

09:00-12:30 Assembly

- 09:00-09:30 Introduction to whole-genome assembly and introduction to the study system and infrastructure
- 09:30-12:25 Assembling the yeast genome
 - GenomeScope2
 - Smudgeplot
 - HiFiAdapterFilt
 - hifiasm
 - YaHS
- 10:45-11:15 Coffee break together with Genetic Diversity Monitoring Workshop
- 12:25-12:30 Summary

12:30-13:30 Lunch together with Genetic Diversity Monitoring Workshop

13:30-14:15 Validation

- 13:30-13:45 Introduction to assembly validation
- 13:45-14:10 Interpreting assembly validation results
 - gfastats
 - BUSCO
 - Merqury
- 14:10-14:15 Summary

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
Norwegian Biodiversity & Genomics Conference 2023

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

09:00-12:30 Assembly

- 09:00-09:30 Introduction to whole-genome assembly and introduction to the study system and infrastructure
- 09:30-12:25 Assembling the yeast genome
 - GenomeScope2
 - Smudgeplot
 - HiFiAdapterFilt
 - hifiasm
 - YaHS
- 10:45-11:15 Coffee break together with Genetic Diversity Monitoring Workshop
- 12:25-12:30 Summary

12:30-13:30 Lunch together with Genetic Diversity Monitoring Workshop

13:30-14:15 Validation

- 13:30-13:45 Introduction to assembly validation
- 13:45-14:10 Interpreting assembly validation results
 - gfastats
 - BUSCO
 - Merqury
- 14:10-14:15 Summary

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?

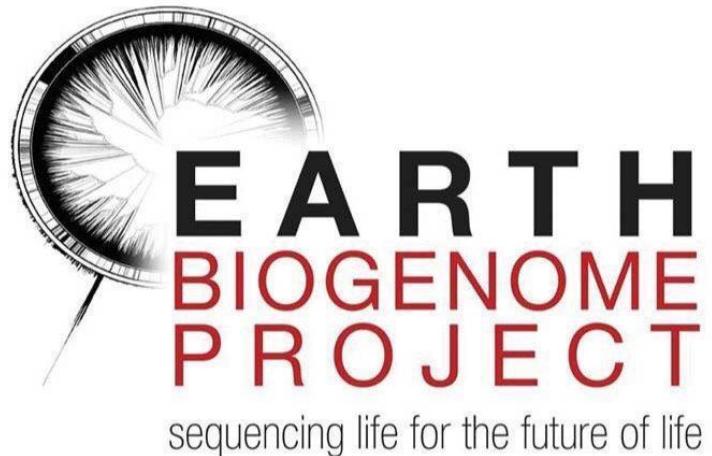
- Do all kinds of different species (DTol, ERGA, EBP)
- Vertebrate Genomes Project is targeted (all vertebrates), maybe not a biodiversity genomics project



