09:00-12:00 Assembly

- 09:00-09:30 Introduction to whole-genome assembly and introduction to the study system and infrastructure
- 09:30-11:55 Assembling the yeast genome
    - GenomeScope2
    - Smudgeplot
    - HiFiAdapterFilt
    - hifiasm
    - YaHS
- 11:55-12:00 Summary

12:00-13:00 Lunch

13:00-14:00 Validation

- 13:00-13:15 Introduction to assembly validation
- 13:15-13:55 Interpreting assembly validation results
    - gfastats
    - BUSCO
    - Merqury
- 13:55-14:00 Summary

14:00-14:15 Break

14:15-16:00 Decontamination and manual curation

- 14:15-14:30 Introduction to decontamination and manual curation
- 14:30-15:50 Decontaminating and curating the yeast genome
    - FCS-GX
    - The GRIT Rapid Curation suite
    - Working in PretextView
- 15:50-16:00 Summary

Genome assembly, curation and validation

Lecturers: Benedicte Garmann-Johnsen & Ole Kristian Tørresen

Oslo Bioinformatics Workshop Week 2022

13th December

Learning outcomes

After attending the workshop learners should:

1. Know about most-used approaches for genome assembly
2. Assess information inherit in sequencing reads
3. Be able to validate genome assemblies
4. Know about manual curation of assemblies

09:00-12:00 Assembly

- 09:00-09:30 Introduction to whole-genome assembly and introduction to the study system and infrastructure
- 09:30-11:55 Assembling the yeast genome
    - GenomeScope2
    - Smudgeplot
    - HiFiAdapterFilt
    - hifiasm
    - YaHS
- 11:55-12:00 Summary

12:00-13:00 Lunch

13:00-14:00 Validation

- 13:00-13:15 Introduction to assembly validation
- 13:15-13:55 Interpreting assembly validation results
    - gfastats
    - BUSCO
    - Merqury
- 13:55-14:00 Summary

14:00-14:15 Break

14:15-16:00 Decontamination and manual curation

- 14:15-14:30 Introduction to decontamination and manual curation
- 14:30-15:50 Decontaminating and curating the yeast genome
    - FCS-GX
    - The GRIT Rapid Curation suite
    - Working in PretextView
- 15:50-16:00 Summary

Introduction - Assembly

- EBP and EBP-Nor
- Why do assemblies?
- What do we want from genome assemblies?
- How do we generate genome assemblies?

What is a biodiversity genomics project?

- Do all kinds of different species (DToL, ERGA, EBP)
- Vertebrate Genomes Project is targeted (all vertebrates), maybe not a biodiversity genomics project

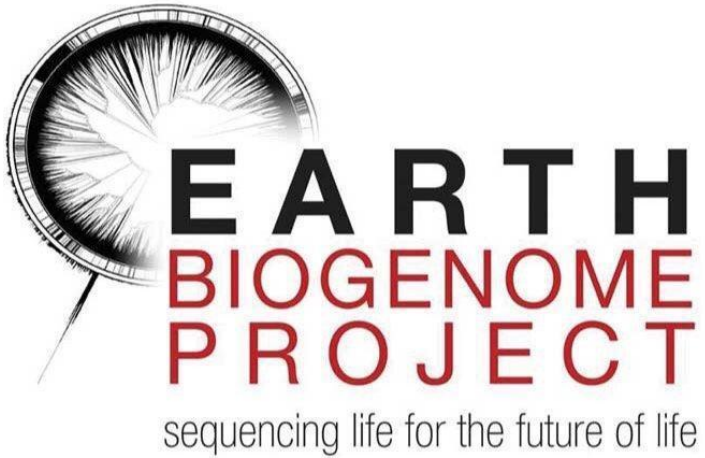
EARTH BIOGENOME PROJECT
sequencing life for the future of life

Darwin TREE of LIFE

What is the Earth Biogenome Project?

- Better understanding of biology and evolution
- Conserve, protect and restore biodiversity
- Create new benefits for society and human welfare

-> sequence all eukaryotes



EBP-Nor

Funded by the Research Council of Norway

- Phase 1 2021-2024 (30 MNOK)
  - Do 100-150 species
  - Norwegian, marine and arctic species
  - Coordination with ERGA, DToL, VGP, EBP and other projects (e.g. https://goat.genomehubs.org/)

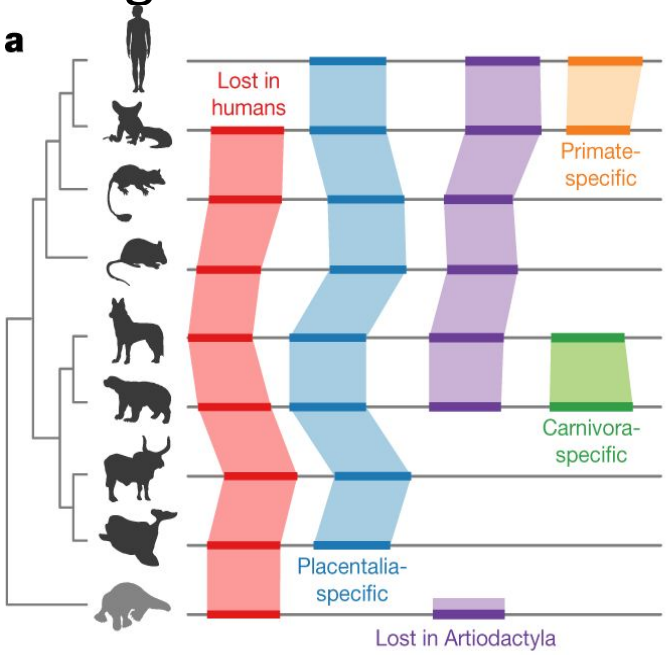
- Preparation for 2 phase has begun

SARS-CoV-2 and host range

- SARS-CoV-2 binds to ACE2
- Primates have very high probability
- Cervid deer and cetacean high probability

Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates

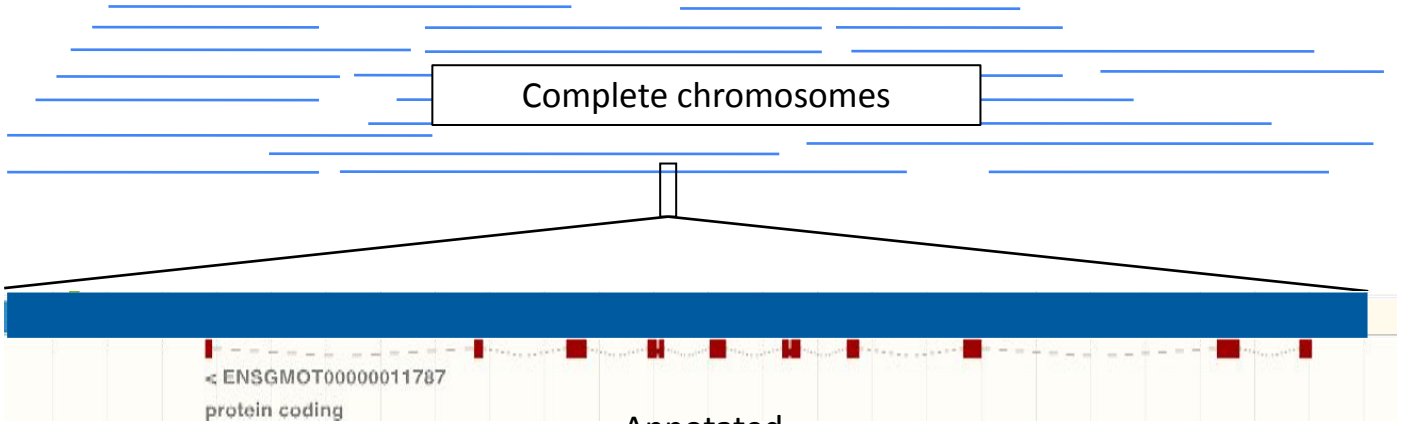
Joana Damas[a,1], Graham M. Hughes[b,1], Kathleen C. Keough[c,d,1], Corrie A. Painter[e,1], Nicole S. Persky[f,1], Marco Corbo[a], Michael Hiller[g,h,i], Klaus-Peter Koepfli[j], Andreas R. Pfenning[k], Huabin Zhao[l,m], Diane P. Genereux[n], Ross Swofford[n], Katherine S. Pollard[d,o,p], Oliver A. Ryder[q,r], Martin T. Nweeia[s,t,u], Kerstin Lindblad-Toh[n,v], Emma C. Teeling[b], Elinor K. Karlsson[n,w,x], and Harris A. Lewin[a,y,z,2]



Damas et al. *PNAS* 2020

Zoonomia Consortium *Nature* 2020

An ideal genome assembly

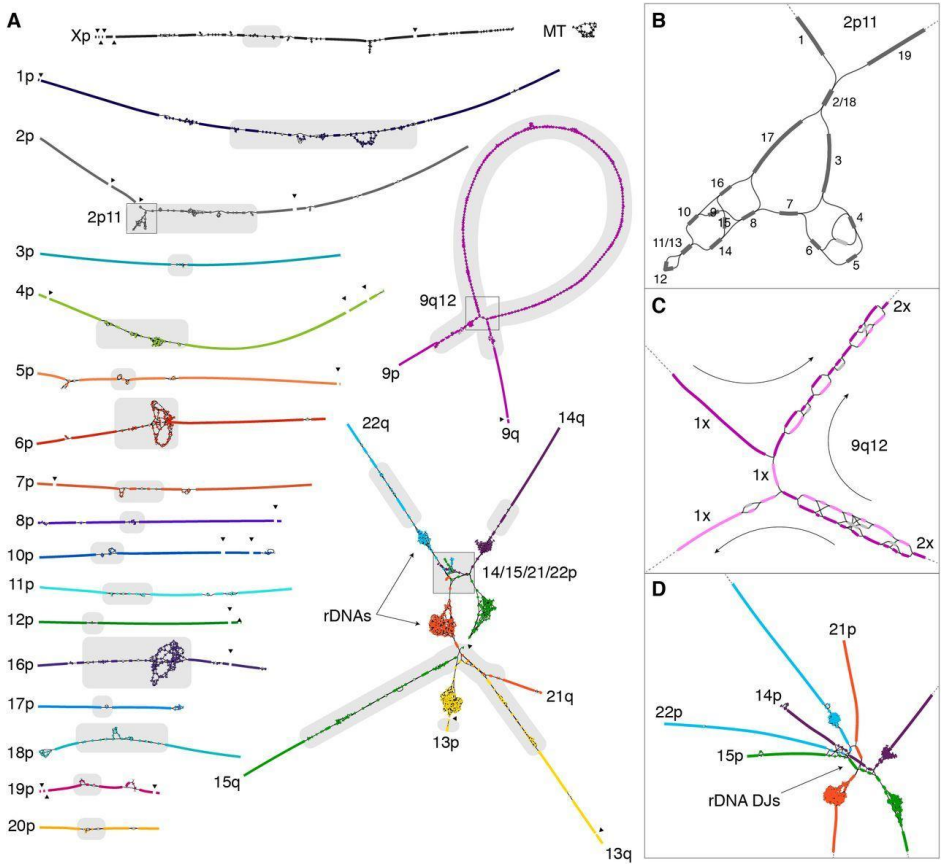


Annotated gene

Slide modified after Lex Nederbragt

Telomer-to-telomer human genome

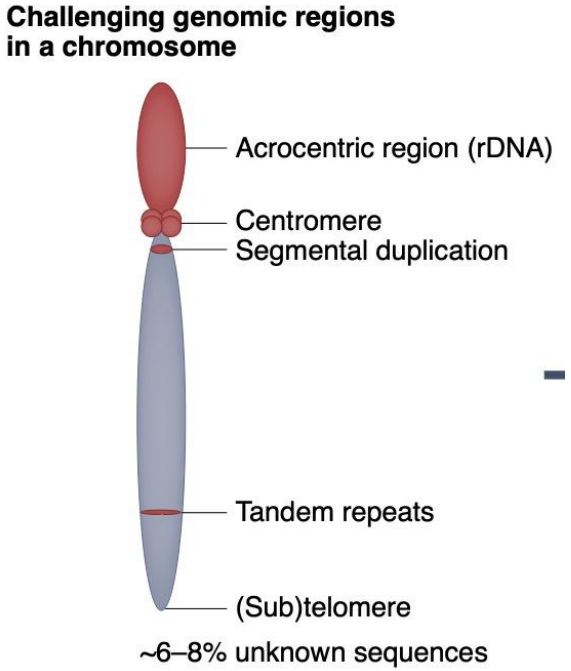
- Uses HiFi reads to create an assembly string graph
- Uses ONT reads to resolve tangles and to close gaps
- Based on a haploid genome
- Called T2T-CHM13



Nurk et al. *Science* 2022

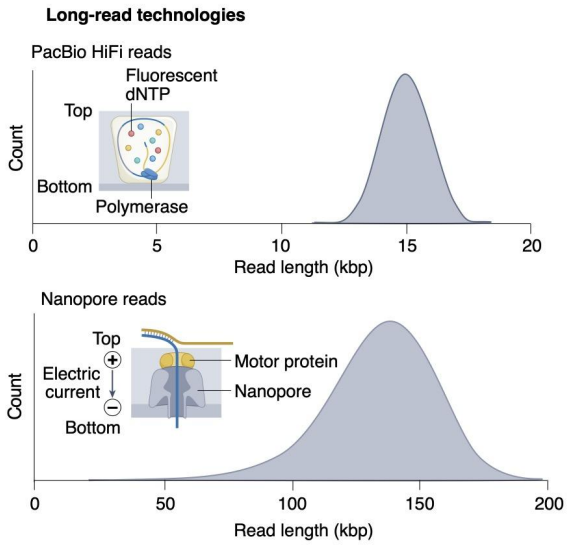
Challenging regions in a genome

- Regions with repetitive sequences are still difficult to sequence/assemble
  - rDNA
  - Centromere
  - Tandem repeats
- No single sequencing technology can handle these easily

