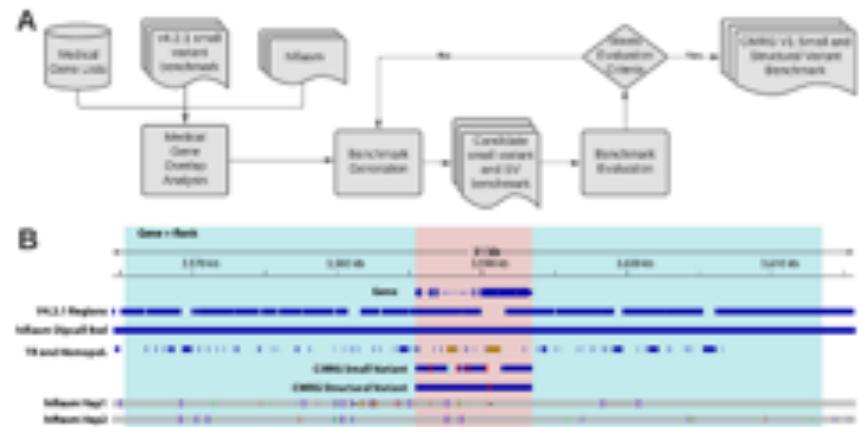


Fritz Sedlazeck

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

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

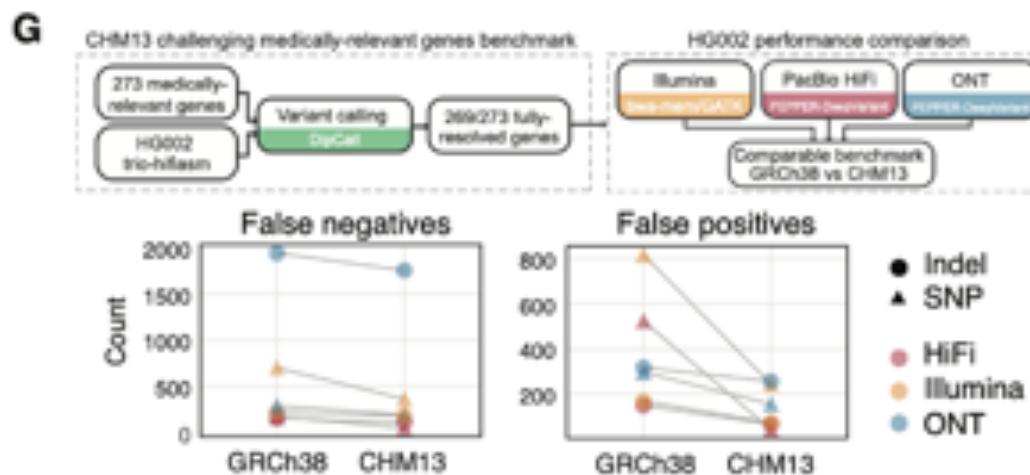
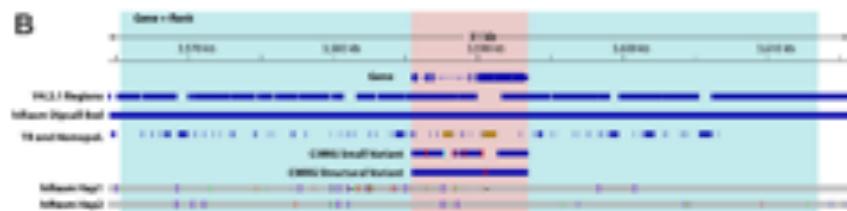
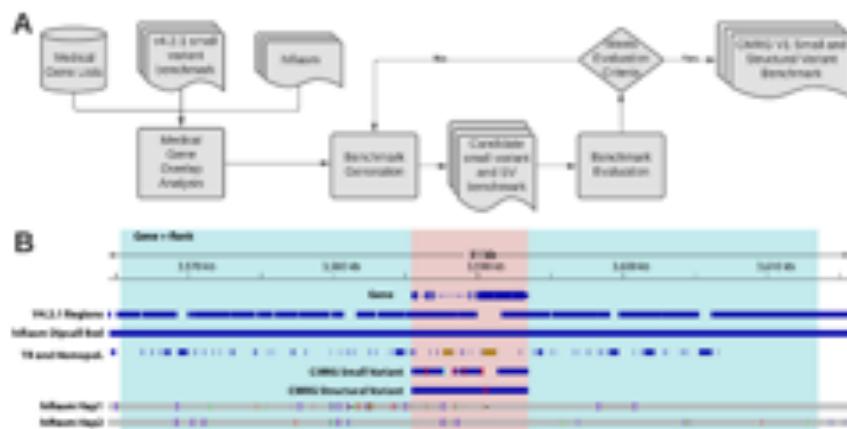
Justin Wagner,¹ Nathan D Olson,¹ Lindsay Harris,¹ Jennifer McDaniel,¹ Haoyu Cheng,² Arkarachai Fungtammasan,³ Yih-Chii Hwang,³ Richa Gupta,³ Aaron M Wenger,⁴ William J Rowell,⁴ Ziad M Khan,⁵ Jesse Farek,⁶ Yiming Zhu,⁶ Aishwarya Pisupati,⁶ Medhat Mahmoud,⁶ Chunlin Xiao,⁶ Byunggi Yoo,⁶ Sayed Mohammad Ebrahim Sahraeian,⁶ Danny E. Miller,⁹ David Jáspež,¹⁰ José M. Lorenzo-Salazar,¹⁰ Adrián Muñoz-Barrera,¹⁰ Luis A. Rubio-Rodríguez,¹⁰ Carlos Flores,¹¹ Giuseppe Narzisi,¹² Uday Shanker Evani,¹² Wayne E. Clarke,¹² Joyce Lee,¹³ Christopher E. Mason¹⁴, Stephen E. Lincoln¹⁵, Karen H. Miga¹⁶, Mark T. W. Ebbert,¹⁷ Alaina Shumate,¹⁸ Heng Li,² Chen-Shan Chin^{*3}, Justin M Zook^{*1}, Fritz J Siedlazeck^{*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,⁷ Byunggil Yoo,⁷ Sayed Mohammad Ebrahim Sahraeian,⁸ Danny E. Miller,⁹ David Jásperz,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. Mason¹⁴, Stephen E. Lincoln¹⁵, Karen H. Miga¹⁶, Mark T. W. Ebbert,¹⁷ Alaina Shumate,¹⁸ Heng Li,² Chen-Shan Chin*,² Justin M Zook*,¹ Fritz J Sedlazeck*,¹



