Towards the detection of all classes of human structural variation

David Porubsky
Eichler lab
University of Washington, Dept. of Genome Sciences

Current Trends in Long-Read Sequencing and Bioinformatics Analysis
Leuven, October 3rd, 2024

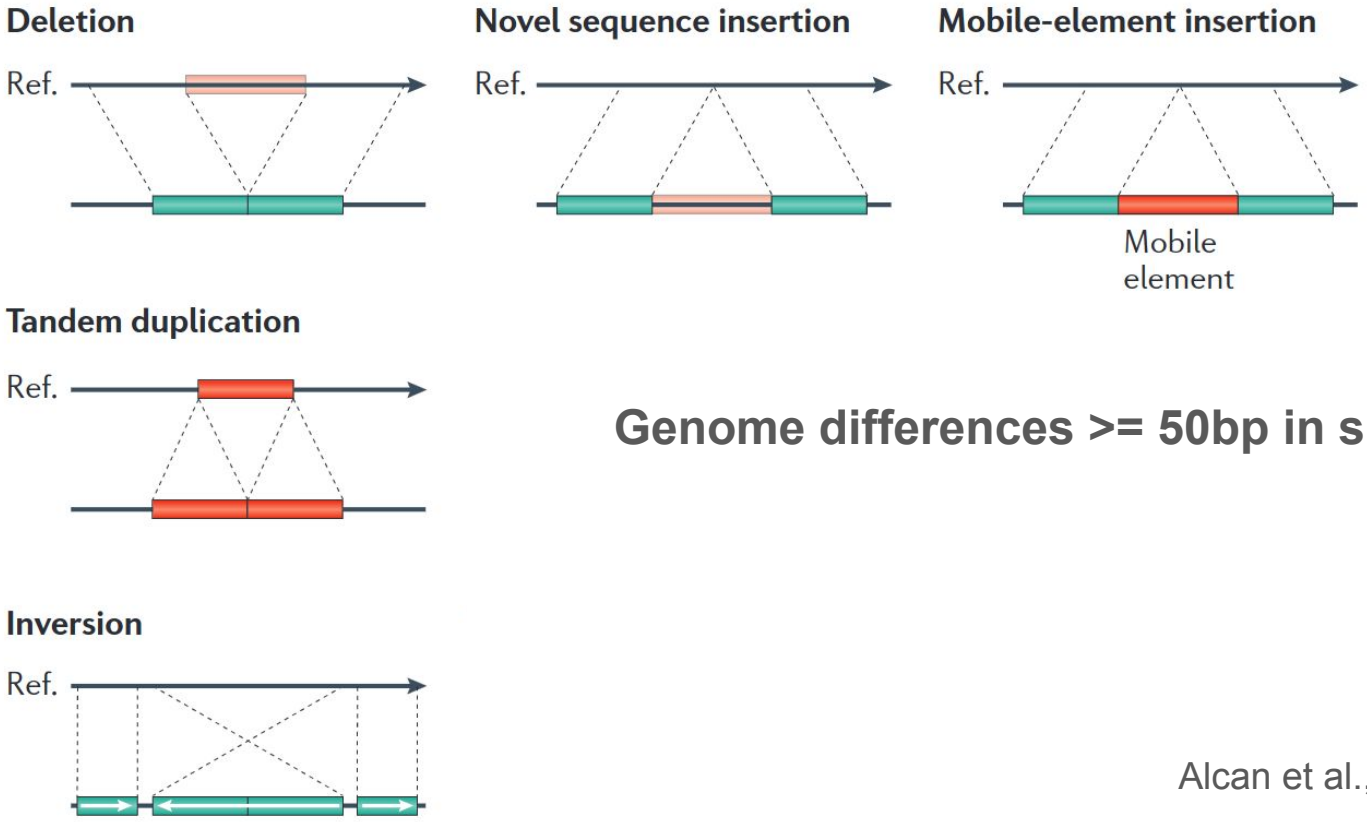
Overview

1. Path to the nearly complete view of human structural variation

2. Using extended human family to probe inheritance of human variation

3. Detailed analysis of a complex region of the human genome at 22q11.21

Classes of human genome structural variation

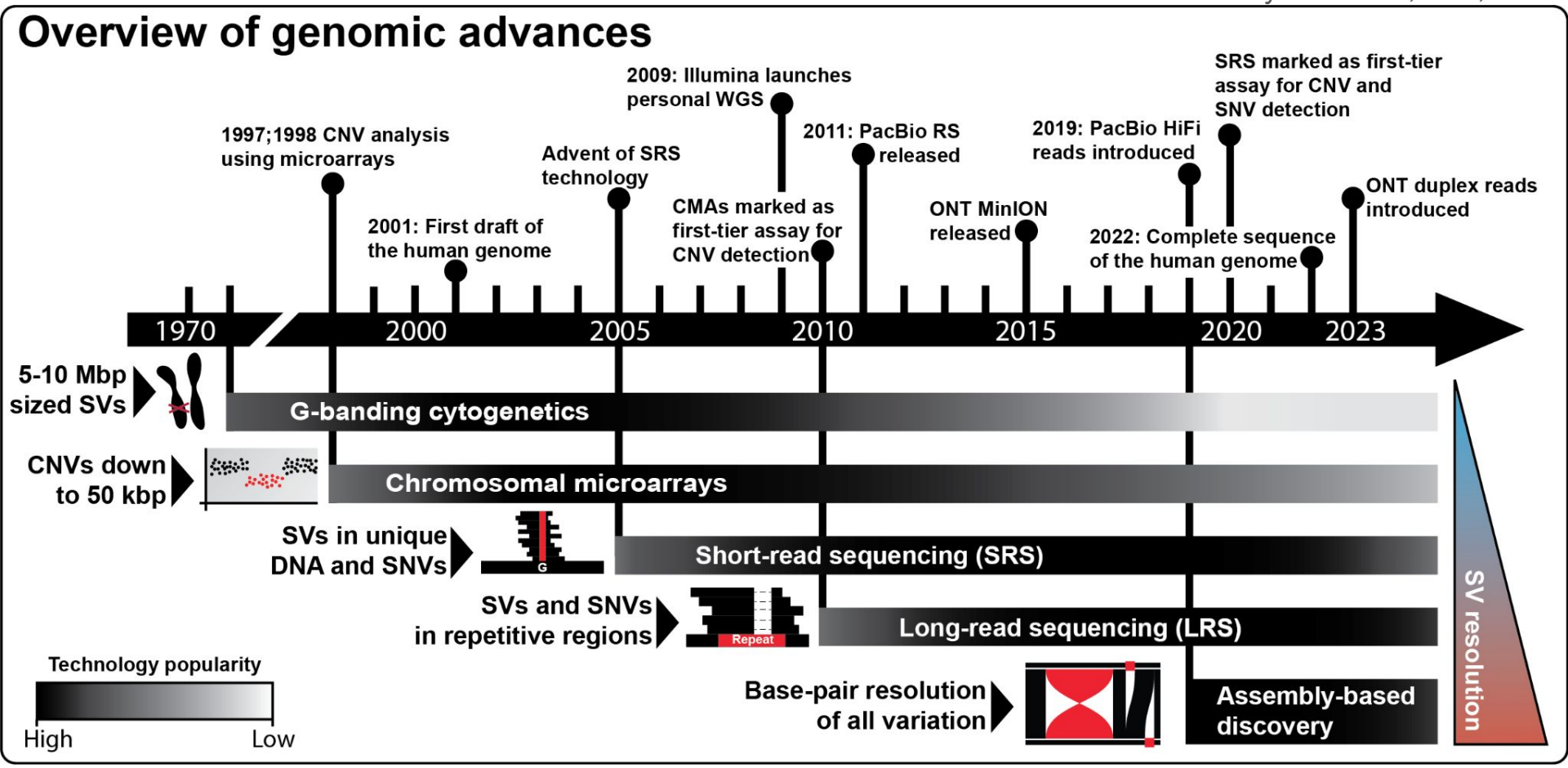


Genome differences >= 50bp in size