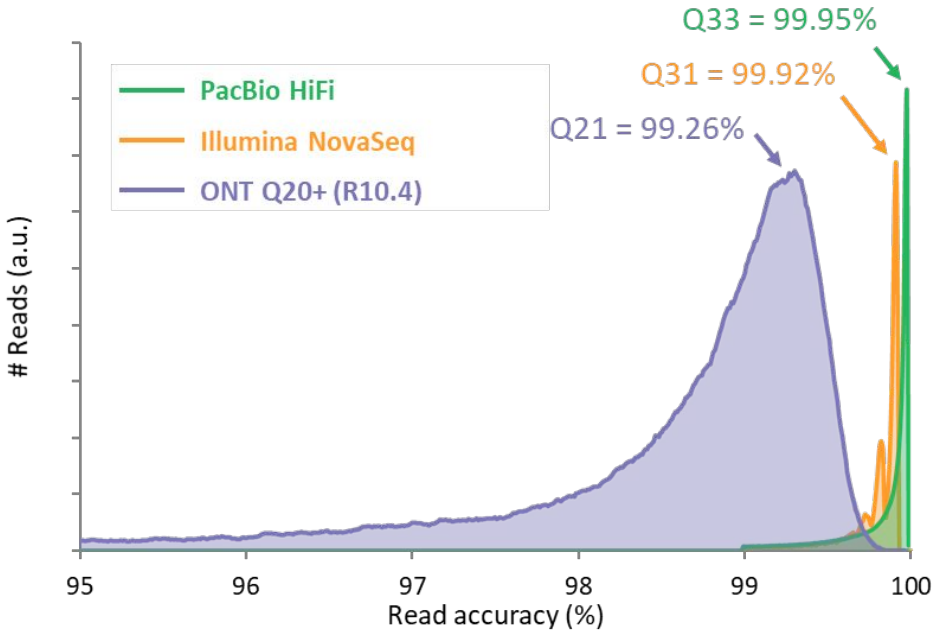


Mao and Zhang *Nature Methods* 2022

Sequencing data

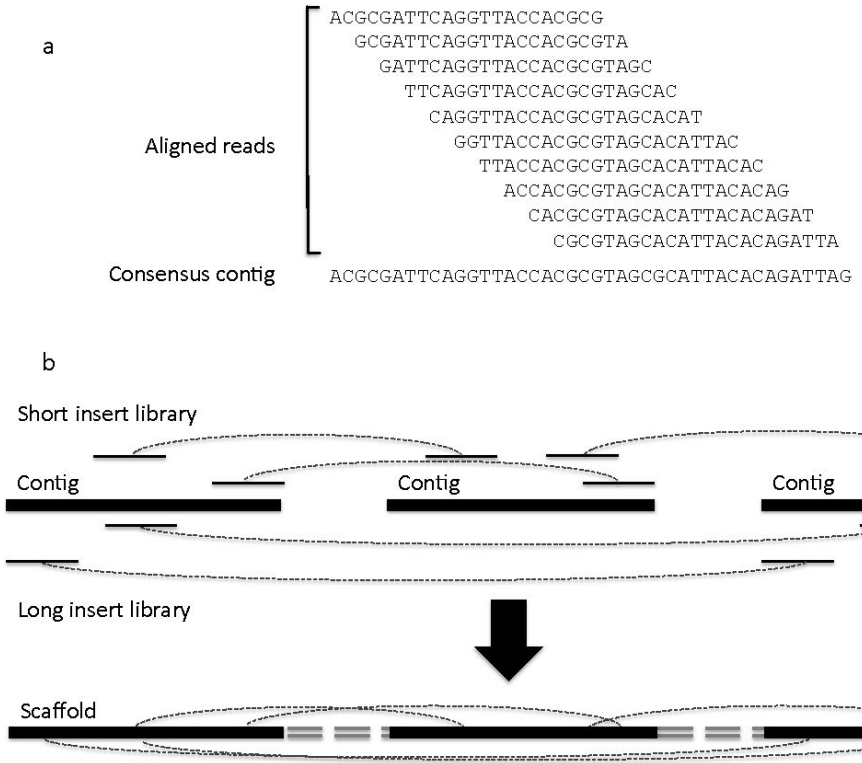


Mao and Zhang *Nature Methods* 2022



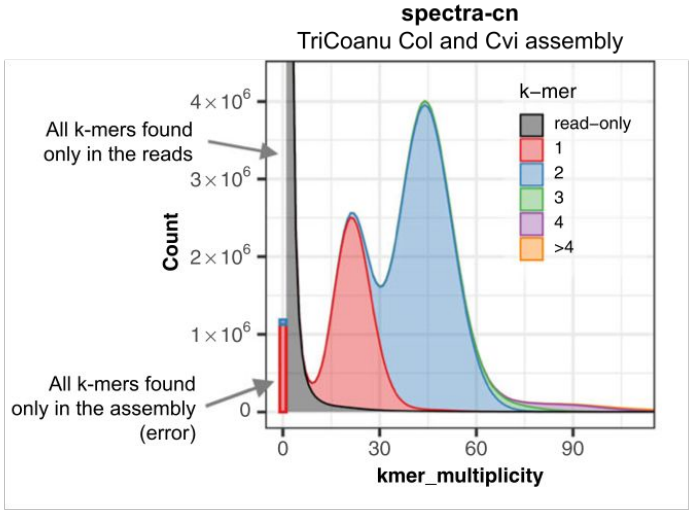
PacBio HiFi: HG003 18 kb library, Sequel II System Chemistry 2.0, precisionFDA Truth Challenge V2
Illumina: HG002 2×150 bp NovaSeq library, precisionFDA Truth Challenge V2
ONT: Q20+ chemistry (R10.4, Kit 12), Oct 2021 GM24385 Dataset Release

An assembly consists of contigs and scaffolds



EBP standards

- **6.C.Q40**
  - 10^6 bp N50 contig
  - Chromosome scale N50 scaffolding
  - Q40 error rate, fewer than 1 error per 10,000 bp
- < 5% false duplications
- > 90% kmer completeness
- > 90% sequence assigned to candidate chromosomal sequences
- > 90% single copy conserved genes (e.g. BUSCO) complete and single copy
- > 90% transcripts from the same organism mappable