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

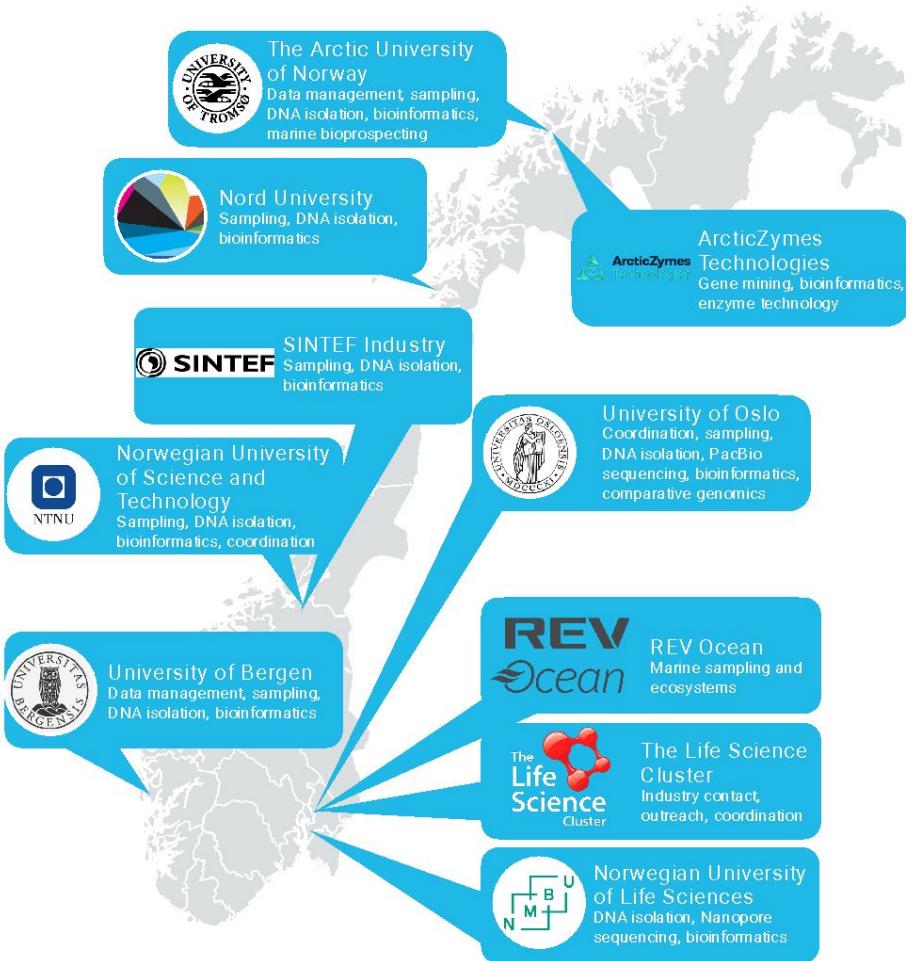
-> sequence all eukaryotes



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

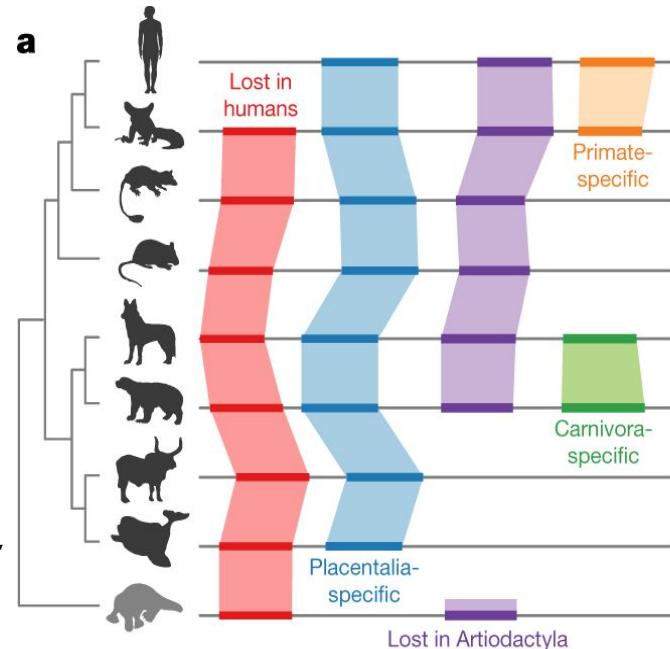


SARS-CoV-2 and host range

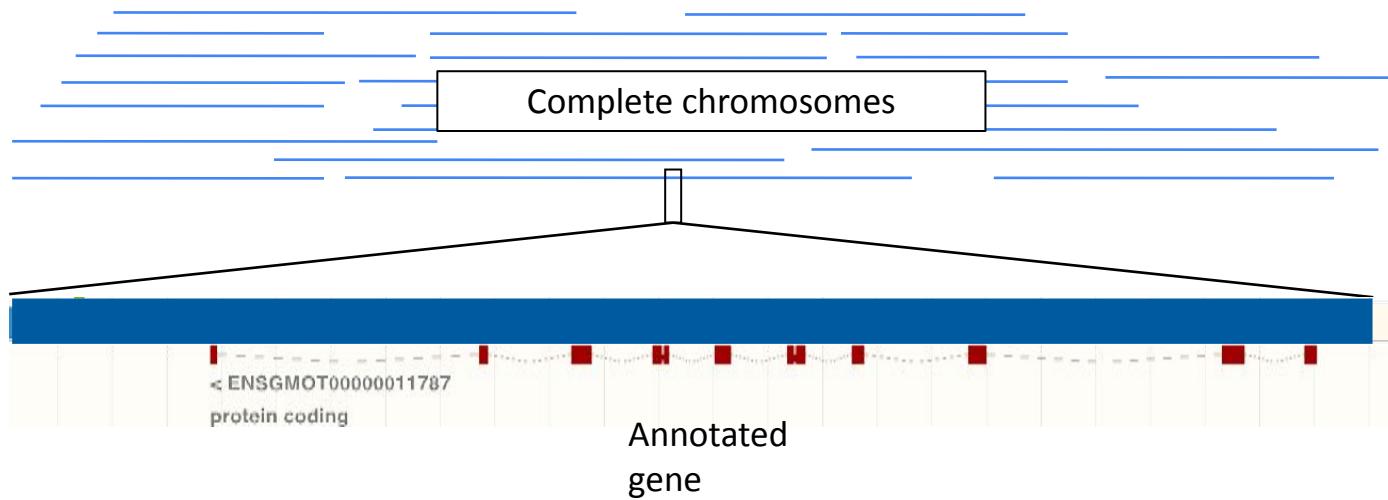
- SARS-CoV-2 binds to ACE2
- Primates have very high probability
- Cervid deer and cetacean high probability

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 , Andreas R. Pfenning^k , Huabin Zhao^{l,m} , Diane P. Genereuxⁿ , Ross Swoffordⁿ , Katherine S. Pollard^{d,o,p} , Oliver A. Ryder^{q,r} , Martin T. Nweiss^{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} 

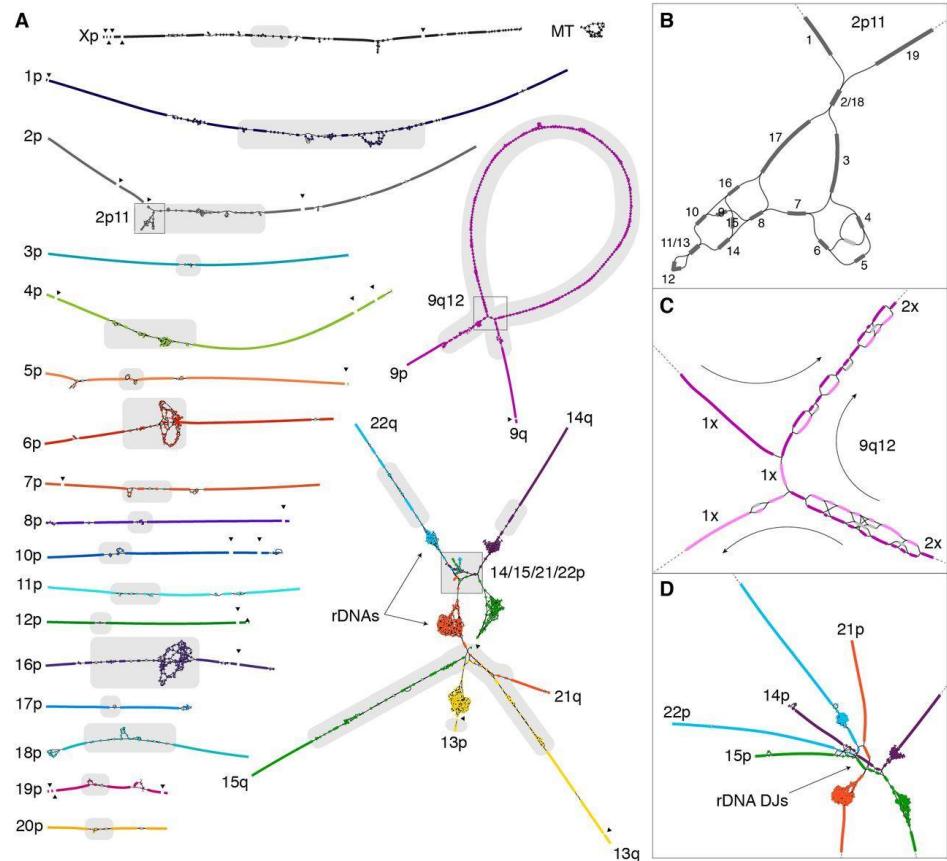


An ideal genome assembly



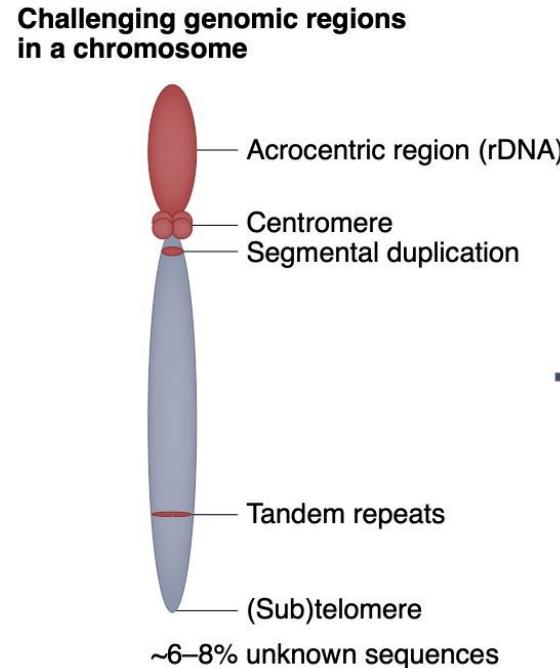
Telomer-to-telomer human genome

- 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

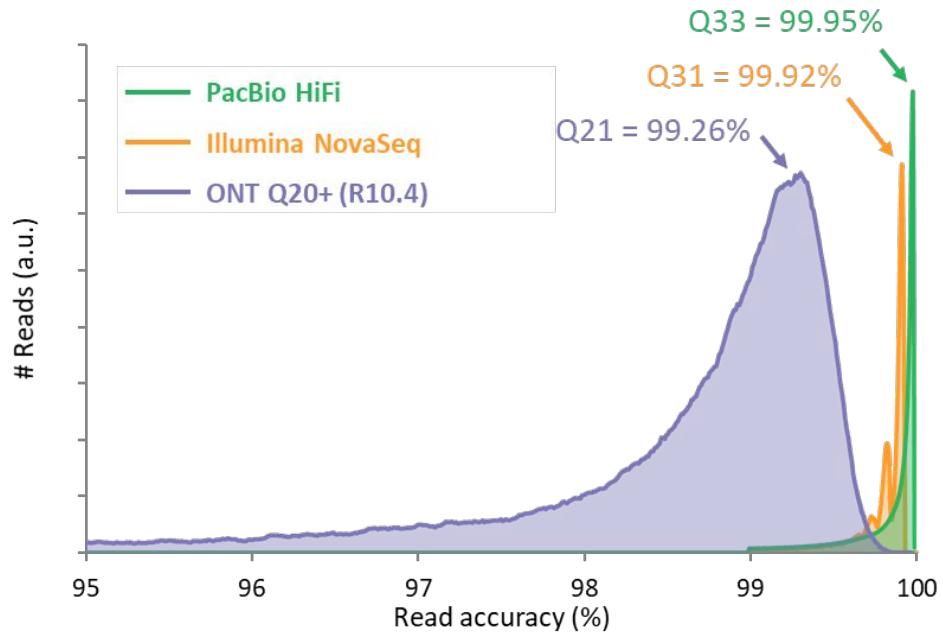
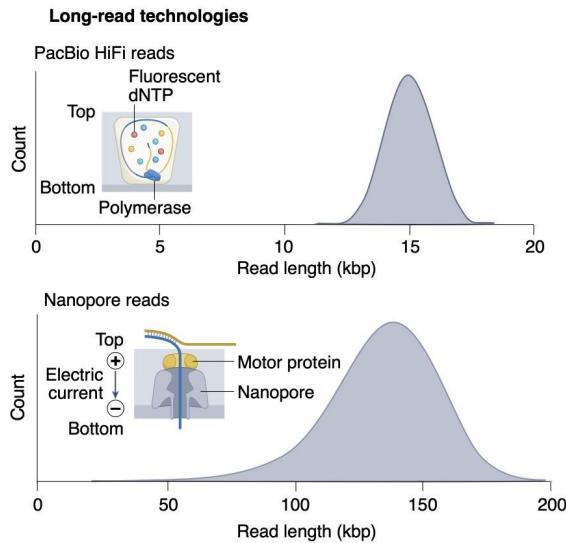