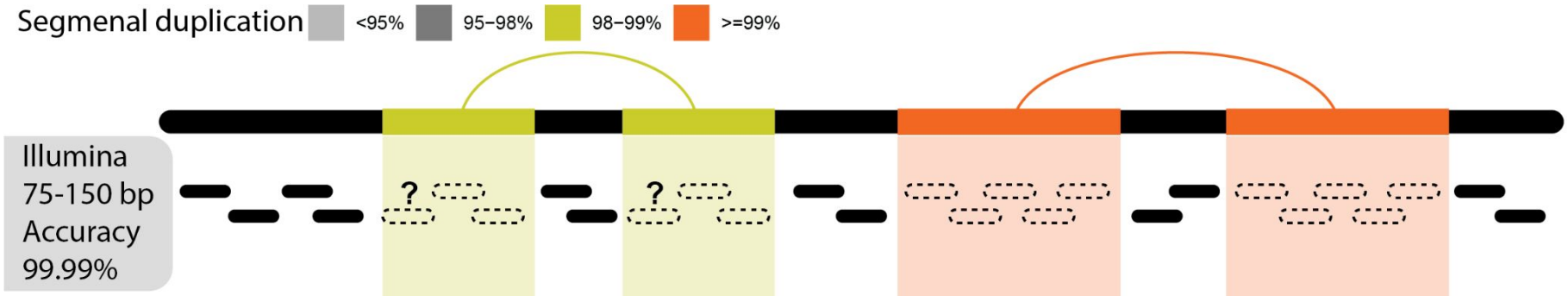
Alcan et al., Nature Reviews, 2011

A 25-years of structural variation detection

Porubsky & Eichler, Cell, 2024

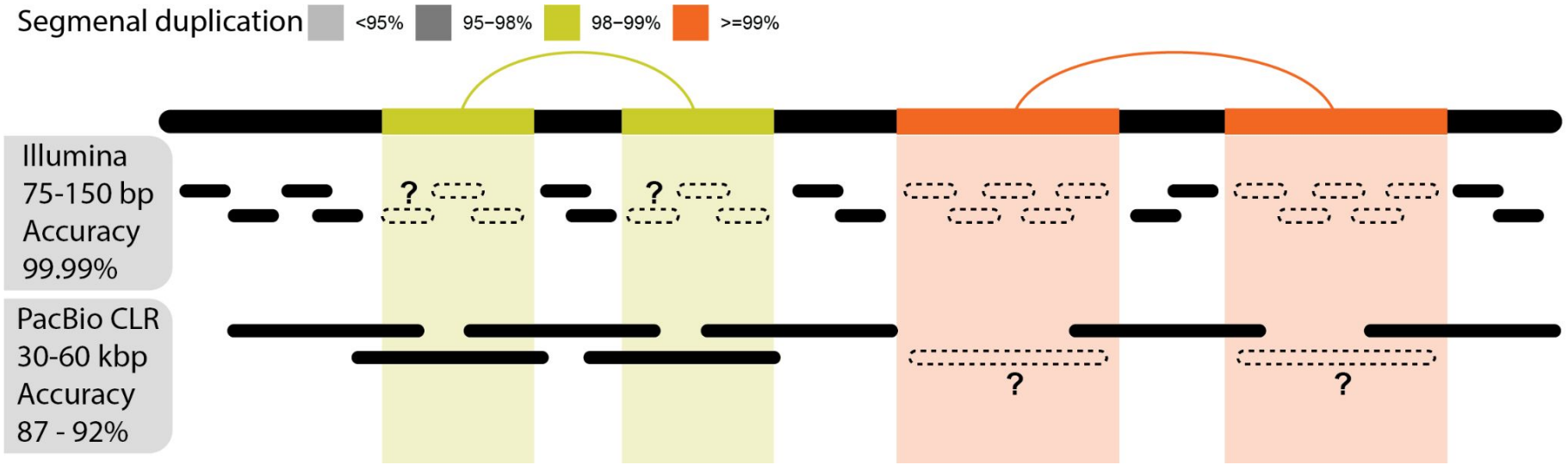


Analysis of human genome is complicated by its repetitive nature

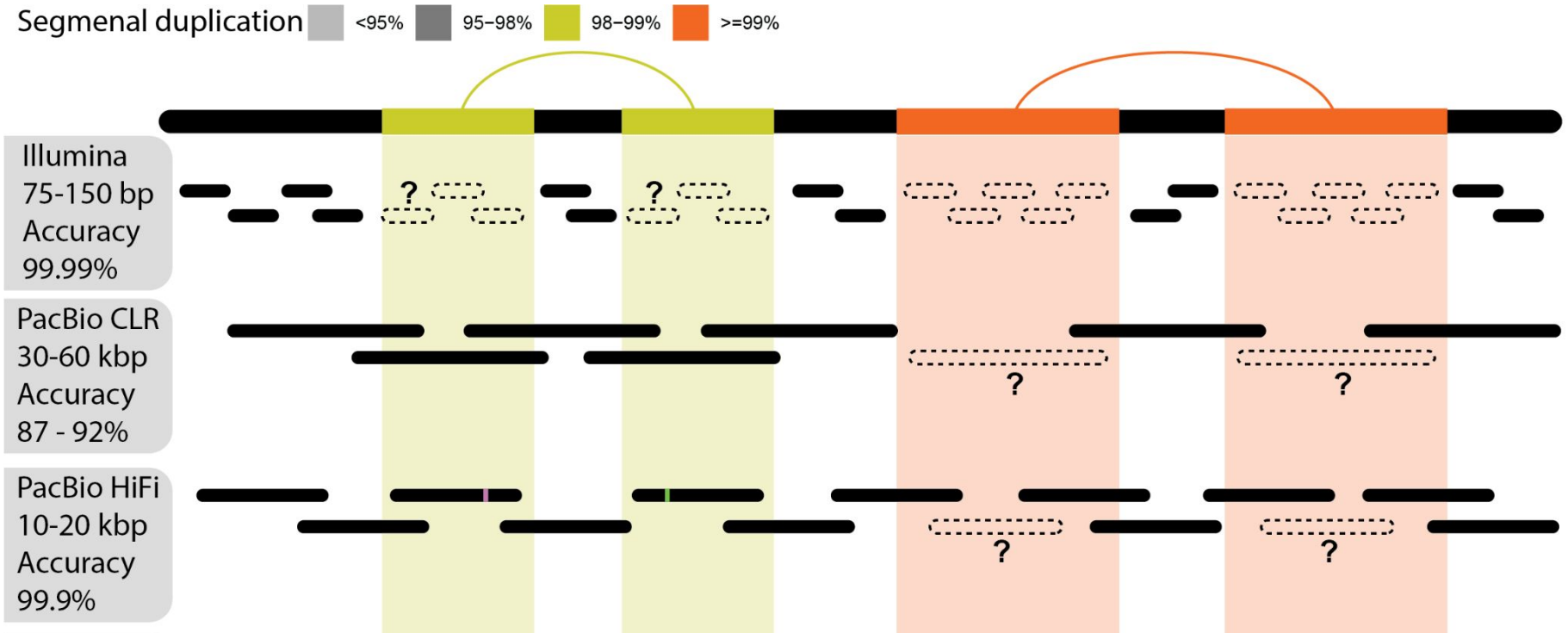


Segmental duplication remained one of the last regions of the human genome to be fully sequenced and assembled

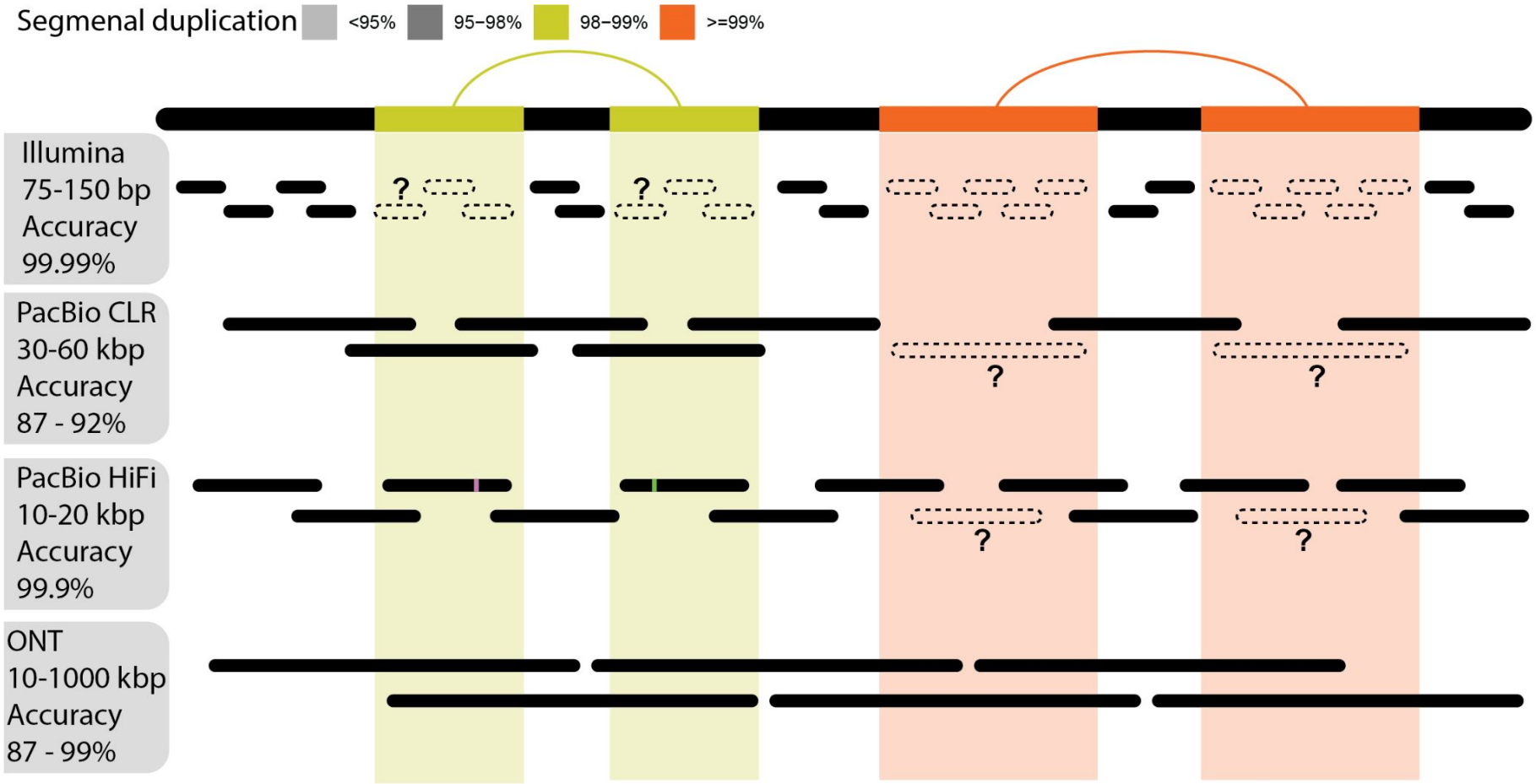
Analysis of human genome is complicated by its repetitive nature



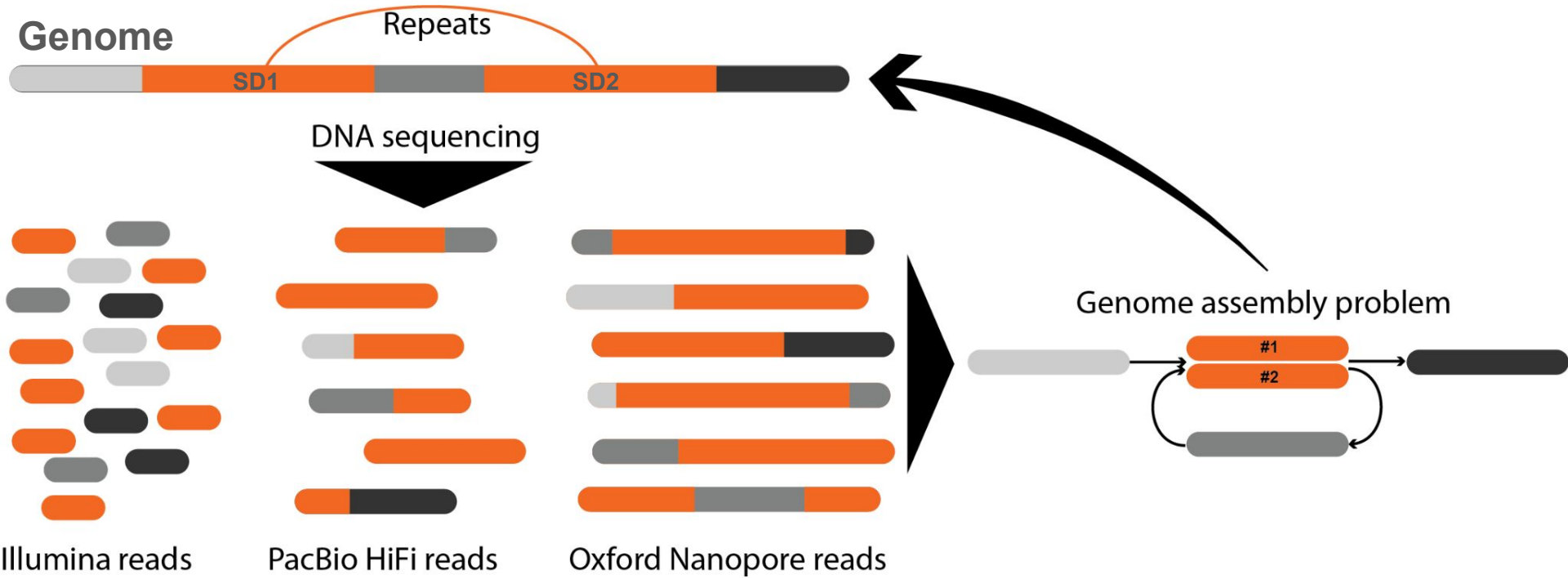
Analysis of human genome is complicated by its repetitive nature