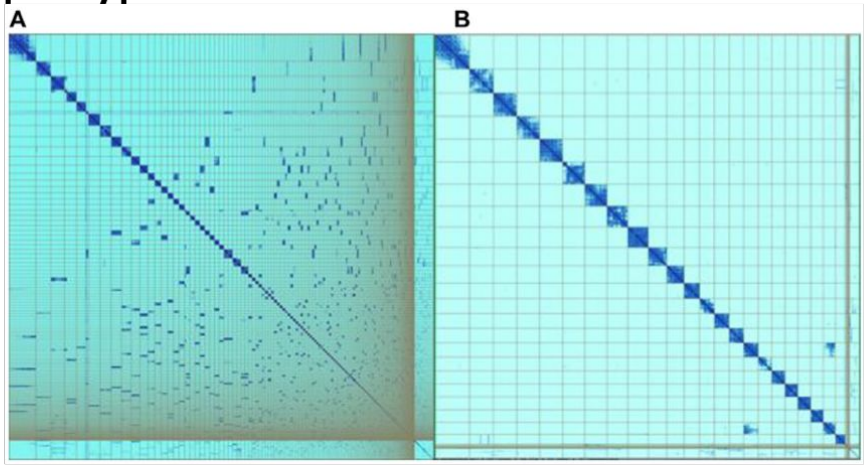


Merqury: Rhie et al. *Genome Biology* 2020

https://www.earthbiogenome.org/assembly-standards

EBP standards (cont'd)

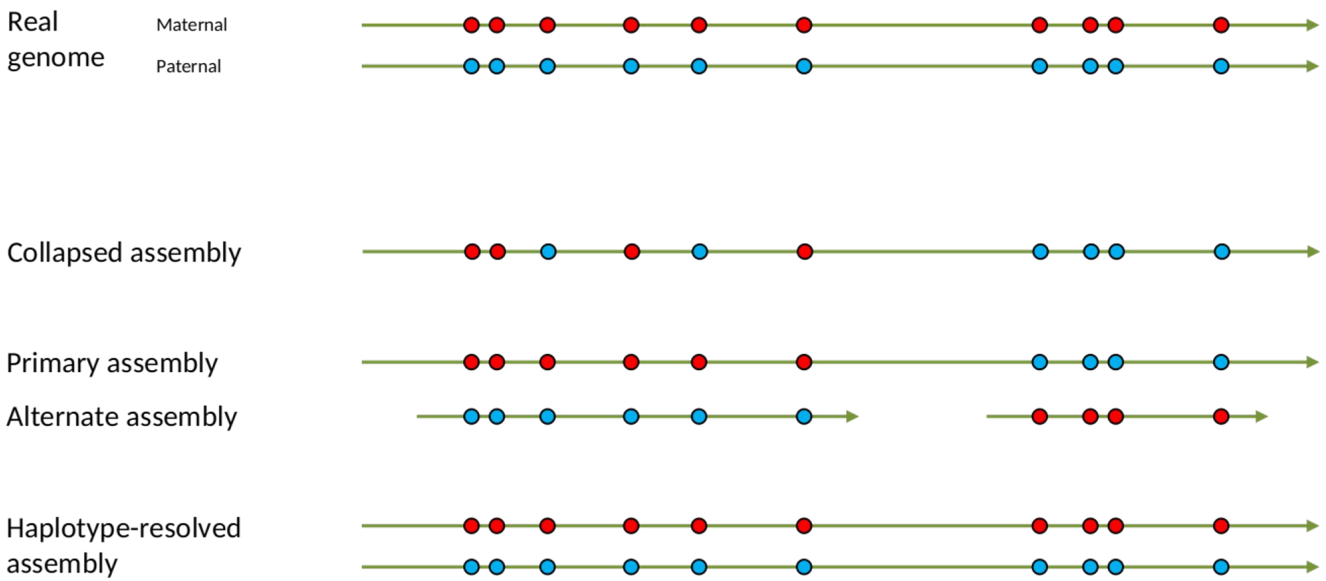
- Separation of target species vs contaminants/symbionts/cobionts
- For diploid species: Identification of primary (haploid) assembly, with secondary assembly with alternate haplotypes
- Organelle genomes
- Manual curation
- Reconciliation with known karyotype



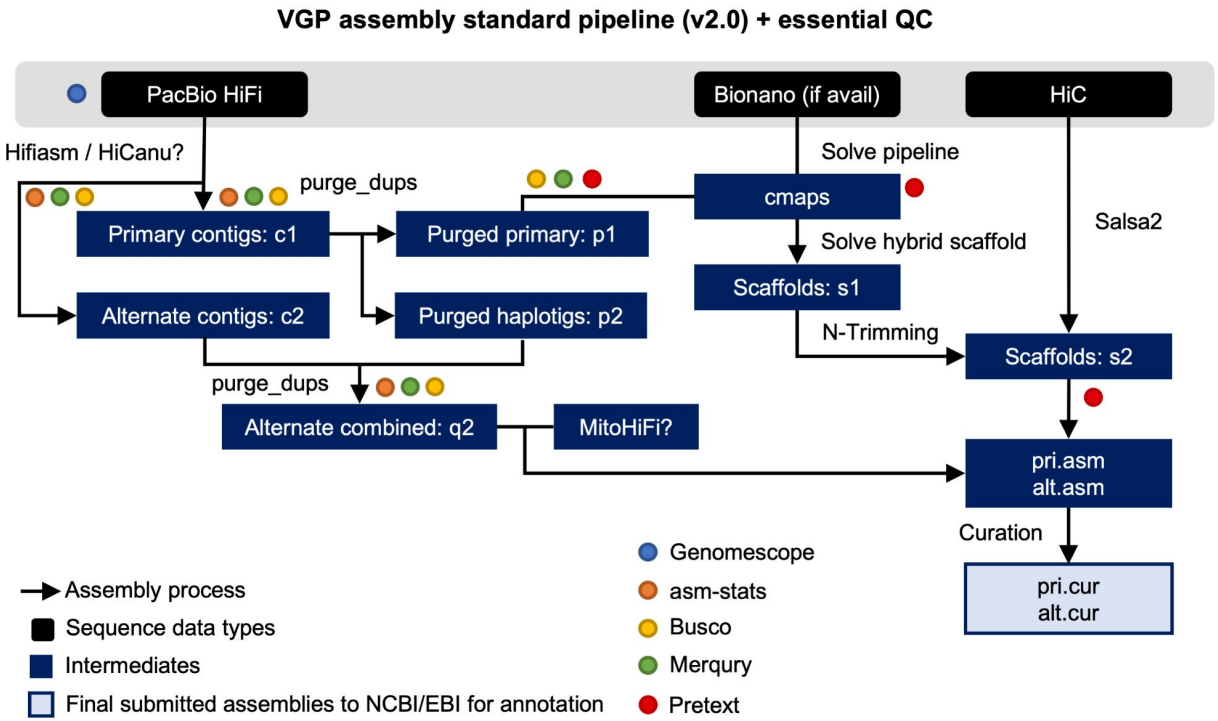
https://www.earthbiogenome.org/assembly-standards