Challenging regions in a genome

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



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

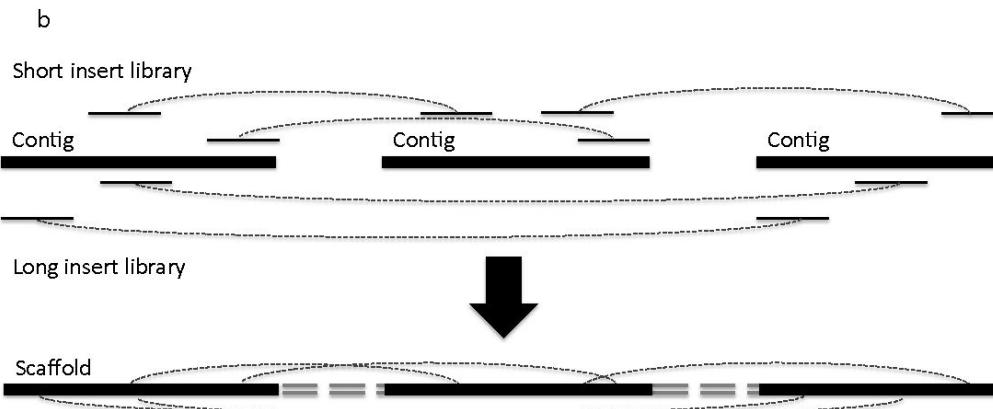
a

Aligned reads

ACCGCGATTCAAGGTTACCACCGC
GCGATTCAAGGTTACCACCGCTA
GATTCAAGGTTACCACCGTAGC
TTCAGGTTACCACCGTAGCAC
CAGGTTACCACCGTAGCACAT
GGTTACCACCGTAGCACATTAC
TTACCACCGTAGCACATTACAC
ACCACCGTAGCACATTACACAG
CACCGTAGCACATTACACAGAT
CCGCGTAGCACATTACACAGATTA

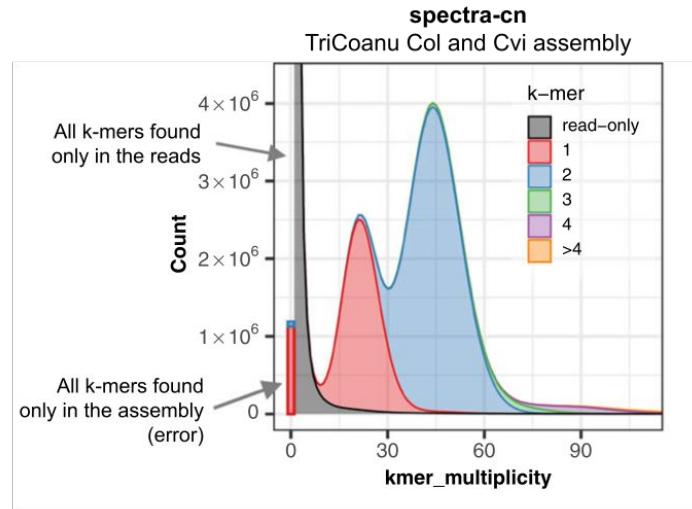
Consensus contig

ACCGCGATTCAAGGTTACCACCGTAGCGCATTACACAGATTAG



EBP standards

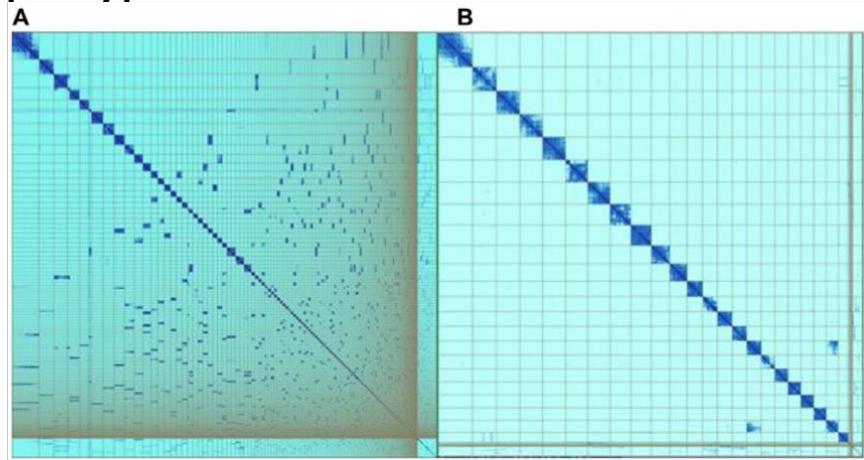
- **6.C.Q40**
 - 10^6 bp N50 contig
 - Chromosome scale N50 scaffolding
 - Q40 error rate, fewer than 1 error per 10,000 bp
- < 5% false duplications
- > 90% kmer completeness
- > 90% sequence assigned to candidate chromosomal sequences
- > 90% single copy conserved genes (e.g. BUSCO) complete and single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



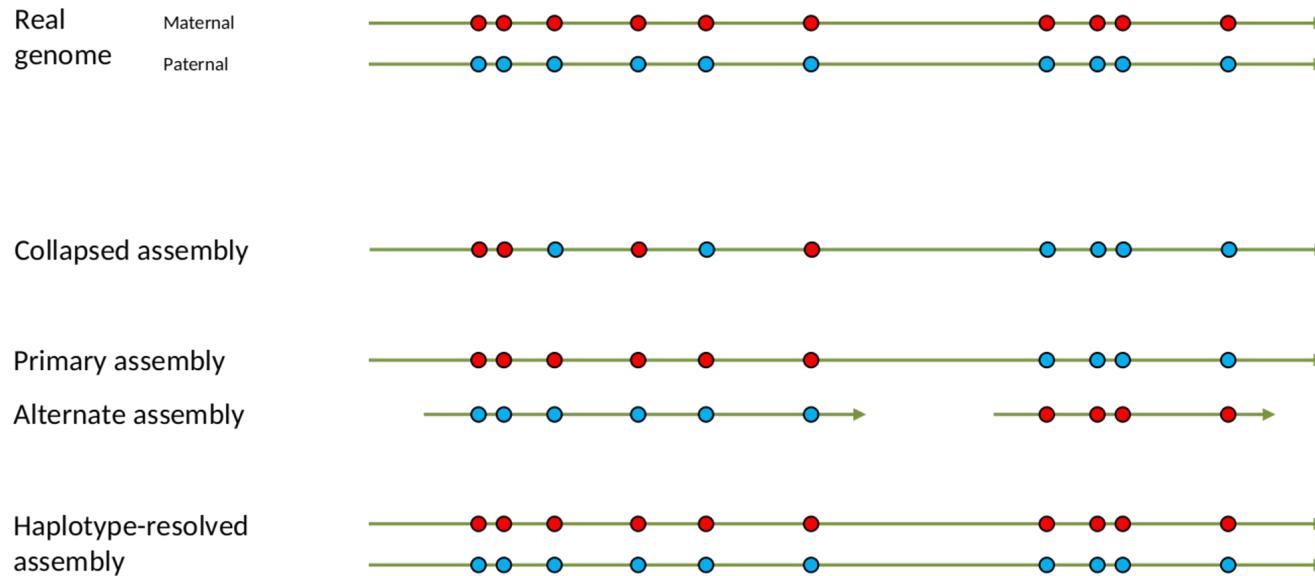
Merqury: Rhie et al. *Genome Biology* 2020

EBP standards (cont'd)