Analysis of human genome is complicated by its repetitive nature



Genome assembly problem we aspire to solve



Combination of long-read technologies enables the best assembly

Common assembly terminology

Contig

Scaftig

N's

Haptig

Maternal

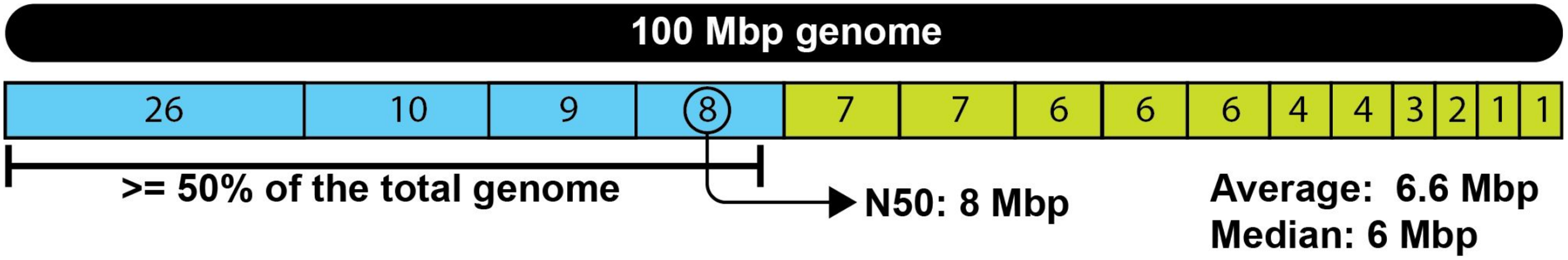
Paternal

T2T contig

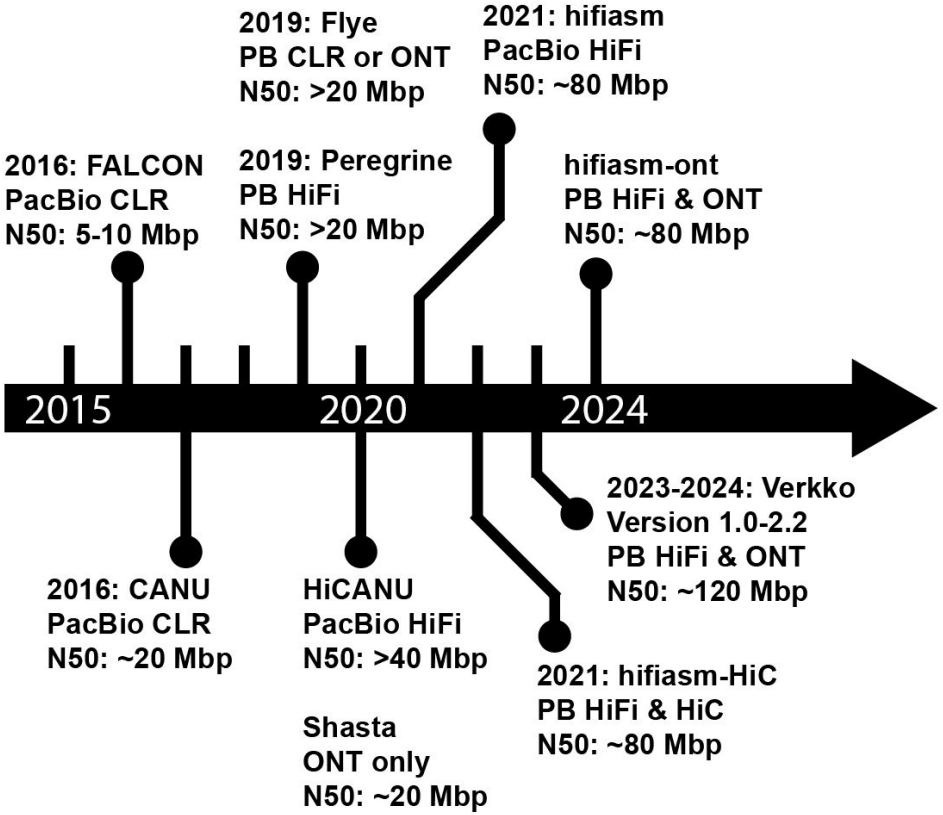
near-T2T contig

Measures of assembly contiguity

N50, NG50, AuN



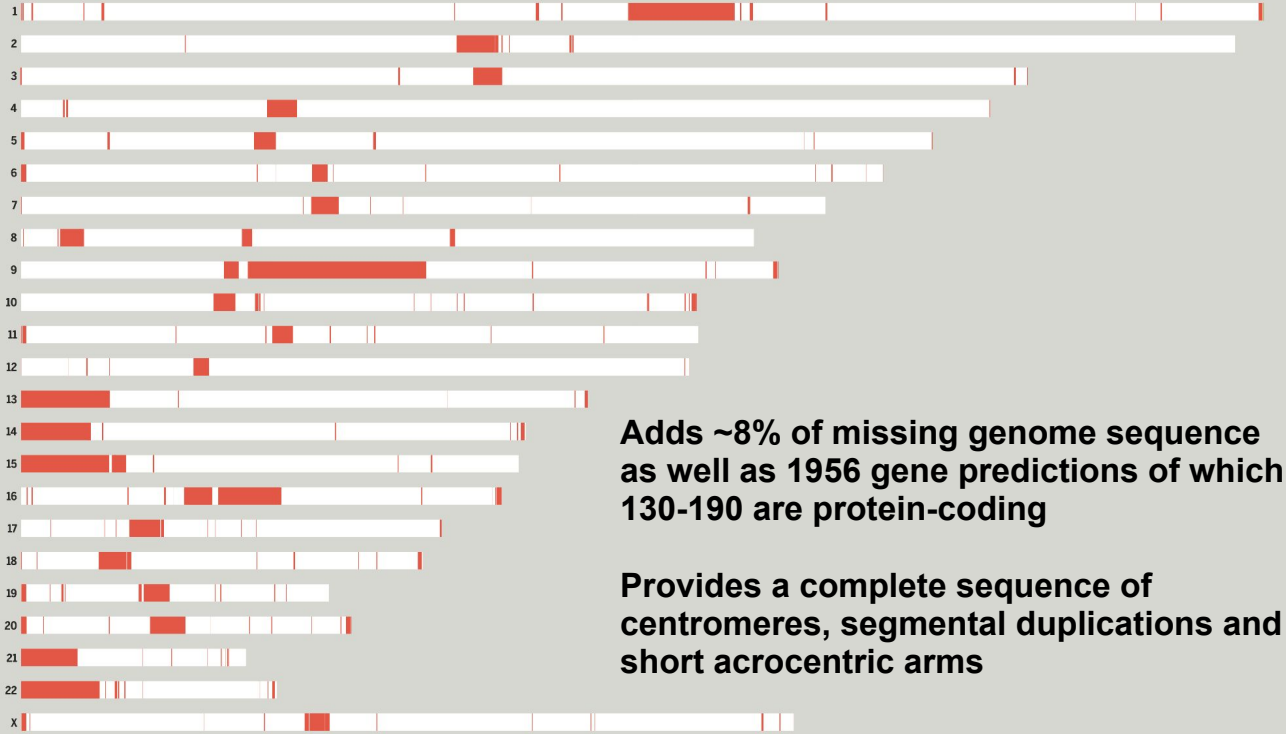
Overview of long-read assemblers



Assembly contiguity has substantially improved in the last 3 years

Completion of the human genome reference

T2T-CHM13 reference genome



Adds ~8% of missing genome sequence as well as 1956 gene predictions of which 130-190 are protein-coding

Provides a complete sequence of centromeres, segmental duplications and short acrocentric arms

Nurk et al, Science, 2022

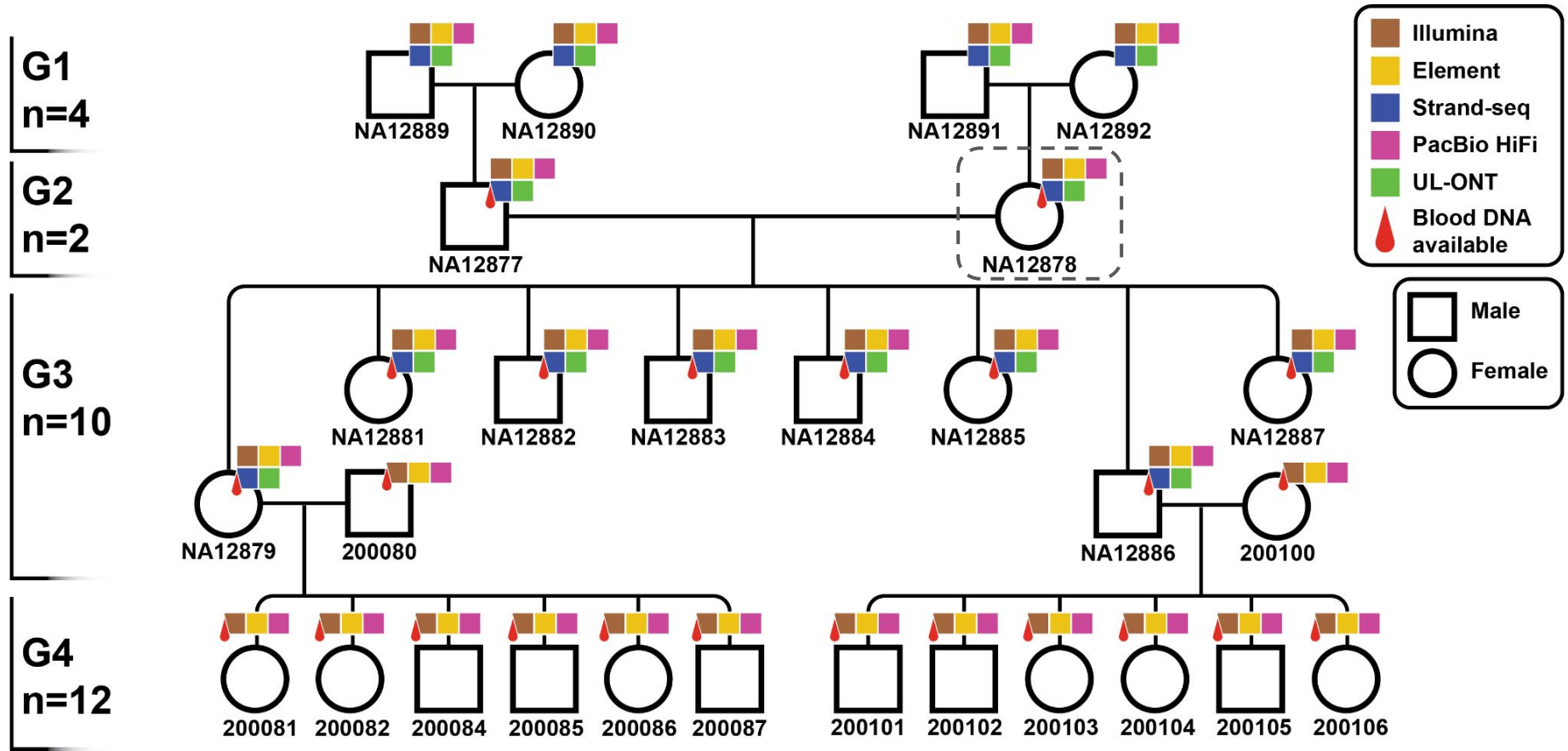
We are now able to generate highly contiguous diploid genome assemblies