Howe et al. *GigaScience* 2021

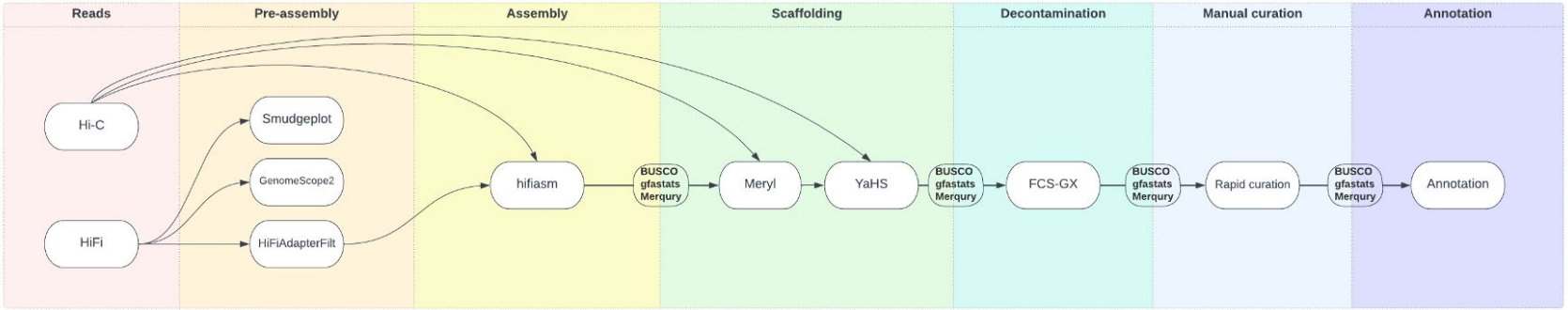
Haplotype-resolved assemblies



Vertebrate Genomes Project pipeline



EBP-Nor pipeline



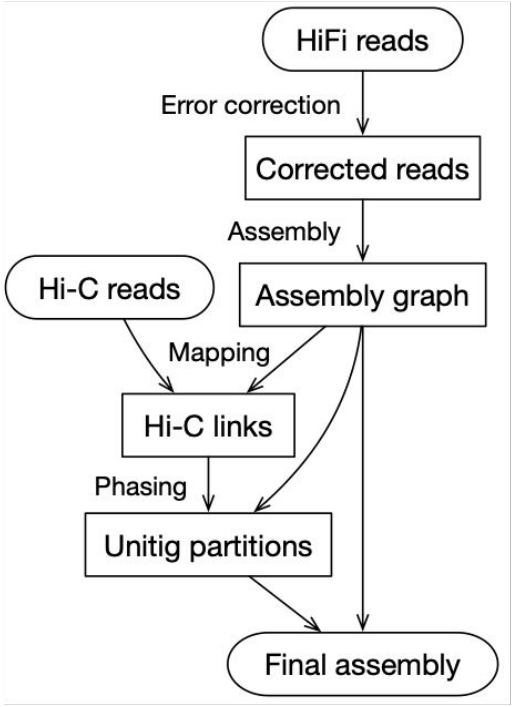
Metschnikowia zobellii: a marine fungi

- Parasite; infects copepods
- Sequenced to a high coverage by Darwin Tree of Life
- PacBio HiFi and Hi-C
- 13.6 Mbp genome and 5 chromosomes
- Subsampled to 30x HiFi and 60x Hi-C
    - Added a surprise to the HiFi reads

Combining Hi-C and HiFi in hifiasm

- Powerful combination
- Used by Human Pangenome Project
- Add scaffolding with Hi-C and manual curation -> easy workflow

- Few switching errors and N50 contig >20 Mbp for mammals (which are easy to assemble)



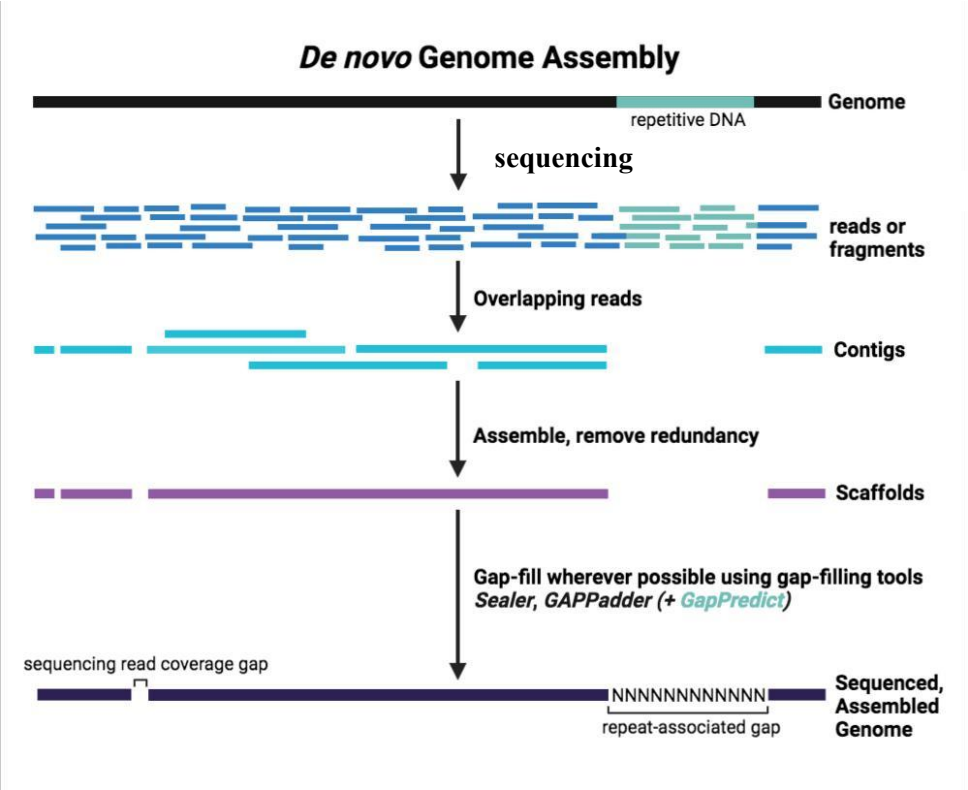
Cheng et al. *Nature Biotechnology* 2022

Fox cluster

- Go to: shorturl.at/ekEG1
- Go to introduction and infrastructure and do what it says there

- We’ll do this together

Genome assembly



Summary - Assembly

Break

See you back here at 13:00 :)

09:00-12:00 Assembly

- 09:00-09:30 Introduction to whole-genome assembly and introduction to the study system and infrastructure
- 09:30-11:55 Assembling the yeast genome
    - GenomeScope2
    - Smudgeplot
    - HiFiAdapterFilt
    - hifiasm
    - YaHS
- 11:55-12:00 Summary