The complete sequence of a human genome

Sergey Nurk, Sergey Koren, Arang Rhee, Mikko Rautiainen, Andrey V. Balakadze, Alia Milkeenko, Mitchell R. Vollger, Nicolas Altemose, Lev Uralksky, Ariel Gershman, Sergey Aganezov, Savannah J. Hoyt, Mark Diekhans, Glennis A. Logsdon, Michael Alonso, Stylianos E. Antonarakis, Matthew Borchers, Gerard G. Bouffard, Sheline 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 Formisano, 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, Milen Jain, Erich D. Jarvis, Peter Karpedjian, Melanie Kirsch, Mikhail Kolmogorov, Jonas Korlach, Milian Kremitzki, Heng Li, Valerie V. Haduro, Tobias Marschall, Ann M. McCarrone, Jennifer McDaniel, Danny E. Miller, James C. Mullikin, Eugene W. Myers, Nathan D. Olson, Benedict Paten, Paul Peluso, Pavel A. Pevzner, David Porubsky, Tatjana Potapova, Evgeny I. Rozaev, Jeffrey A. Rosenfeld, Steven L. Salzberg, Valerie A. Schneider, Fritz J. Sedlacek, Kishwar Shafin, Colin J. Shew, Alaina Shumate, Yumi Sims, Arian F.A. Smit, Daniela C. Soto, Ivan Sovic, Jessica M. Storer, Aaron Streets, Beth A. Sullivan, Francoise Thibaud-Nissen, James Torrance, Justin Wagner, Brian P. Walenz, Aaron Wagner, Jonathan M. D. Wood, Chunlin Xiao, Stephanie M. Yan, Alice C. Young, Samantha Zarate, Urvashi Suri, 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 Rhee, Melissa Meredith, Skylar Martin, Joyce Lee, Sergey Koren, Jeffrey A. Rosenfeld, Benedict Paten, Ryan Layer, Chen-Shan Chin, Fritz J. Sedlacek, 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. Balakadze, Pragna Sidhwani, Sasha A. Langley, Gina V. Caldas, Savannah J. Hoyt, Lev Uralksky, Fedor D. Ryabov, Colin J. Shew, Michael E.G. Sauria, Matthew Borchers, Ariel Gershman, Alia Milkeenko, Valery A. Shepelev, Tatjana Dvorkina, Olga Kurnasikaya, Mitchell R. Vollger, Arang Rhee, Ann M. McCartney, Mobiin Asri, Ryan Lorig-Roach, Kishwar Shafin, Sergey Aganezov, Daniel Olson, Leonardo Gomes de Lima, Tatjana Potapova, Gabrielle A. Hartley, Marina Haukness, Peter Karpedjian, Fedor Gusarov, Kristoff Tigray, Sheline Brooks, Alice Young, Sergey Nurk, Sergey Koren, Sofie R. Salama, Benedict Paten, Evgeny I. Rozaev, 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 Rhee, 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

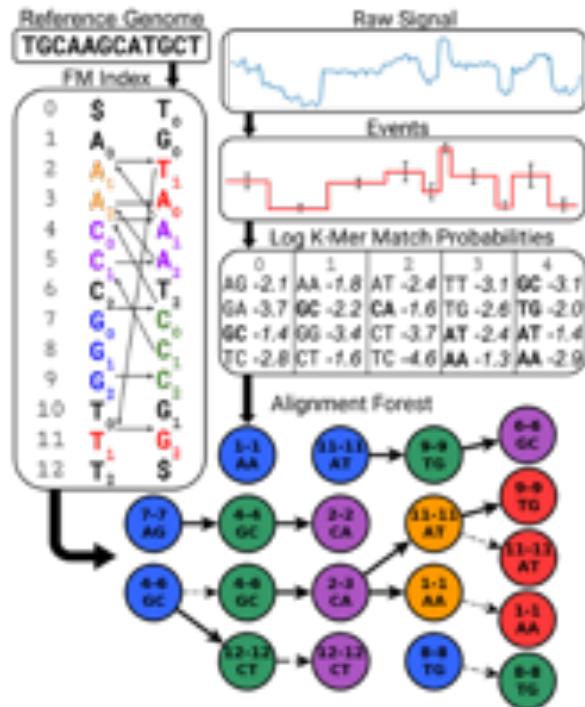
Mitchell R. Vollger, Xavi Guitart, Philip C. Dishuck, Ludovica Mercuri, William T. Harvey, Ariel Gershman, Mark Diekhans, Arvis Sulovari, Katherine M. Munson, Alexandra M. Lewis, Kendra Hoekzema, David Porubsky, Ruiyang Li, Sergey Nurk, Sergey Koren, Karen H. Miga, Adam M. Phillippy, Winston Timp, Mario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