Trio-phased assembly of HG002

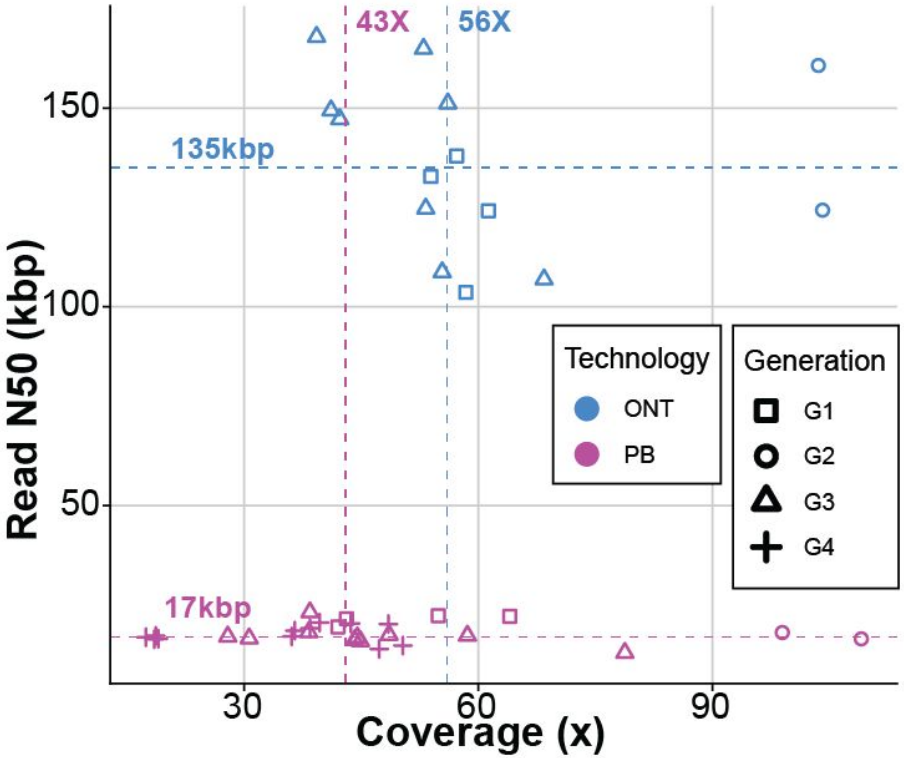
A familial, telomere-to-telomere (T2T) reference genome for human variation, recombination, and mutation

Porubsky et al.,bioRxiv, 2024

Orthogonal sequencing datasets generated for 28 members of a single 4-generational family (1463 CEPH)

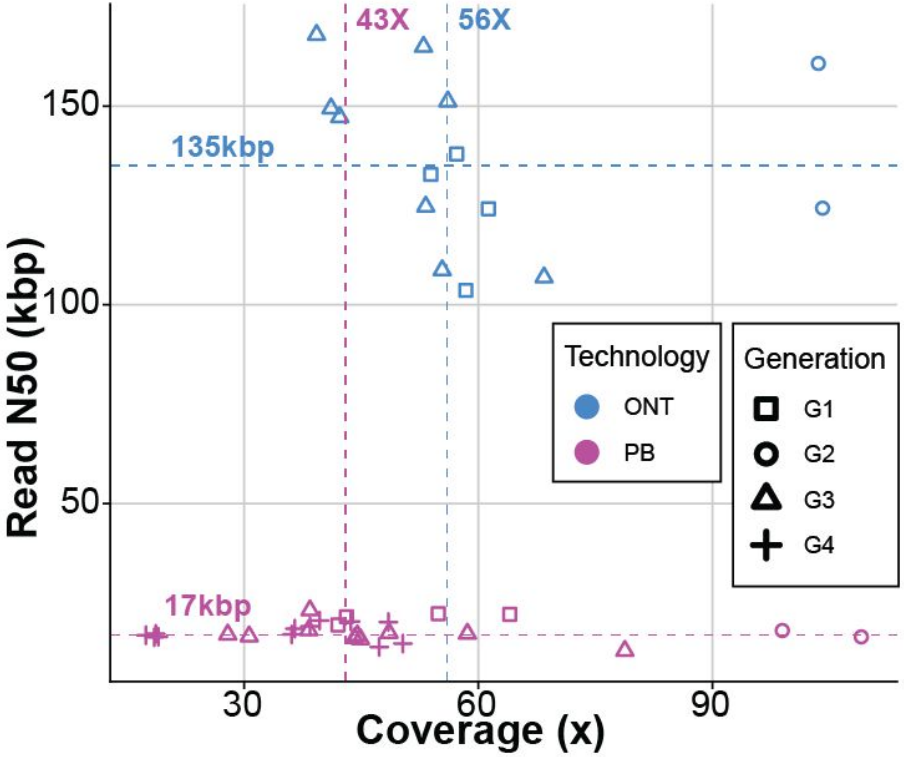


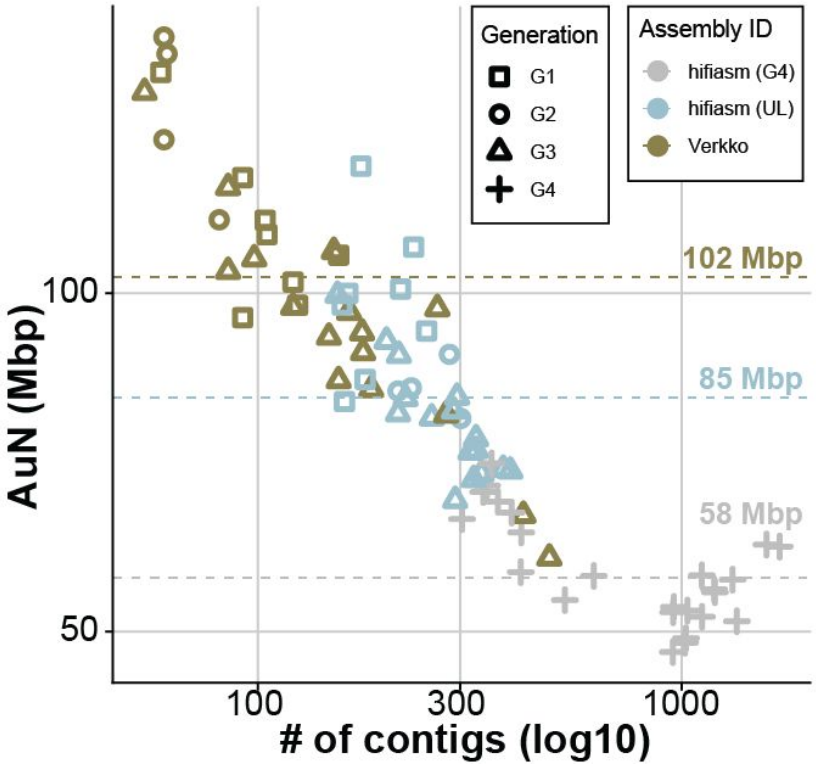
Generated long-read datasets and phased assemblies



- **Overall quality of each assembly is QV 54 [range: 47-57].**

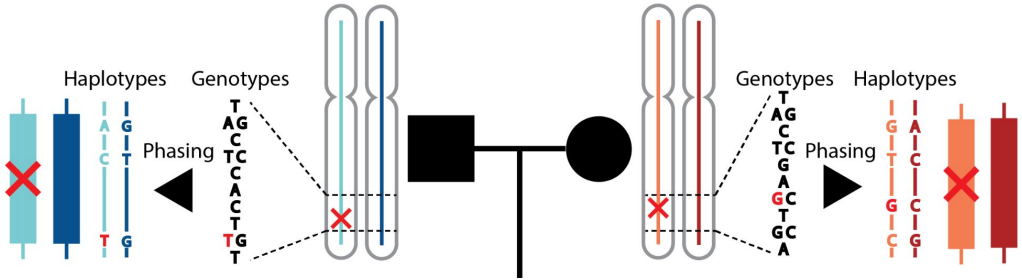
Generated long-read datasets and phased assemblies



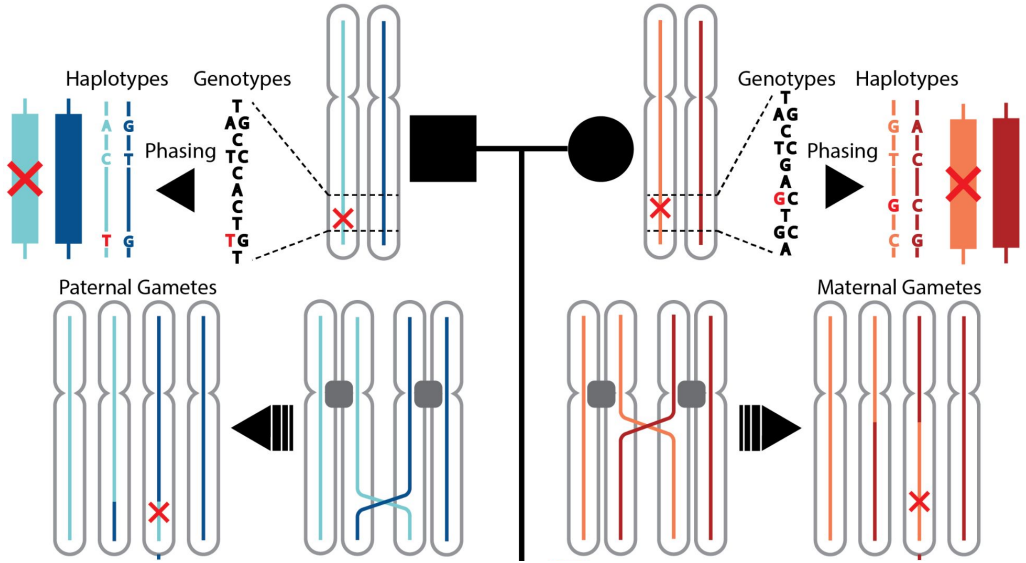


- **Overall quality of each assembly is QV 54 [range: 47-57].**

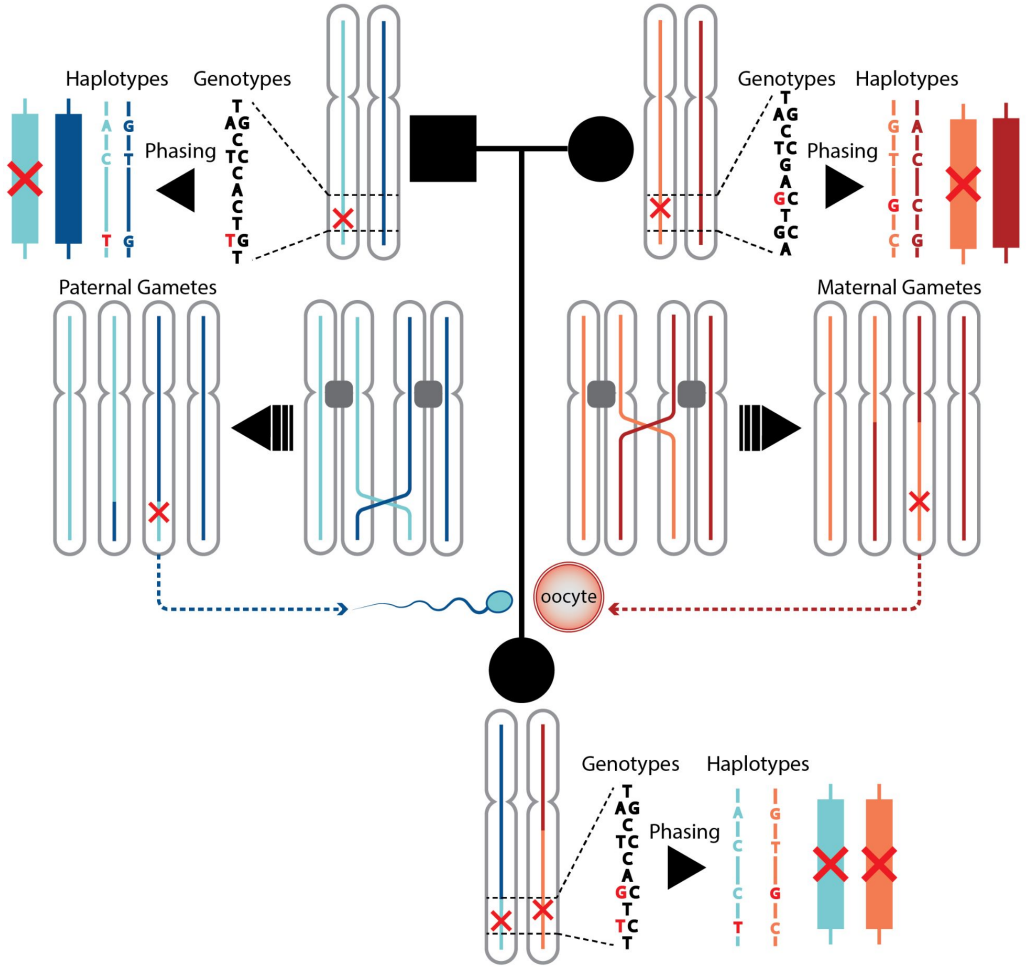
Phasing and detection of meiotic recombination events