12:00-13:00 Lunch

13:00-14:00 Validation

- 13:00-13:15 Introduction to assembly validation
- 13:15-13:55 Interpreting assembly validation results
    - gfastats
    - BUSCO
    - Merqury
- 13:55-14:00 Summary

14:00-14:15 Break

14:15-16:00 Decontamination and manual curation

- 14:15-14:30 Introduction to decontamination and manual curation
- 14:30-15:50 Decontaminating and curating the yeast genome
    - FCS-GX
    - The GRIT Rapid Curation suite
    - Working in PretextView
- 15:50-16:00 Summary

Introduction - Validation

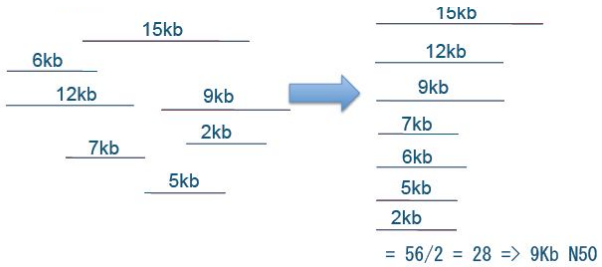
gfastats

BUSCO

Merqury

N50 and assembly statistics

- Length of contig such that 50 % of total bases are in contigs of this length or longer

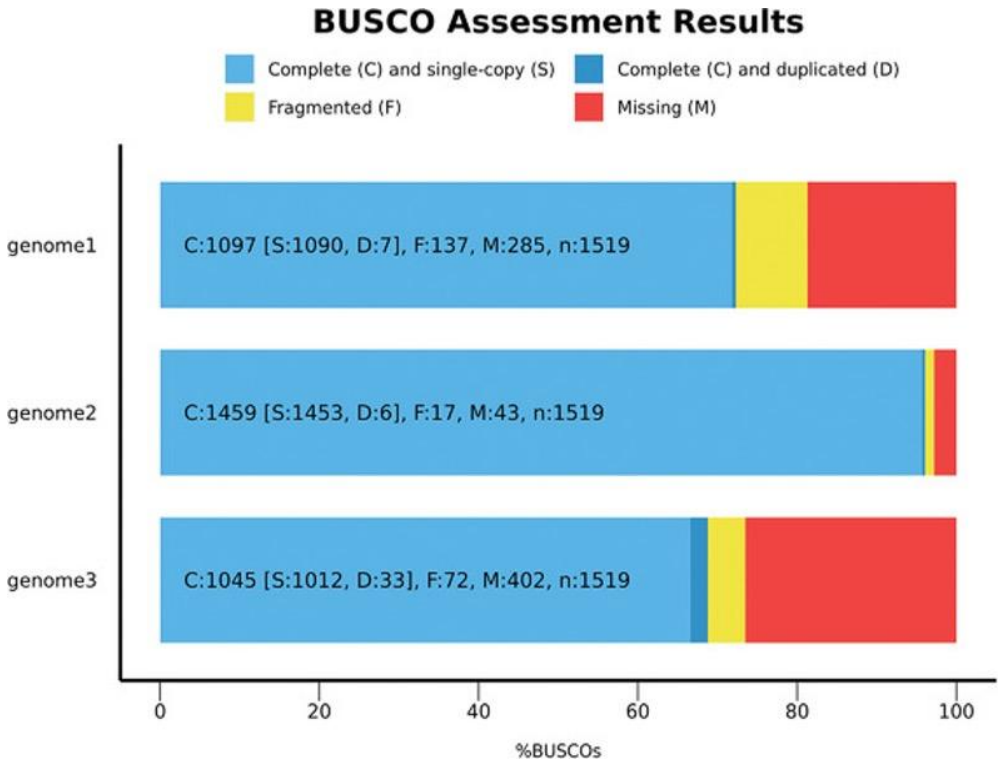


Sum of lengths: 56 kbp

- gfastats gives N50 and several other statistics such as average, longest, etc

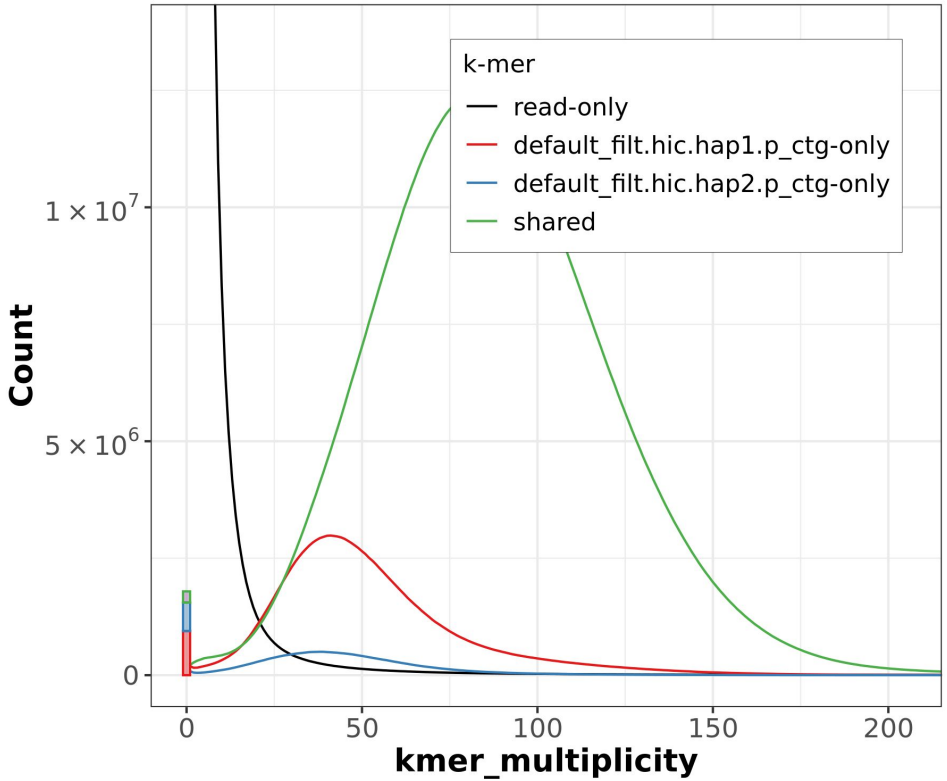
BUSCO

- Searches for conserved genes in genomes, transcriptomes and protein datasets
- Gives complete (single and duplicated), fragmented and missing status
- Which genome is best?