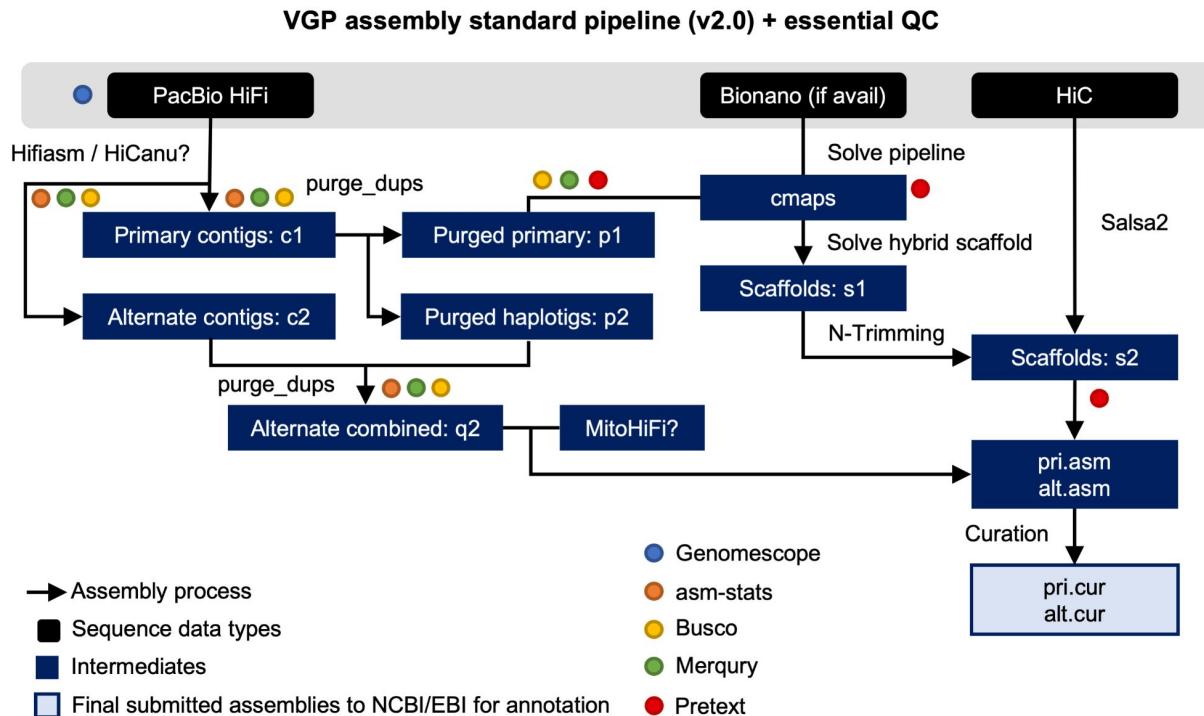
- Separation of target species vs contaminants/symbionts/cobionts
- For diploid species: Identification of primary (haploid) assembly, with secondary assembly with alternate haplotypes
- Organelle genomes
- Manual curation
- Reconciliation with known karyotype



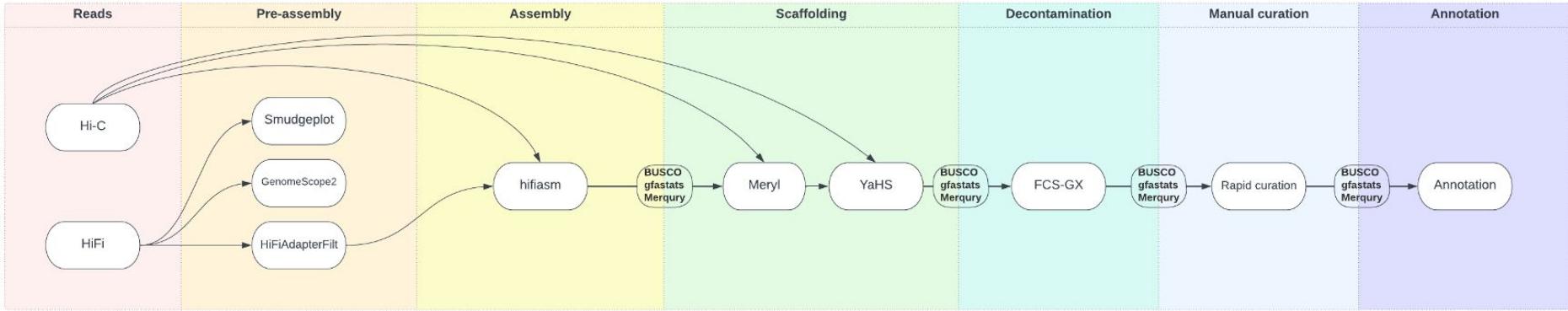
Haplotype-resolved assemblies



Vertebrate Genomes Project pipeline



EBP-Nor pipeline



Athalia rosae: the coleseed sawfly

Also known as the turnip sawfly

Sequenced to a high coverage by Darwin Tree of Life

PacBio HiFi and Hi-C (also 10X)

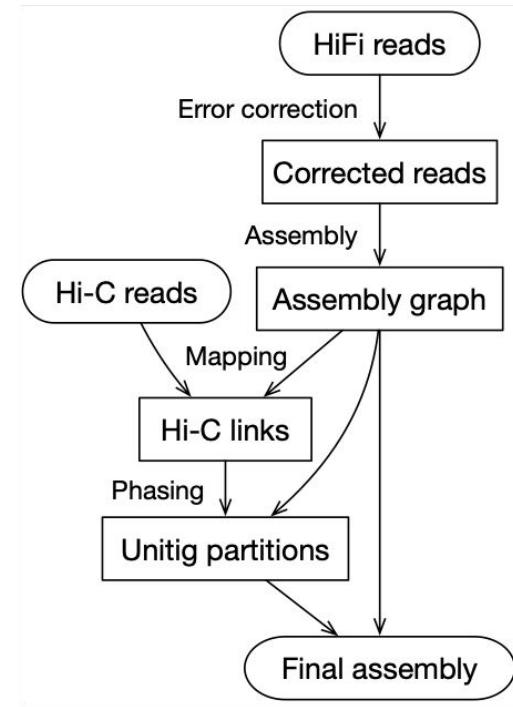
168.69 Mbp genome with 8 chromosomes

Subsampled to 30x HiFi and 60x Hi-C for this workshop



Combining Hi-C and HiFi in hifiasm

- Powerful combination
- Used by Human Pan-genome 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)

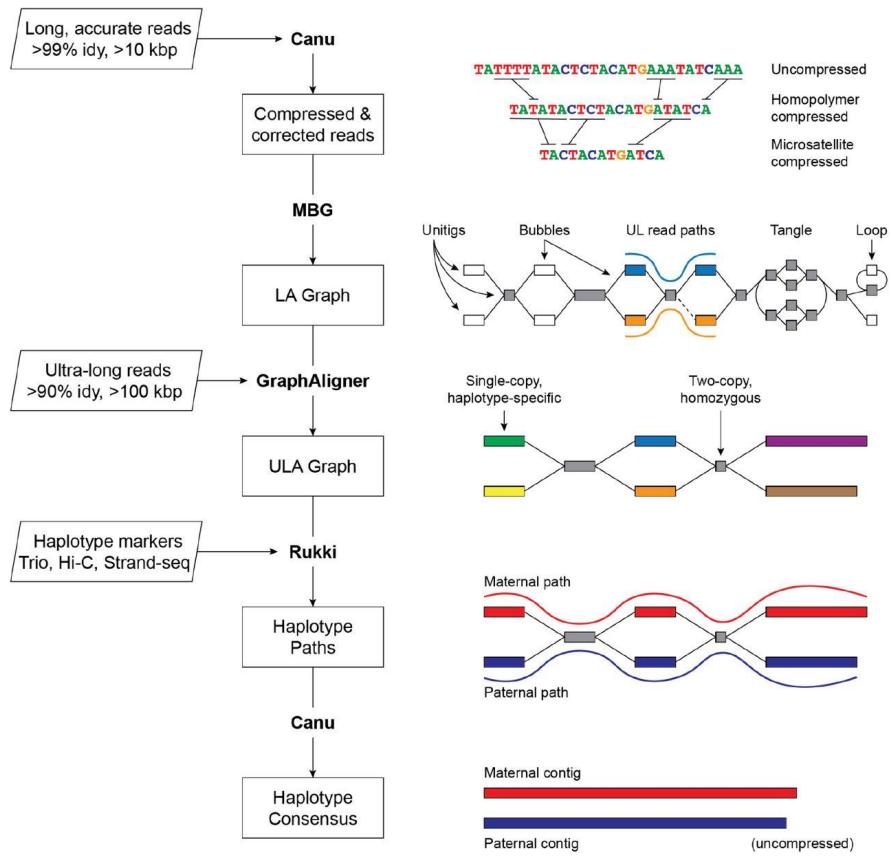


Summary - Assembly

- A genome assembly is a model of the genome, it most likely contain differences to what is in the cells of a species
- Genome assemblies from EBP associated projects are made to certain standards to be able to be used for all kinds of different purposes (not just “good enough”)
- Balancing sequencing costs, computing costs and manpower costs
- The data can be utilized better than today (better than hifiasm + YaHS)

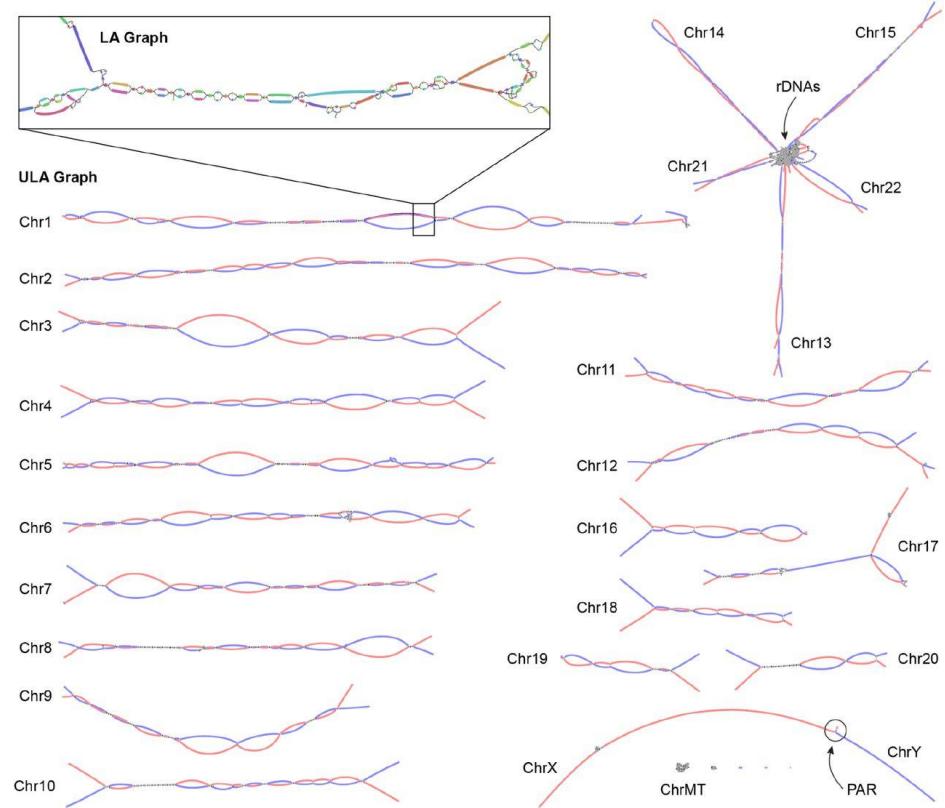
Assembly - the future

Workflow of verkko



Assembly - the future

Diploid graph human genome
from verkko



Rautiainen et al. *bioRxiv* 2022

Practicals - Genome assembly



<https://github.com/ebp-nor/genome-assembly-workshop-2023>

Break

See you back here at 13:30 :)

09:00-12:30 Assembly

- 09:00-09:30 Introduction to whole-genome assembly and introduction to the study system and infrastructure
- 09:30-12:25 Assembling the yeast genome
 - GenomeScope2
 - Smudgeplot
 - HiFiAdapterFilt
 - hifiasm
 - YaHS
- 10:45-11:15 Coffee break together with Genetic Diversity Monitoring Workshop
- 12:25-12:30 Summary

12:30-13:30 Lunch together with Genetic Diversity Monitoring Workshop

13:30-14:15 Validation

- 13:30-13:45 Introduction to assembly validation
- 13:45-14:10 Interpreting assembly validation results
 - gfastats
 - BUSCO
 - Merqury
- 14:10-14:15 Summary

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

BUSCO

merqury

marbl/merqury

K-mer based assembly evaluation

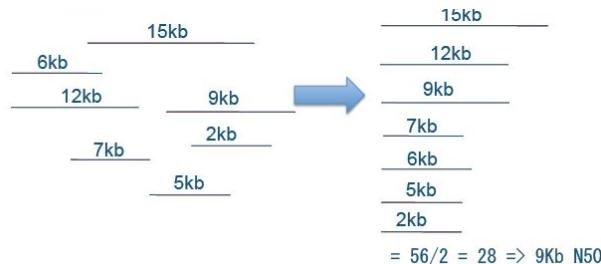


5 Contributors 17 Issues 185 Stars 16 Forks



N50 and assembly statistics

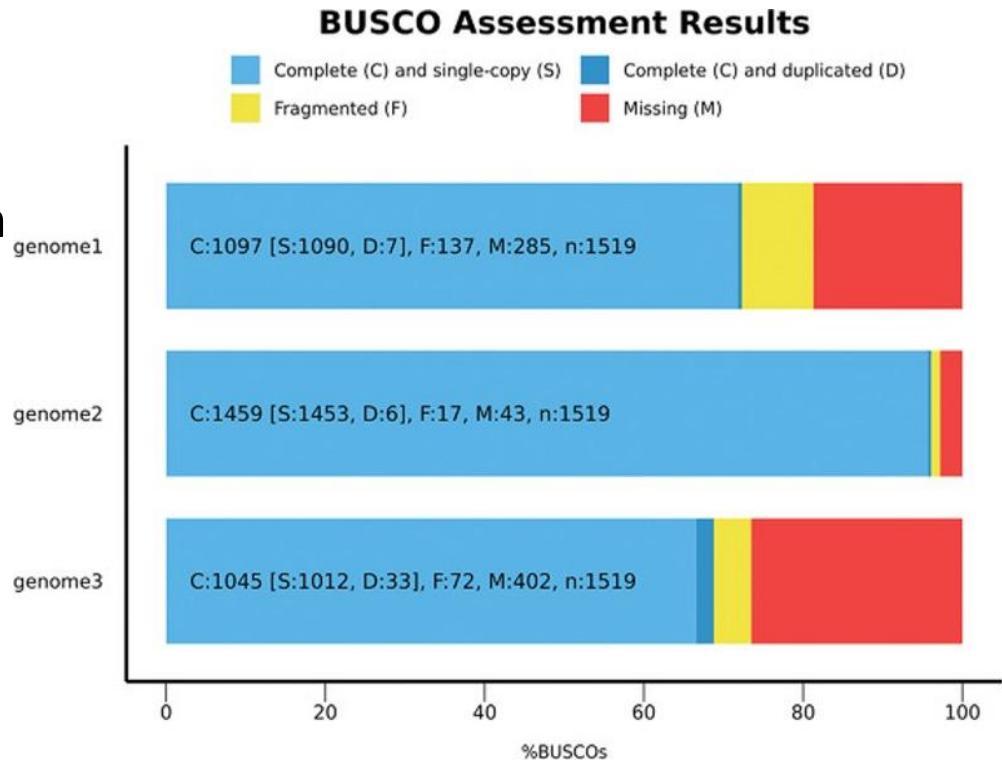
- Length of contig such that 50 % of total bases are in contigs of this length or longer
- gfastats gives N50 and several other statistics such as average, longest, etc