Phasing and detection of meiotic recombination events

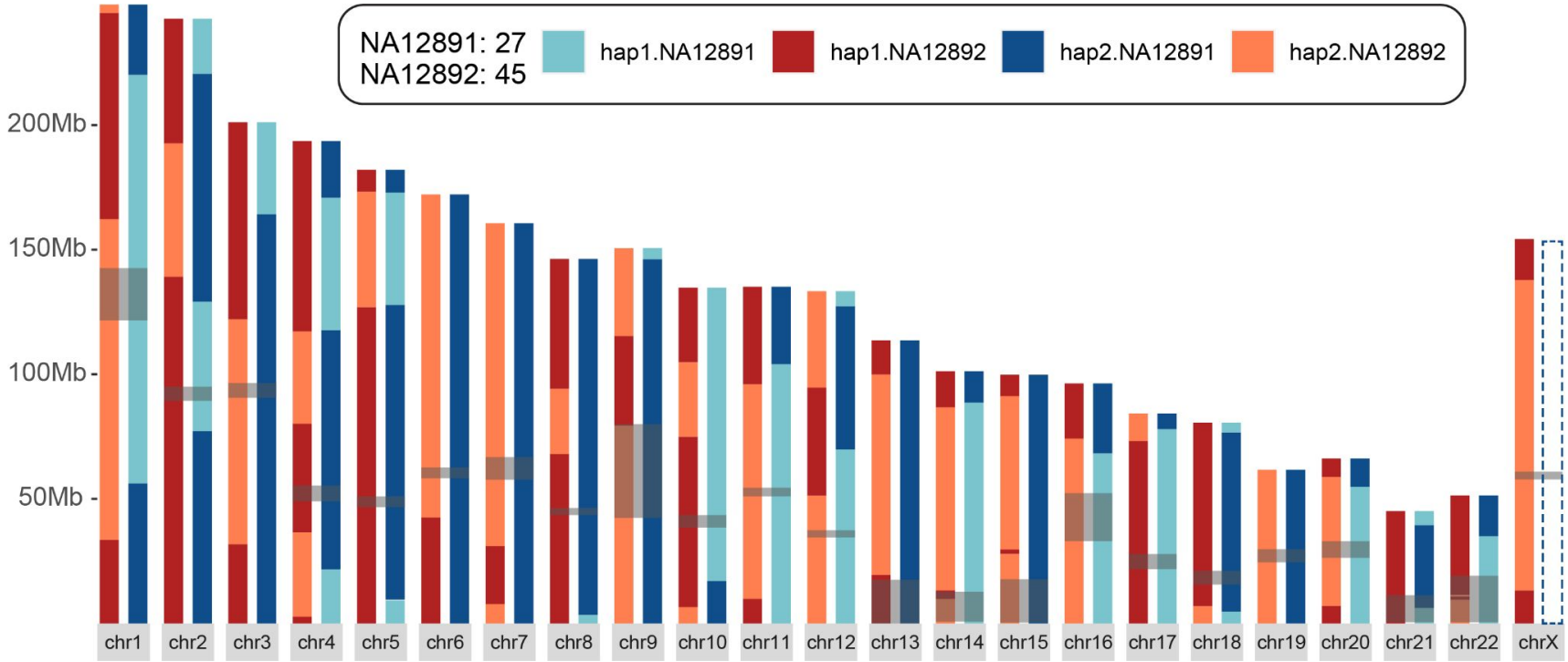


Phasing and detection of meiotic recombination events

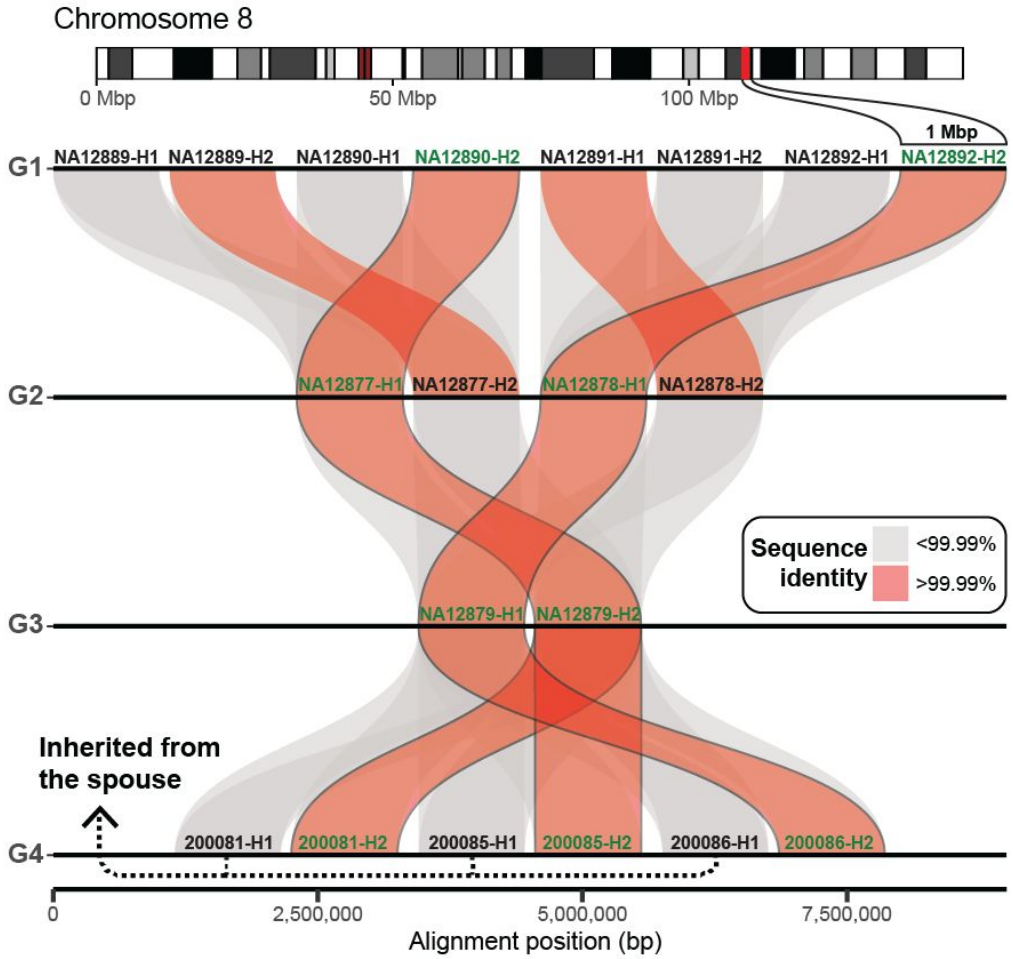


Detecting meiotic recombination breakpoints at high resolution (G2-NA12878)

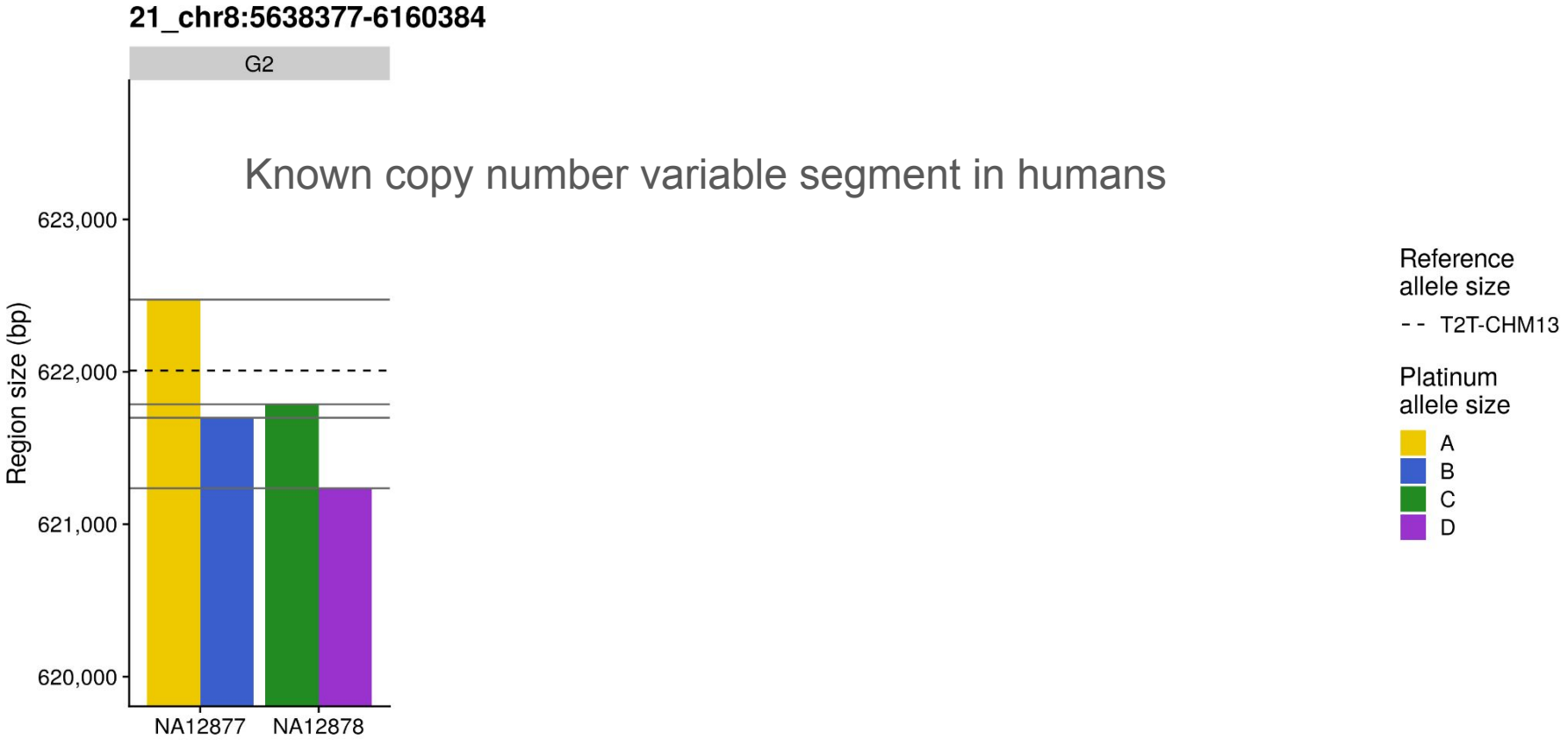
- **Strand-seq provide chromosome-length phasing for G1-G3**
- **Trio-aware phasing of G2-G4**