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

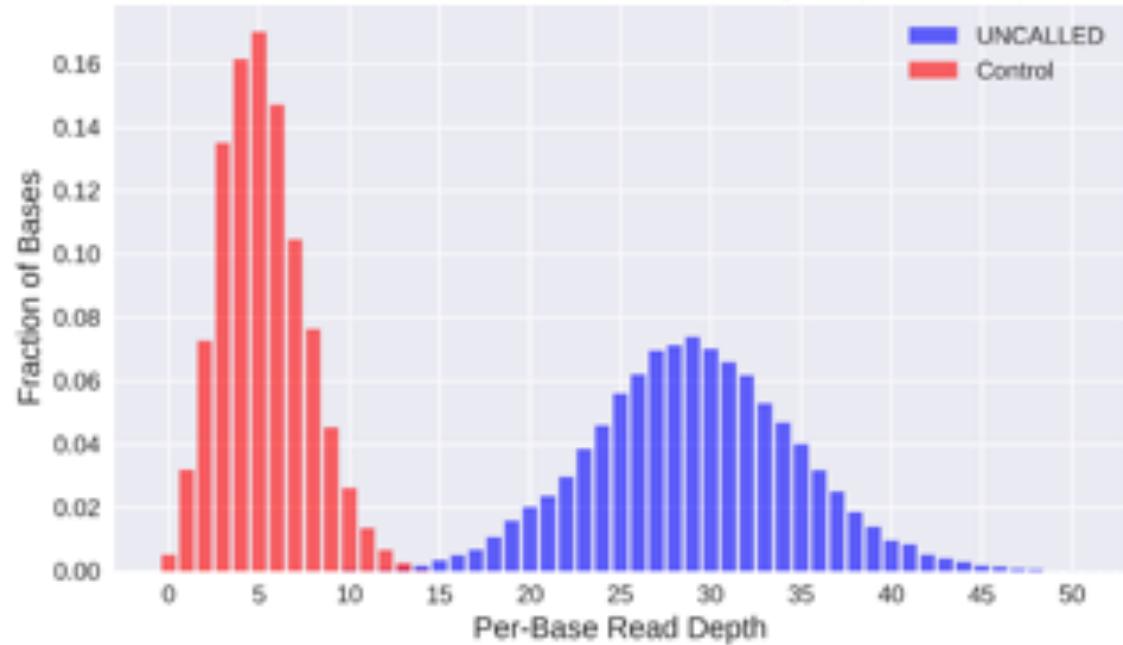
What's next?