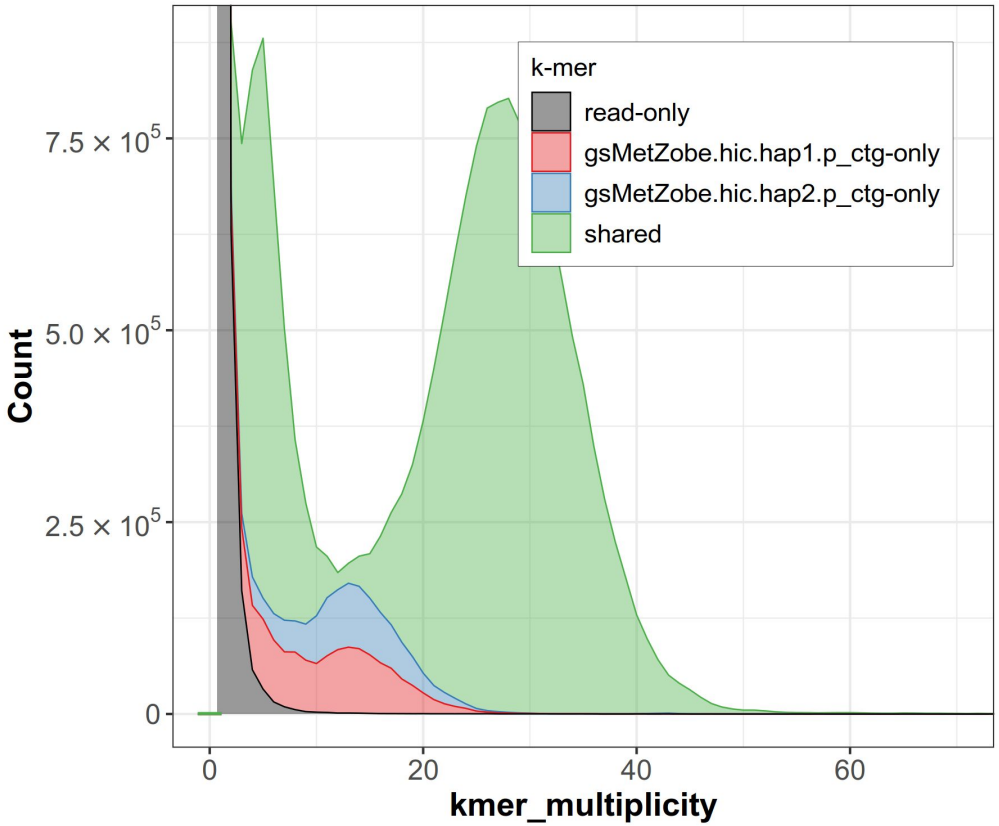


Merqury

- Compares k-mers in assemblies towards k-mers from reads
- Gives completeness and quality scores
- Plots



Summary - Validation



Break

See you back here at 14:15 :)

09:00-12:00 Assembly

- 09:00-09:30 Introduction to whole-genome assembly and introduction to the study system and infrastructure
- 09:30-11:55 Assembling the yeast genome
    - GenomeScope2
    - Smudgeplot
    - HiFiAdapterFilt
    - hifiasm
    - YaHS
- 11:55-12:00 Summary

12:00-13:00 Lunch

13:00-14:00 Validation

- 13:00-13:15 Introduction to assembly validation
- 13:15-13:55 Interpreting assembly validation results
    - gfastats
    - BUSCO
    - Merqury
- 13:55-14:00 Summary

14:00-14:15 Break

14:15-16:00 Decontamination and manual curation

- 14:15-14:30 Introduction to decontamination and manual curation
- 14:30-15:50 Decontaminating and curating the yeast genome
    - FCS-GX
    - The GRIT Rapid Curation suite
    - Working in PretextView
- 15:50-16:00 Summary

Introduction - Decontamination and manual curation

Why do we need to decontaminate our assemblies?

- Contamination during handling of samples
  - Human DNA
  - E-coli or other bacteria and microorganisms
- Contamination during sequencing
  - Other organisms
  - Adaptor sequences
- Contamination from the sample itself
  - Guts, gills, etc
  - Symbionts

fcs_gx_report.txt contamination summary:

TOTAL	7281	293427901
anml:rotifers	4280	162181695
prok:a-proteobacteria	1848	85469633
prok:bacteria	741	29848528
prok:g-proteobacteria	133	6650516
prok:b-proteobacteria	161	6137181
prok:CFB group bacteria	48	1890442
prok:high GC Gram+	9	298963
prok:firmicutes	7	293653
fung:ascomycetes	6	192444
prok:d-proteobacteria	3	132369
prst:algae	26	83253
fung:chytrids	2	78873
prok:proteobacteria	1	41214
arch:archaea	1	32871
prst:alveolates	1	31606
anml:insects	2	29647
anml:fishes	7	22622
fung:fungi	2	5177
fung:basidiomycetes	2	5009
anml:basal metazoans	1	2205

Many different tools available

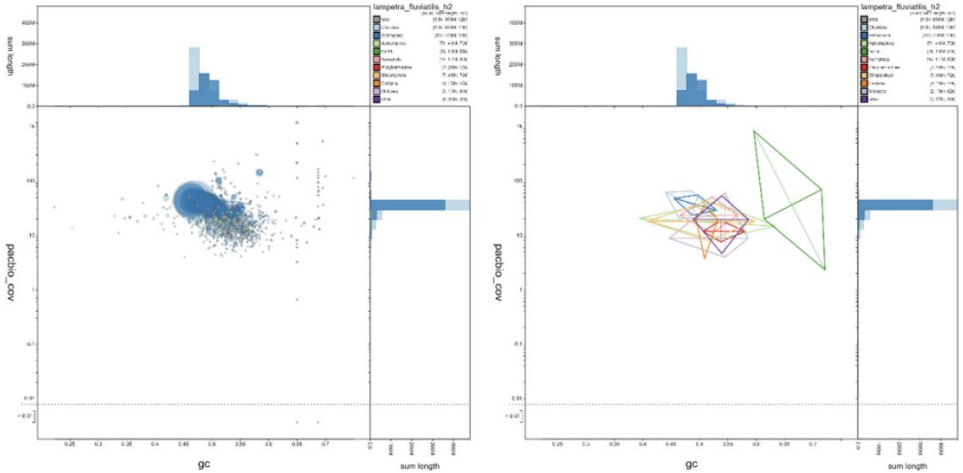
JOURNAL ARTICLE

BlobToolKit – Interactive Quality Assessment of Genome Assemblies

Richard Challis ✉, Edward Richards, Jeena Rajan, Guy Cochrane, Mark Blaxter

G3 Genes|Genomes|Genetics, Volume 10, Issue 4, 1 April 2020, Pages 1361–1374,
https://doi.org/10.1534/g3.119.400908

Published: 01 April 2020 **Article history** ▾



Methodology Article | Open Access | Published: 01 December 2017

Decontaminating eukaryotic genome assemblies with machine learning

Janna L. Fierst ✉ & Duncan A. Murdock

BMC Bioinformatics **18**, Article number: 533 (2017) | Cite this article

Article

CleanSeq: A Pipeline for Contamination Detection, Cleanup, and Mutation Verifications from Microbial Genome Sequencing Data

Caiyan Wang [1], Yang Xia [2], Yunfei Liu [2], Chen Kang [1], Nan Lu [2], Di Tian [2], Hui Lu [2], Fuhai Han [2], Jian Xu [2,*] and Tetsuya Yomo [2,*]

We use FCS-GX because it is…

Fast

- Compared to BlobToolKit
- Runtime on Fox

Easy to use

- You´ll see when you look at the tutorial ;)

Gives great results

- *Sphagnum troendelagicum* example:

Assembly	# contigs	Total assembly size
Hap 1	7831	708484278
Hap2	2232	521020003

Why do we need to manually curate our assemblies?

- Mistakes during assembly or scaffolding
- Reads with ambiguous contact signals

REVIEW

Significantly improving the quality of genome assemblies through curation

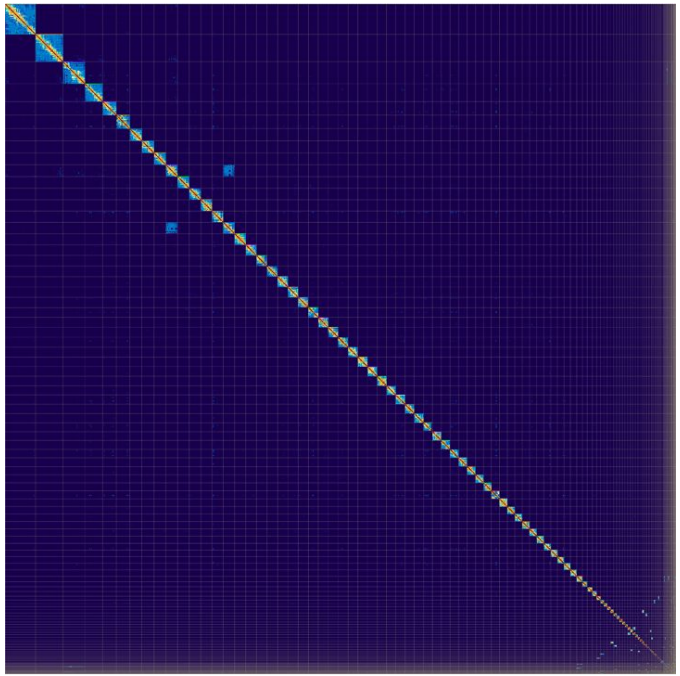
Kerstin Howe*, William Chow, Joanna Collins, Sarah Pelan, Damon-Lee Pointon, Ying Sims, James Torrance, Alan Tracey and Jonathan Wood

Tree of Life, Wellcome Sanger Institute, Cambridge CB10 1SA, UK

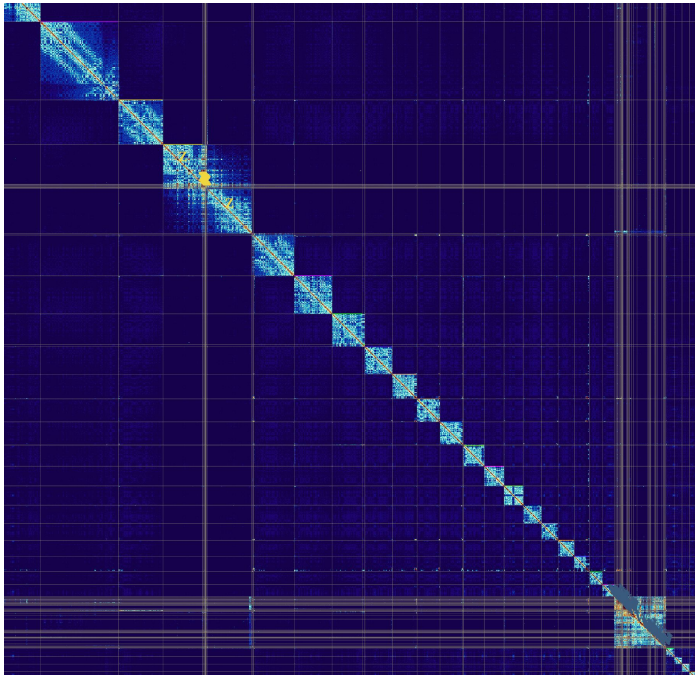
*Correspondence address. Kerstin Howe, Tree of Life, Wellcome Sanger Institute, Cambridge CB10 1SA, UK.
E-mail: kerstin@sanger.ac.uk http://orcid.org/0000-0003-2237-513X

Examples

River lamprey (*Lampetra fluviatilis*)



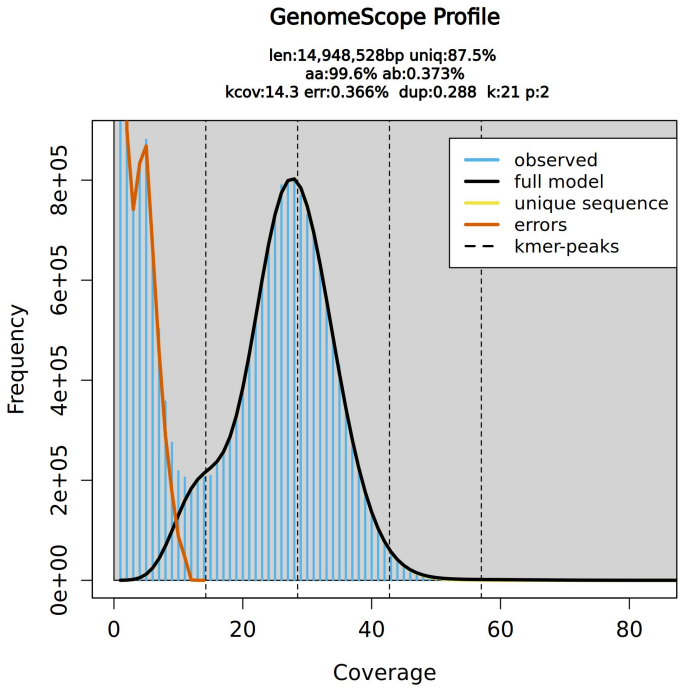
Snowy owl (*Bubo scandiacus*)



Summary - Decontamination and manual curation

Where you started:

- over 420 megabases of HiFi reads
- over 840 megabases of Hi-C reads
- over 25 megabases of contaminant reads



Where you are at now:

- Haplotype resolved, chromosome level assemblies
- Decontaminated and manually curated
- N50 of 3.15 Mbp and BUSCO completeness of 94%.

End of workshop summary

After attending the workshop learners should:

1. Know about most-used approaches for genome assembly
    a. Filtration with **HiFiAdapterFilt**
    b. Assembly with **hifiasm**
    c. Scaffolding with **YaHS**
2. Assess information inherit in sequencing reads
    a. Pre-assembly checks with **GenomeScope2** and **Smudgeplot**
3. Be able to validate genome assemblies
    a. Assembly validation with **gfastats**, **BUSCO** and **Merqury**
4. Know about manual curation of assemblies
    a. The **Rapid curation suite** and **PretextView**

Evaluation form