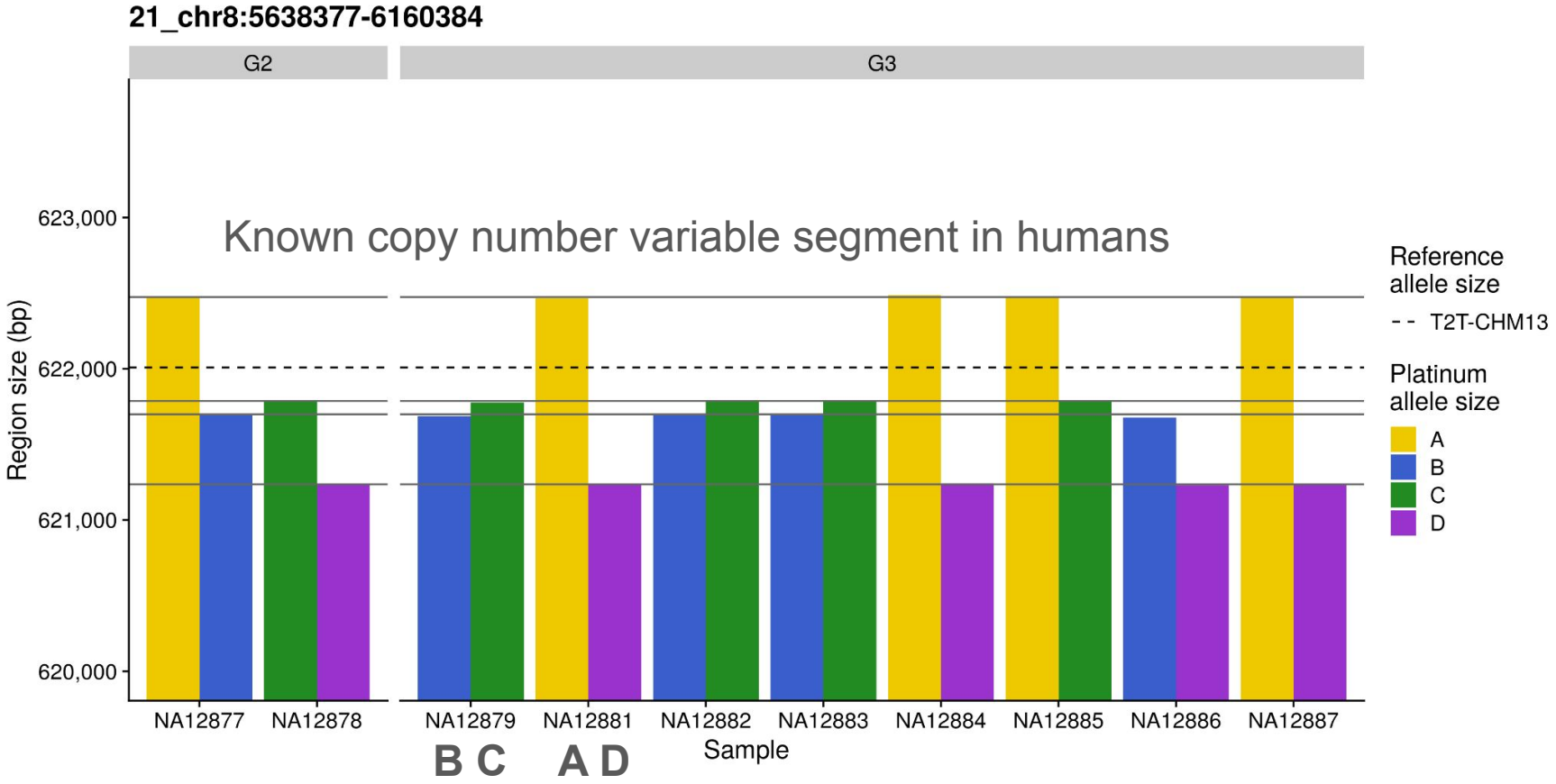
Tracking inheritance of any DNA segment across generations



Tracking inheritance of any DNA segment across generations

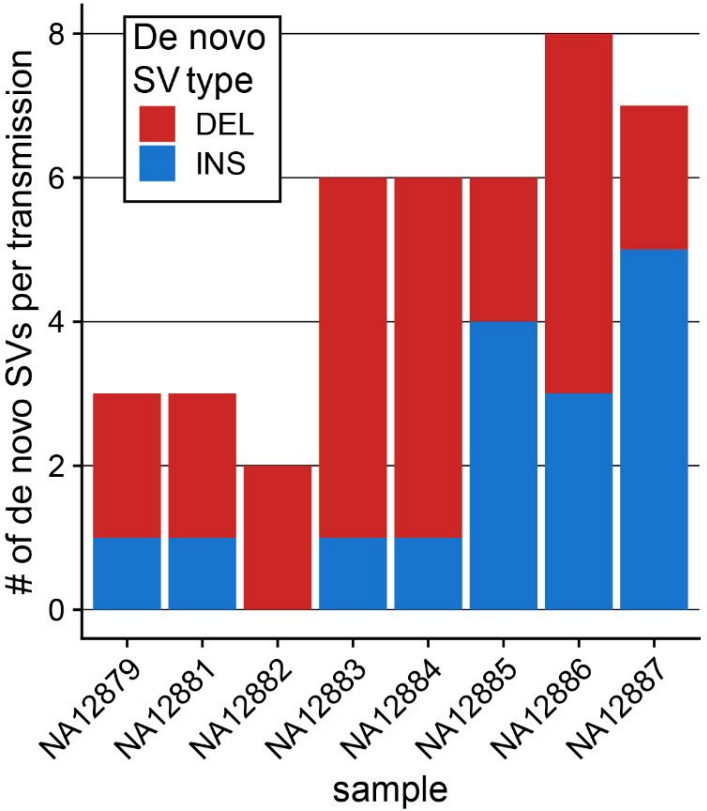


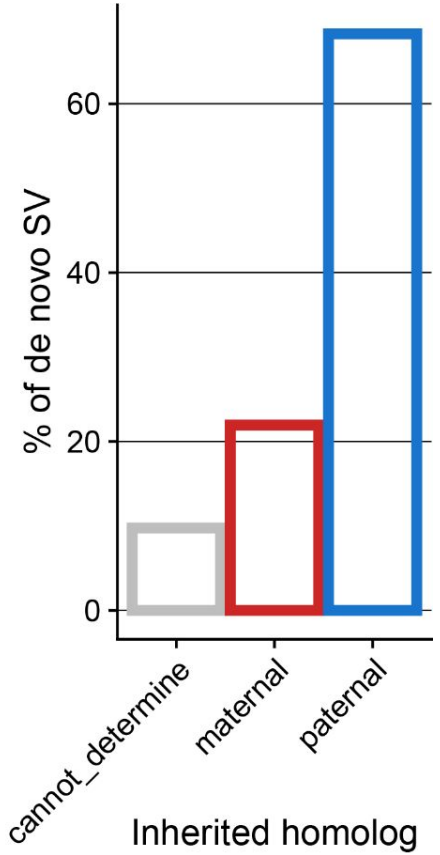
Tracking inheritance of any DNA segment across generations



Valuable information for validation of various SV classes

de novo SVs detected in G3 (n=8)