UNCALLED

Utility for Nanopore Current Alignment to Large Expanses of DNA

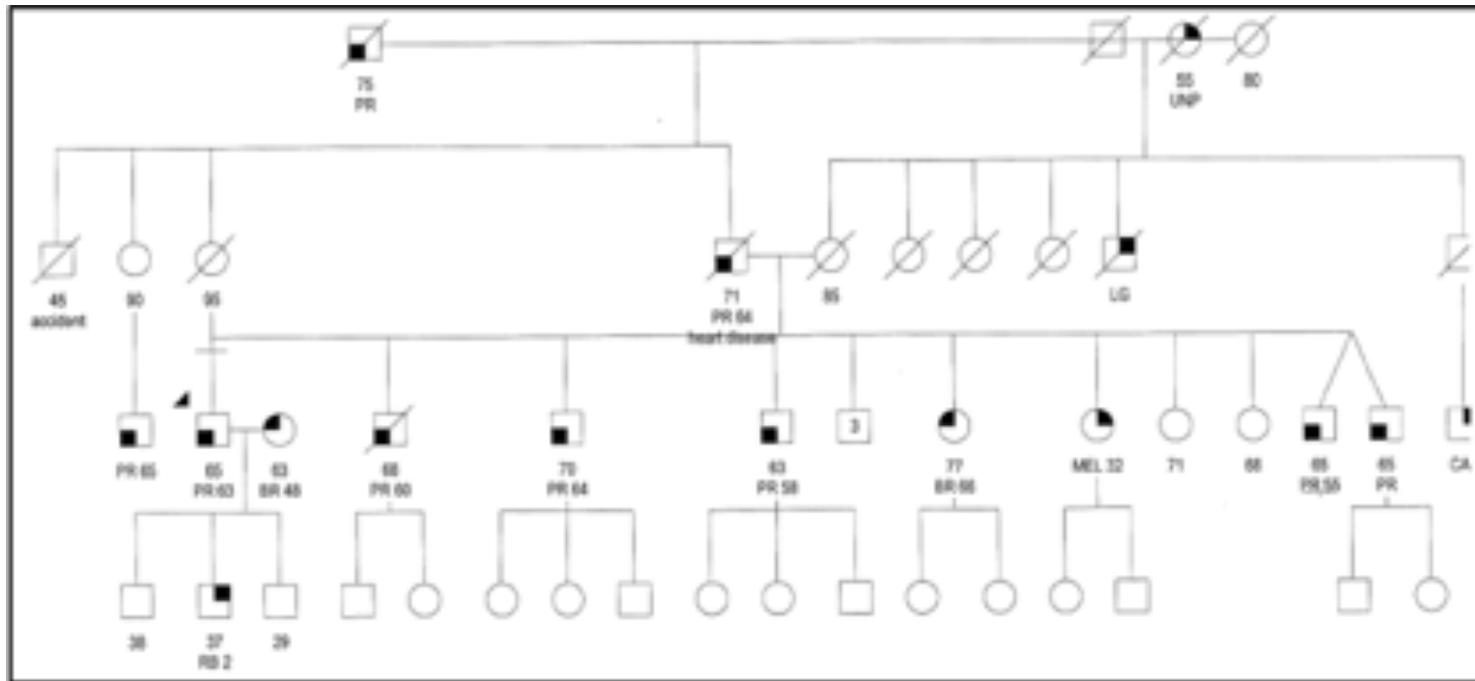


Invitae pan-cancer hereditary risk genes (148 genes)



Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED
Kovaka et al. (2020) Nature Biotechnology doi: <https://doi.org/10.1038/s41587-020-0731-9>

What causes “outlier” families?



The Mark Foundation®
for Cancer Research

Eli Van Allen,
Dana-Farber Cancer Institute

T2T Across the Tree of Life

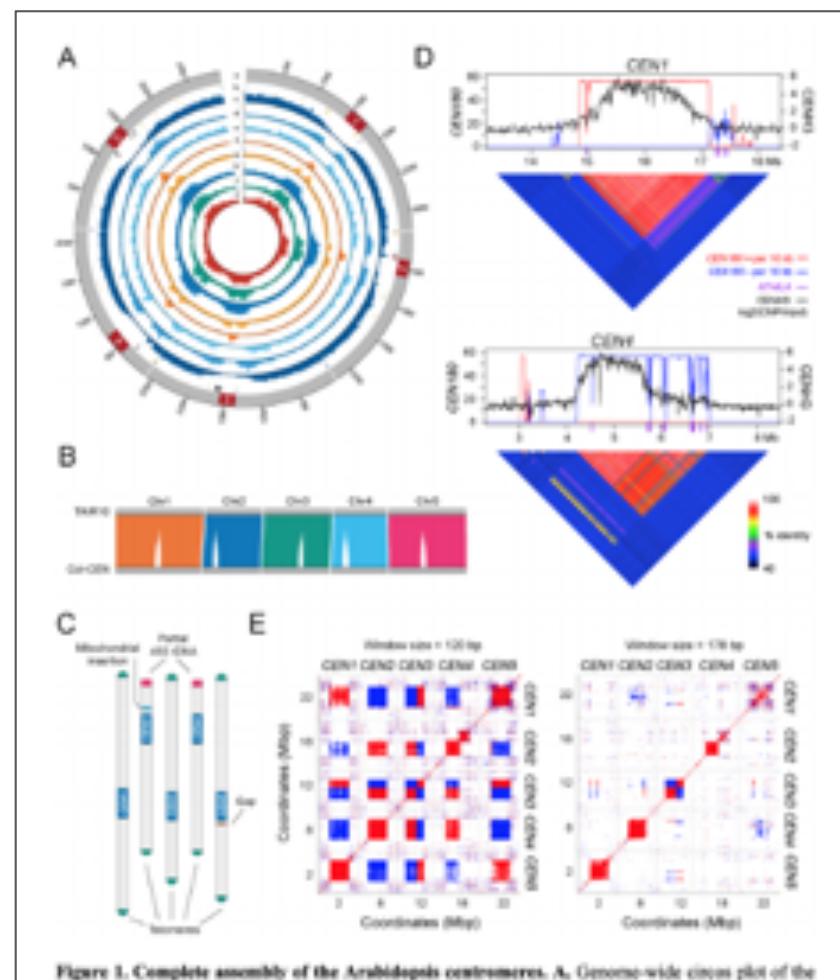
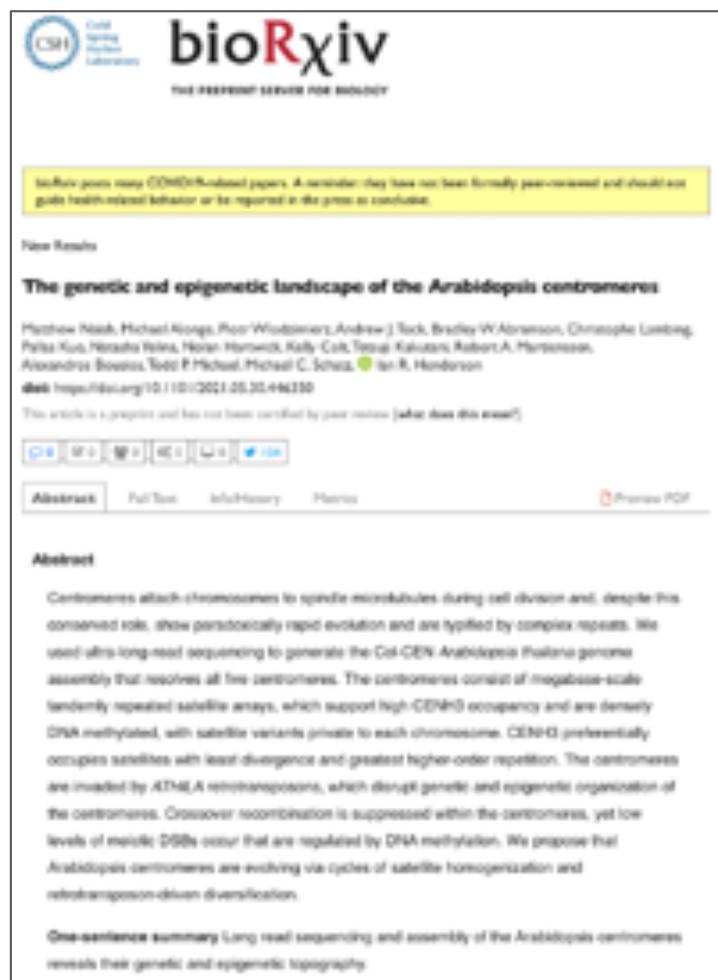


Figure 1. Complete assembly of the *Arabidopsis* centromeres. A, Genome-wide circos plot of the

Summary & Future Work

More accurate and high throughput long read sequencing is unlocking the universe of structural variations and other complex variants

- Enabling the complete sequencing of human genomes
- Discovering tens of thousands of variants previously missed, as well as clarifying tens of thousands of false positives per sample
- Throughput & accuracy rapidly improving, realtime direct alignment of nanopore signal data, highly scalable analysis over tens of thousands of genomes

Beyond mere variant identification

..... towards “Rules of Life” interaction maps and beyond

- Identify the specific pathways for many important traits across the tree of life
- Major opportunities to improve disease prevention and control

The future lies in data integration at tremendous scales!



Acknowledgements

Schatz Lab

Mike Alonge
Sergey Aganezov
Arun Das
Katie Jenike
Melanie Kirsche
Sam Kovaka
Bohan Ni
Gautam Prabhu
Srividya
Ramakrishnan
T. Rhyker
Ranallo-Benavide
Rachel Sherman
Margaret Starostik
Samantha Zarate
Your Name Here

JHU

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

Cold Spring Harbor Laboratory

McCombie Lab
Lippman Lab

Baylor College of Medicine

Sedlazeck Lab

Dana Farber Cancer Institute

Van Allen Lab

Telomere-to-Telomere Consortium

Adam Phillippy & Karen Miga et al.





Thank you!

@mike_schatz
<http://schatz-lab.org>