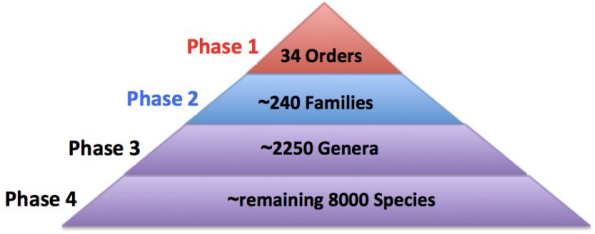
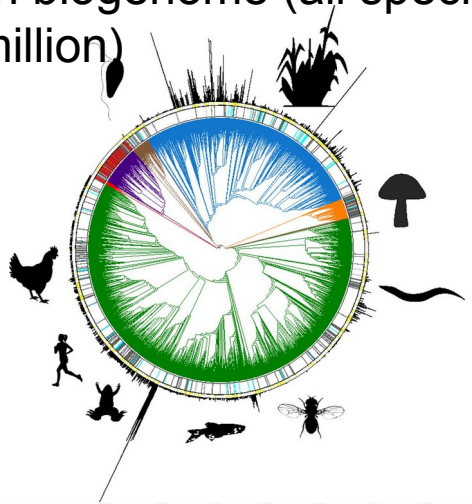
# De novo assembly

Figure 1. New reconstruction of the avian family tree, based on whole genome sequences of 48 bird species from 34 orders. See here for more. Painting by Jon Fjeldså.



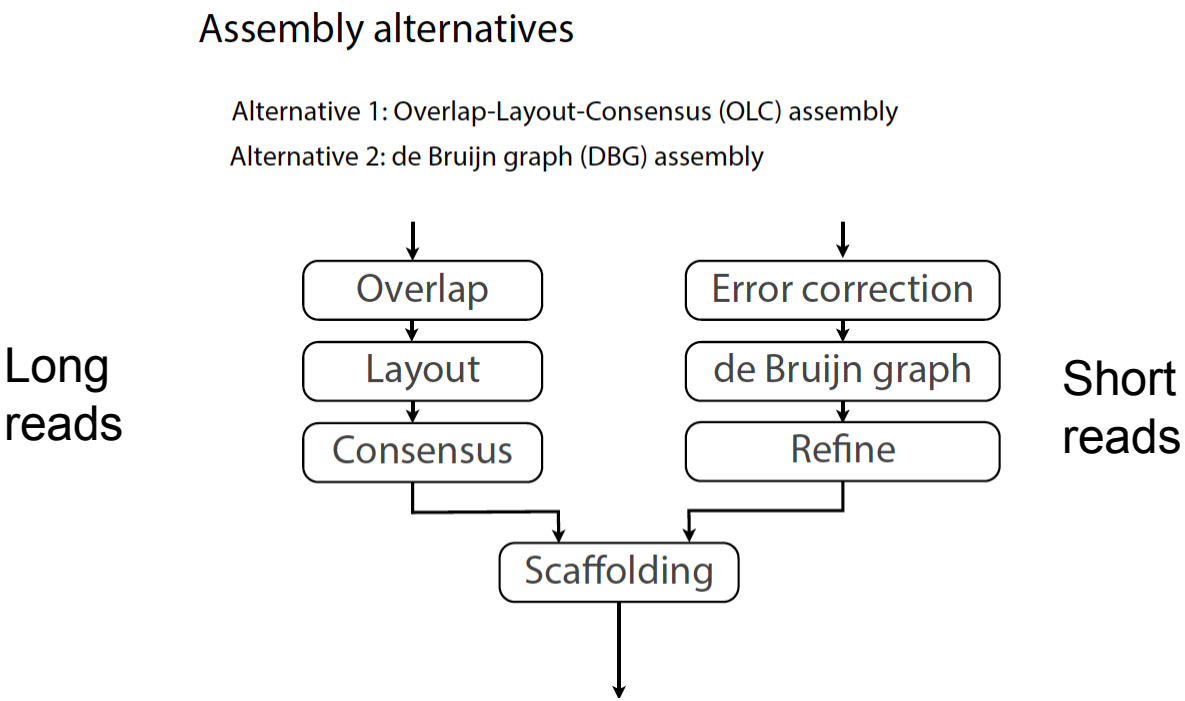

Earth biogenome (all species, >5 million)



# Basic idea

- Assembly by overlap
- Contigs constructed
- Scaffolds constructed by tying contigs together using other information
  - Mate pair information (until 2020)
  - Hi-C

# Two different ways of assembling



# Short read challenges

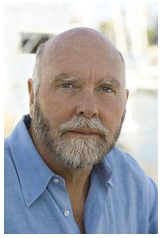
Computational

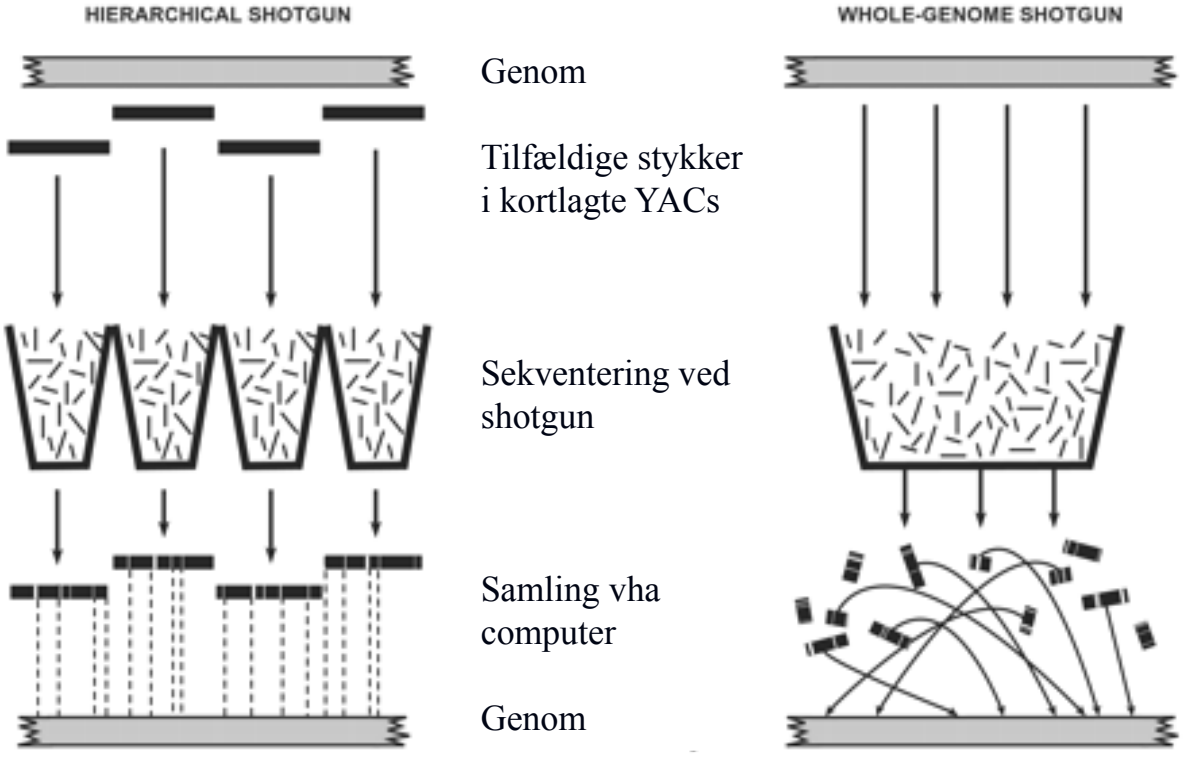
- Memory requirements
- Computational demand, in principle comparing all reads to all reads, 1 billion reads, 10 million trillion comparisons

Biological

- Repeats
- Duplications
- Uneven coverage

# Historical fight on the human genome 2001 – HUGO versus Craig Venter

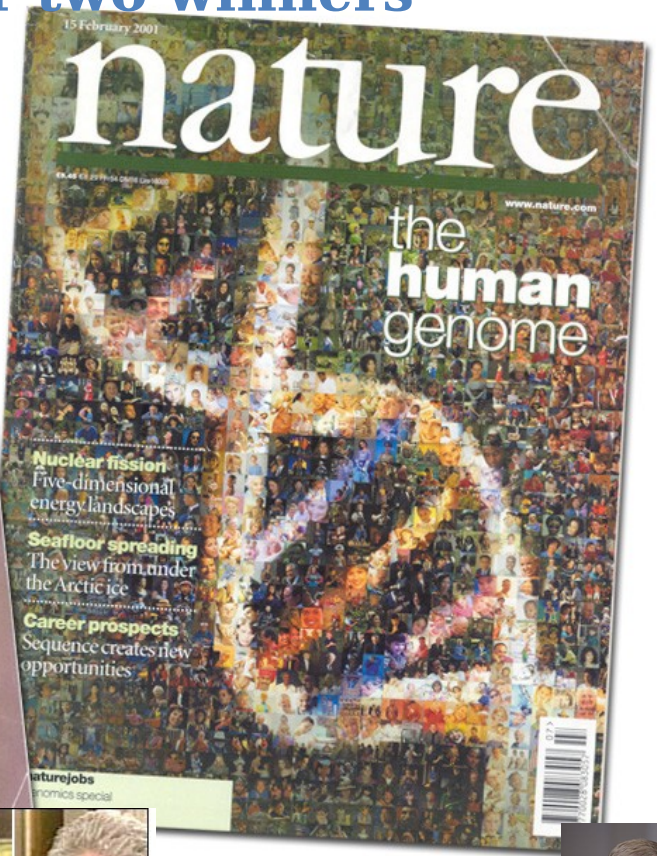


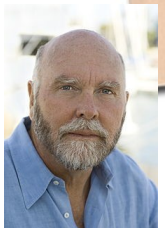



HUGO

Celera

# No winner or two winners

Celera

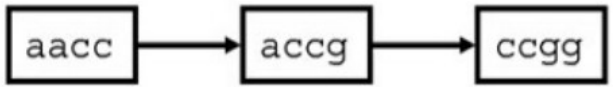
)1

# de Novo assembly algorithms

- Most popular short read assemblers are based on de Bruijn graphs of k-mers

aaccgg

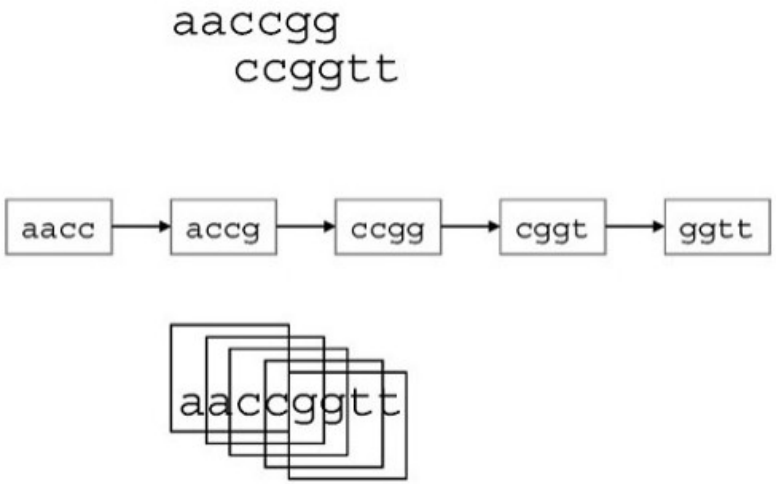
4-mers



# de Novo assembly algorithms

All popular assemblers are based on de Bruijn graphs of k-mers

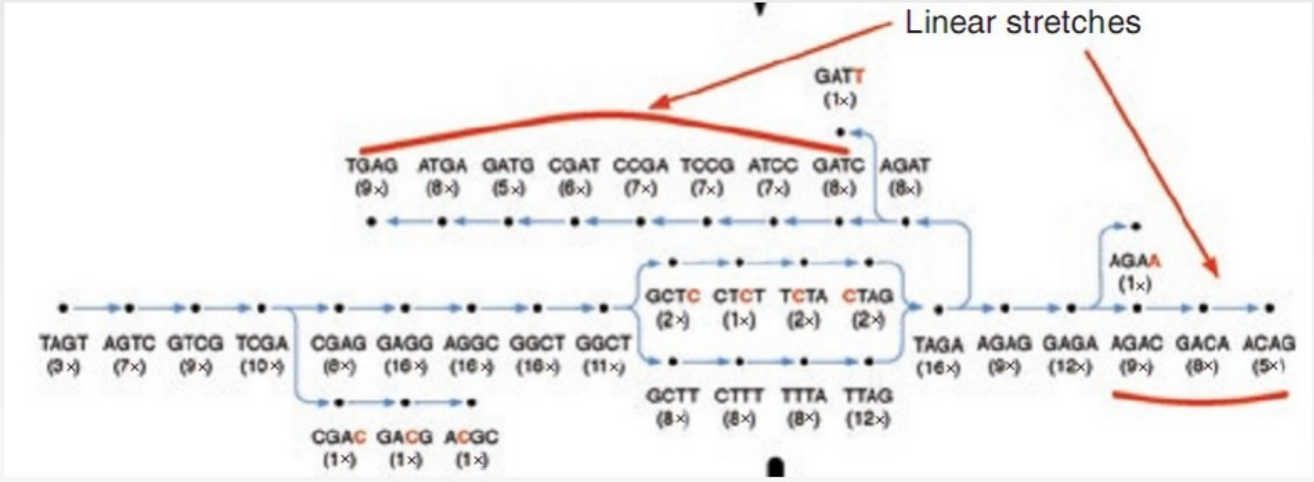
Combining two reads



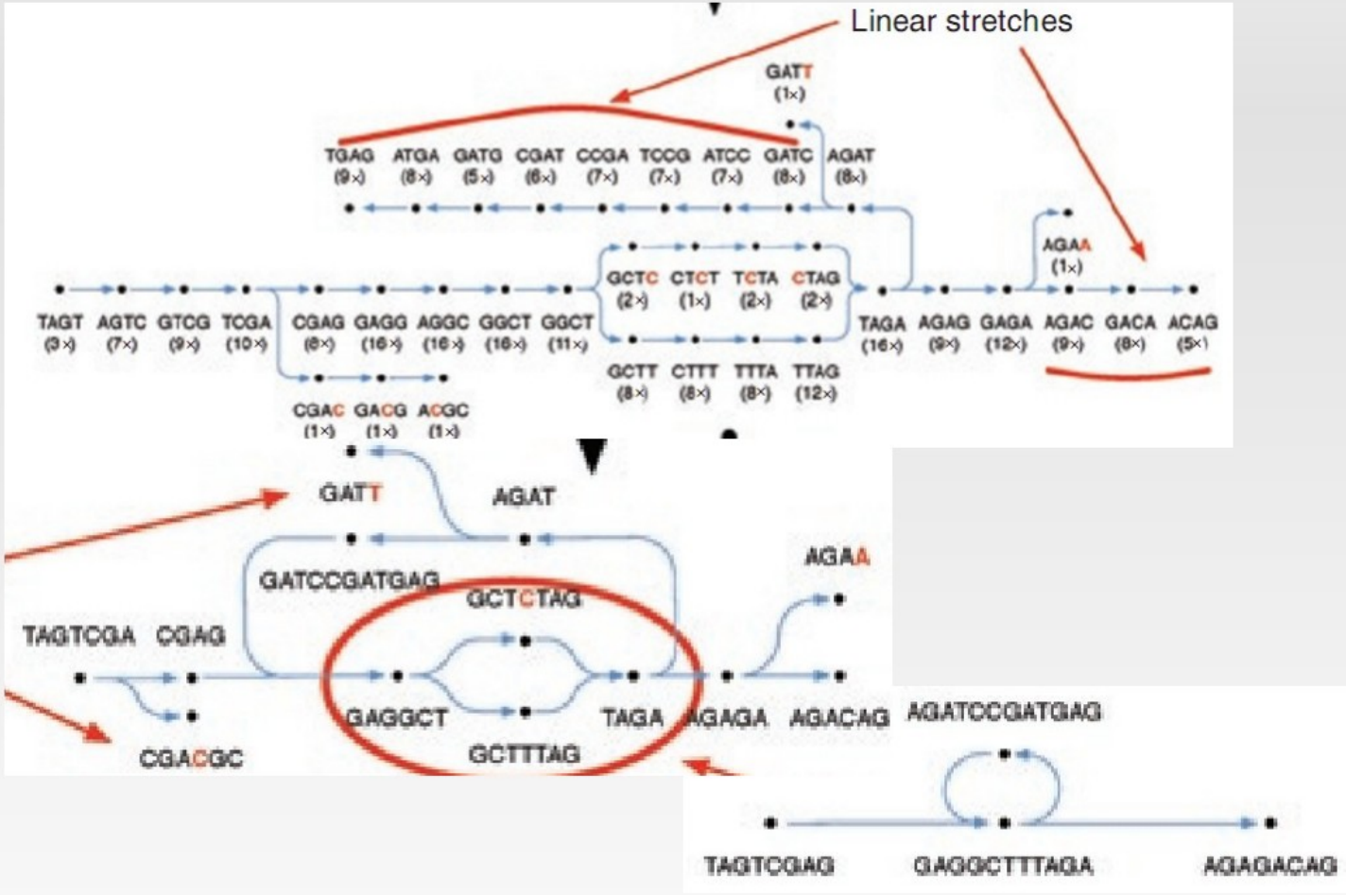
# Putting it all together … 1/2


Sequencing with errors

4-mers collected as a graph with linear stretches, bubbles & spurs

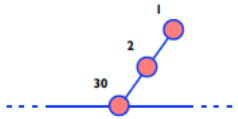


# Putting it all together … 2/2

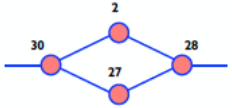


# Simplify graphs and making contigs

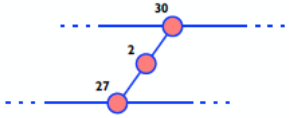
Clip tips
(seq err, end)



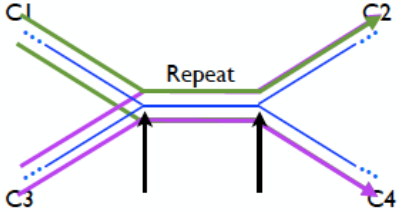

Pinch bubbles
(seq err, middle, SNP)



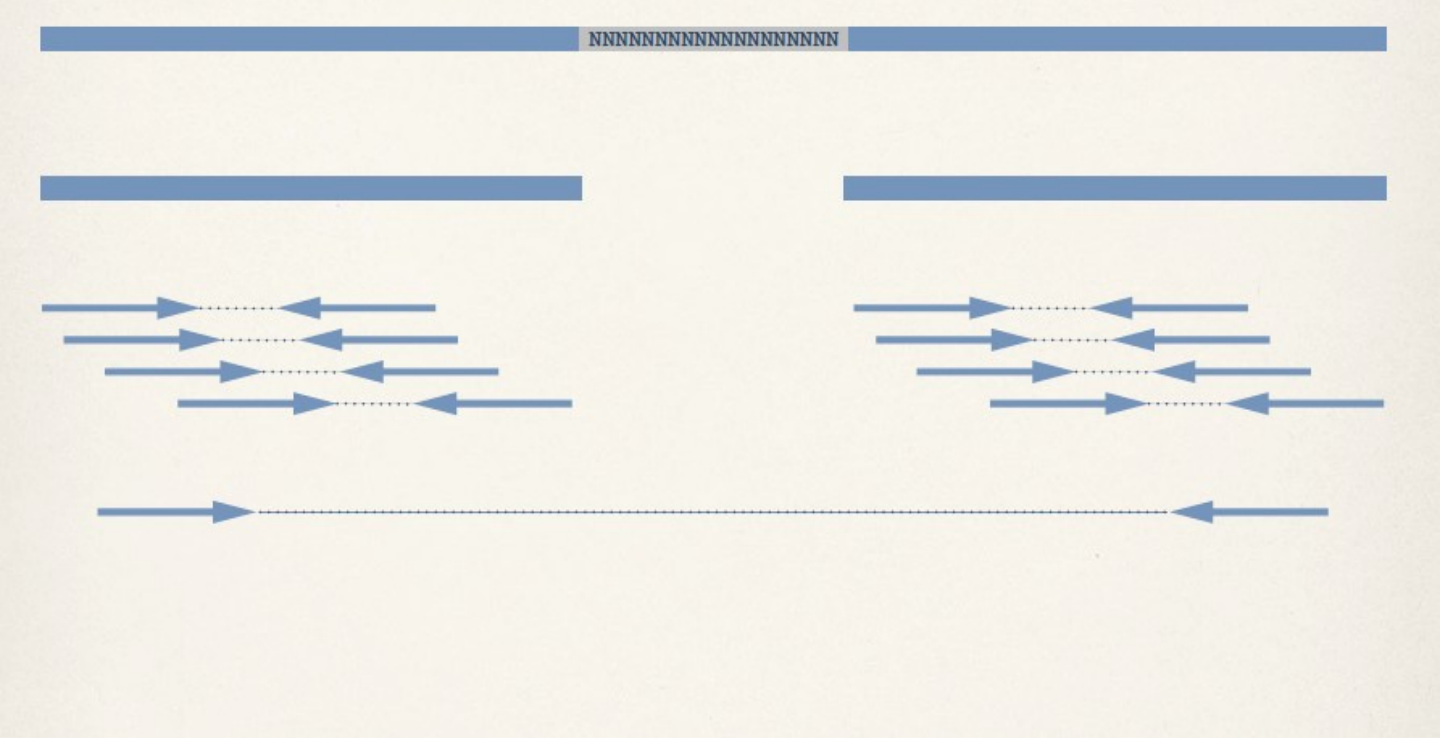

Remove low cov. links

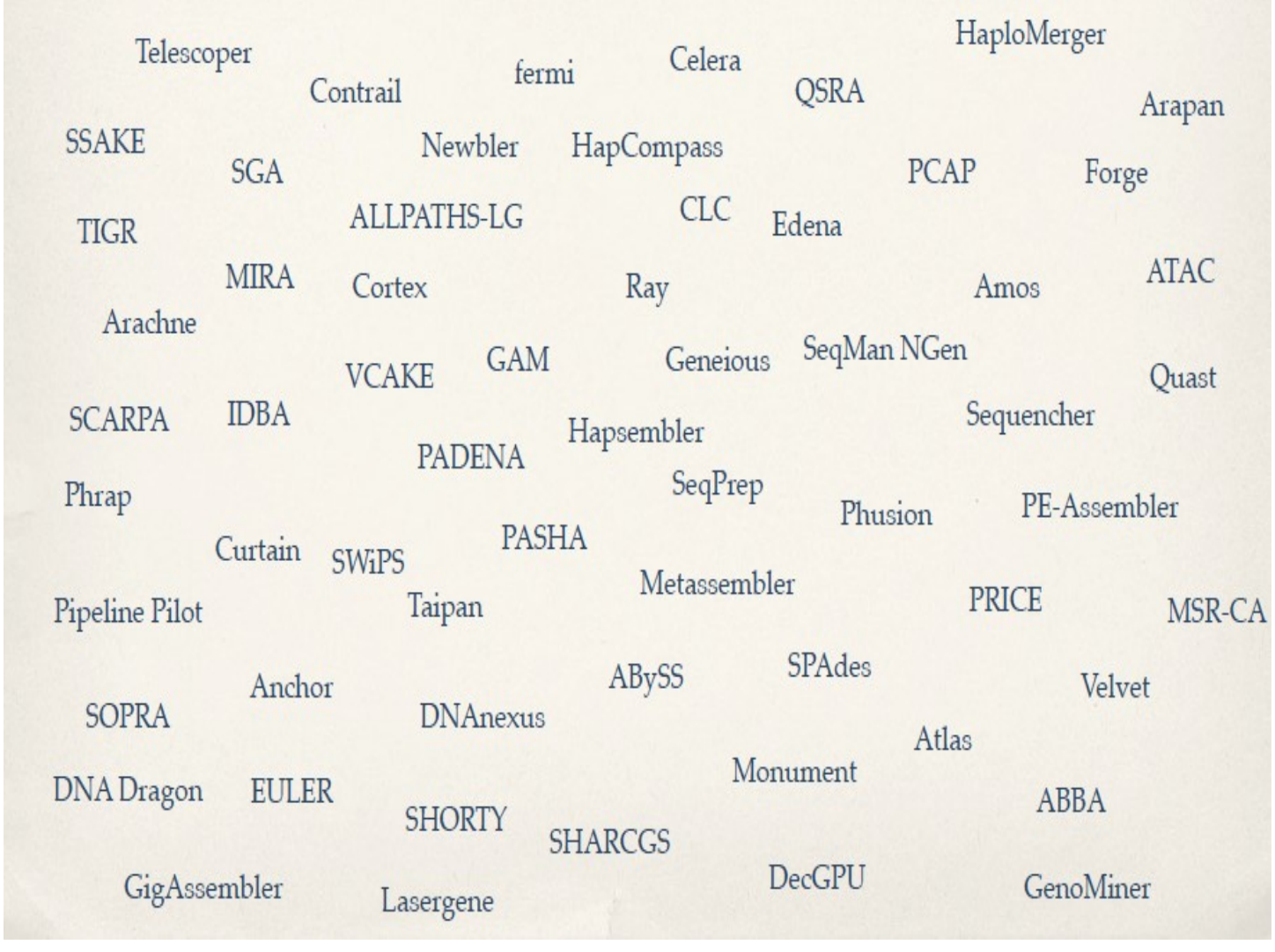


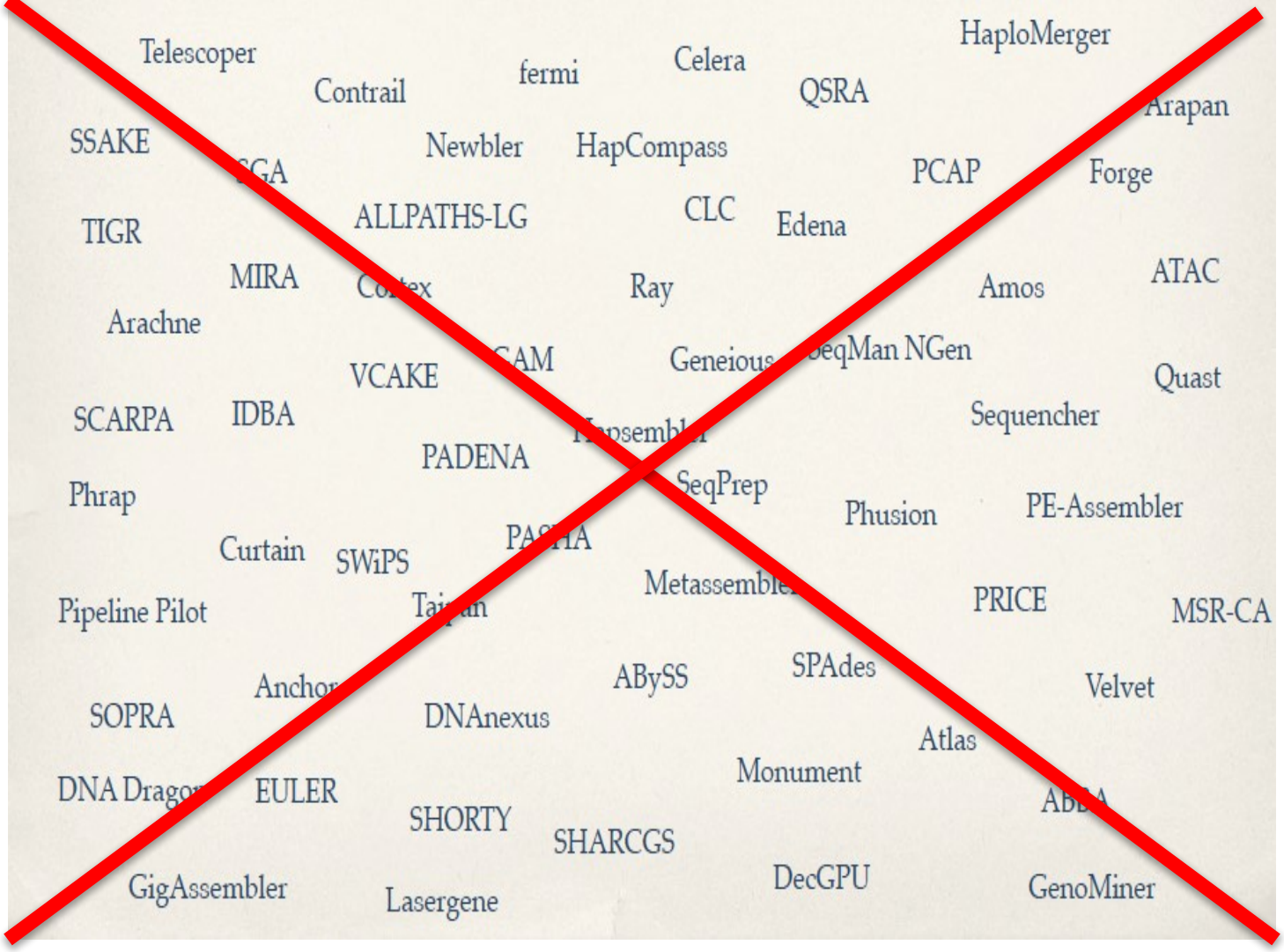

Cut graph at repeat boundaries to create contigs



# Scaffolding using paired or mate pair reads




HaploMerger
Telescoper
fermi
Celera
Contrail
QSRA
Arapan
SSAKE
Newbler
HapCompass
SGA
PCAP
Forge
TIGR
ALLPATHS-LG
CLC
Edena
MIRA
Cortex
Ray
Amos
ATAC
Arachne
VCAKE
GAM
Geneious
SeqMan NGen
Quast
SCARPA
IDBA
Sequencher
Hapsembler
PADENA
SeqPrep
Phrap
Phusion
PE-Assembler
Curtain
SWiPS
PASHA
Metassembler
PRICE
Pipeline Pilot
Taipan
MSR-CA
Anchor
ABySS
SPAdes
Velvet
SOPRA
DNAnexus
Atlas
Monument
DNA Dragon
EULER
ABBA
SHORTY
SHARCGS
DecGPU
GigAssembler
Lasergene
GenoMiner

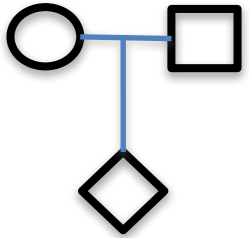

Telescoper
Contrail
fermi
Celera
HaploMerger
QSRA
Arapan
SSAKE
Newbler
HapCompass
PCAP
Forge
TIGR
ALLPATHS-LG
CLC
Edena
MIRA
Ray
Amos
ATAC
Arachne
VCAKE
Geneious
Quast
SCARPA
IDBA
Sequencher
PADENA
SeqPrep
Phrap
Phusion
PE-Assembler
Curtain
SWiPS
Pipeline Pilot
PRICE
MSR-CA
ABySS
SPAdes
Velvet
SOPRA
DNAnexus
Atlas
Monument
EULER
SHORTY
SHARCGS
DecGPU
GenoMiner
GigAssembler
Lasergene

# A Danish pan genome

Cataloguing **all** variation in a genome

# Design of the Danish pan genome 2012-2017

- 50 trios (father, mother, child), 150 individuals



- Sequence deeper and better than before
  - Base all analysis of denovo assembly

- Make a Danish reference for future (and previous) studies

# Computational challenges

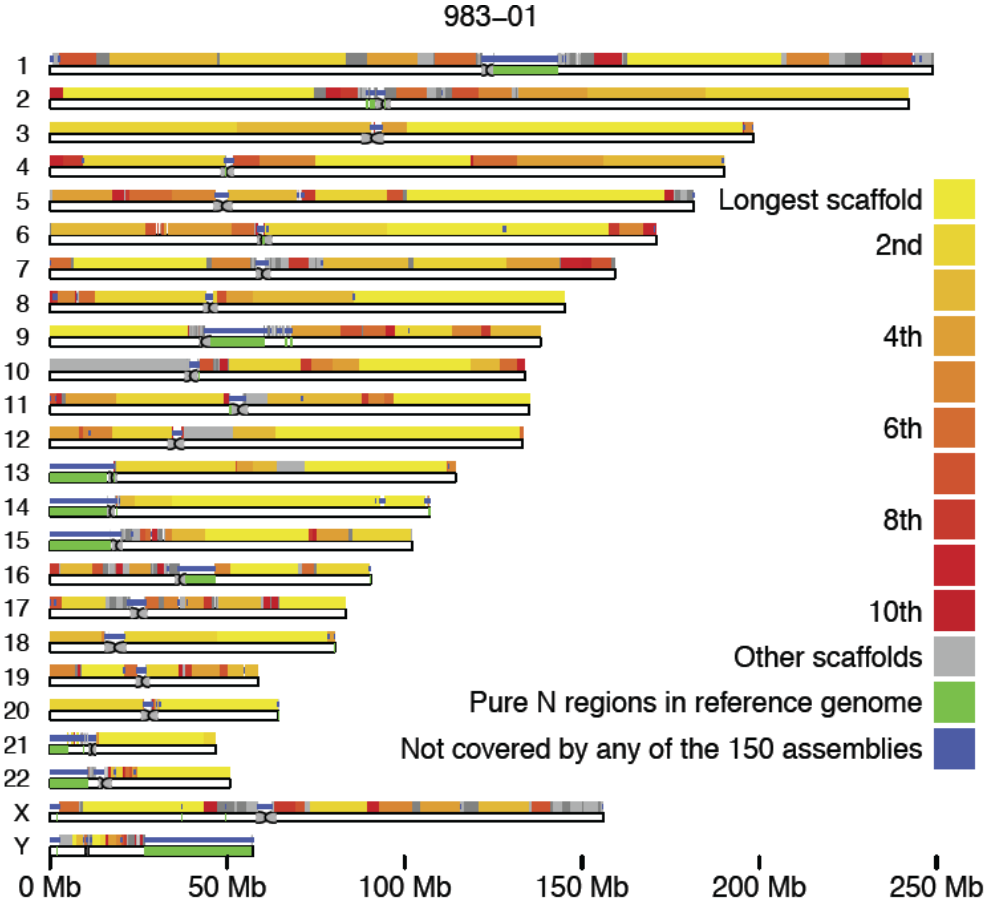
- 150 individuals * 3 Gb * 78 coverage
  = 351 billion sequences or 35.1 trillion base pairs

- Storage (peak of > 2 PB storage)
- Computation (>20 million CPU hours)
- IO intensive

# The genomeDK HPC cluster

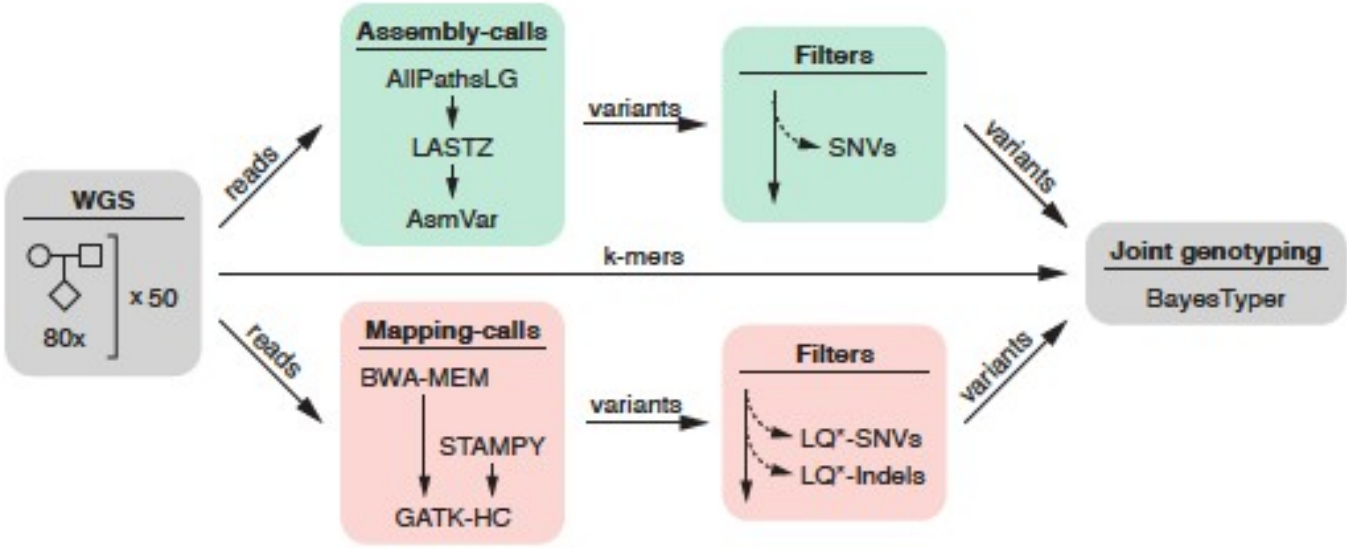
- genome.au.dk
- Opened in 2012
- Three full time administrators
- Wide range of bioinformatics software
- 8000 computing cores
- 12 PB very fast storage
- In 2017 joint datacenter with Region Midt

# Assembling genomes individually removes reliance on reference genome



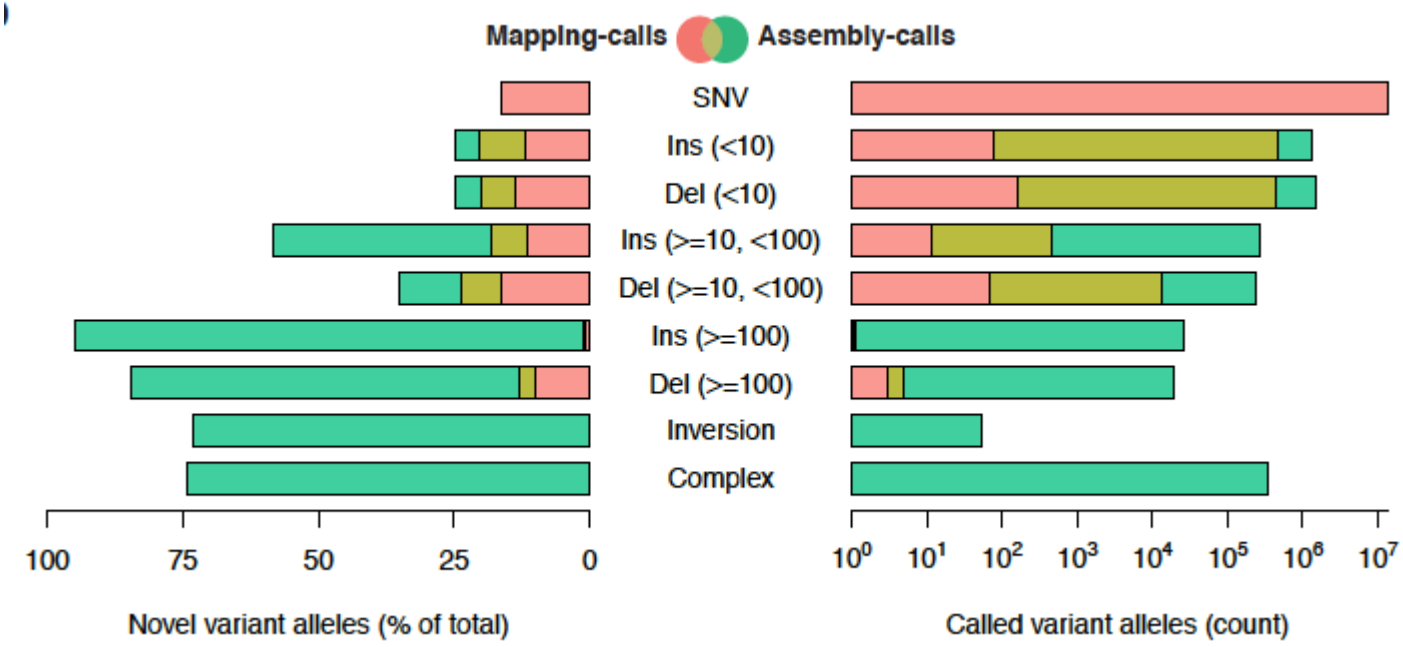

Maretty et al., *Nature*, 2017

# Variant calling pipeline




Maretty et al., *Nature*, 2017

# Lots of indel variation previously inknown




Maretty et al., *Nature* 2017

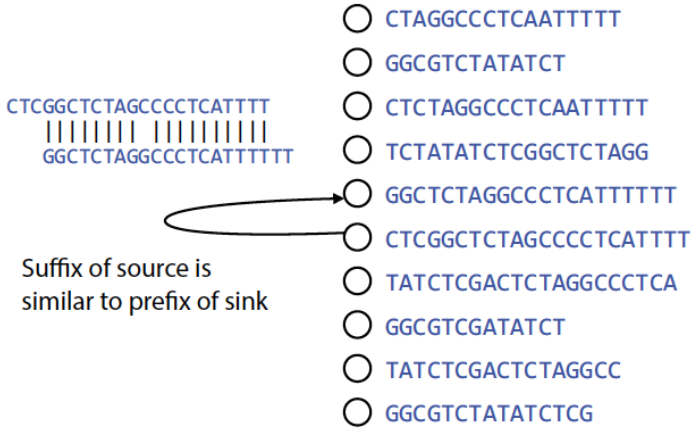
# Long reads change everything

- Pacbio:
  - 2018: 10-100kb, 5-10% errors, indels in particular
  - 2019: 30-200kb, <1% error
  - 2020: 10-20 kb, <0.1% error
  - 2022: 12-18 kb, <0.001% indel error, <0.0000001% SNP error
- Nanopore:
  - 2018: 50-1000kb, 10-15% errors, deletions in particular, some systematic
  - 2019: 50-2000kb, <5% error
  - 2021: 100-4000kb, 1% error
  - 2022: 20-4000kb, 0.1% error (duplex)
- Assembly by overlap (kmers discard far too much information)
  - SPADES (bacteria)
  - Canu (Celera update after 20 years)
  - Falcon (Pacbio)
  - Minimap and miniasm (PACbio, nanopore)
  - HiCanu and Hifiasm

# Principles of overlap graphs I

## Overlaps

Finding all overlaps is like building a *directed graph* where directed edges connect overlapping nodes (reads)

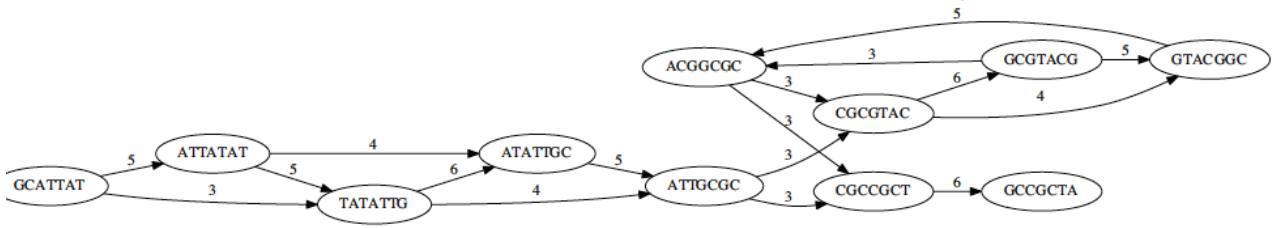


# Principles of overlap graphs II

## Finding overlaps

Example overlap graph with $l = 3$
$k = 7$

Edge label is overlap length




Original string: GCATTATATATTGCGCGTACGGCGCCGCTACA

# The principles of trio-binned Hifiasm

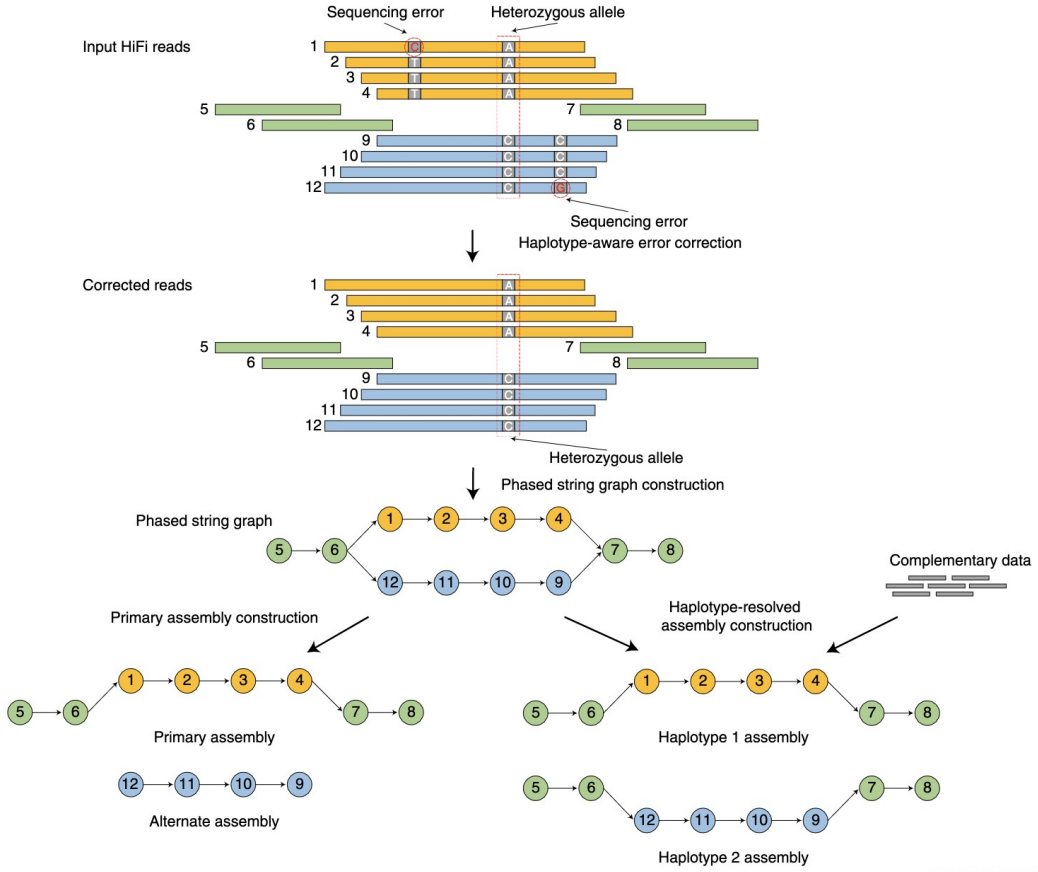


**Table 1 | Statistics of nonhuman assemblies**

| Dataset | Assembler | Size (Gb) | N50 (Mb) | NG50 (Mb) | Alternate size (Gb) | Completeness (asmgene or BUSCO) | |
|---|---|---|---|---|---|---|---|
| | | | | | | Complete (%) | Duplicated (%) |
| *M. musculus* (25×) | Hifiasm | 2.610 | 21.1 | 20.6 | 0.044 | 99.73 | 0.23 |
| | HiCanu | 2.594 | 16.0 | 14.8 | 0.077 | 99.68 | 0.22 |
| | Peregrine | 2.578 | 17.9 | 17.0 | 0.029 | 99.56 | 0.21 |
| | Falcon | 2.559 | 19.3 | 16.7 | 0.025 | 99.49 | 0.14 |
| *Z. mays* (22×) | Hifiasm | 2.190 | 37.5 | 37.5 | 0.095 | 99.85 | 0.17 |
| | HiCanu | 2.145 | 27.1 | 24.1 | 0.040 | 99.84 | 0.13 |
| | Peregrine | 2.205 | 10.1 | 10.2 | 0.038 | 99.88 | 0.26 |
| | Falcon | 2.132 | 9.5 | 9.3 | 0.016 | 99.77 | 0.17 |
| *F.×ananassa* (36×) | Hifiasm (purge) | 0.829 | 17.6 | 17.6 | 0.458 | 98.45 | 93.43 |
| | HiCanu | 1.044 | 8.4 | 9.8 | 0.295 | 98.08 | 92.94 |
| | HiCanu (purge) | 0.411 | 10.5 | 0.0 | 0.928 | 96.78 | 55.08 |
| | Peregrine | 0.930 | 5.5 | 6.7 | 0.260 | 98.33 | 91.70 |
| | Falcon | 0.971 | 5.4 | 7.3 | 0.213 | 98.27 | 92.81 |
| *R. muscosa* (~29×) | Hifiasm (purge) | 9.664 | 9.1 | | 7.208 | 66.61 | 1.70 |
| | HiCanu | 9.645 | 5.2 | | 6.361 | 65.54 | 3.92 |
| | Peregrine | 9.415 | 0.9 | | 2.936 | 66.84 | 1.72 |
| *S. sempervirens* (~33×) | Hifiasm (purge) | 35.310 | 5.5 | | 15.757 | 61.31 | 39.42 |
| | Peregrine | 35.662 | 0.8 | | | 63.20 | 35.93 |

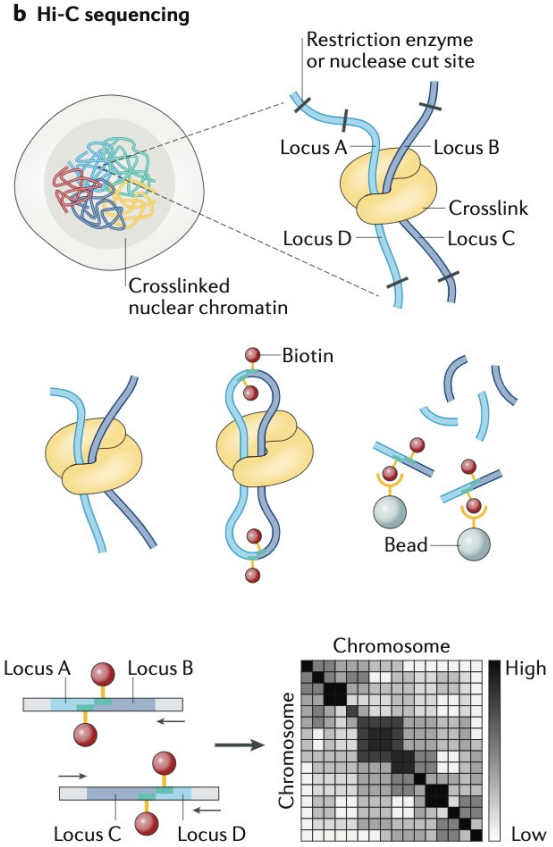
**Table 2 | Statistics of human primary assemblies**

| Dataset | Assembly | Size (Gb) | NG50 (Mb) | NGA50 (Mb) | QV | Multicopy genes retained (%) | Resolved BACs (%) | Gene completeness (asmgene) | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Complete (%) | Duplicated (%) |
| CHM13 (HiFi 32×) | Hifiasm | 3.052 | 88.9 | 86.7 | 54.2 | 99.7 | 98.8 | 99.97 | 0.05 |
| | HiCanu | 3.037 | 69.7 | 67.9 | 54.1 | 98.9 | 97.6 | 99.97 | 0.04 |
| | Peregrine | 2.990 | 37.8 | 33.4 | 43.8 | 51.1 | 39.7 | 99.64 | 0.16 |
| | Falcon | 2.862 | 27.1 | 21.8 | 50.1 | 30.2 | 34.2 | 99.47 | 0.03 |
| CHM13 (ONT 120×) | Canu | 2.936 | 80.0 | 47.3 | 32.7 | 76.9 | 86.7 | 99.30 | 0.10 |
| | Flye | 2.900 | 37.5 | 34.0 | 33.5 | 54.7 | 60.6 | 99.22 | 0.11 |
| | Shasta | 2.820 | 41.3 | 33.4 | 30.4 | 26.7 | 27.9 | 98.05 | 0.01 |
| HG00733 (HiFi 33×) | Hifiasm (purge) | 3.043 | 68.3 | 55.3 | 49.9 | 74.6 | 80.4 | 99.07 | 0.39 |
| | HiCanu (purge) | 2.921 | 40.5 | 34.2 | 50.5 | 55.2 | 65.9 | 98.47 | 0.32 |
| | Peregrine | 3.035 | 30.1 | 30.1 | 40.5 | 37.2 | 38.5 | 98.70 | 0.31 |
| | Falcon | 2.861 | 24.4 | 23.2 | 46.3 | 33.6 | 38.0 | 96.51 | 0.15 |
| HG00733 (ONT 50×) | Canu | 2.923 | 41.1 | 36.6 | 29.5 | 54.6 | 69.3 | 98.32 | 0.66 |
| | Flye | 2.890 | 26.7 | 25.4 | 29.9 | 34.2 | 44.7 | 97.88 | 0.20 |
| | Shasta | 2.805 | 21.2 | 20.8 | 30.0 | 17.0 | 22.9 | 97.19 | 0.05 |
| HG002 (HiFi 36×) | Hifiasm (purge) | 3.067 | 98.2 | 64.1 | 51.5 | 75.8 | | 99.26 | 0.32 |
| | HiCanu (purge) | 2.953 | 48.3 | 39.4 | 52.1 | 59.7 | | 98.71 | 0.18 |
| | Peregrine | 3.081 | 33.4 | 32.5 | 41.3 | 42.5 | | 99.14 | 0.36 |
| | Falcon | 2.955 | 30.4 | 29.0 | 46.7 | 36.6 | | 99.00 | 0.20 |

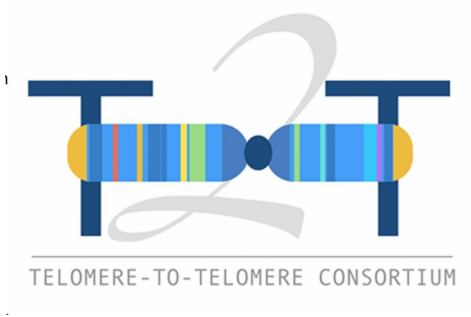
**Table 3 | Statistics of haplotype-resolved human assemblies**

| Dataset | Assembly | Size (Gb) | QV | NG50 (Mb) | Multicopy genes retained (%) | Resolved BACs (%) | Switch error (%) | Hamming error (%) | FNR (%) | FDR (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| HG00733 | Hifiasm (trio) | 6.071 | 49.9 | 34.9 | 84.0 | 95.5 | 0.08 | 0.22 | 2.43 | |
| | HiCanu (trio) | 6.079 | 49.2 | 10.6 | 84.3 | 90.5 | 0.04 | 0.04 | 4.78 | |
| | Peregrine (trio) | 5.938 | 42.2 | 19.1 | 37.6 | 39.7 | 0.10 | 0.23 | 12.34 | |
| | Peregrine (Hi-C) | 5.867 | 41.6 | 26.1 | 33.2 | 35.2 | 0.12 | 0.67 | 3.31 | |
| | Peregrine (Strand-seq) | 5.805 | 45.8 | 26.6 | 33.0 | 46.9 | 0.18 | 0.72 | 3.99 | |
| HG002 | Hifiasm (trio) | 5.967 | 51.6 | 43.0 | 80.6 | | 0.79 | 0.34 | 0.88 | 0.26 |
| | HiCanu (trio) | 6.003 | 50.4 | 12.1 | 80.4 | | 0.75 | 0.19 | 1.57 | 0.32 |
| | Peregrine (trio) | 5.888 | 42.7 | 25.8 | 38.7 | | 0.70 | 0.18 | 4.42 | 4.18 |

# Hi-C can string large contigs to whole chromosomes



# A complete human genome

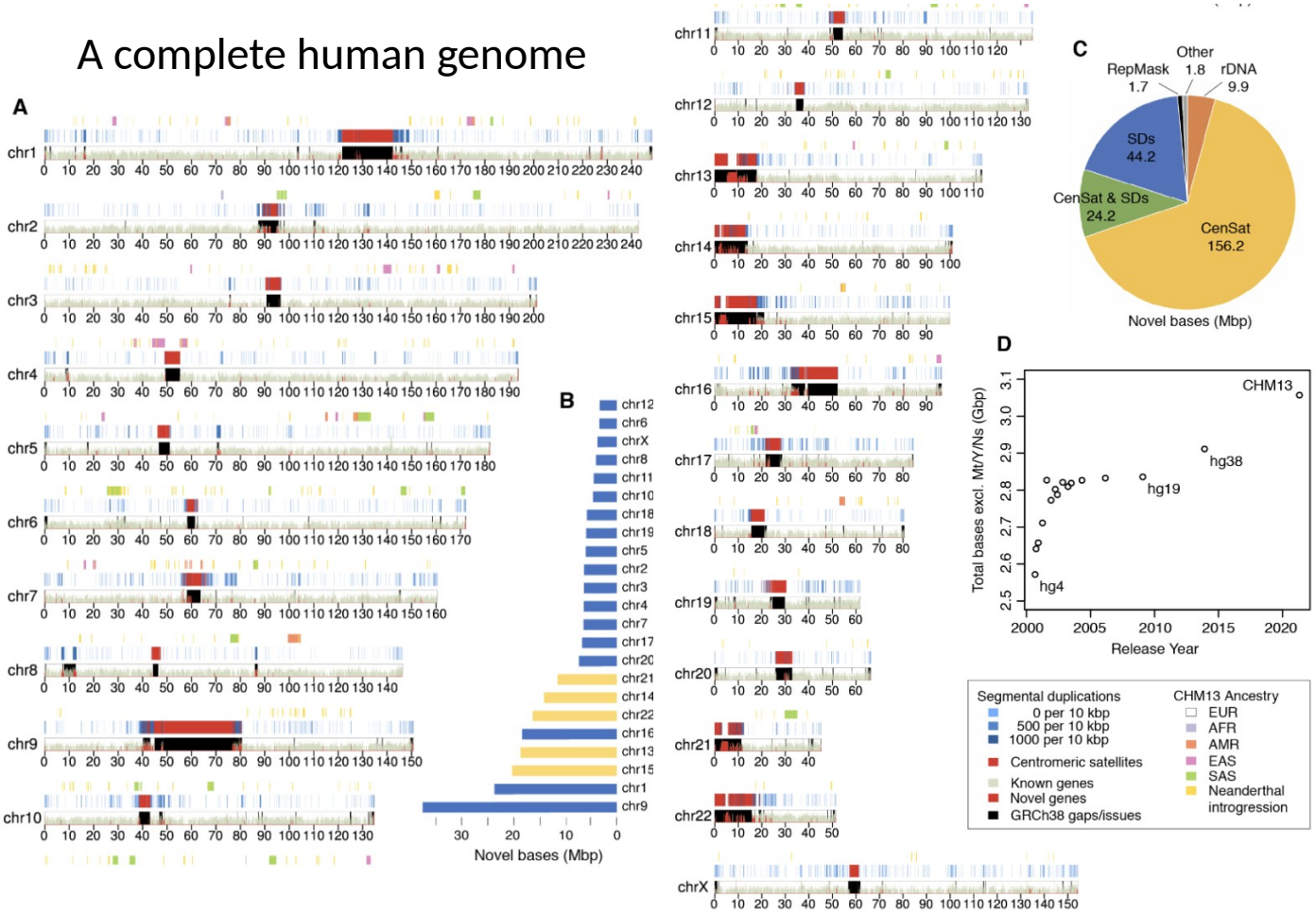



Sequencing a hydatidiform mole
ie a haploid genome

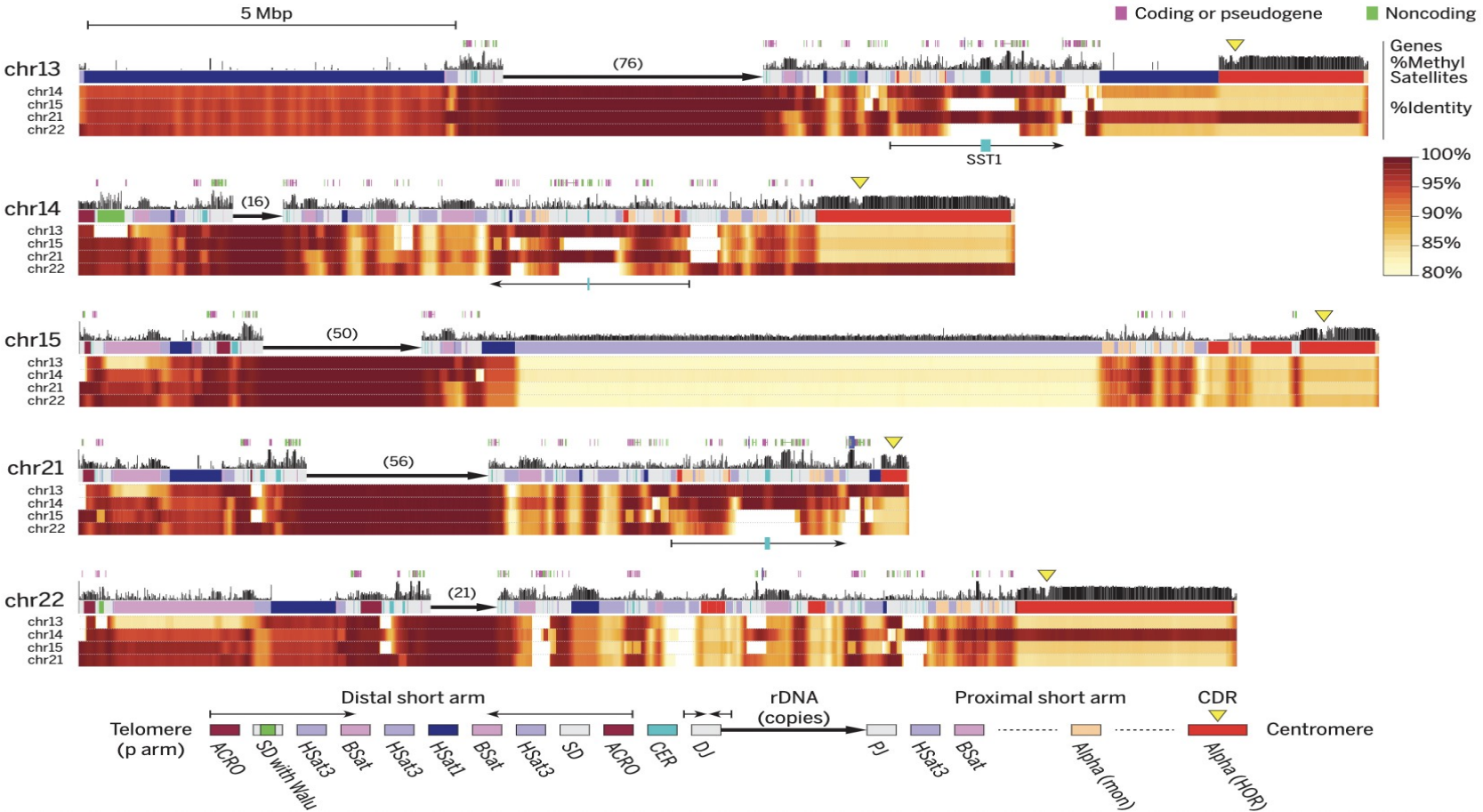
Using Nanopore and Pacbio HIFI

| STATISTICS | GRCH38 | T2T-CHM13 | DIFFERENCE (±%) |
|---|---|---|---|
| **Summary** | | | |
| Assembled bases (Gbp) | 2.92 | 3.05 | +4.5 |
| Unplaced bases (Mbp) | 11.42 | 0 | −100.0 |
| Gap bases (Mbp) | 120.31 | 0 | −100.0 |
| Number of contigs | 949 | 24 | −97.5 |
| Contig NG50 (Mbp) | 56.41 | 154.26 | +173.5 |
| Number of issues | 230 | 46 | −80.0 |
| Issues (Mbp) | 230.43 | 8.18 | −96.5 |
| **Gene annotation** | | | |
| Number of genes | 60,090 | 63,494 | +5.7 |
| Protein coding | 19,890 | 19,969 | +0.4 |
| Number of exclusive genes | 263 | 3,604 | |
| Protein coding | 63 | 140 | |
| Number of transcripts | 228,597 | 233,615 | +2.2 |
| Protein coding | 84,277 | 86,245 | +2.3 |
| Number of exclusive transcripts | 1,708 | 6,693 | |
| Protein coding | 829 | 2,780 | |

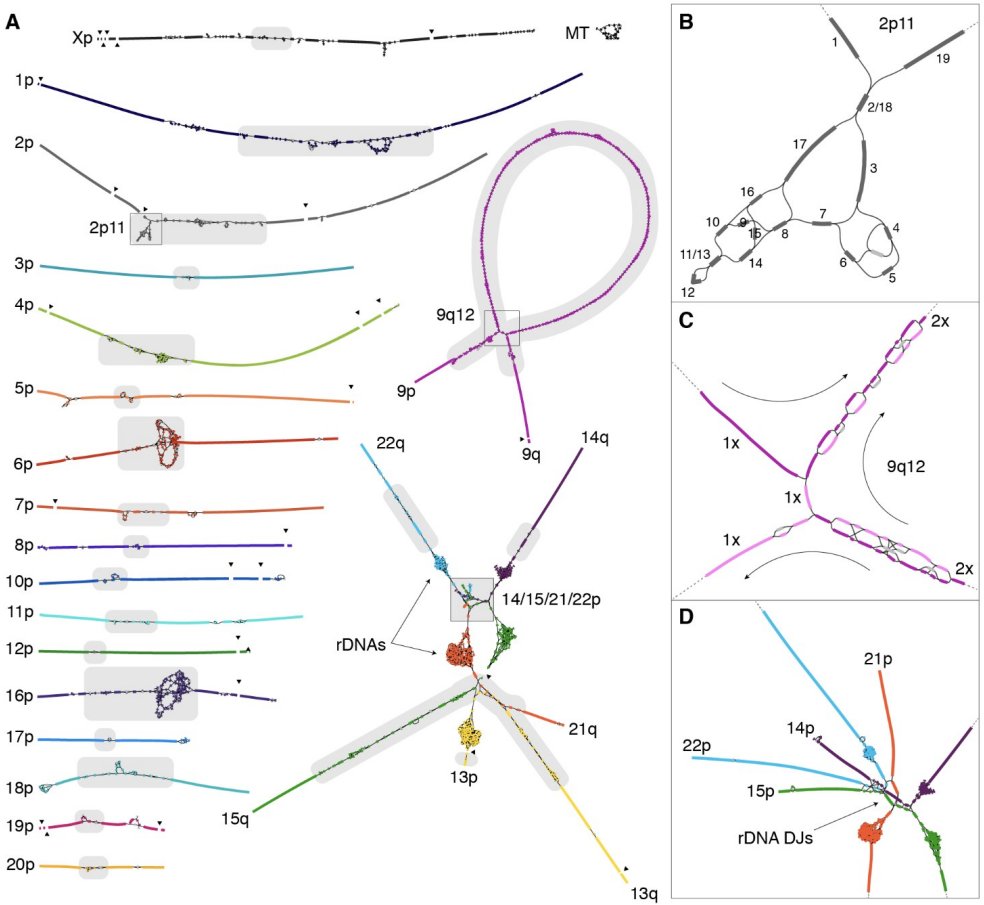
# A complete human genome



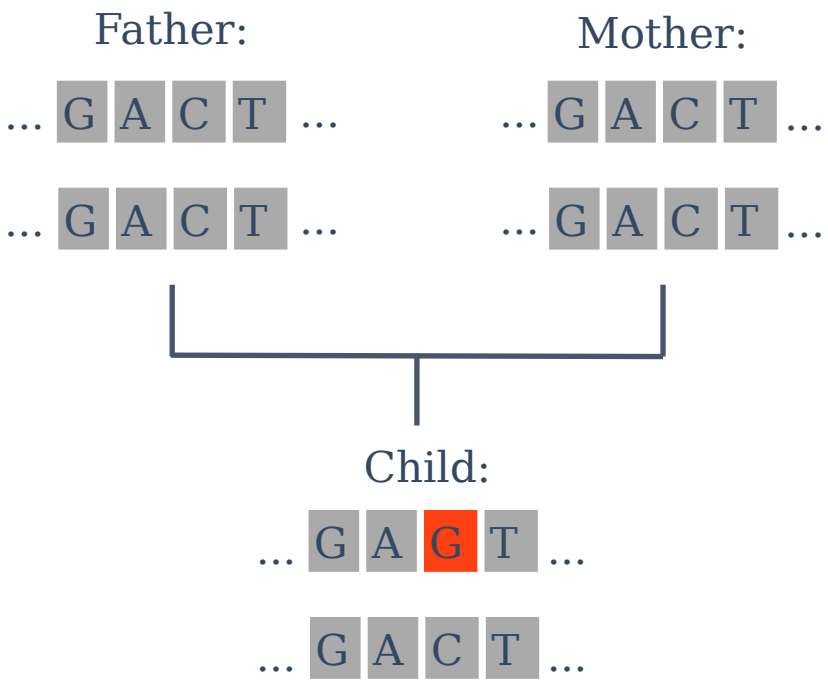
# Even telomeres and centromeres resolved



# Complete genome graph



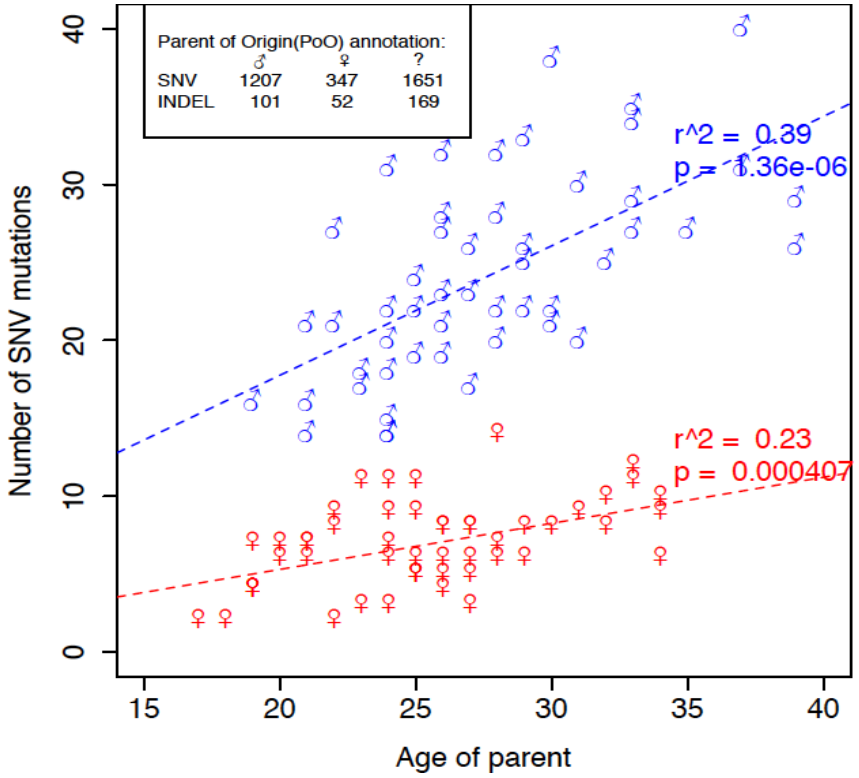
# Direct estimation of mutations in trios



# Mutations as a function of parental age at conception in Danish pangenome data set

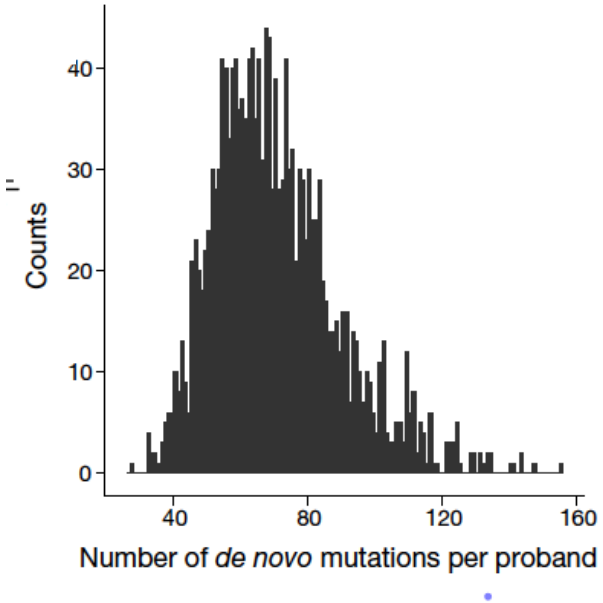
Use read backed phasing to find out which parent mutation came from

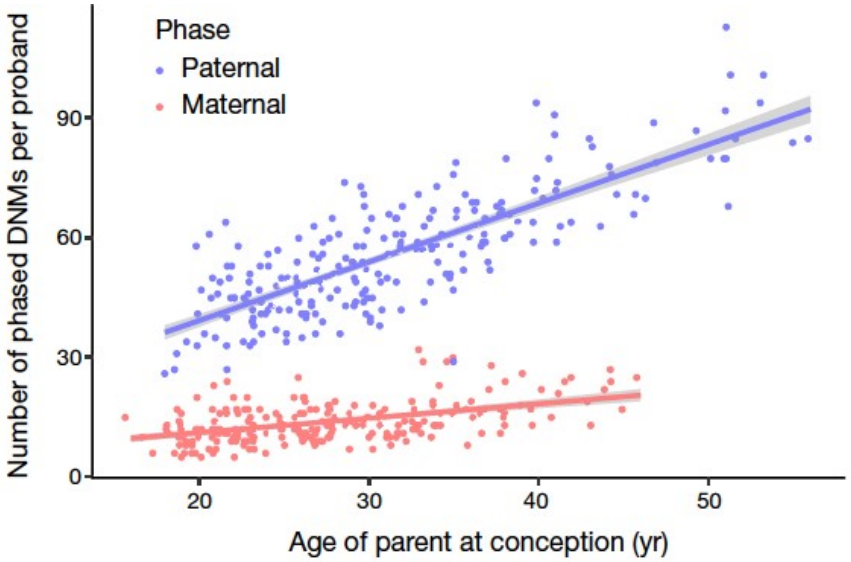
Number of new maternal mutations from also increase with maternal age




Maretty et al, Nature 2017

# Results from 1500 trios

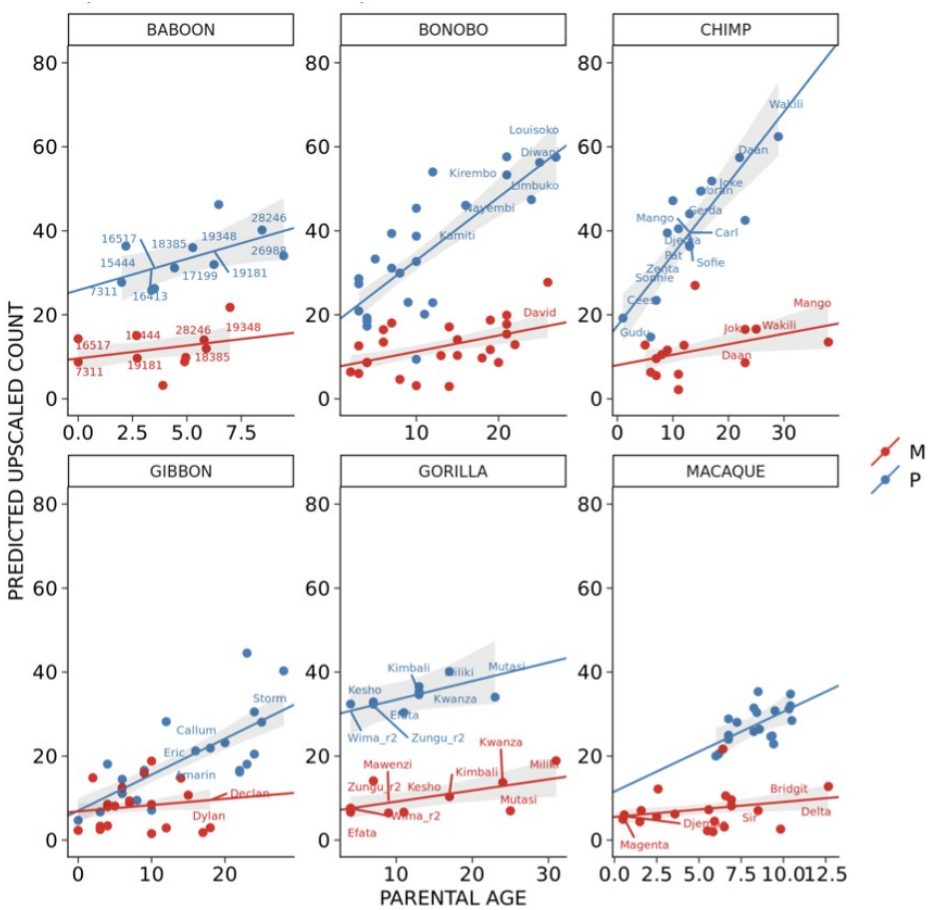



Expected rate = 1.77e-09 + 7.26e-11 * maternal age + 2.87e-10 * paternal age.

Jonsson et al. 2017, Nature

# Different effect of paternal age in primates

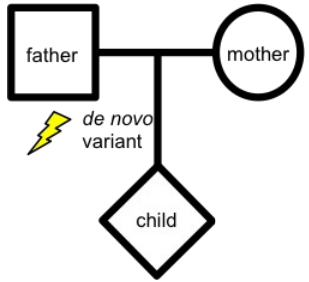



Baboons and Gorillas have less sperm competition

Riera, …., Schierup, Stefansson, unpubl

# Next step: Understanding the mutational process in testis

”the likely largest monogenic burden on pediatric health … remains due to de novo mutations”
Breuss, Yang and Gleeson. *Trends in Genetics* 2021 Oct;37(10):890-902.




*De novo* mutations significant cause of

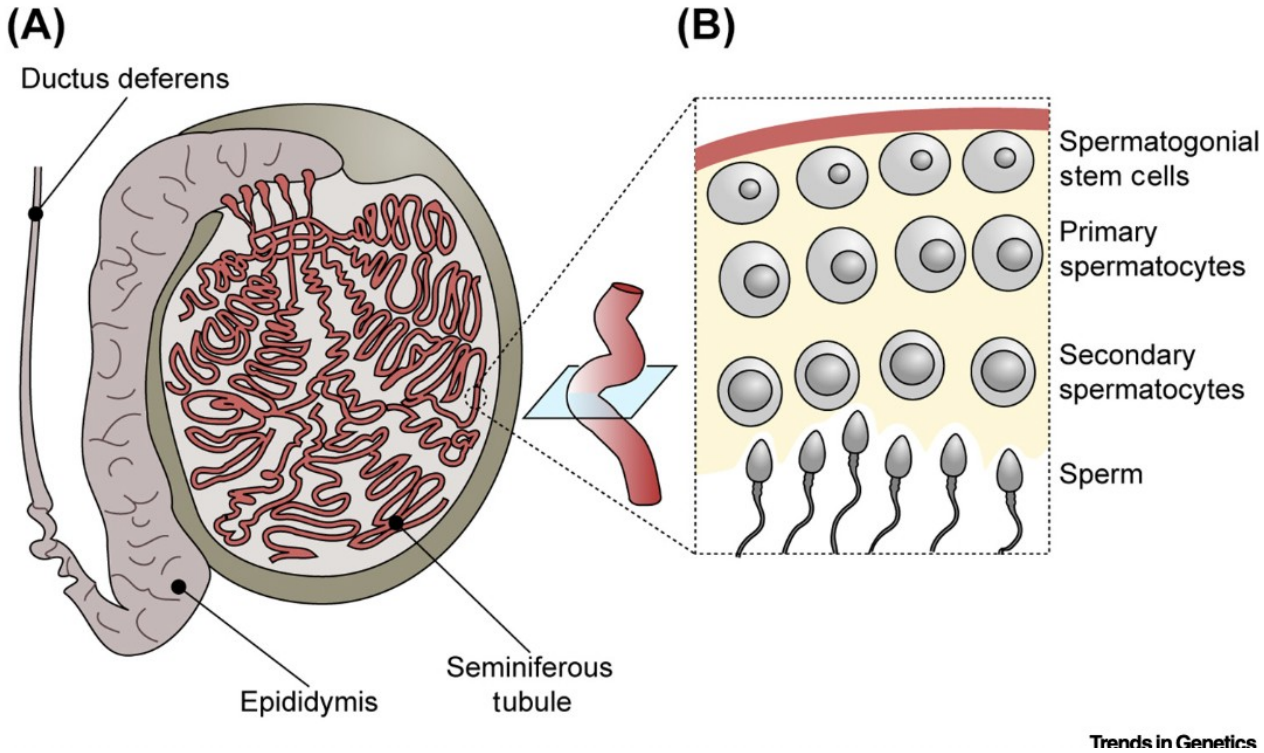
- severe intellectual disability (IQ < 50)
- severe autism
- male infertility

Most *de novo* mutations arise in the testis of old fathers

An unexplored opportunity to prevent serious disease

Sequencing is now accurate and cheap

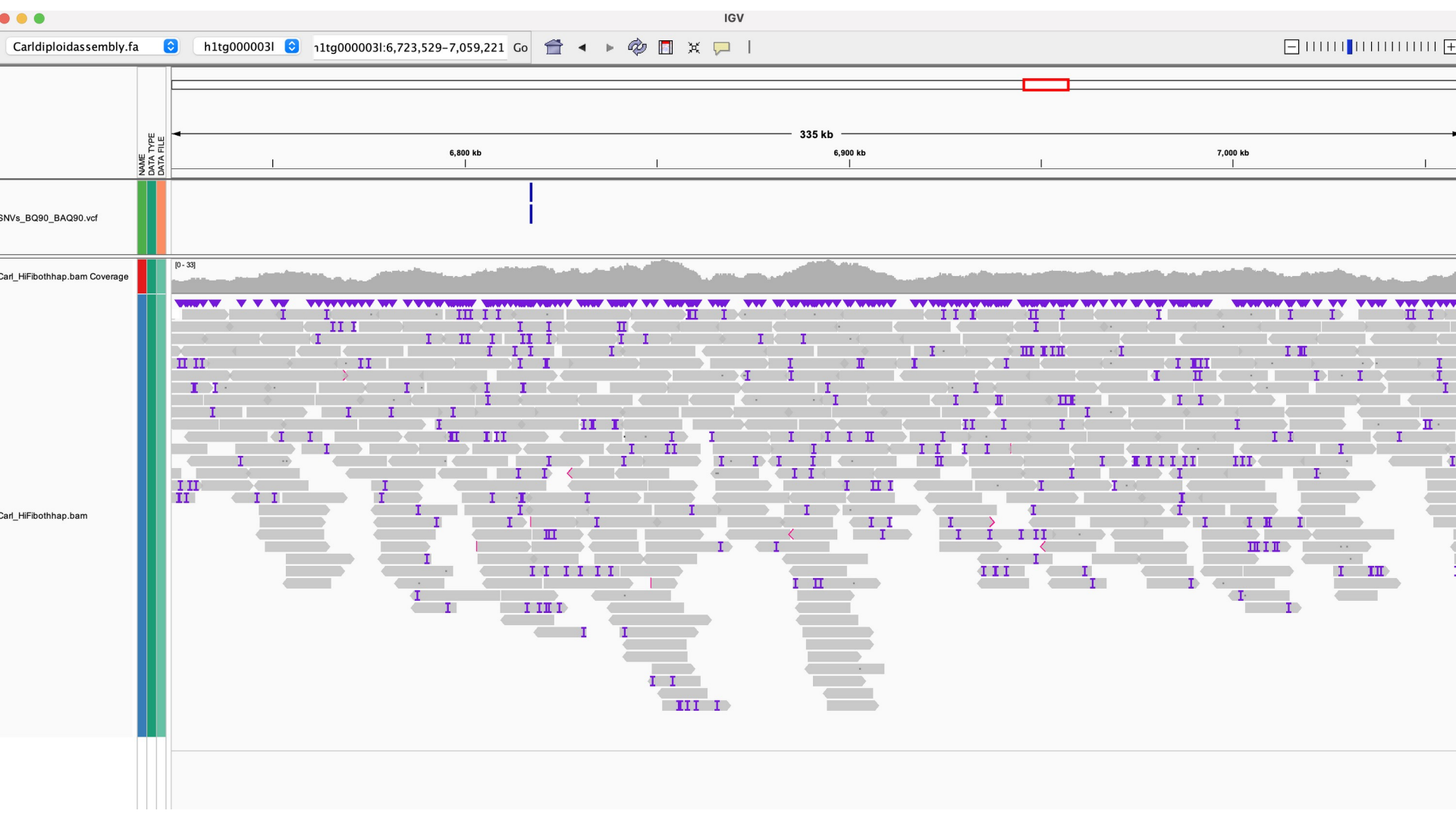
# Understanding the male mutational process




Breuss et al. 2021

# Diploid assembly from 27X Hifi Pacbio of the chimp Carl

| file | num_seqs | sum_len | min_len | avg_len | max_len | N50 |
|---|---|---|---|---|---|---|
| Carldiploid | 1,532 | 5,672,691,091 | 499,180.8 | 65,919,327 | 11,364 | 5,981,461 |


IGV
Carldiploidassembly.fa
h1tg000003l
h1tg000003l:6,723,529-7,059,221
Go
335 kb
6,800 kb
6,900 kb
7,000 kb
NAME
DATA TYPE
DATA FILE
SNVs_BQ90_BAQ90.vcf
Carl_HiFibothhap.bam Coverage
[0 - 33]
Carl_HiFibothhap.bam

# Even the Y chromosome can now be accessed