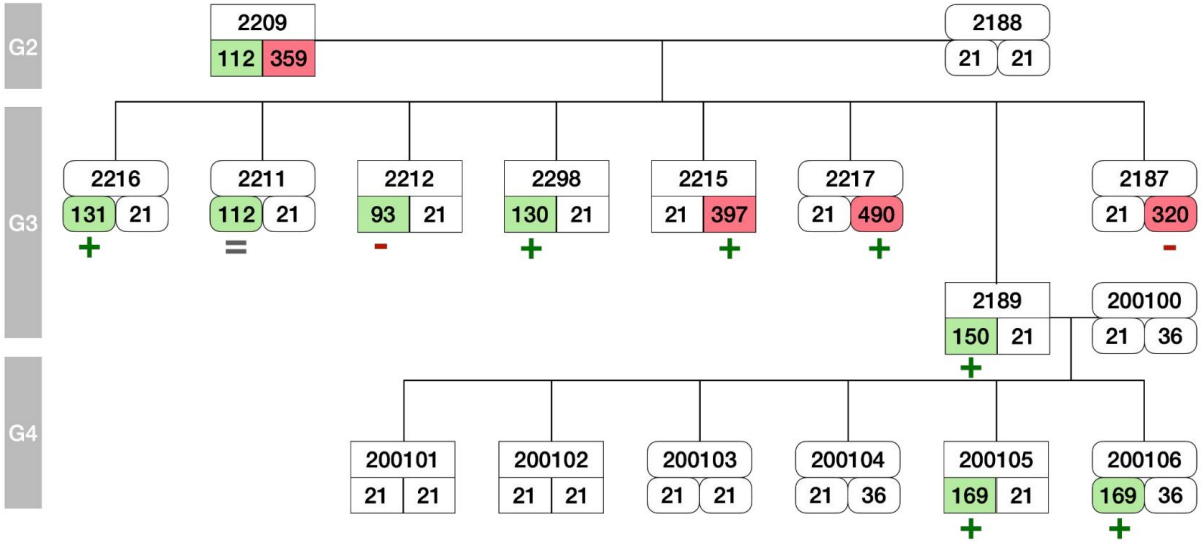




Vast majority of these are change in VNTR lengths

Detection of *de novo* VNTR variants

Example: *de novo* calls at the exact same site in almost all children

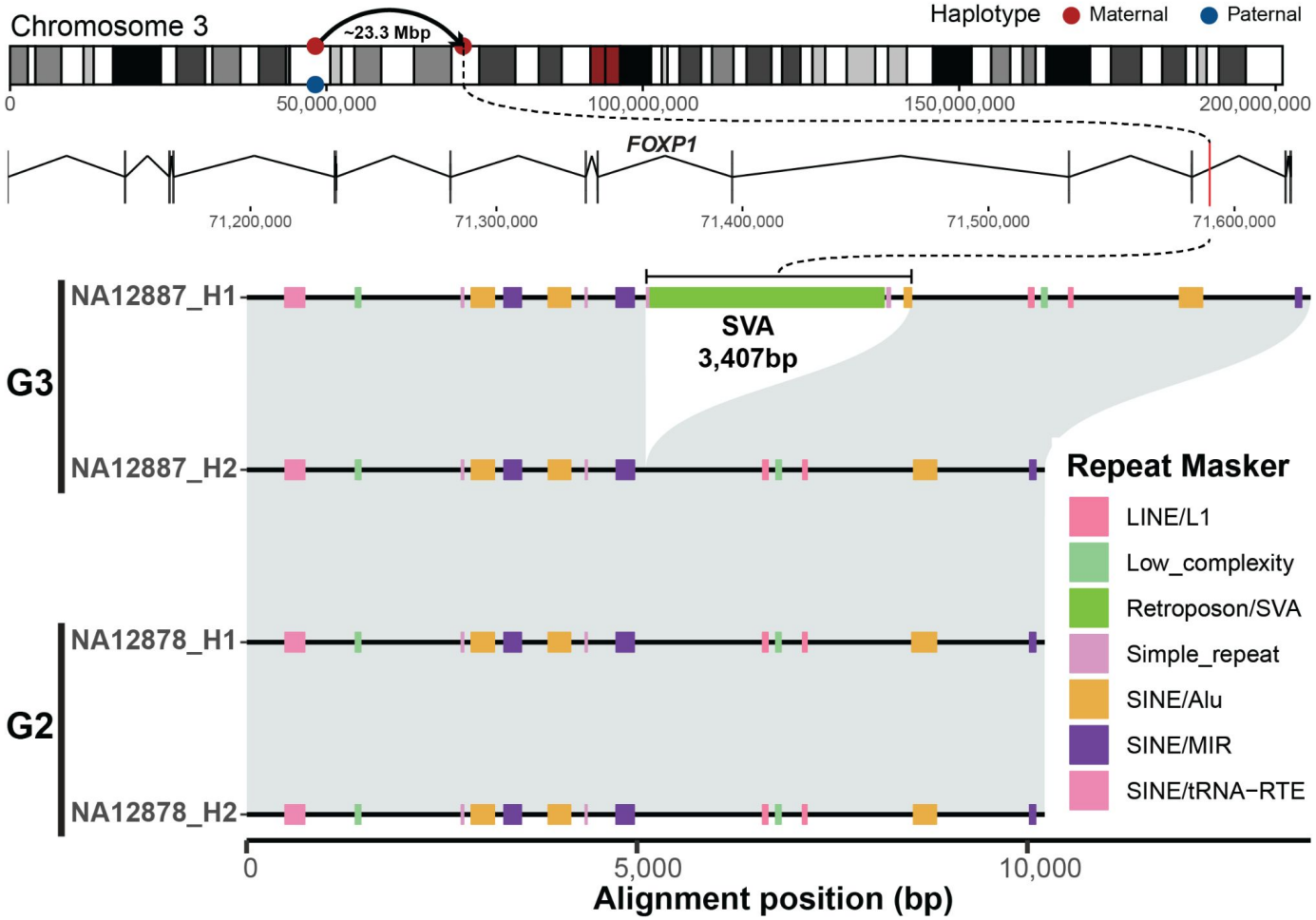


Site: chr8:2,623,322-2,623,462

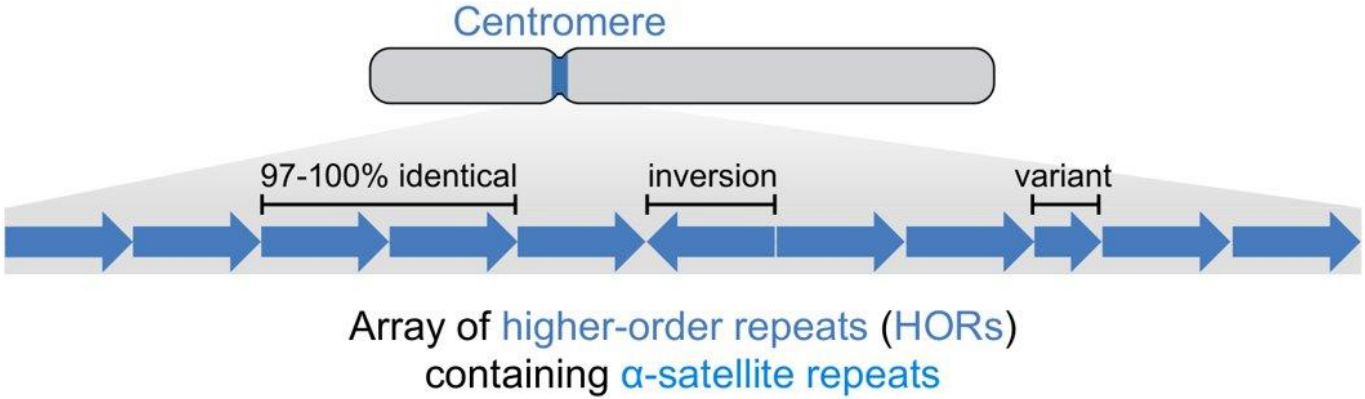
Motif: GAGGCGCCAGGAGAGAGCGCT, Length: 21

We have detected 32 loci with recurrent VNTR changes.
Of those 16 show three and more de novo mutations.

Example of an fully assembled *de novo* insertion in an intron of *FOXP1* gene

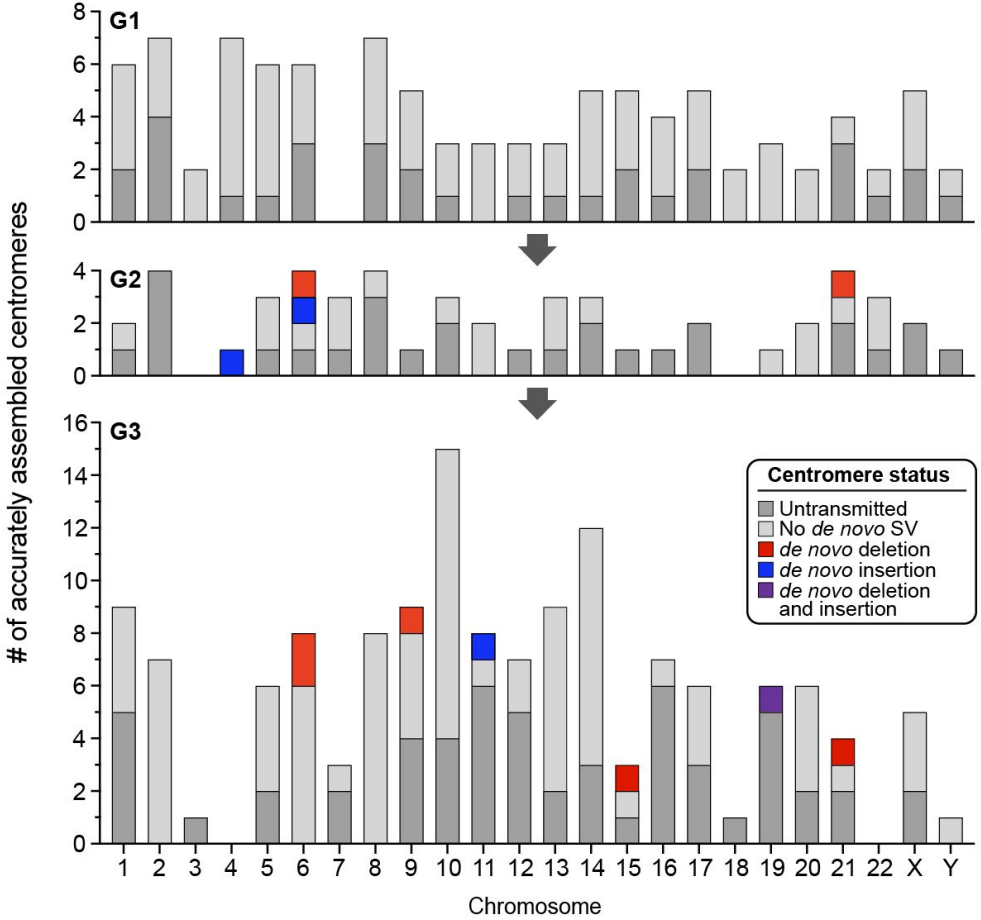


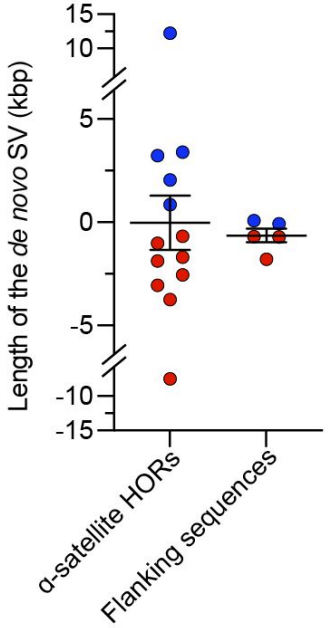
Centromeres are complex highly repetitive regions of the genome



- We have assembled completely sequenced and assembled 288 centromeres.
- Of those, we were able to assess 150 transmissions (33 from G1 to G2 and another 117 transmissions from G2 to G3) for de novo SV detection.

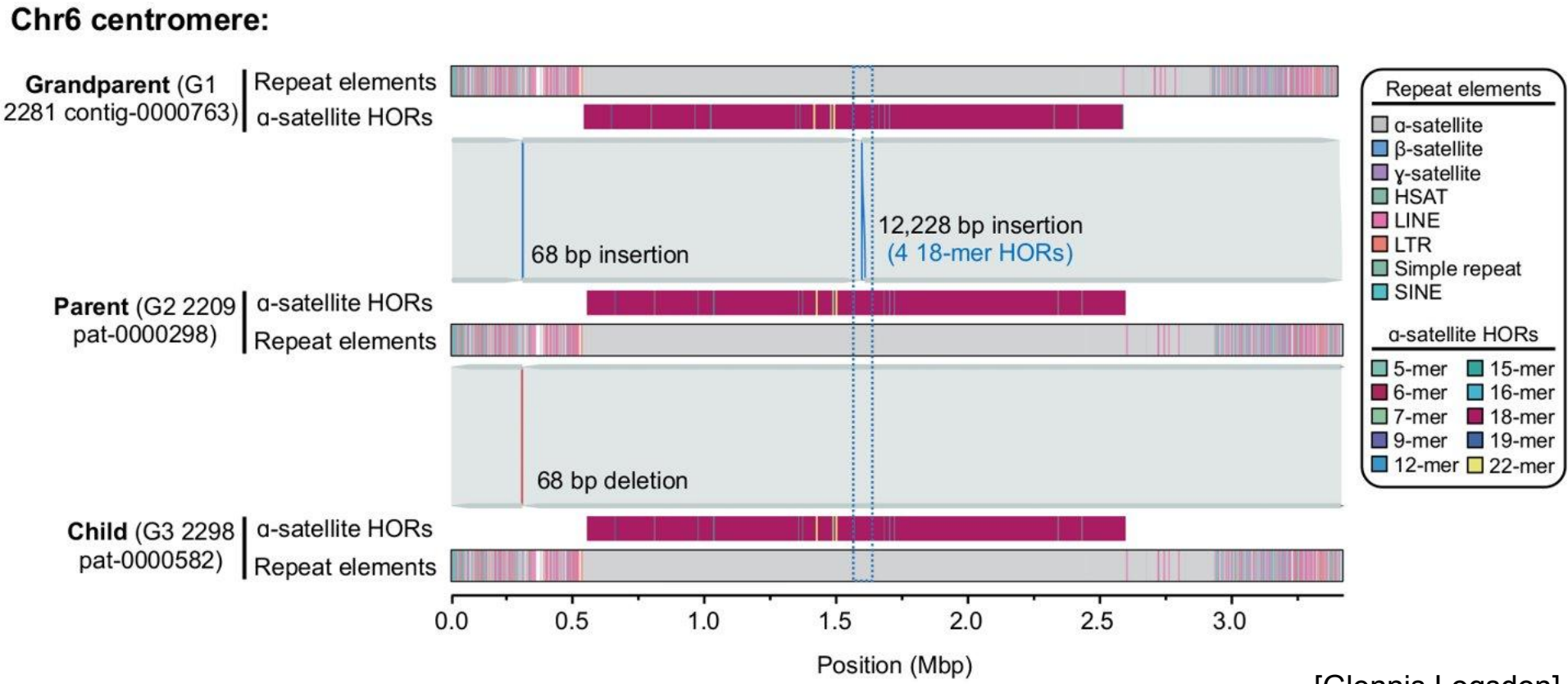
We detected 18 *de novo* SVs within centromeres





