Complete, telomere-to-telomere assembly of diploid human genomes and beyond

Sergey Koren

Biodiversity Genomics Academy
October 2nd, 2024

The **Forefront** of **Genomics**®

NIH National Human Genome Research Institute

Telomere-to-Telomere

- The human genome is finally finished!
- 8% was left after HGP
- Solved with combination of HiFi + ultra-long ONT

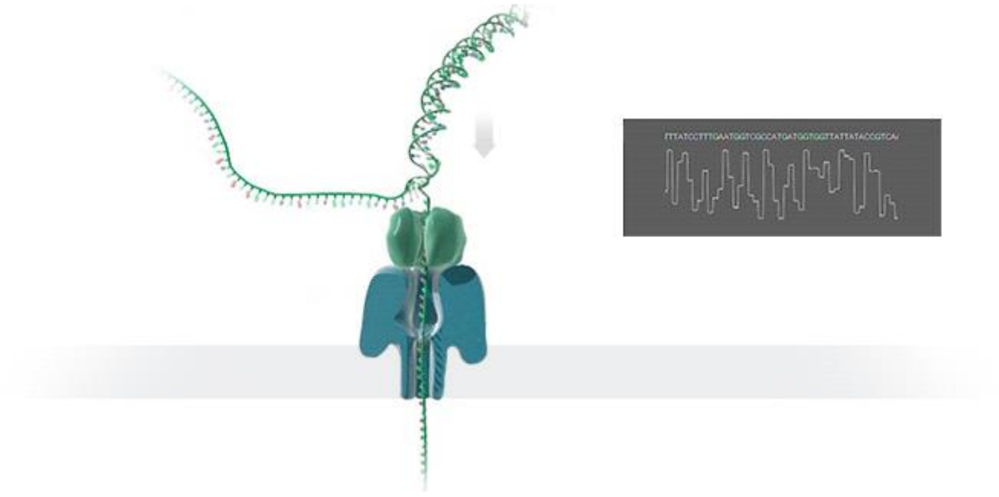
Earth's heart of iron begins to yield its secrets p. 18
Microglia in chronic pain recovery and relapse pp. 33 & 86
Particle acceleration in a nova explosion p. 77
Science
$15
1 APRIL 2022
SPECIAL ISSUE
science.org
AAAS
FILLING THE GAPS
Closing in on a complete human genome p. 42

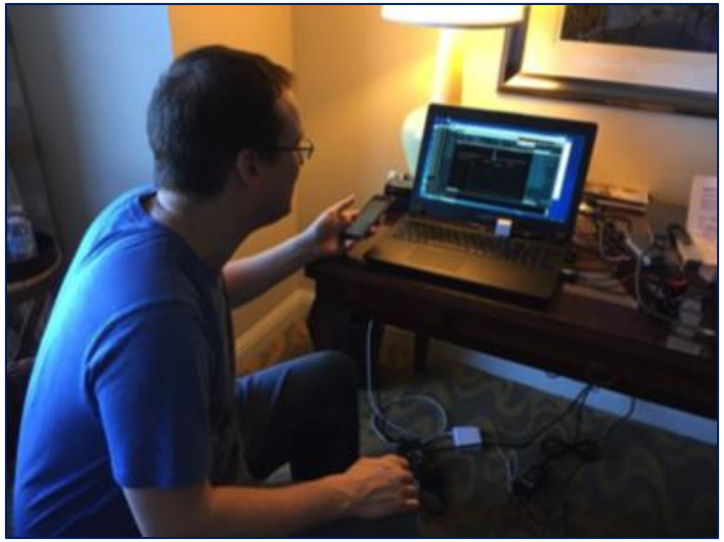
The complete sequence of a human genome.
Nurk, Koren, Rhie, Rautiainen, et al. *Science* (2022)

A new era of sequencing

Nanopore ultra-long sequencing

- **Nanopore UL**
    - >100 kb reads, up to 1 Mb
    - 95% (Q13) read quality
    - 99.9% (Q30+) assembly quality
- **Pros**
    - Length and throughput
    - Reads *span* repeats
- **Cons**
    - Lower base quality



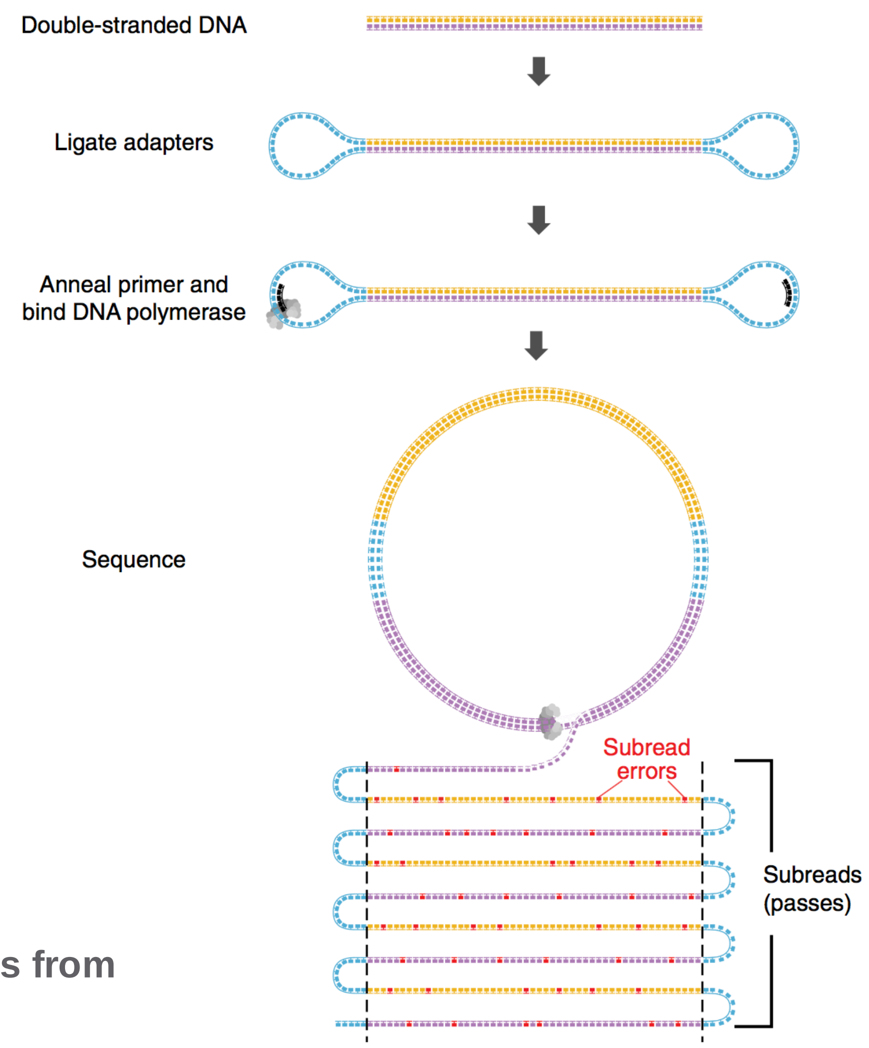


Nanopore sequencing and assembly of a human genome with ultra-long reads. Jain et al. *Nature Biotechnology* (2018)

Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Shafin et al. *Nature Biotechnology* (2020)

PacBio circular consensus sequencing

- **PacBio HiFi**
  - 20 kb reads
  - 99.9% (Q30) read quality
  - 99.9999% (Q60+) assembly quality
- **Pros**
  - Near-perfect accuracy
  - Reads *distinguish* repeats
- **Cons**
  - Limited length and coverage

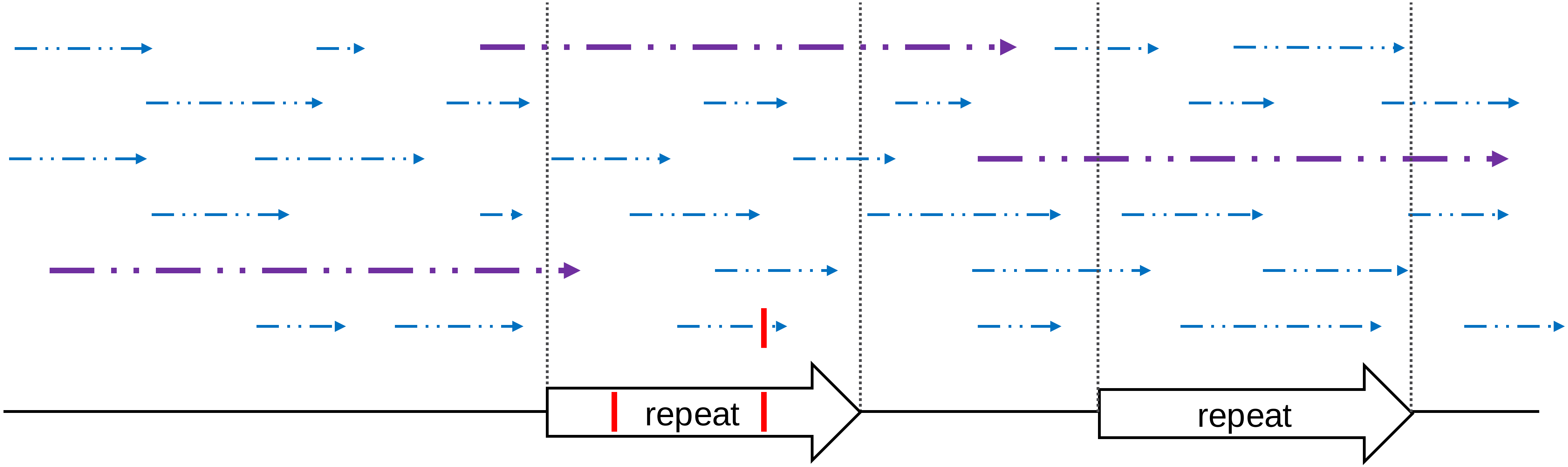


Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Wenger et al. *Nature Biotechnology* (2019)

HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Nurk et al. *Genome Research* (2020)

Two ways to resolve repeats: length

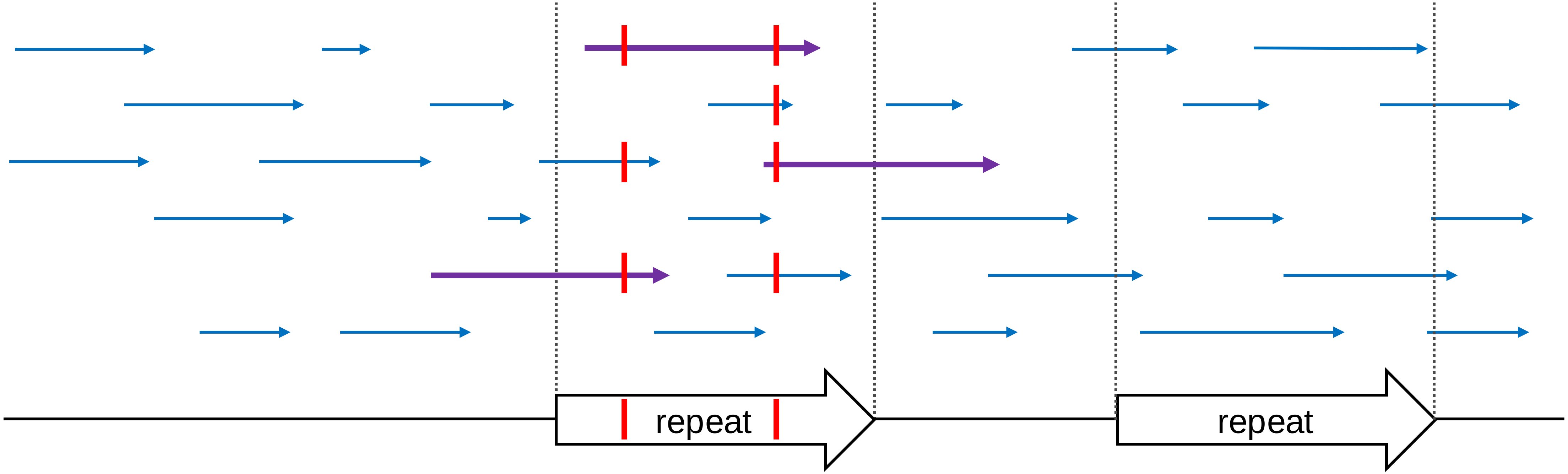
- Nanopore UL read length distribution is long tailed



NIH NHGRI

Two ways to resolve repeats: accuracy

- HiFi reads are accurate



Best of both worlds

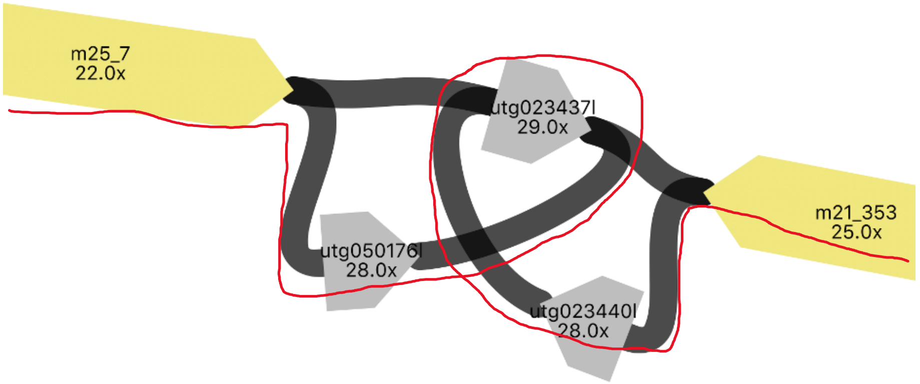
Integrating HiFi and ONT

- **HiFi accurate assembly graph**
  - Homopolymer compression (CAAAAT → CAT)
  - Alignment-based read cleaning and correction
  - Assembly graph from long *perfect* overlaps
- **Nanopore long repeat resolution**
  - Nanopore reads aligned to the graph
  - Correctly count, order, and orient the repeats
  - HiFi-based consensus minimizes error-prone polishing

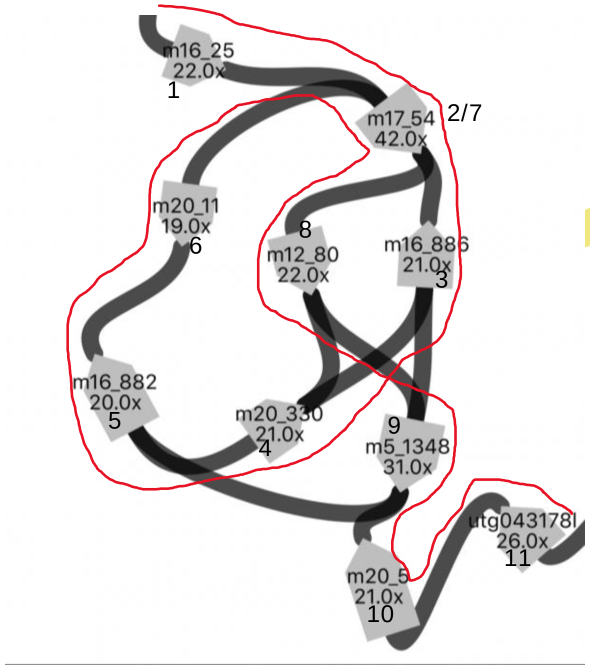
The complete sequence of a human genome.
Nurk, Koren, Rhie, Rautiainen, et al. *Science* (2022)

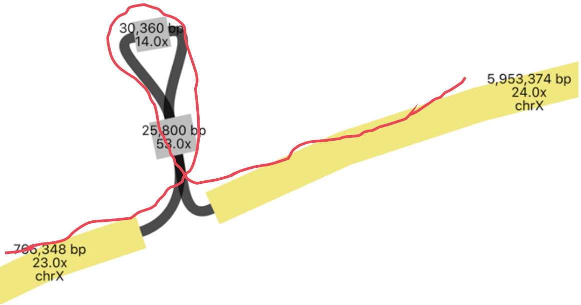
Simplifying the HiFi graph

- Walking simple paths

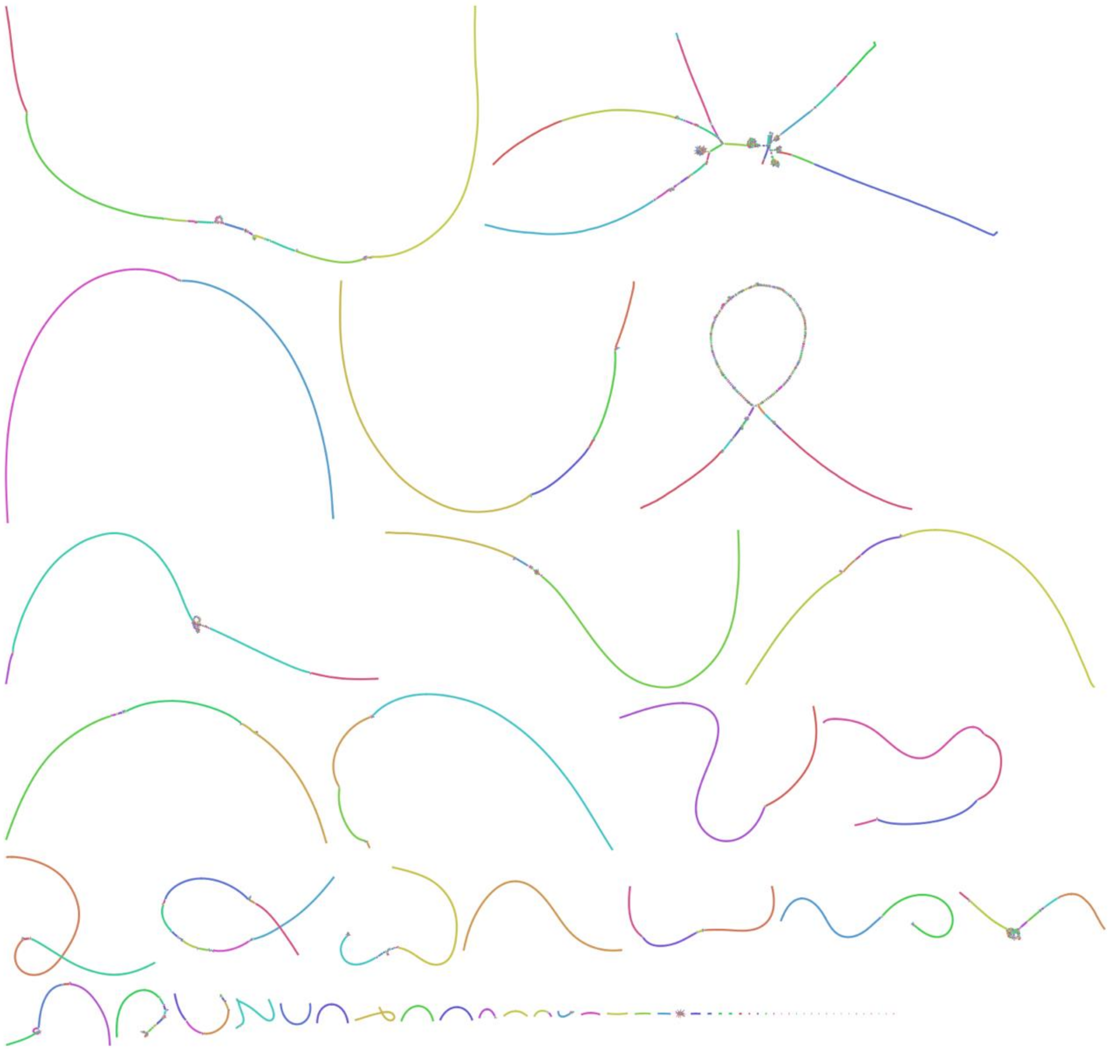


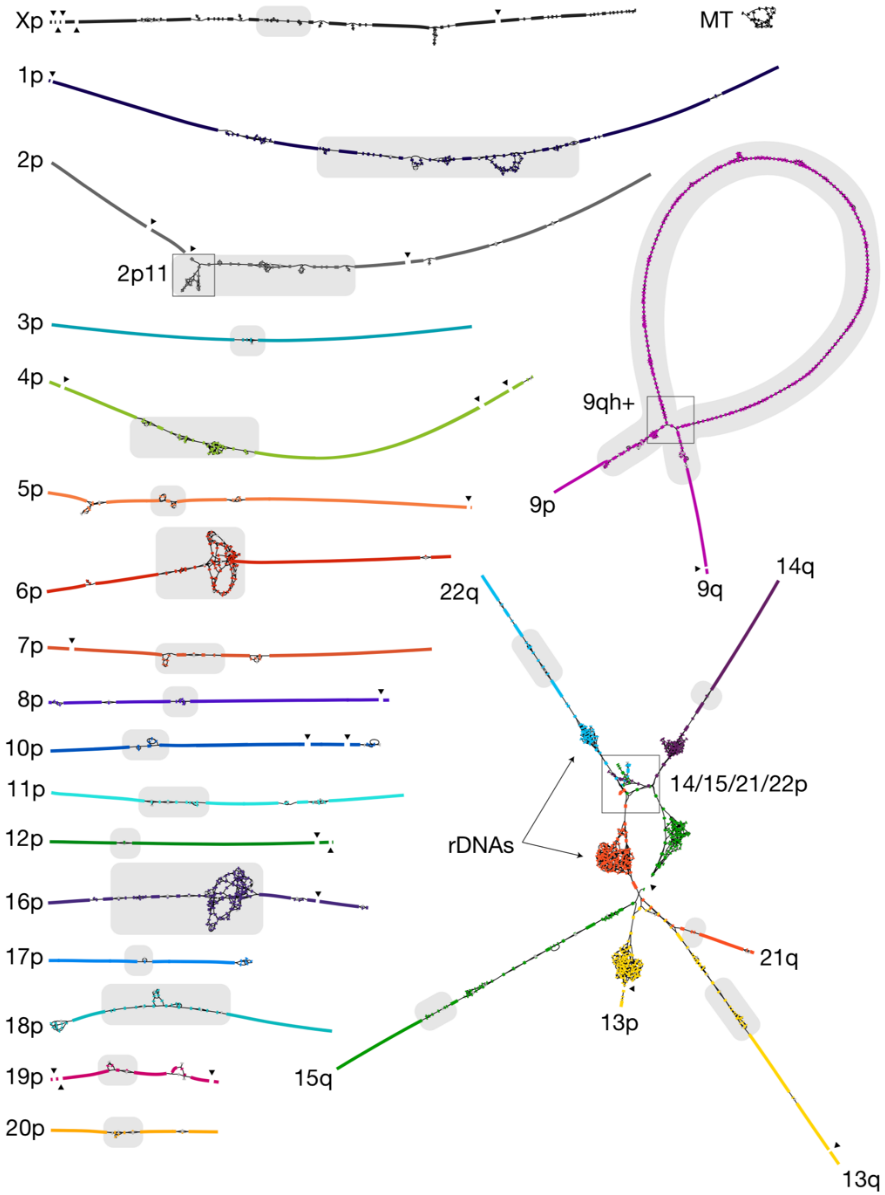
- Aligning Nanopore reads



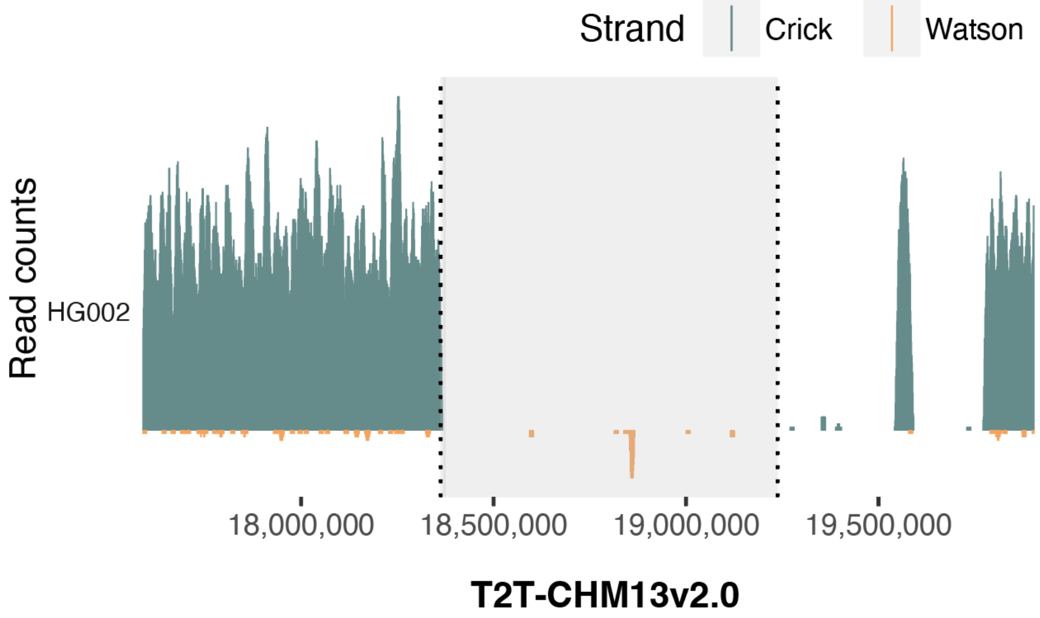


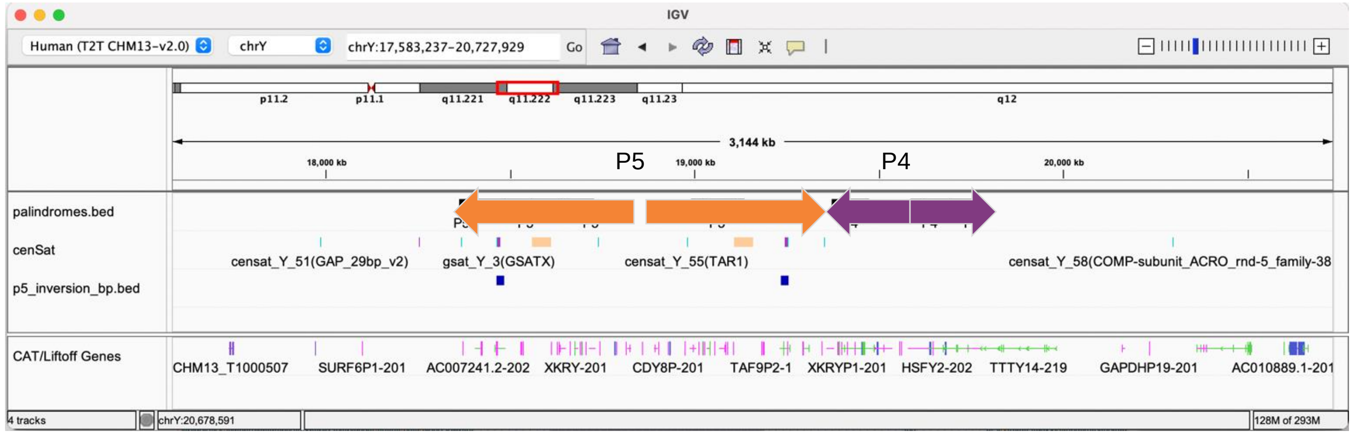
Manual analysis is time consuming

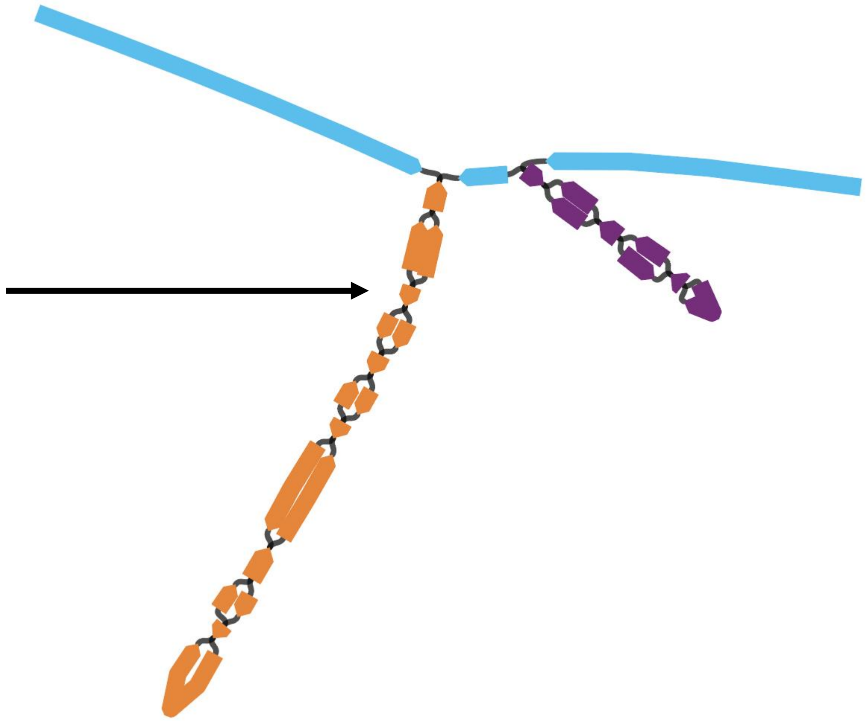




…and error prone





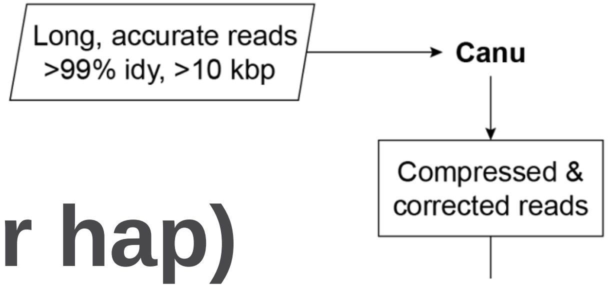


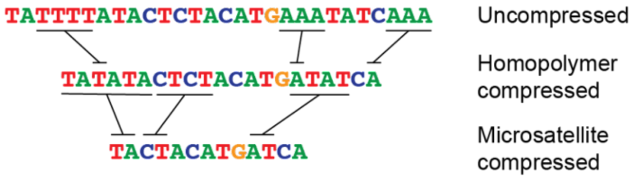
P5 mis-assembly (~800 kb)

~20 kbp 100% identical

Can we automate this?

Verkko!





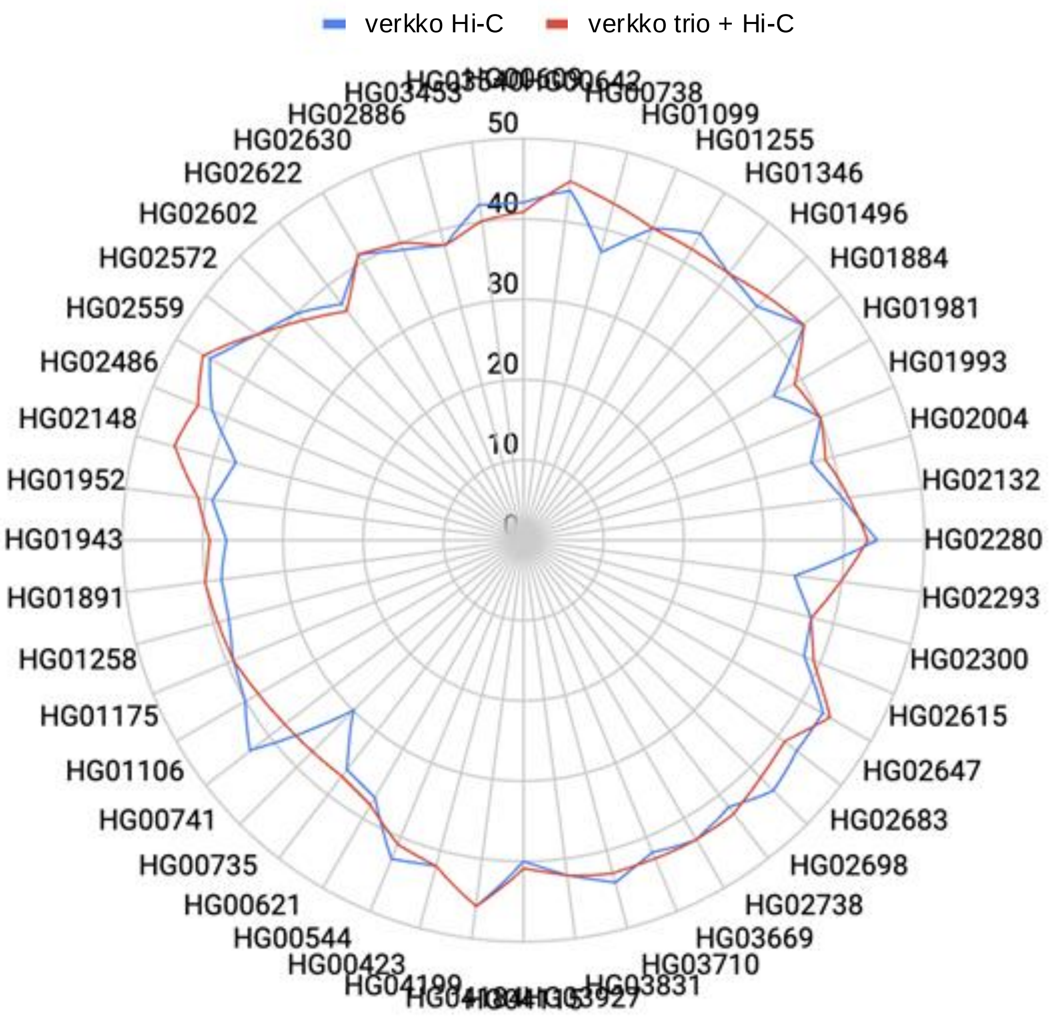
- **Sequencing recipe (per hap)**
  - 20-25x high accuracy
    - (Pac Bio HiFi, Duplex, HERRO)
  - 15-20x ONT ultra-long (>100 kb)
  - 20x Illumina Trio or Hi-C
  - Available from conda
- **Verkko pipeline**
  - Read correction
  - Sparse multiplex DBG
  - ONT graph simplification
  - Walk haplotypes
  - Haplotype consensus

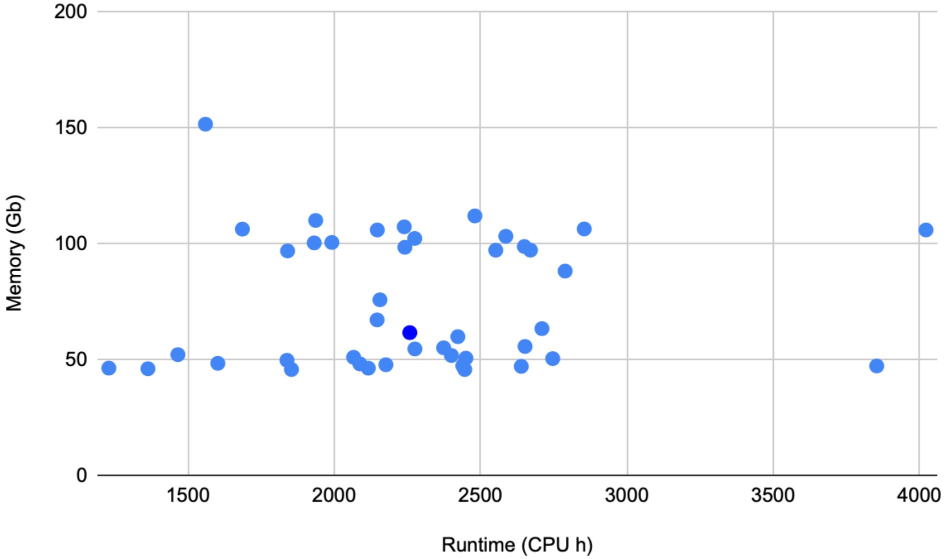
Verkko: telomere-to-telomere assembly of diploid chromosomes
Rautiainen, *et al.* bioRxiv (2022)

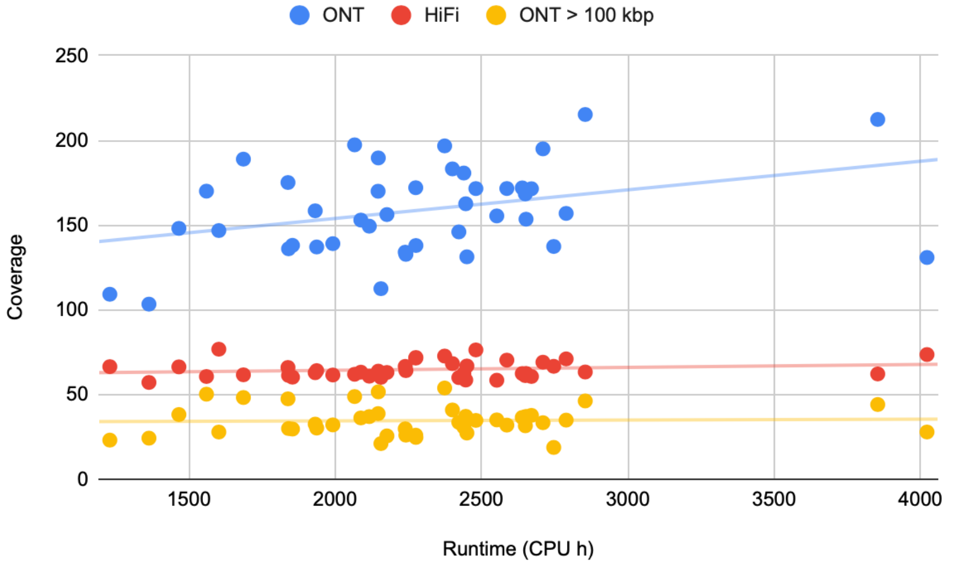
LJA: Assembling Long and Accurate Reads Using Multiplex de Bruijn Graphs
Bankevich, *et al.* Nat Biotech (2021)

State of the assembly, Sept 2024

- Our assemblies are strong, 40/46 T2T scaffold average on 101 human samples
    - 62x HiFi, 34x >100 kbp ONT (158x total), 68x Hi-C

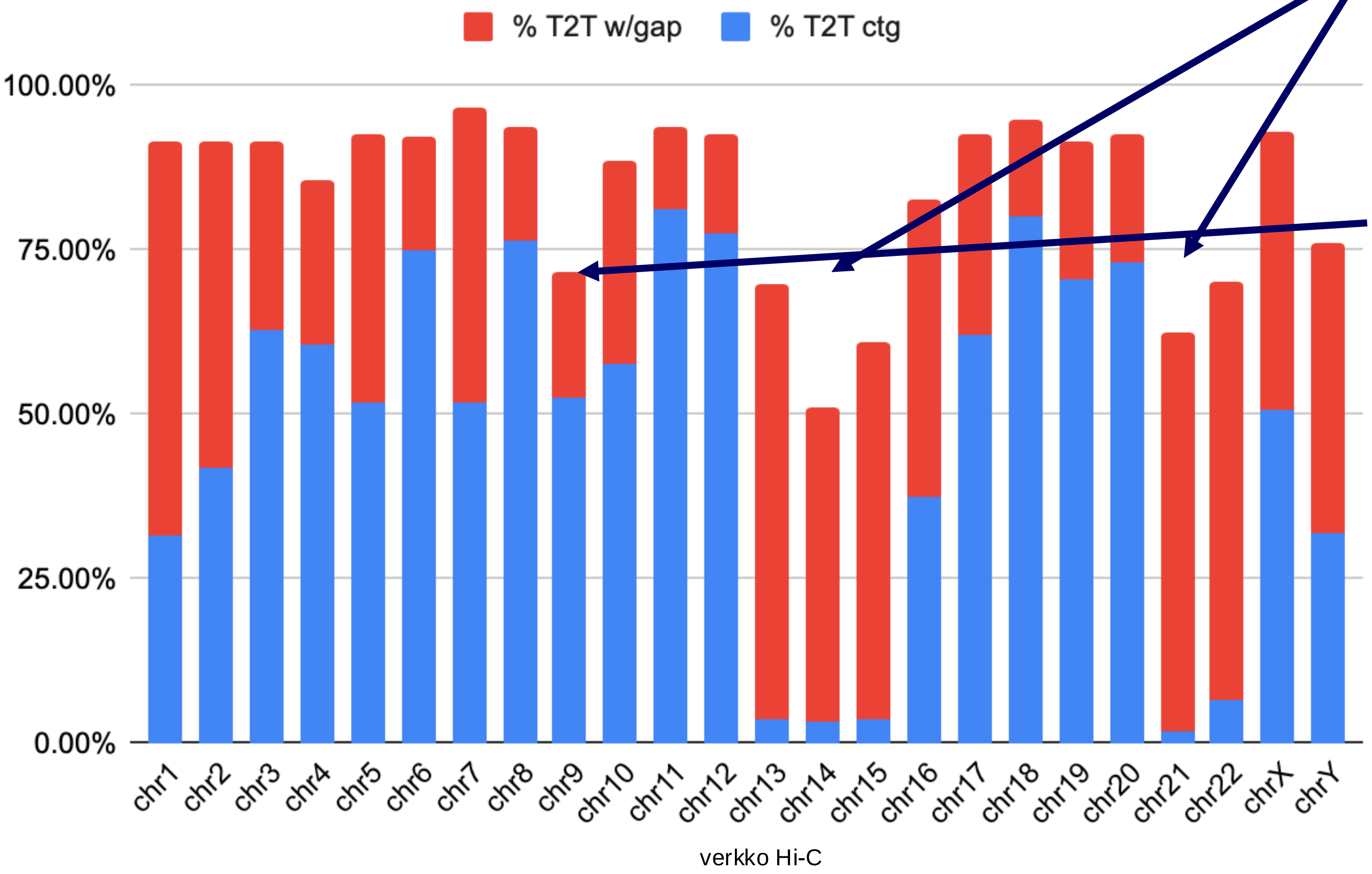






State of the assembly, Sept 2024

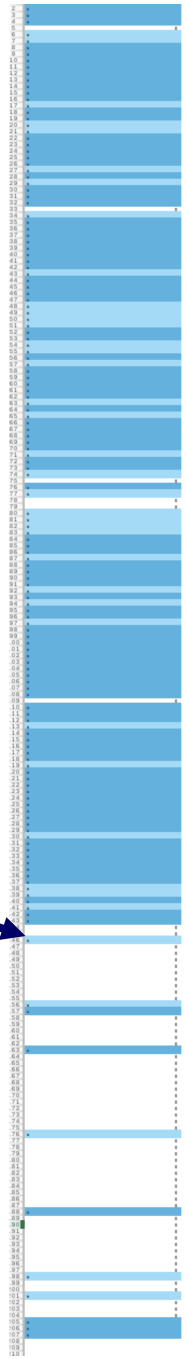
- Another view, by chromosome, 52% T2T contig, 91% T2T scaffold



Acrocentrics

Unexpected

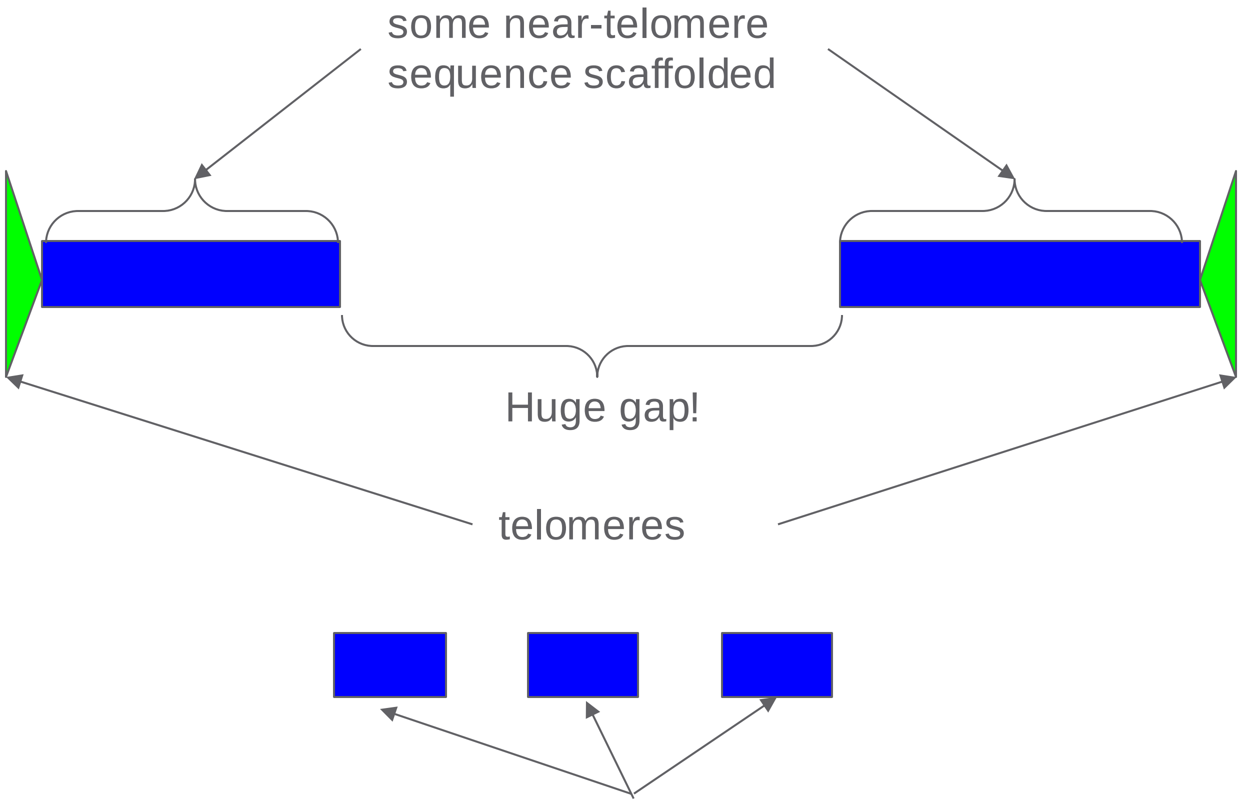
Switch to Revio 24hr movie (30hr is OK)



How do we know it’s any good?

T2T QC

- T2T contigs and scaffolds
- QV
  - Merqury, yak
- Hamming & switch error rate
  - If trio data available
- Missing/duplicated core genes
  - Compleasm, busco
- non-T2T contiguity metrics
  - N50, L50
- Alignment-based evaluation
  - NucFreq, Flagger, VerityMap



Contigs from the middle of the chromosome

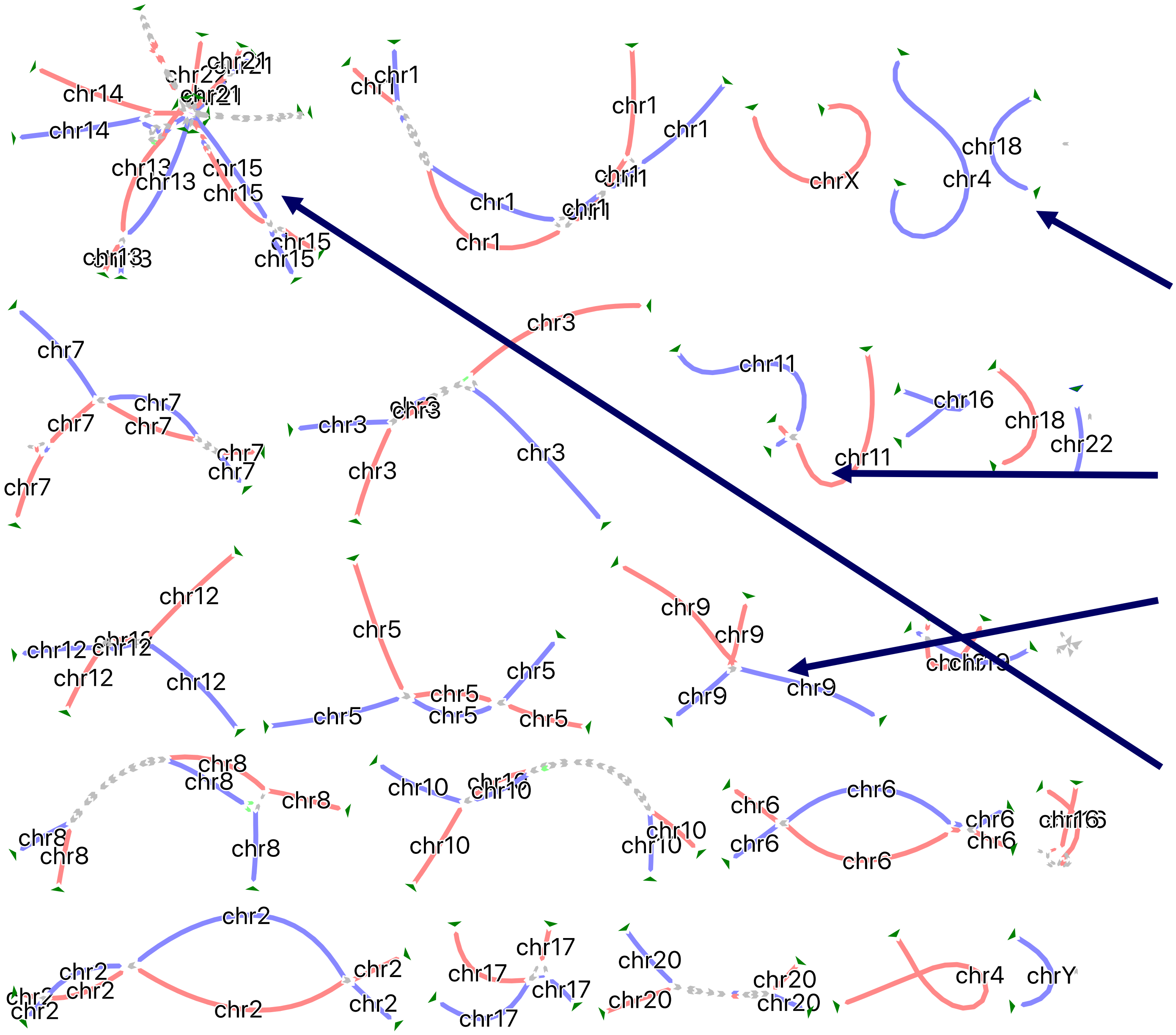
None of the metrics on the left helps to see that something is not right here!

QC: Not only summary metrics!

- Typically include detailed locations of problematic regions
    - yak trioeval: which contig has most switch errors? Are there lots of “small” switches causing hamming error or one big one?
    - compleasm: on which chromosome are the missing genes? On which scaffold are the duplicated genes?
- With verkko-generated *assembly.scfmap* and *assembly.paths.tsv* you can locate those problematic places in graph and sometimes see something interesting

Genome graphs are our friends

Graph, the "Good, the meh and the ugly"



Q20 Herro-correct, ≅300 nodes
- 44 T2T scfs (36 ctgs)

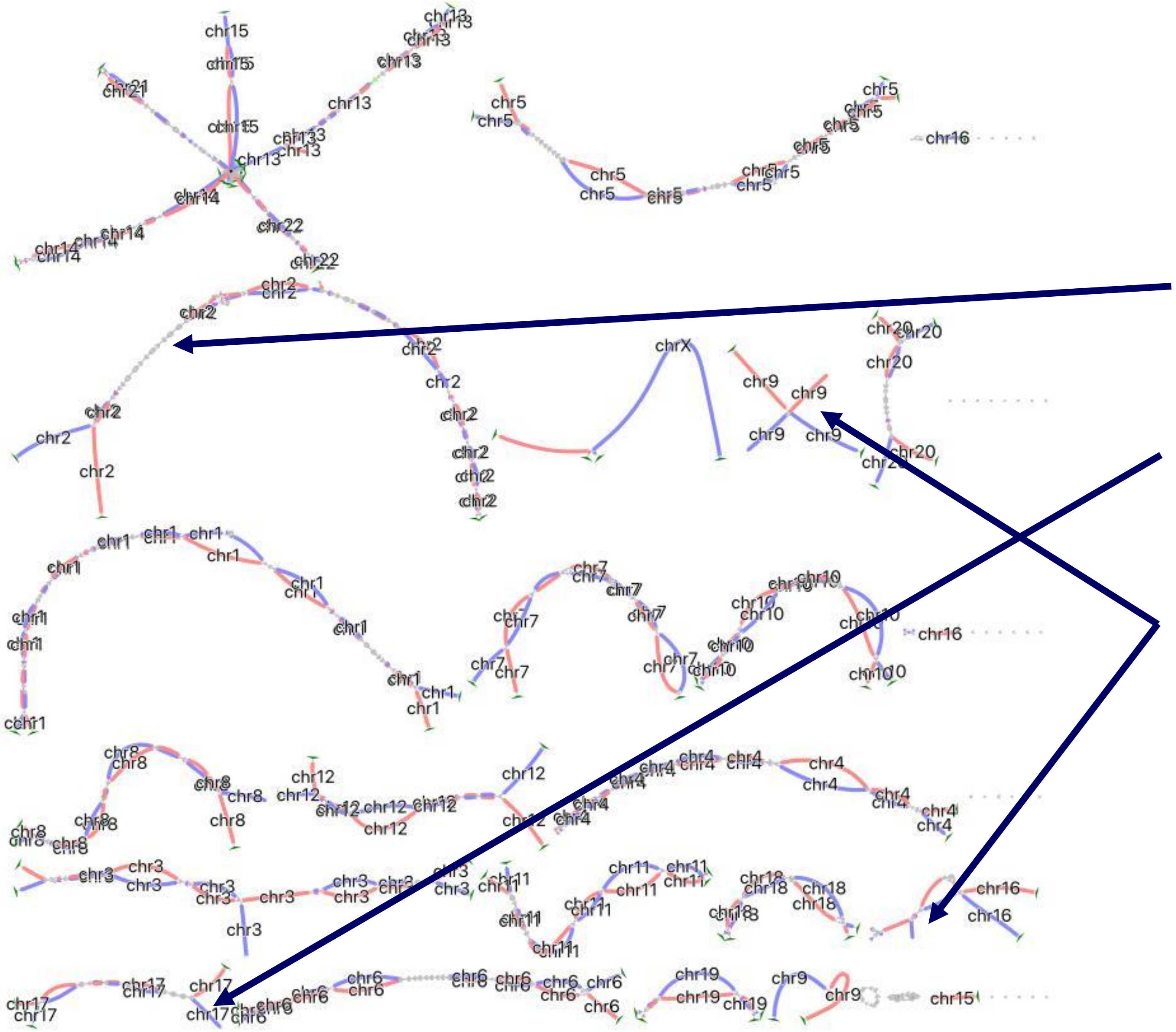
Large nodes, little gray (hom) or yellow (switch)

Telomeres visible in all components (4 in diploid/2 in haploid)

No hanging edges, all nodes have neighbors or go to telomere

Clear rDNA tangle (species specific)

Graph, the "Good, the meh and the ugly"



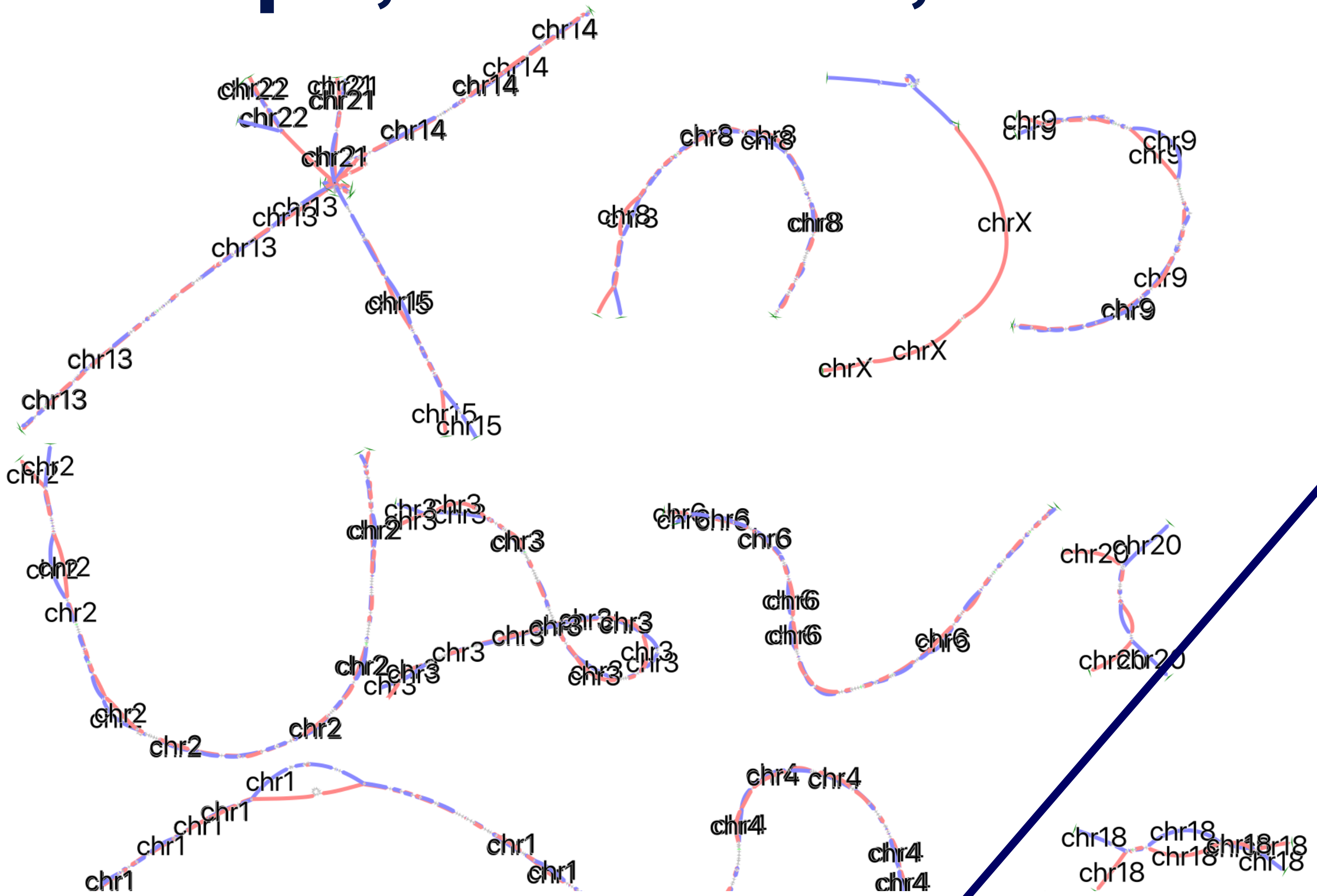
Revio/R10 YR4, ≅900 nodes
- 37 T2T scfs (25 ctgs)

Shorter nodes, more gray (hom) or yellow (switch)

Telomeres missing when expected

Hanging edges, missing connection in one or both haplotypes

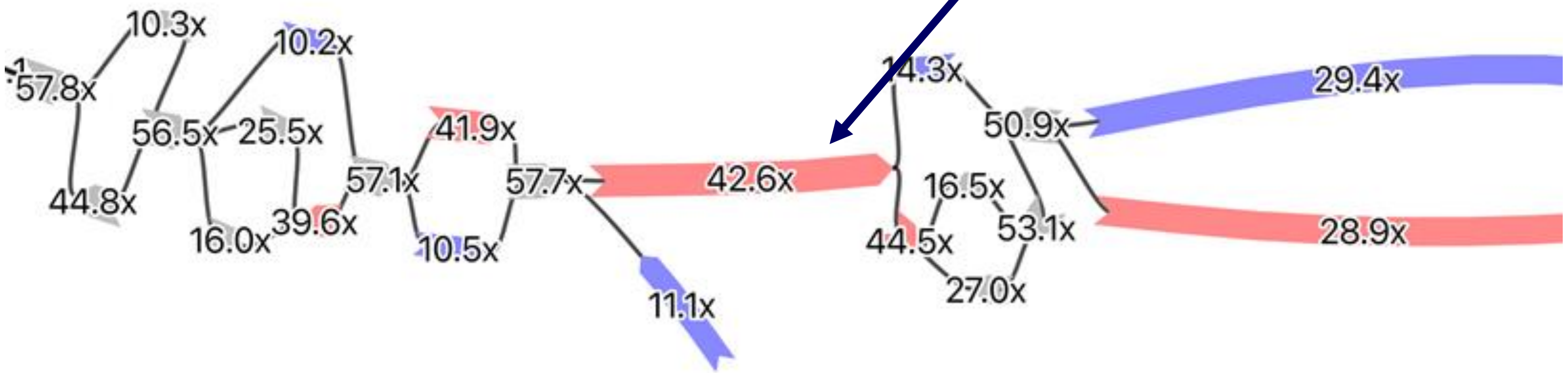
Graph, the “Good, the meh and the ugly”



Sequel II, R9 ≅2000 nodes
- 40 T2T scfs (31 ctgs)

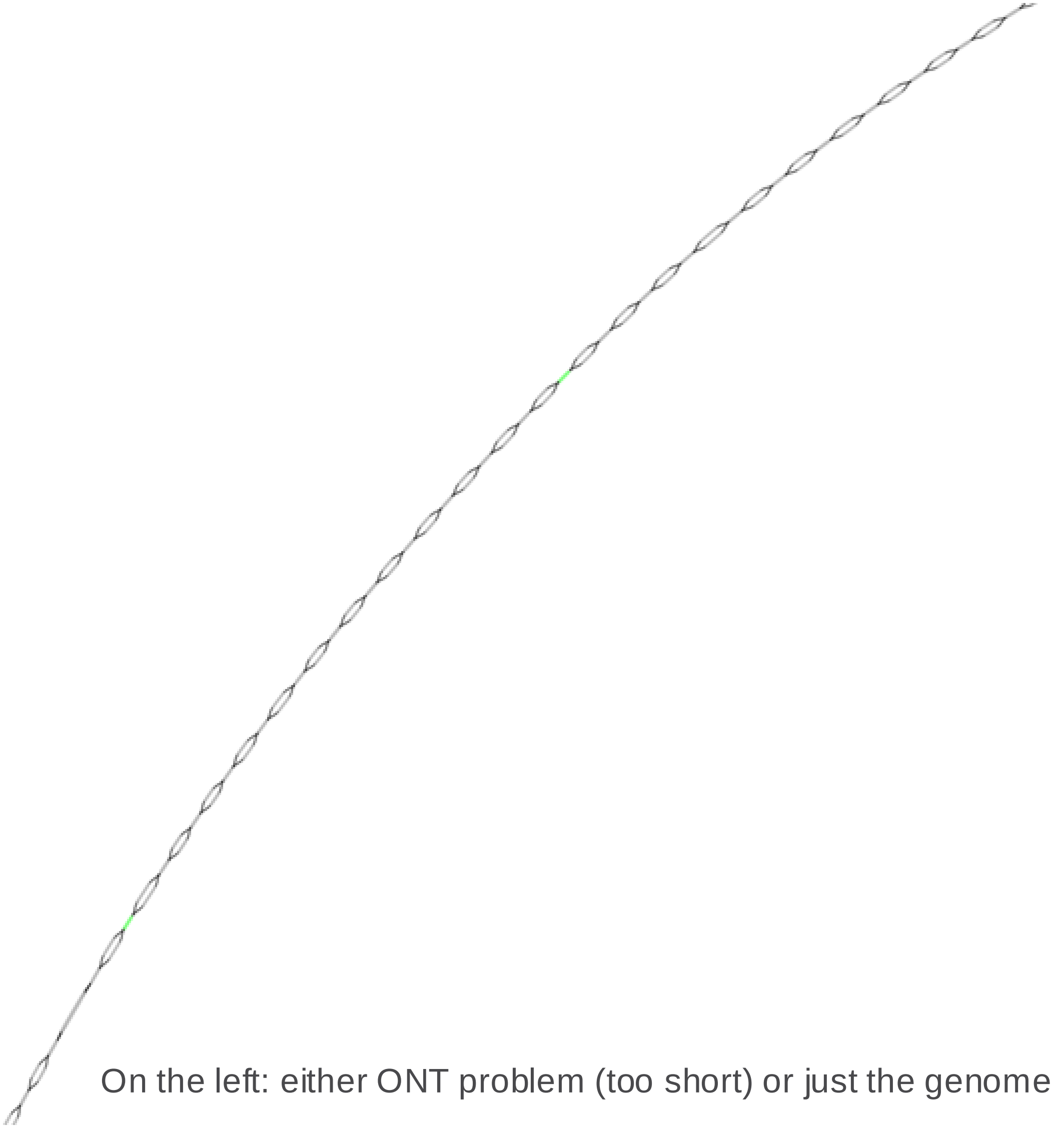
Shorter nodes, more gray (hom) or yellow (switch)

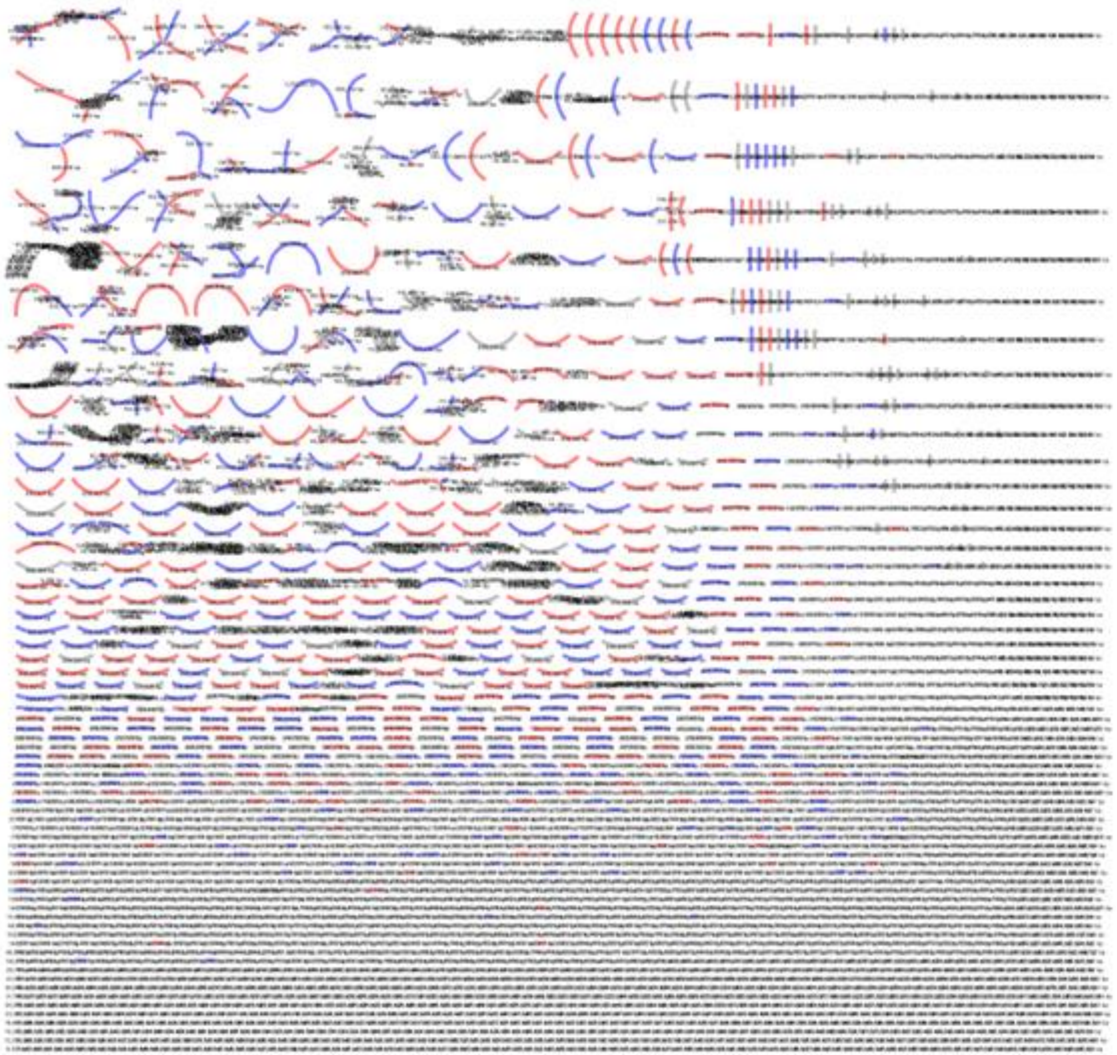
Non-diploid structure, cell line issue?



What can go wrong?

Too few T2T, fragmented assembly

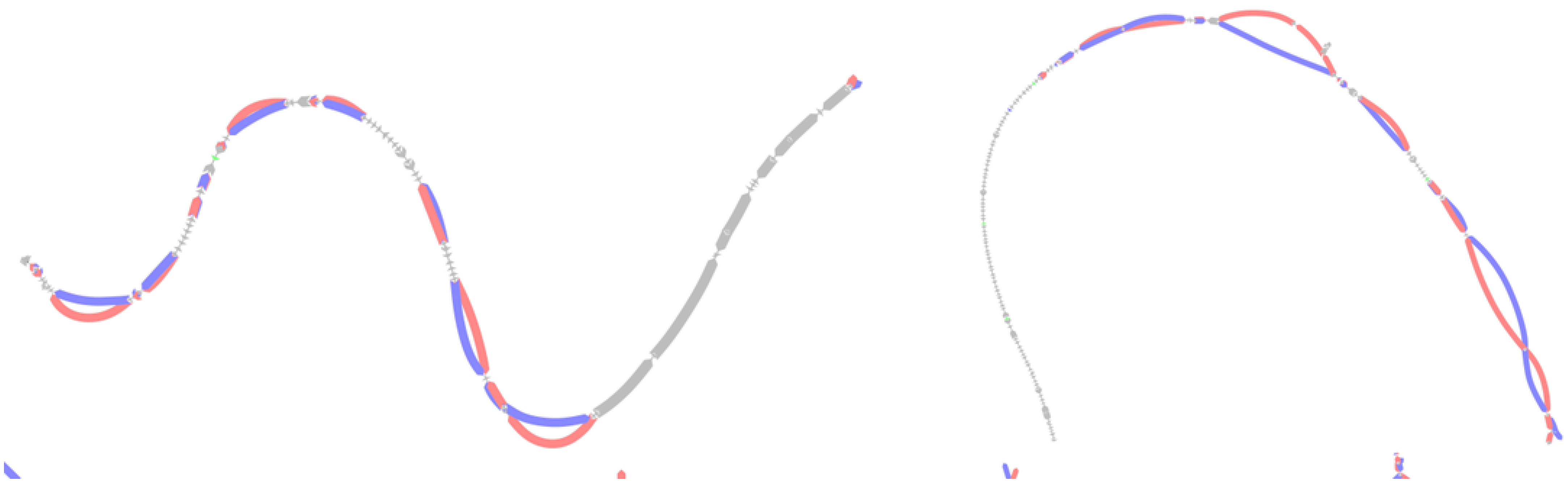




On the left: either ONT problem (too short) or just the genome structure

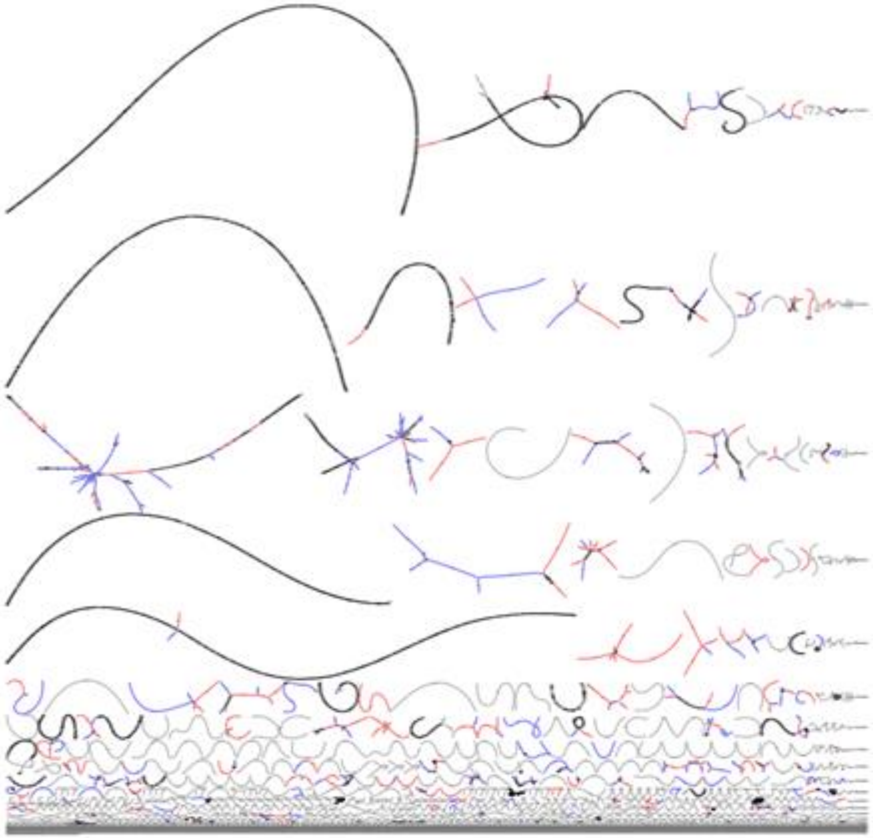
On the right: HiFi ultra-low input protocol problems

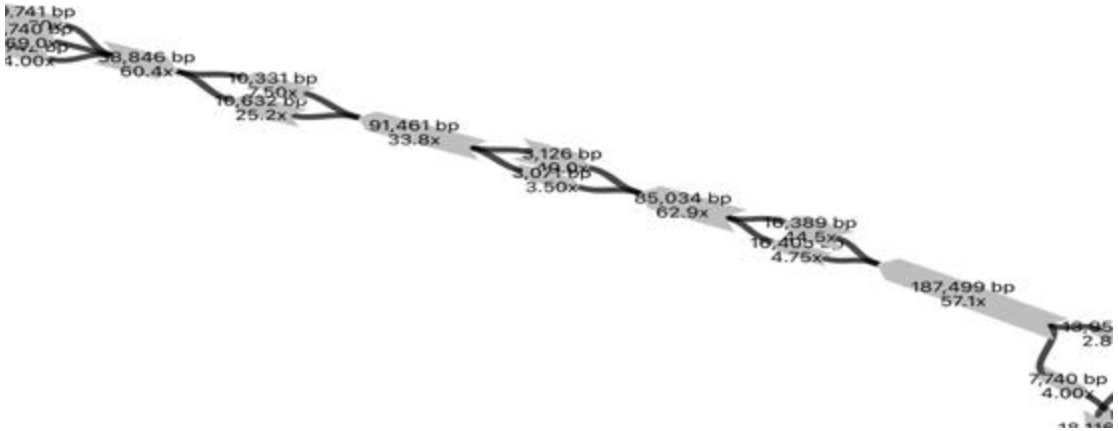
Phasing issues: large homozygous regions



- Different chromosomes of same bonobo sample
- Left is phased correctly (long nodes), right - lots of unassigned (and so missing genes)

Heterozygosity level matters!

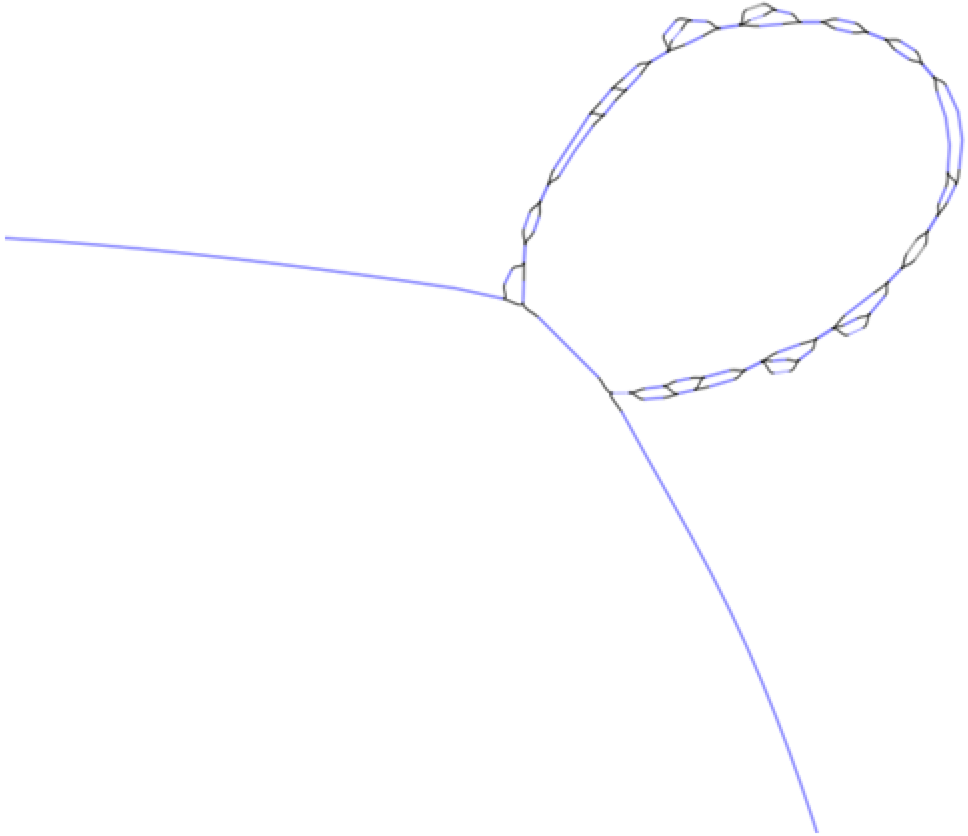




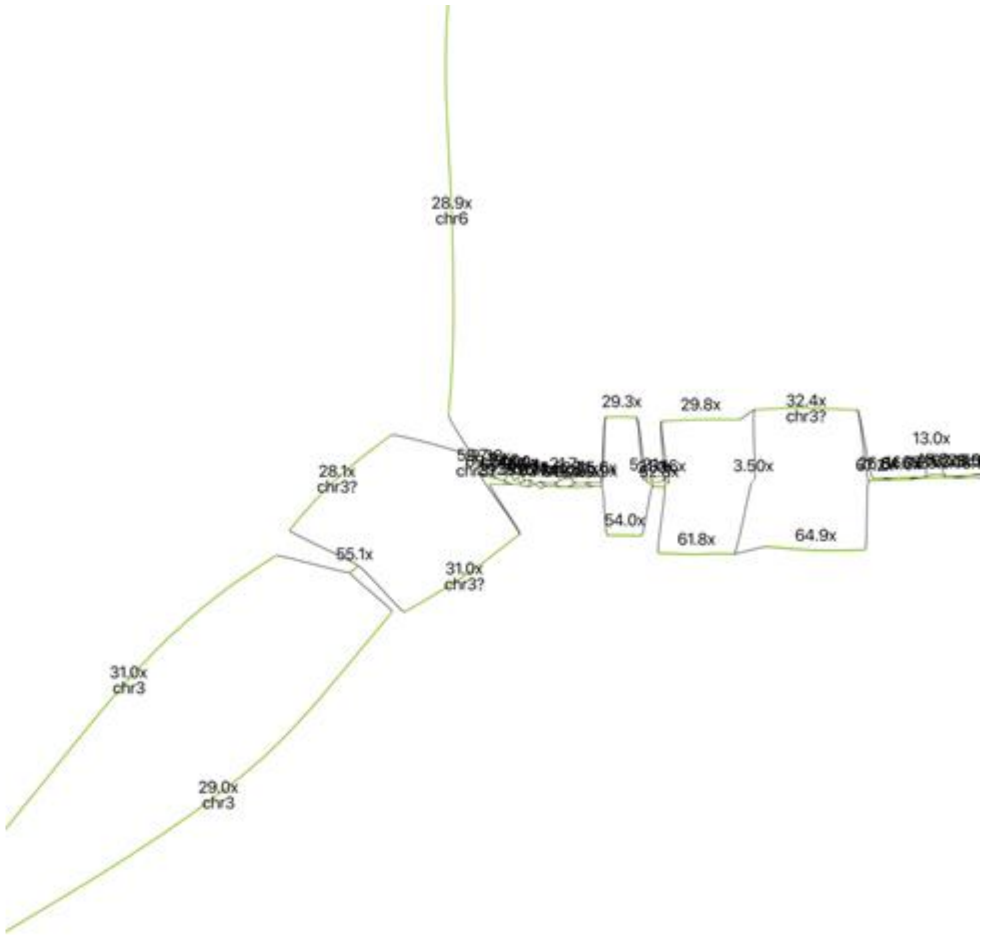
- Verkko can have problems with both very high and very low heterozygosity
- Sometimes this may even happen in the same sample!

Large tandem repeats

- Large (few Mb) tandem repeats is quite typical issue preventing verkko from T2T.
- Verkko/rukki heuristics stops because there are multiple large “blue” extensions for a large blue node here.
- Usually random walk will not add many errors here



“Biological” conclusions from graphs



- Part of chr6 on one of the haplotypes is partially replaced with chr3!
- Coverage confirms “triploidy” for half of chr3
- Still can be a cell line issue

Team T2T (…and many more)

T2T
TELOMERE-TO-TELOMERE CONSORTIUM

DNAnexus

amazon web services

PACBIO®

Oxford NANOPORE Technologies

ARIMA GENOMICS

NIH NHGRI