Sum of lengths: 56 kbp

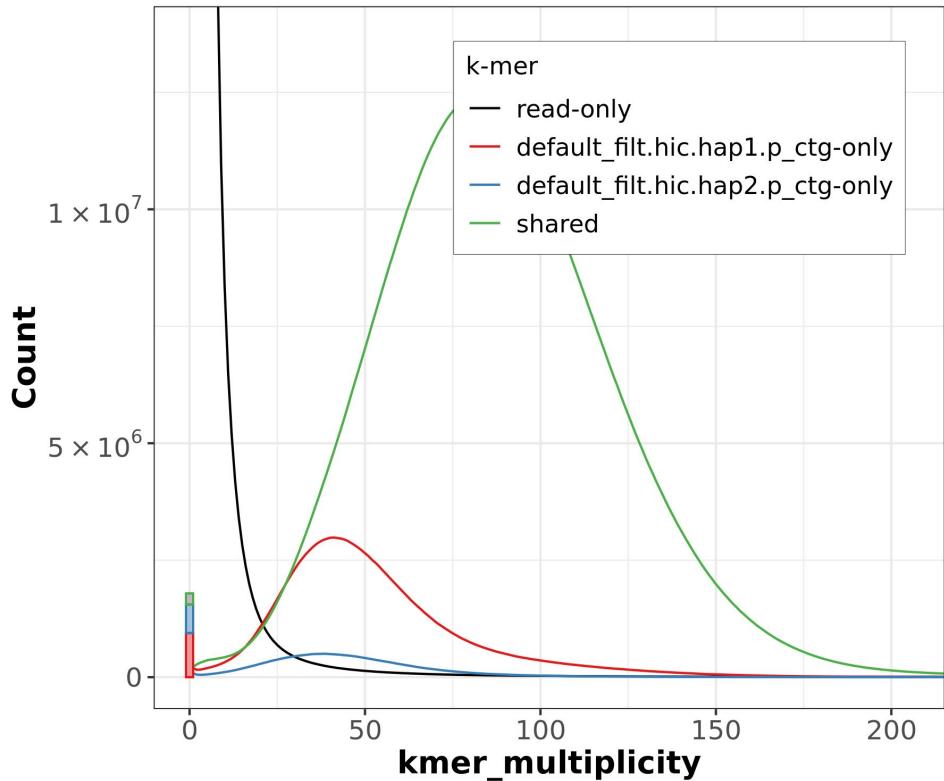
BUSCO

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



mercury

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



Summary - Validation

Results from gfastats:

```
+++Assembly summary+++
# scaffolds: 276
Total scaffold length: 183264929
Average scaffold length: 664003.37
Scaffold N50: 18416673
Scaffold auN: 18824422.95
Scaffold L50: 4
Largest scaffold: 31887225
Smallest scaffold: 1000
# contigs: 419
Total contig length: 183236329
Average contig length: 437318.21
Contig N50: 2248278
Contig auN: 2429466.60
Contig L50: 26
Largest contig: 8063521
Smallest contig: 1000
# gaps in scaffolds: 143
Total gap length in scaffolds: 28600
Average gap length in scaffolds: 200.00
Gap N50 in scaffolds: 200
Gap auN in scaffolds: 200.00
Gap L50 in scaffolds: 72
Largest gap in scaffolds: 200
Smallest gap in scaffolds: 200
Base composition (A:C:G:T): 54014066:37547201:37528614:54146448
GC content %: 40.97
# soft-masked bases: 0
# segments: 419
Total segment length: 183236329
Average segment length: 437318.21
# gaps: 143
# paths: 276
```

Summary - Validation

Results from gfastats:

Results from BUSCO:

***** Results: *****

C:95.6%[S:95.3%,D:0.3%],F:1.6%,M:2.8%,n:5991
5722 Complete BUSCOs (C)
5707 Complete and single-copy BUSCOs (S)
15 Complete and duplicated BUSCOs (D)
96 Fragmented BUSCOs (F)
173 Missing BUSCOs (M)
5991 Total BUSCO groups searched

Assembly Statistics:

276 Number of scaffolds
419 Number of contigs
183264929 Total length
0.016% Percent gaps
18 MB Scaffold N50
2 MB Contigs N50

Dependencies and versions:

hmmsearch: 3.1
bbtools: 39.01
metaeuk: 6.a5d39d9
busco: 5.4.3

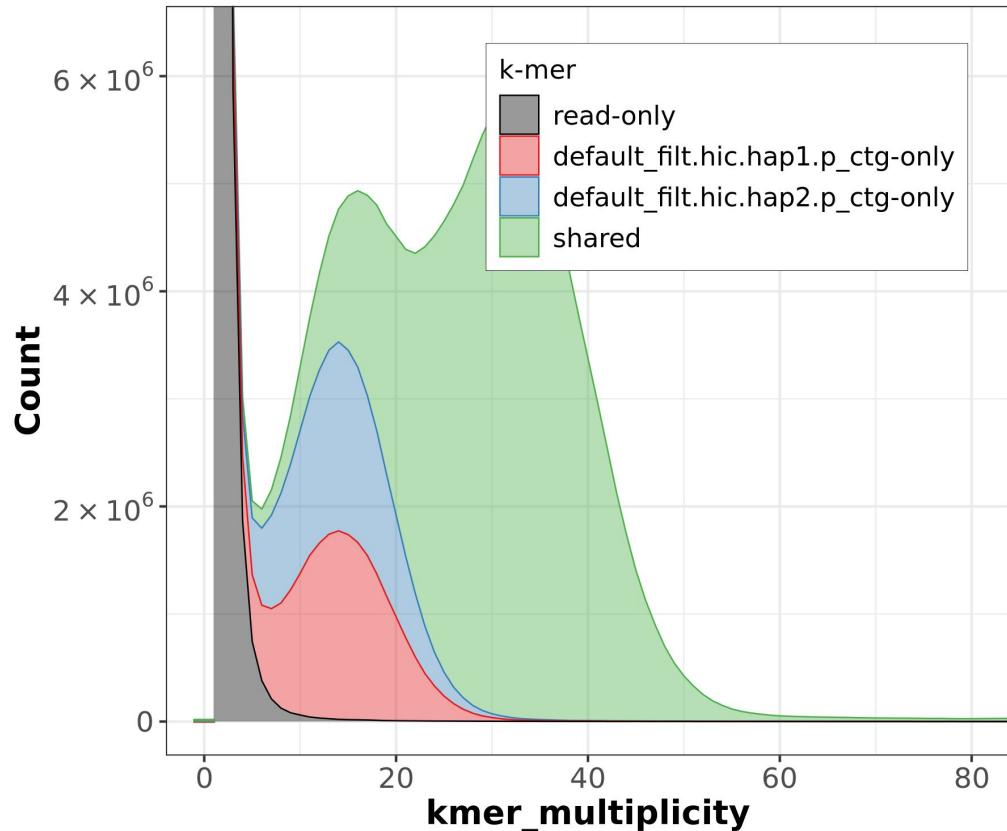
Summary - Validation

Results from gfastats:

Results from BUSCO:

Results from merquery:

***** Results: *****



QV-values and K-mer completeness

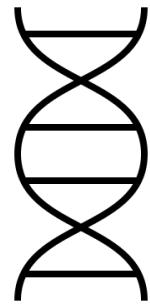
```
-- -- -- -- --  
(base) 11:45:50-benedga@login-5:/cluster/projects/nn8013k/results/workshop_2023/  
default_filt.hic.hap1.p_ctg    7565    183228249      57.0639 1.9661e-06  
default_filt.hic.hap2.p_ctg    5999    178770625      57.9643 1.59798e-06  
Both    13564   361998874      57.4853 1.7843e-06  
(base) 11:45:57-benedga@login-5:/cluster/projects/nn8013k/results/workshop_2023/  
default_filt.hic.hap1.p_ctg    all     152457415      175308467      86.9652  
default_filt.hic.hap2.p_ctg    all     152335386      175308467      86.8956  
both    all     174547611      175308467      99.566  
-- -- -- -- --
```

Decontamination and manual curation

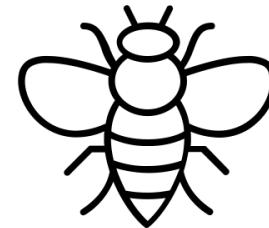
Why do we need to decontaminate our assemblies?



Contamination during
handling of samples



Contamination during
sequencing



Contamination from
the sample itself

How do we decontaminate our assemblies?

JOURNAL ARTICLE

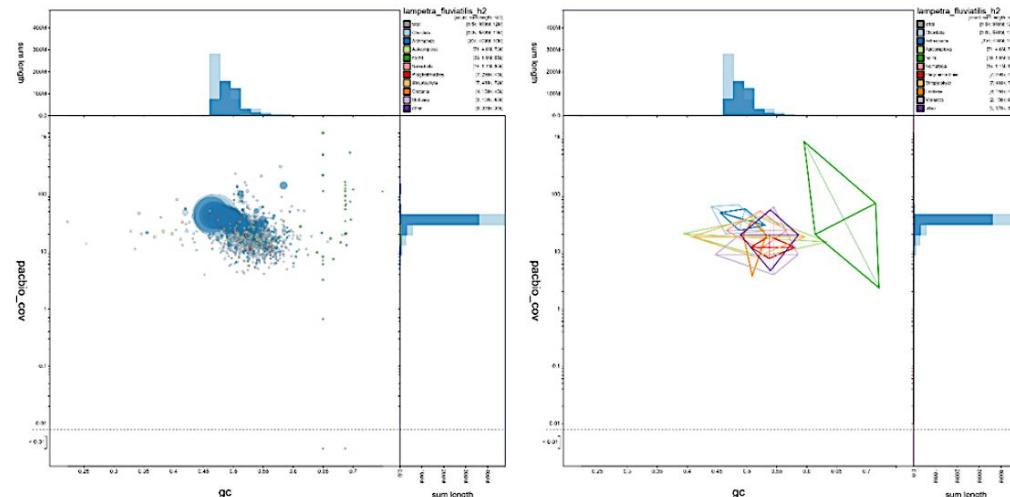
BlobToolKit – Interactive Quality Assessment of Genome Assemblies ⓘ

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 ▾



How do we decontaminate our assemblies?

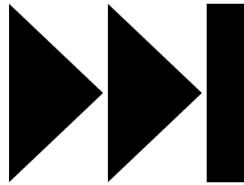
FCS-adaptor and FCS-GX

Adaptor: searches for adaptor sequences

GX: searches for sequences from a wide range of organisms from the RefSeq prokaryotes, RefSeq eukaryotes, ReSeq viruses, RefSeq plasmids and GenBank fungi, nematodes, protists and algea databases



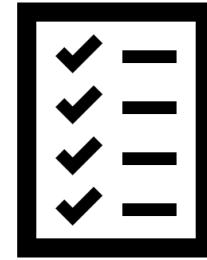
In EBP-Nor we use FCS-GX



Fast: Runtime
of 20 minutes
and 11 seconds



Easy: Quick
set-up and
easy to run on
the cluster

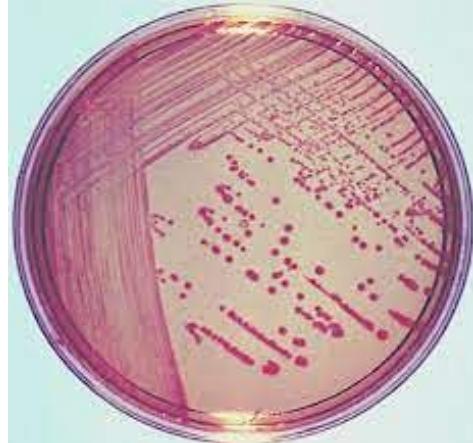


Great results:
.txt reports,
actionable

What is the end result after running FCS-GX?

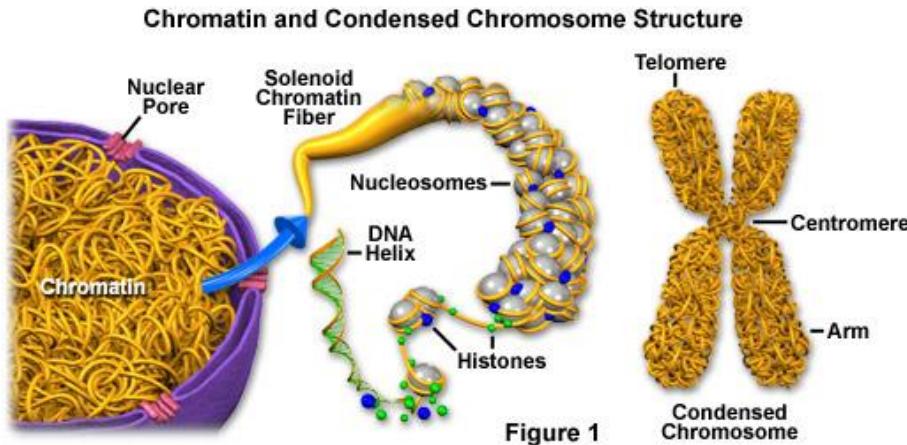
```
##[["FCS genome report", 2, 1], {"git-rev": "0.2.1-14-ga7e5602", "run-date": "Sat Jan 28 10:40:28 2023", "dvg": 0.948, "asserted-div": "anml:insects", "primary-divs": ["anml:insects"]}]}
```

#seq_id	start_pos	end_pos	seq_len	action	div	agg_cont_cov	top_tax_name
scaffold_10	1	3321559	3321559	EXCLUDE	prok:g-proteobacteria	89	Entomonas moraniae
scaffold_72	1	64173	64173	EXCLUDE	prok:g-proteobacteria	88	Moellerella wisconsensis
scaffold_75	1	60705	60705	EXCLUDE	prok:g-proteobacteria	91	Moellerella wisconsensis
scaffold_79	1	58736	58736	EXCLUDE	prok:g-proteobacteria	95	Moellerella wisconsensis
scaffold_84	1	55917	55917	EXCLUDE	prok:g-proteobacteria	100	Moellerella wisconsensis
scaffold_85	1	55405	55405	EXCLUDE	prok:g-proteobacteria	95	Moellerella wisconsensis
scaffold_90	1	53128	53128	EXCLUDE	prok:g-proteobacteria	100	Moellerella wisconsensis
scaffold_94	1	51577	51577	EXCLUDE	prok:g-proteobacteria	100	Moellerella wisconsensis
scaffold_102	1	48896	48896	EXCLUDE	prok:g-proteobacteria	95	Moellerella wisconsensis
scaffold_105	1	46793	46793	EXCLUDE	prok:g-proteobacteria	89	Moellerella wisconsensis
scaffold_112	1	45685	45685	EXCLUDE	prok:g-proteobacteria	100	Moellerella wisconsensis
scaffold_115	1	44639	44639	EXCLUDE	prok:firmicutes	100	Lactococcus lactis
scaffold_116	1	44156	44156	EXCLUDE	prok:g-proteobacteria	100	Moellerella wisconsensis
scaffold_124	1	42653	42653	EXCLUDE	prok:g-proteobacteria	100	Moellerella wisconsensis
scaffold_128	1	41243	41243	EXCLUDE	prok:g-proteobacteria	100	Moellerella wisconsensis
scaffold_130	1	40926	40926	EXCLUDE	prok:g-proteobacteria	100	Moellerella wisconsensis
scaffold_131	1	40423	40423	EXCLUDE	prok:g-proteobacteria	93	Moellerella wisconsensis
scaffold_133	1	40206	40206	EXCLUDE	prok:g-proteobacteria	100	Moellerella wisconsensis

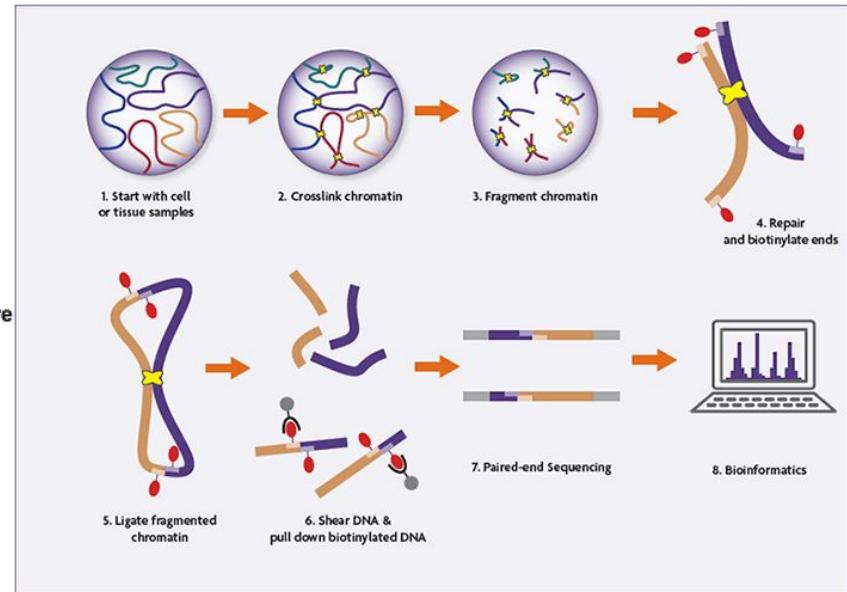


Seyman, D. et al. (2015) "First case of primary bacteremia caused by *Moellerella wisconsensis*: A case report and literature review," *Klinik Dergisi/Klinik Journal*, 26(3), pp. 119–121. Available at: <https://doi.org/10.5152/kd.2013.34>.

Re-introducing Hi-C sequencing

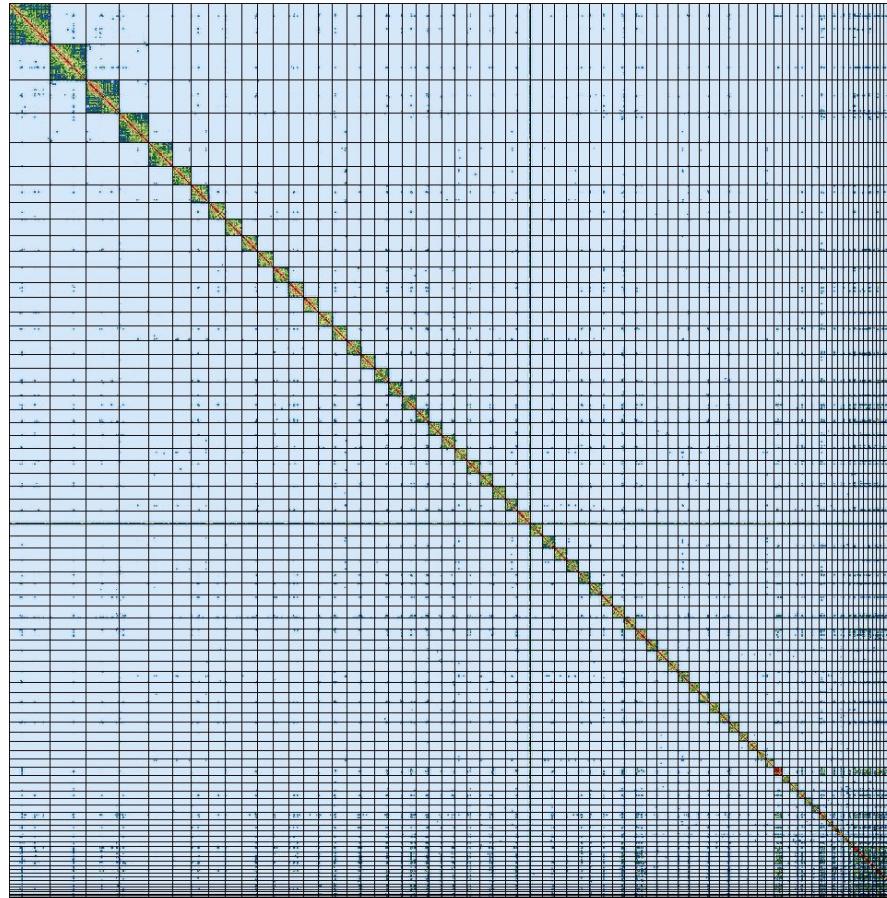


<https://micro.magnet.fsu.edu/cells/nucleus/chromatin.html>

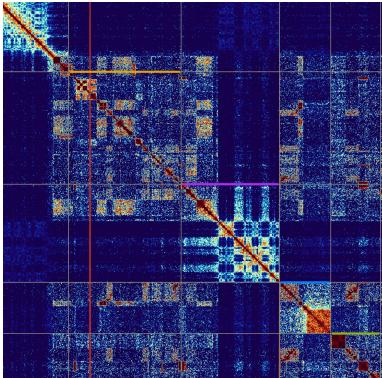


<https://www.activemotif.com/blog-hi-c>

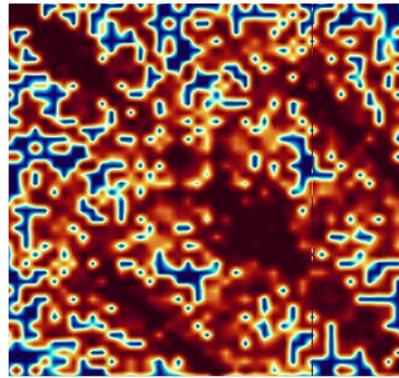
What is a Hi-C contact map?



Some situations you may encounter



Ambiguous
contact signals



Haplotypic
duplications

Extensions
Graph: PB coverage
Graph: gaps
Graph: repeat density

Lack of data

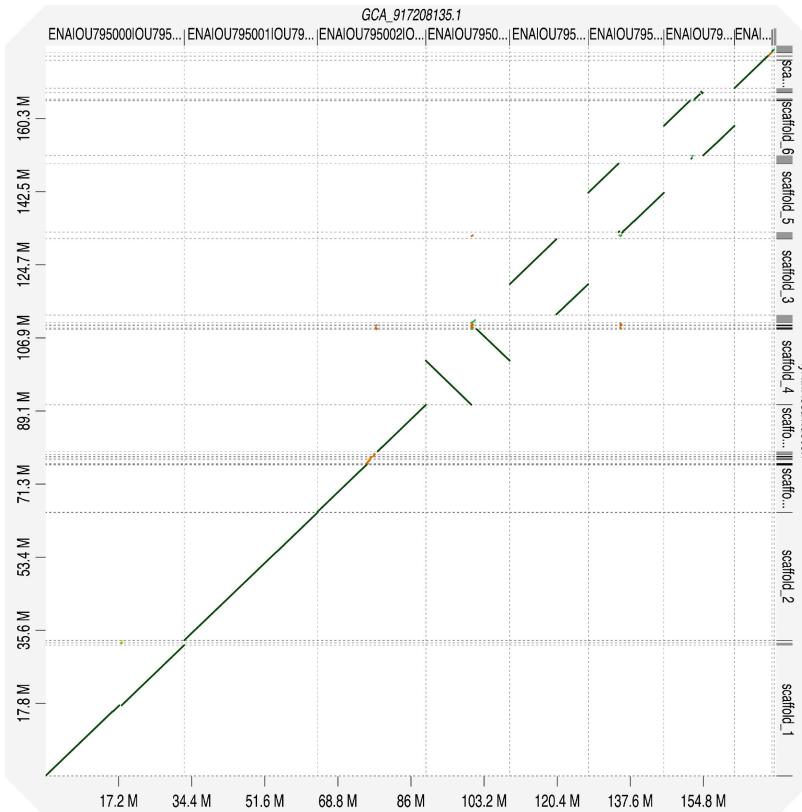
Now it's time for you to do this!

Go to the FCS-GX tutorial on our github repository, and follow along until you have finished decontaminating.

When you are done with the decontamination step, I will curate the coleseed sawfly assembly with you.

After we are done curating, we will go through a final summary of the session and the entire workshop.

Working with a simplified dataset



Where you started

5.7 Gbp of pacbio HiFi data at 30X coverage

9 Gbp of Hi-C data at 60X coverage

Where you are at now

Haplotype resolved, decontaminated and curated whole-genome assemblies of EBP standards

Statistics summary:

Assembly length *decontamination	Number of pseudochromosomes	Longest scaffold *decontamination	Scaffold N50 *decontamination	BUSCO completeness *YaHS
178,145,017	?	31,887,225	18,416,673	95.6%

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