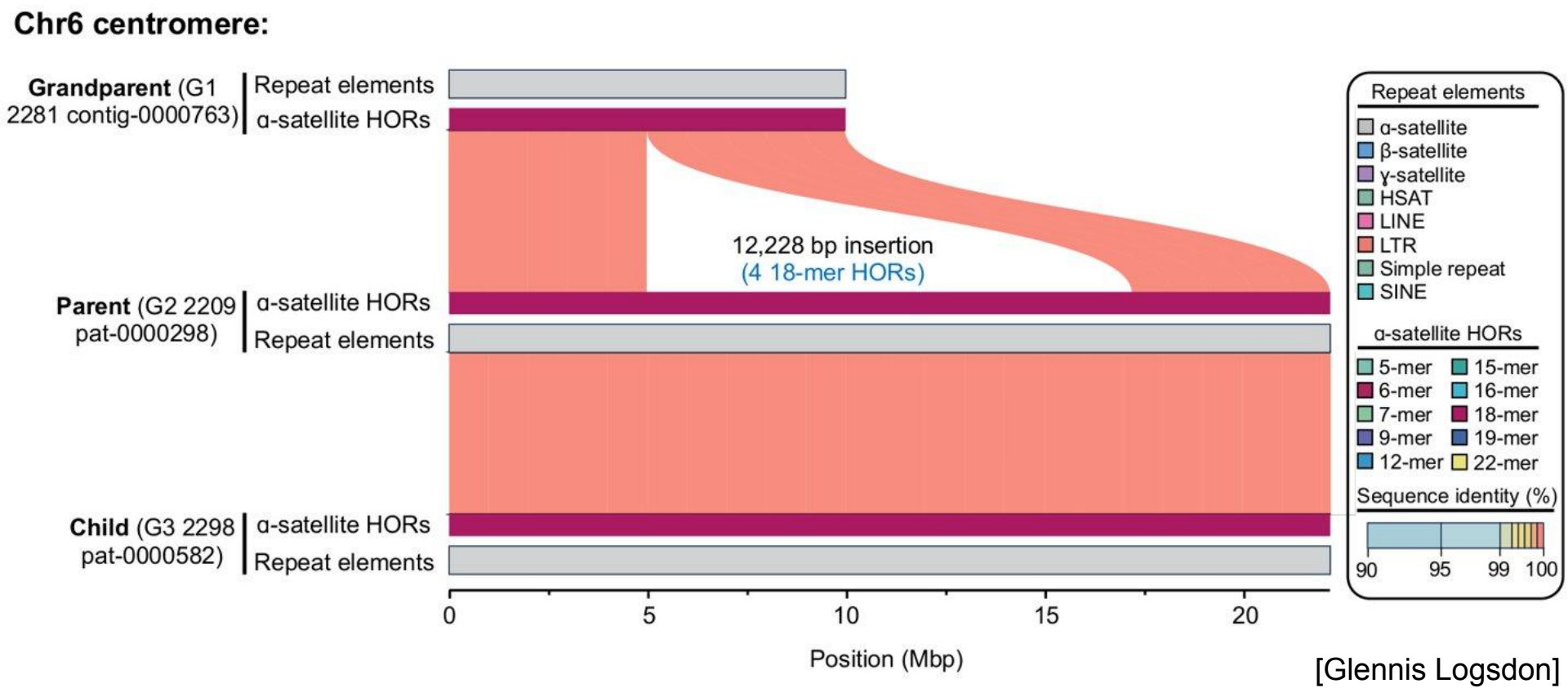
[Glennis Logsdon]

Evidence of inherited *de novo* variation in the next generation



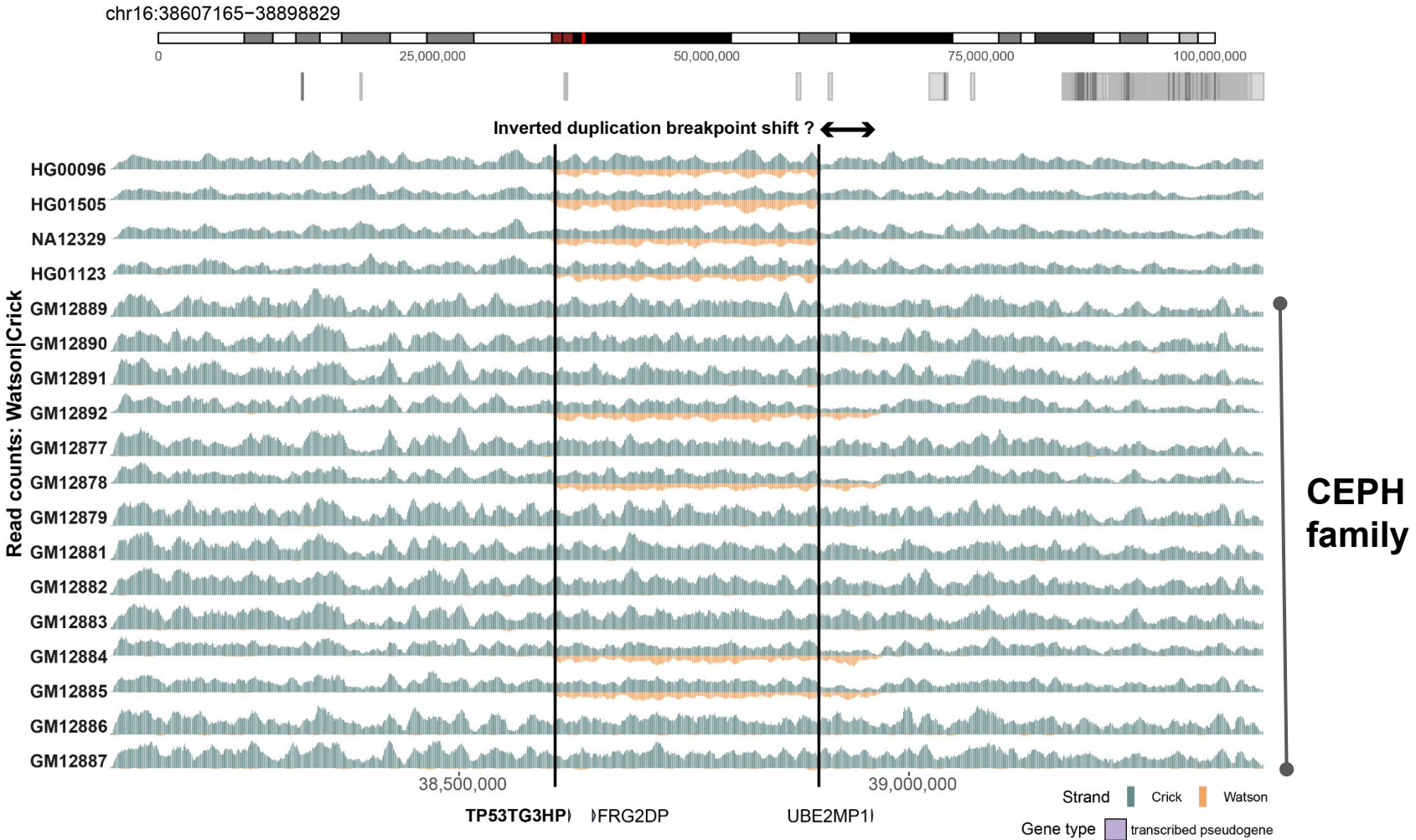
[Glennis Logsdon]

Sequences immediately surround the *de novo* insertion are >99.9% identical

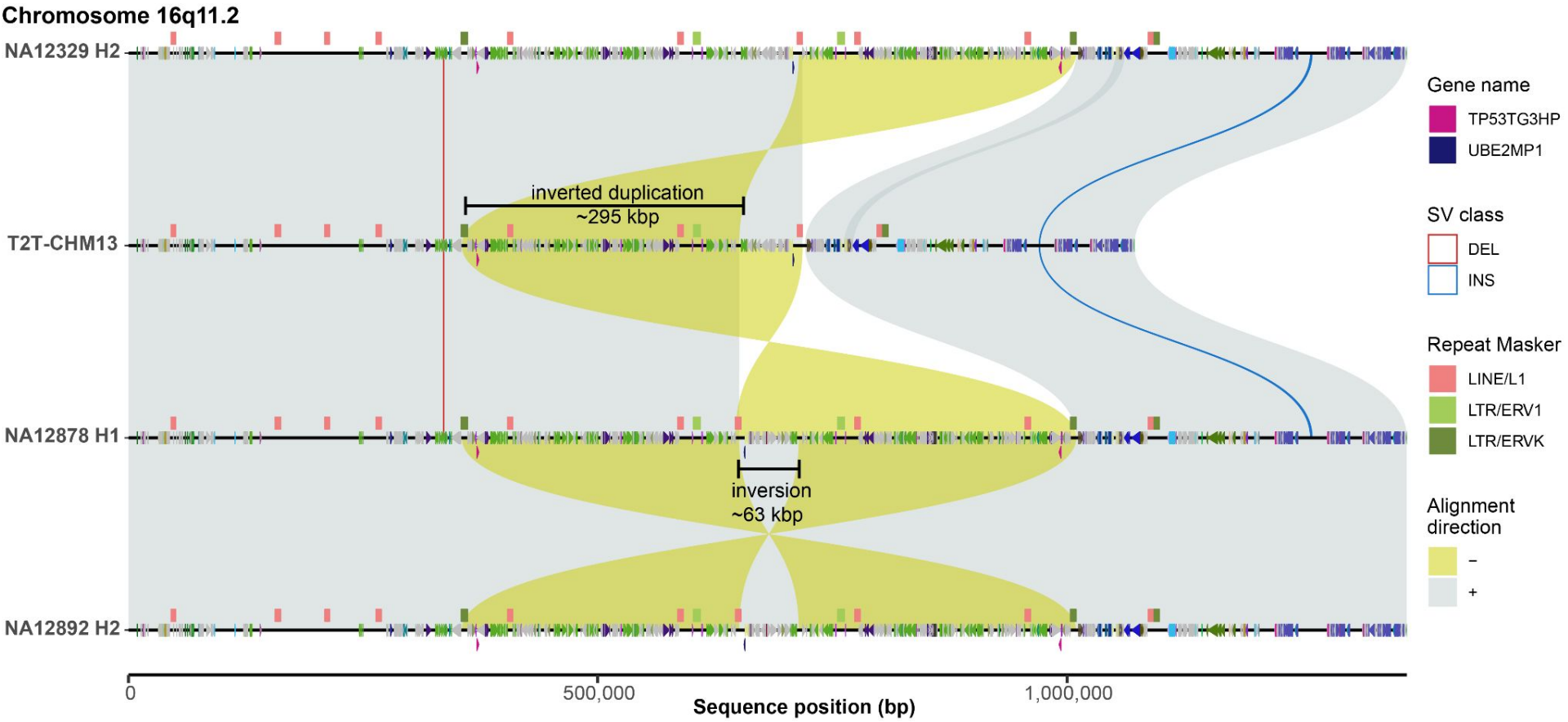


[Glennis Logsdon]

Deciphering the structure of human specific inverted duplication



Deciphering the structure of human specific inverted duplication



Conclusions

- Long-read sequencing is necessary to start looking into structure of the most complex regions of the genome.
- Large family pedigrees are important to track SV inheritance and estimate *de novo* variants frequencies
- Detection of *de novo* mutations such as large insertion in an intron of *FOXP1* gene and *de novo* mutations within centromeric HORs.

Acknowledgements

Eichler lab:
Evan Eichler
Glennis Logsdon
William Harvey
Nidhi Koundinya
Katy Munson
Kendra Hoekzema
Gage Garcia

PacBio:
Michael Eberle
Cillian Nolan
Cairbre Fanslow
William Rowell
Christine Lambert
Egor Dolzhenko
Tom Mokveld

Jorde lab:
Lynn Jorde
Scott Hawkins
Cody Steely

Lansdorp lab:
Peter Lansdorp
Vincent Hanlon

Quinlan lab:
Aaron Quinlan
Harriet Dashnow
Brent Pedersen

Marschall lab:
Tobias Marschall
Peter Ebert
Mir Henglin

Korbel lab:
Jan Korbel
Patrick Hassenfeld

Vermeesch lab:
Joris Vermeesch
Erika Souche

HGSVC: