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 sawfly genome
    - GenomeScope2
    - Smudgeplot
    - HiFiAdapterFilt
    - hifiasm
    - YaHS
- 10:00-10:15 Coffee break
- 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

14:00-16:00 Decontamination and manual curation

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

Genome assembly, annotation and comparative genomics

Day 1

Teachers: Ole K. Tørresen, Bram Danneels and Helle T. Baalsrud

Norwegian Biodiversity & Genomics Conference 2024

8th April

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

Practicals - Genome assembly

https://github.com/ebp-nor/workshop-2024

Or follow link in e-mail.

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













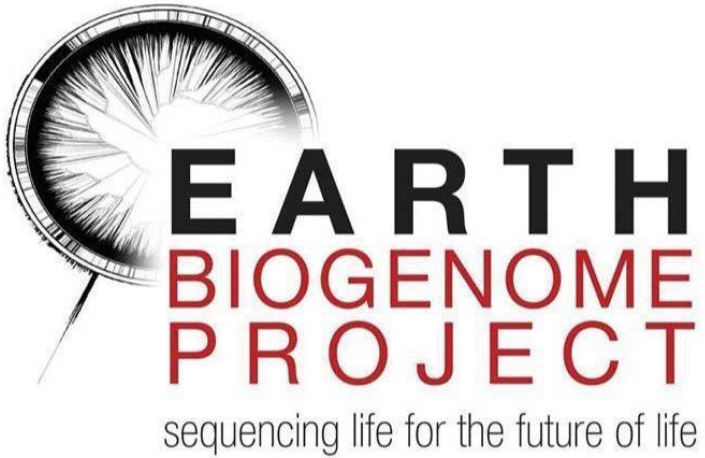

EARTH BIOGENOME PROJECT

NOR

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/)

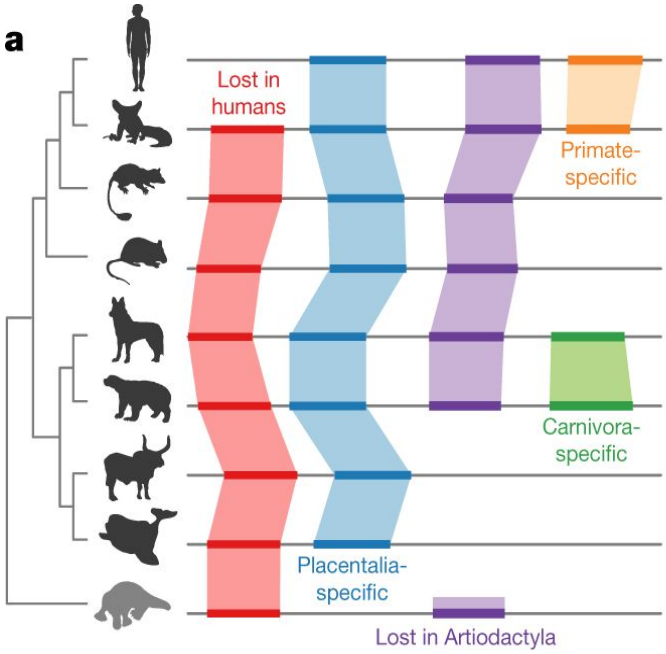
- Preparation for 2 phase has begun

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

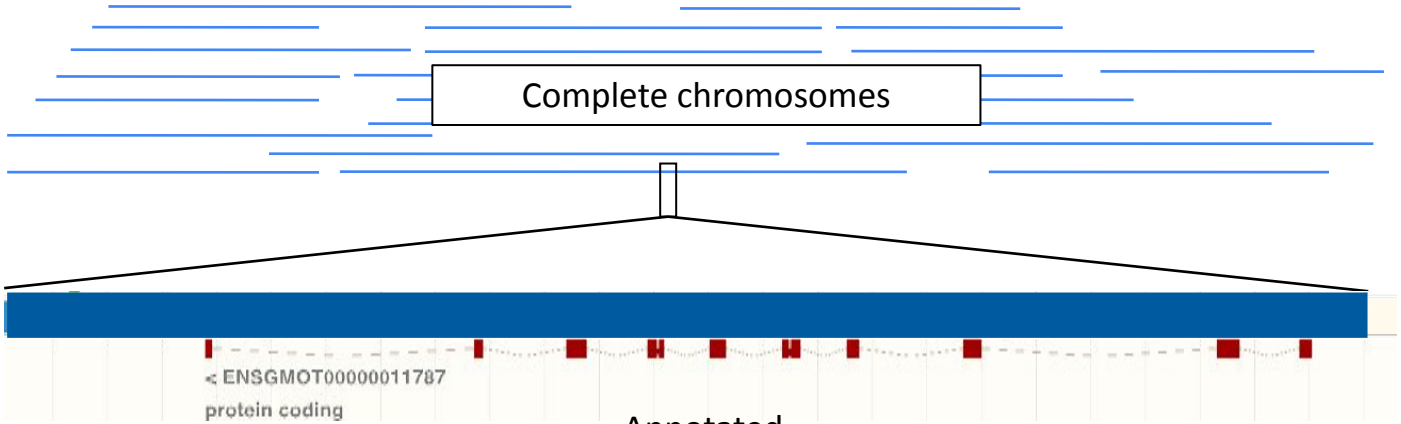
Joana Damas[a,1], Graham M. Hughes[b,1], Kathleen C. Keough[c,d,1], Corrie A. Painter[e,1], Nicole S. Persky[f,1], Marco Corbo[a], Michael Hiller[g,h,i], Klaus-Peter Koepfli[j], Andreas R. Pfenning[k], Huabin Zhao[l,m], Diane P. Genereux[n], Ross Swofford[n], Katherine S. Pollard[d,o,p], Oliver A. Ryder[q,r], 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

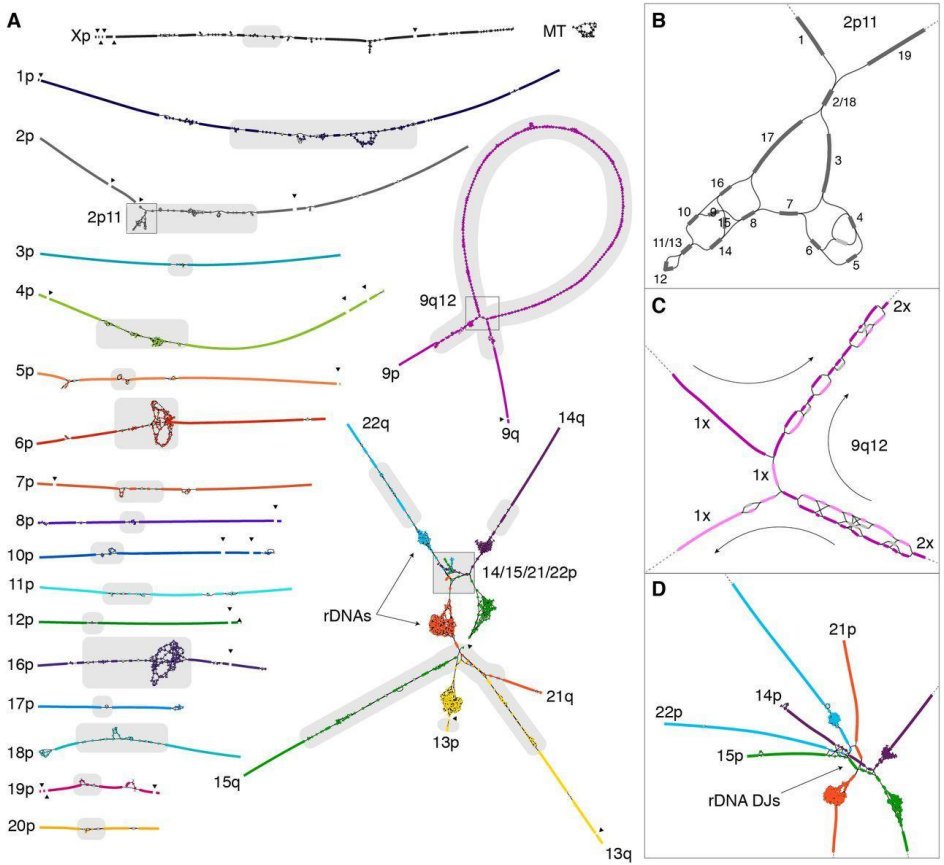


Annotated gene

Slide modified after Lex Nederbragt

Telomere-to-telomere 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

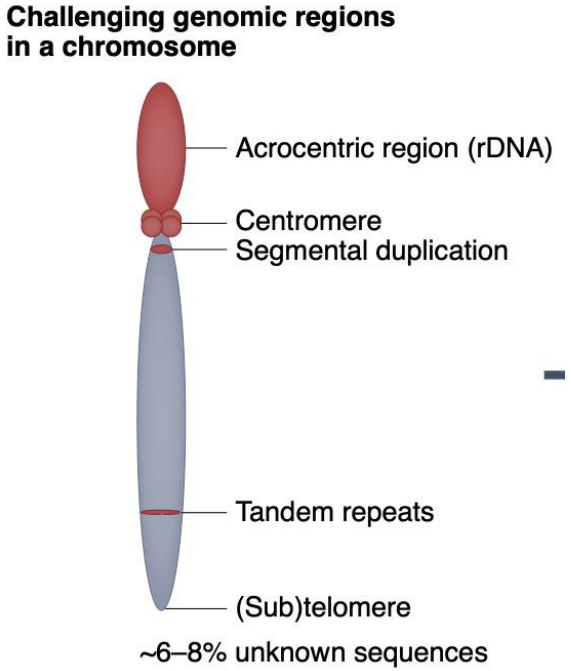


Nurk et al. *Science* 2022

Challenging regions in a genome

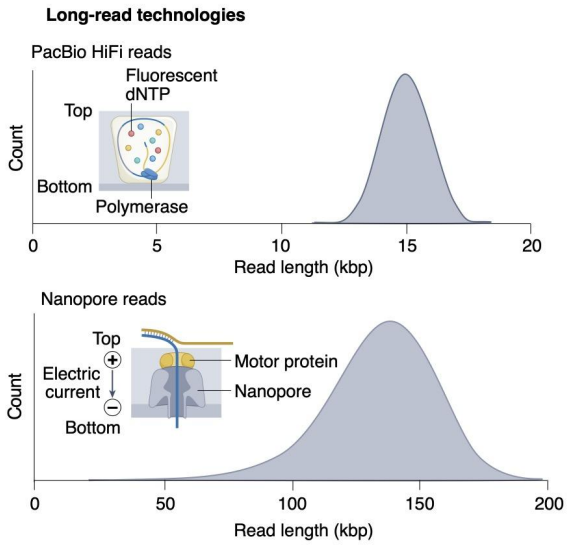
- Regions with repetitive sequences are still difficult to sequence/assemble
  - rDNA
  - Centromere
  - Tandem repeats

- No single sequencing technology can handle these easily

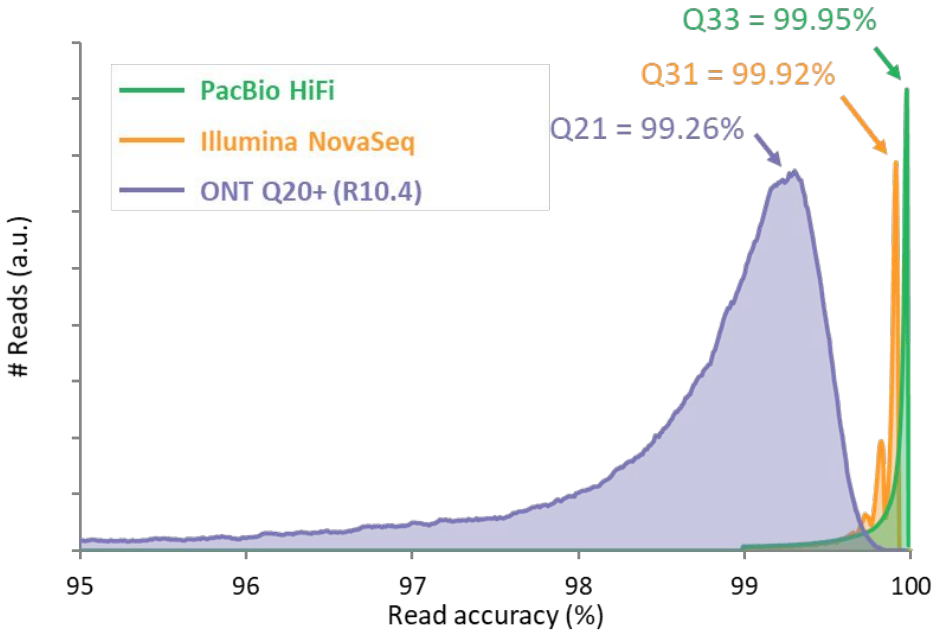


Mao and Zhang *Nature Methods* 2022

Sequencing data

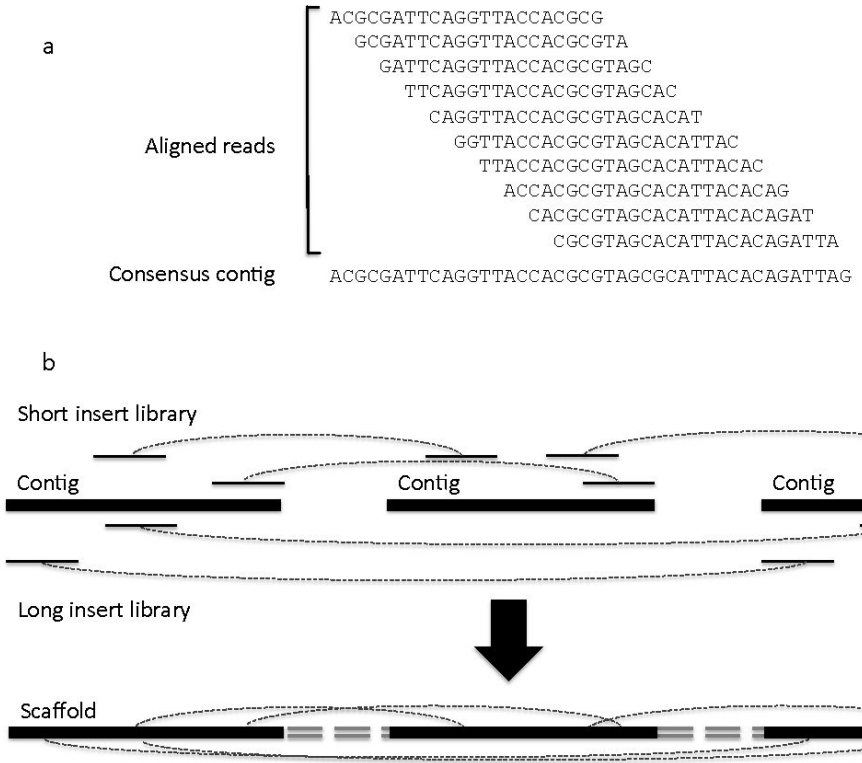


Mao and Zhang *Nature Methods* 2022



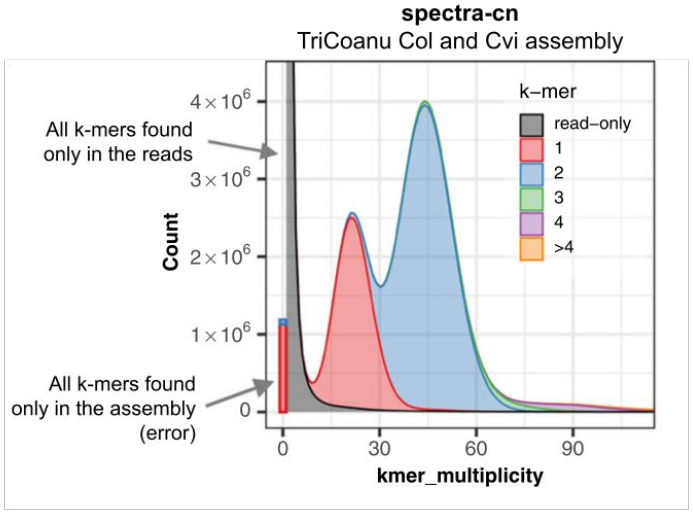
PacBio HiFi: HG003 18 kb library, Sequel II System Chemistry 2.0, precisionFDA Truth Challenge V2
Illumina: HG002 2×150 bp NovaSeq library, precisionFDA Truth Challenge V2
ONT: Q20+ chemistry (R10.4, Kit 12), Oct 2021 GM24385 Dataset Release

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

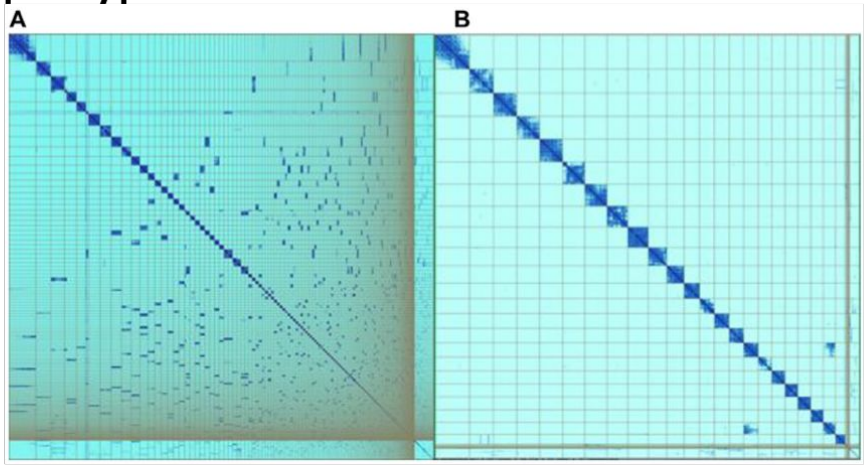


Merqury: Rhie et al. *Genome Biology* 2020

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

EBP standards (cont'd)

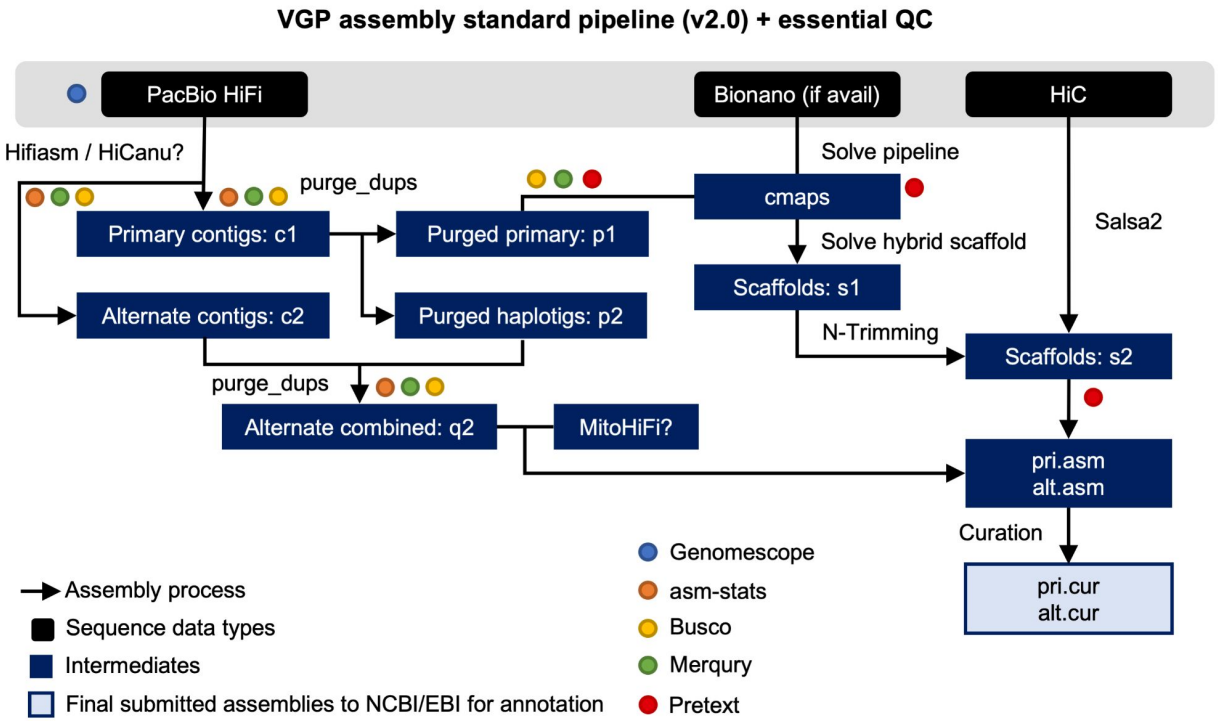
- Separation of target species vs contaminants/symbionts/cobionts
- For diploid species: Identification of primary (haploid) assembly, with secondary assembly with alternate haplotypes
- Organelle genomes
- Manual curation
- Reconciliation with known karyotype



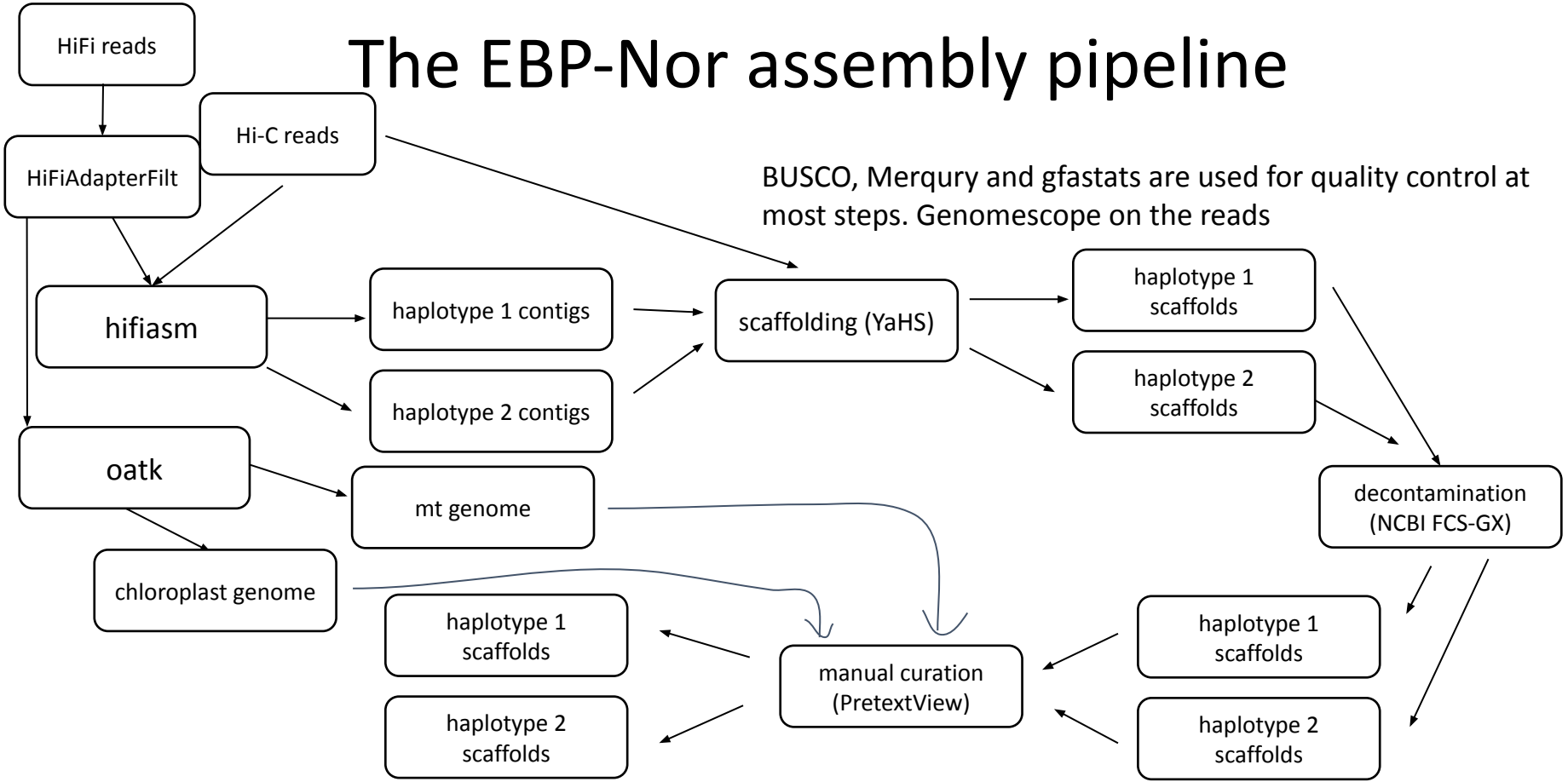
https://www.earthbiogenome.org/assembly-standards

Howe et al. *GigaScience* 2021

Vertebrate Genomes Project pipeline



The EBP-Nor assembly pipeline



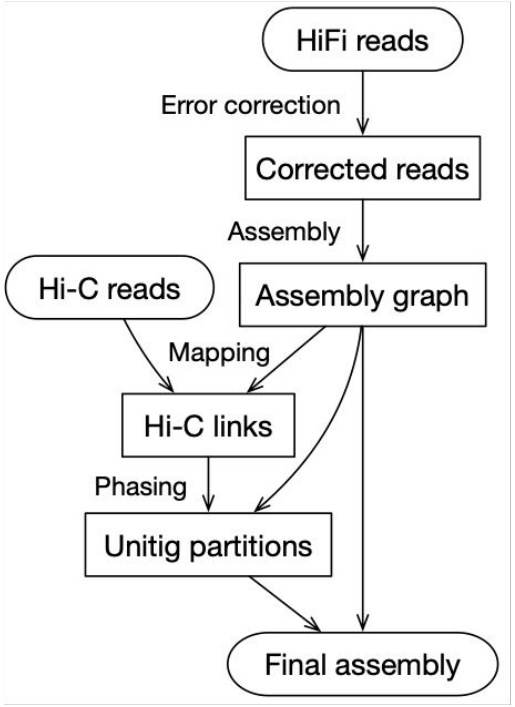
BUSCO, Merqury and gfastats are used for quality control at most steps. Genomescope on the reads

https://github.com/ebp-nor/GenomeAssembly

Combining Hi-C and HiFi in hifiasm

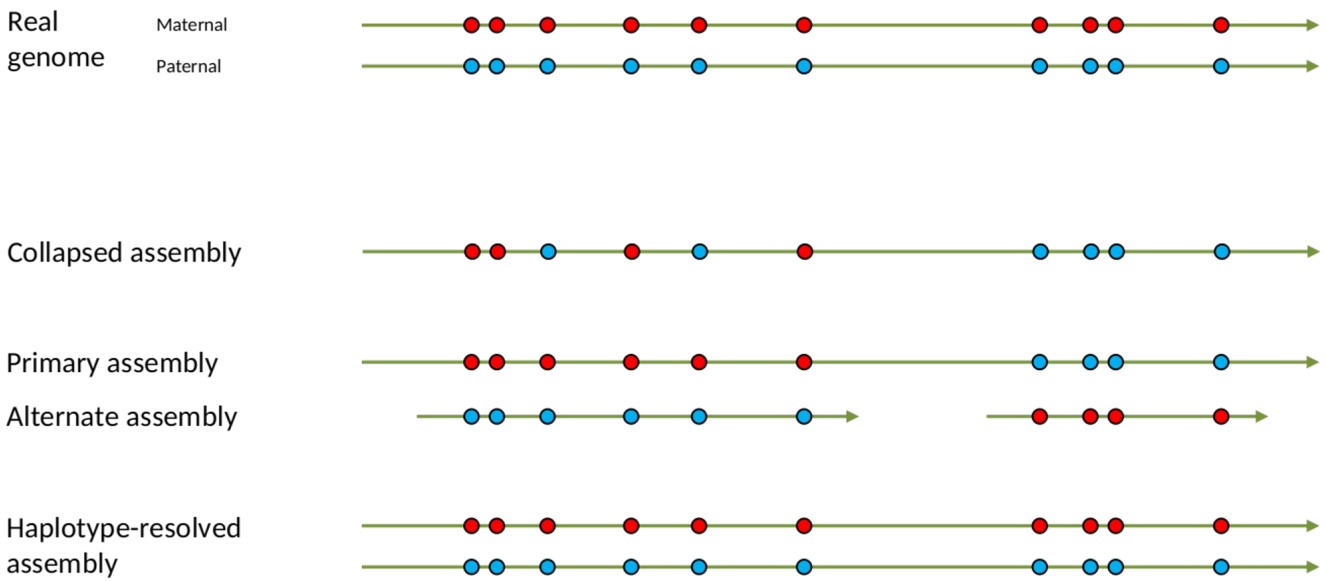
- Powerful combination
- Used by Human Pangenome Project
- Add scaffolding with Hi-C and manual curation -> easy workflow

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



Cheng et al. *Nature Biotechnology* 2022

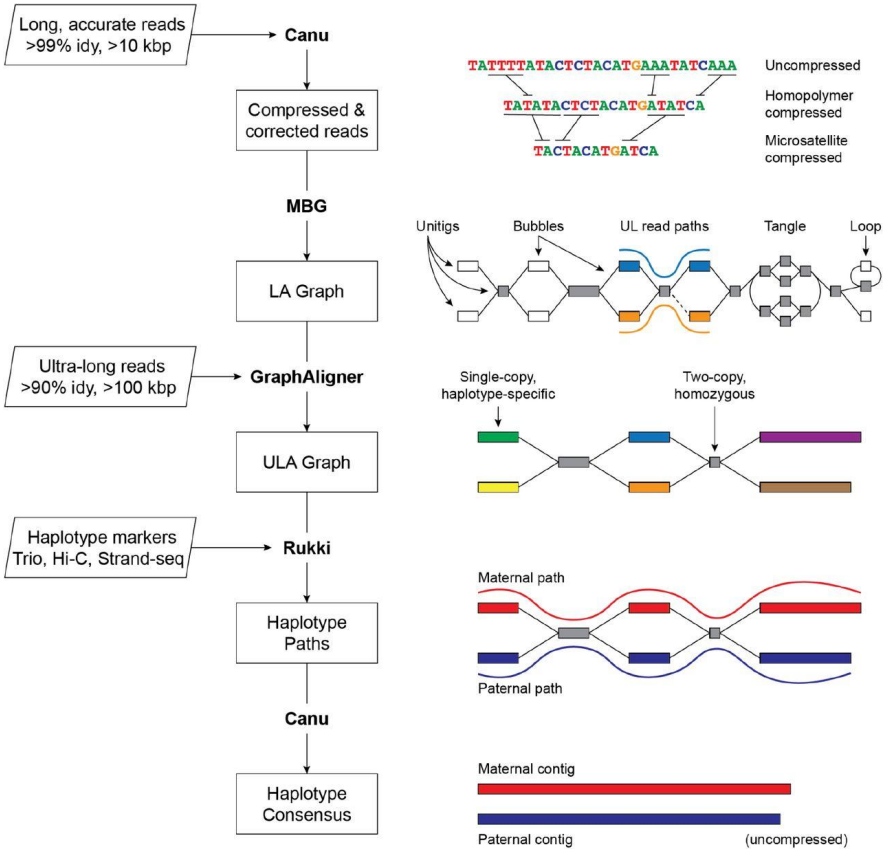
Haplotype-resolved assemblies



Verkko assembler

Combining PacBio HiFi, ONT ultra-long and Hi-C

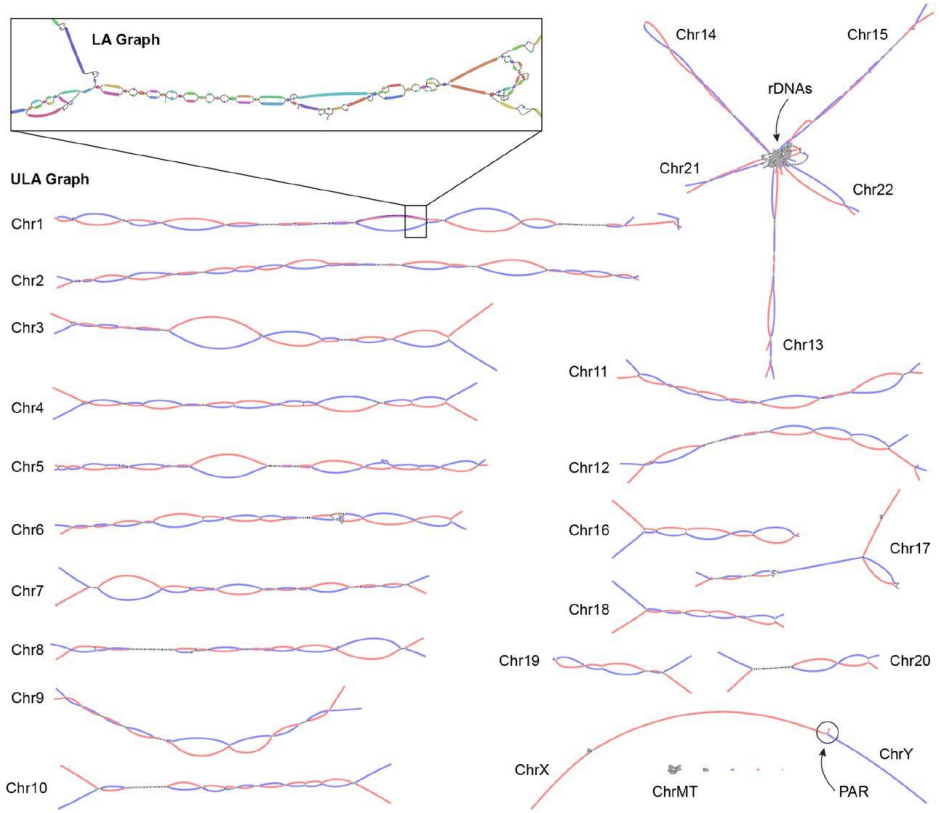
Require a lot of data: ~35x HiFi, ~60x UL, >50x Hi-C (expensive)



Rautiainen et al. *Nature biotechnology* 2023

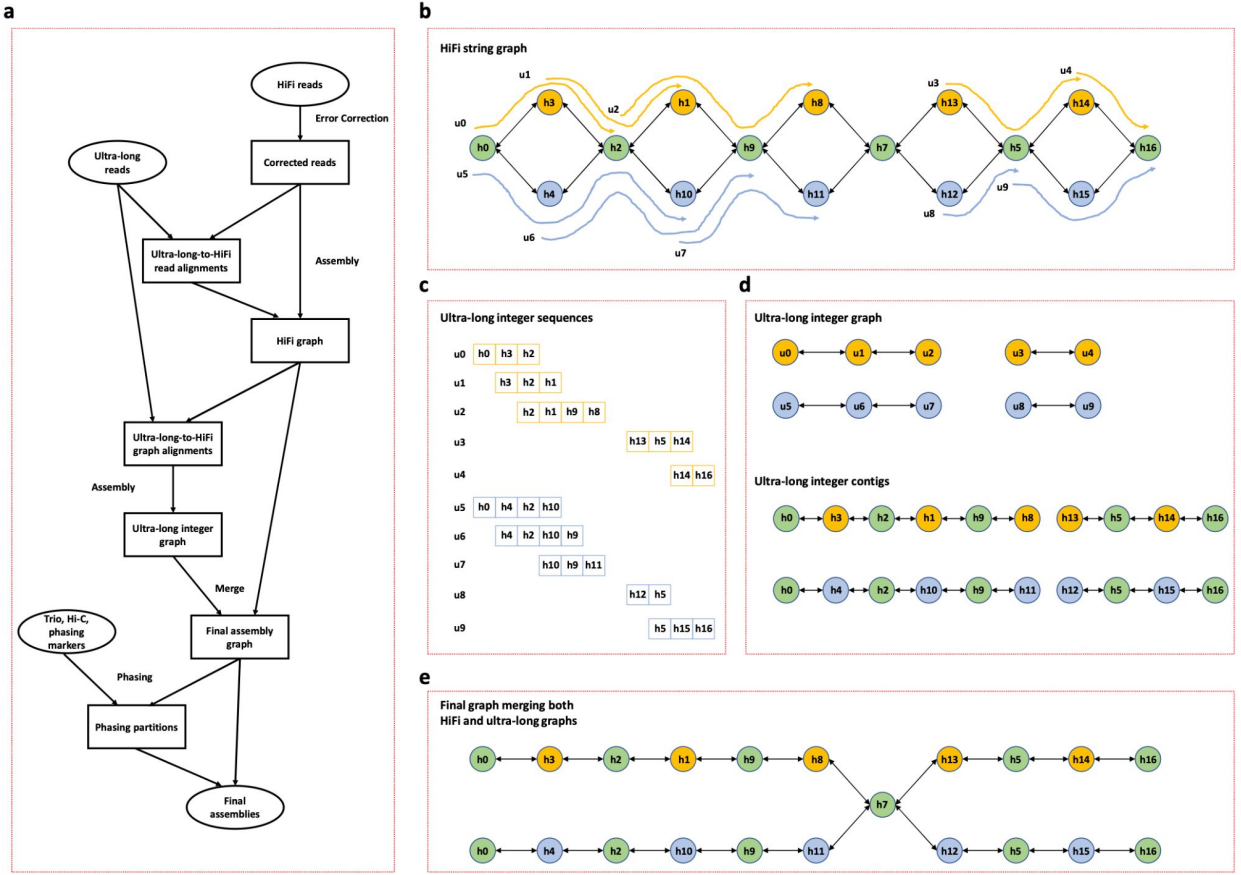
Verkko assembler

Diploid graph human genome from verkko



Rautiainen et al. *Nature biotechnology* 2023

Hifiasm with ONT ultra-long reads

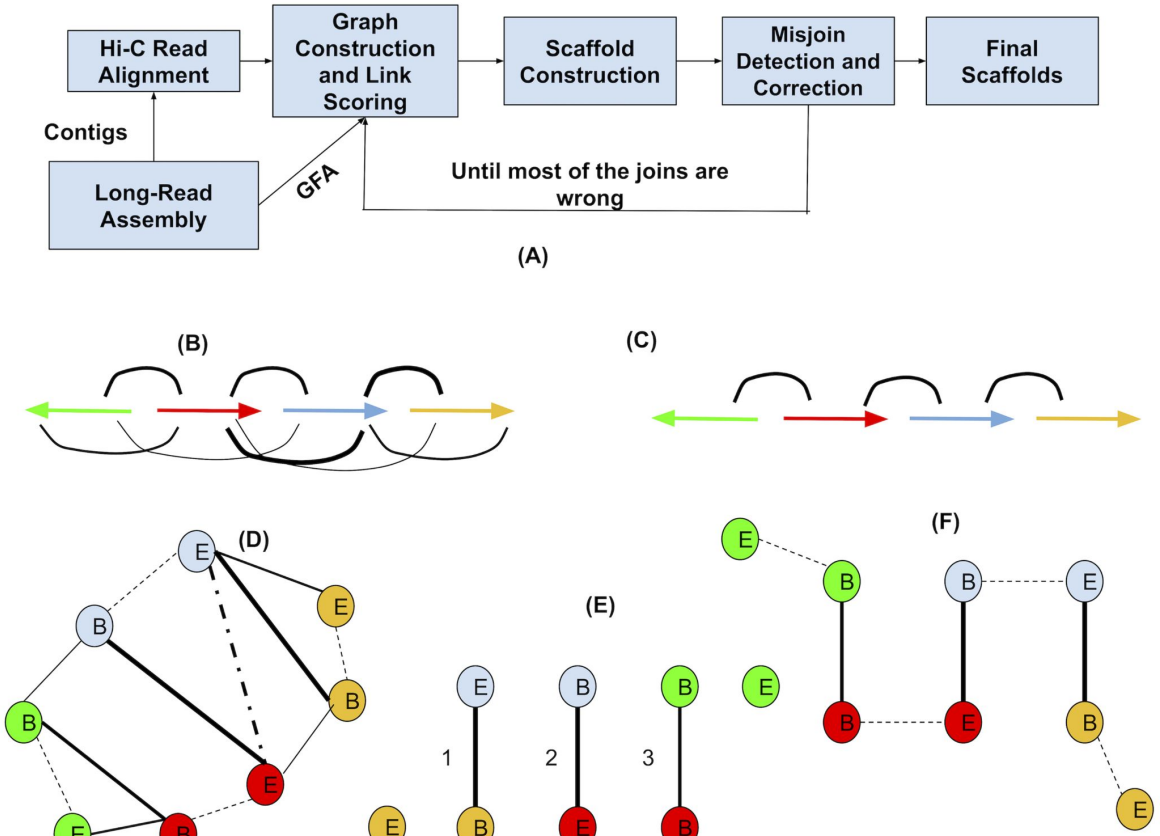


Cheng et al. *arXiv* 2024

Scaffolding with Hi-C

Example from SALSA2.

- Align Hi-C reads to contigs
- Construct graph based on connections between congtigs
- Detect misjoins and correct
- Final scaffolds



Ghurye et al. *PLoS Comp. Biol* 2019

A couple of thoughts

- Too much graph simplification:
  - Assembly creates a graph structure, which is then simplified into FASTA
  - Scaffolding creates a graph structure, which is then simplified into FASTA
- Phasing and scaffolding use the same data twice for quite similar purposes (some kind of clustering in both cases). Combining these processes would be ideal

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 22x HiFi and 50x Hi-C for this workshop

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 sawfly genome
    - GenomeScope2
    - Smudgeplot
    - HiFiAdapterFilt
    - hifiasm
    - YaHS
- 10:45-11:15 Coffee break 12:25-12:30 Summary

12:30-13:30 Lunch

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 sawfly genome
    - FCS-GX
    - The GRIT Rapid Curation suite
    - Working in PretextView
- 15:50-16:00 Summary

Summary - Assembly

- A genome assembly is a model of the genome, it most likely contains 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)

Practicals - Genome assembly

https://github.com/ebp-nor/workshop-2024

Or follow link in e-mail.

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 sawfly genome
    - GenomeScope2
    - Smudgeplot
    - HiFiAdapterFilt
    - hifiasm
    - YaHS
- 10:00-10:15 Coffee break
- 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

14:00-16:00 Decontamination and manual curation

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

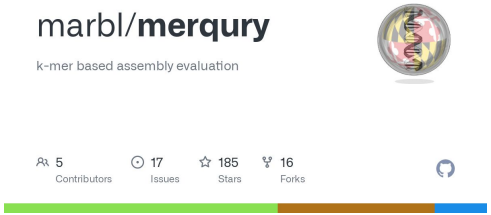
Introduction - Validation

gfastats

BUSCO

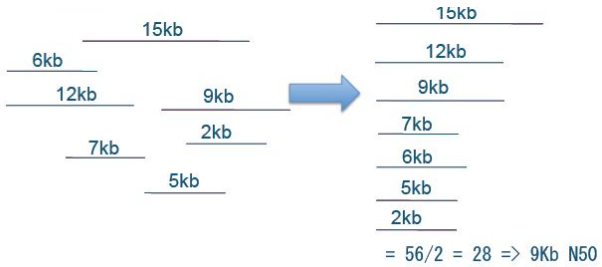
BUSCO

merqury



N50 and assembly statistics

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

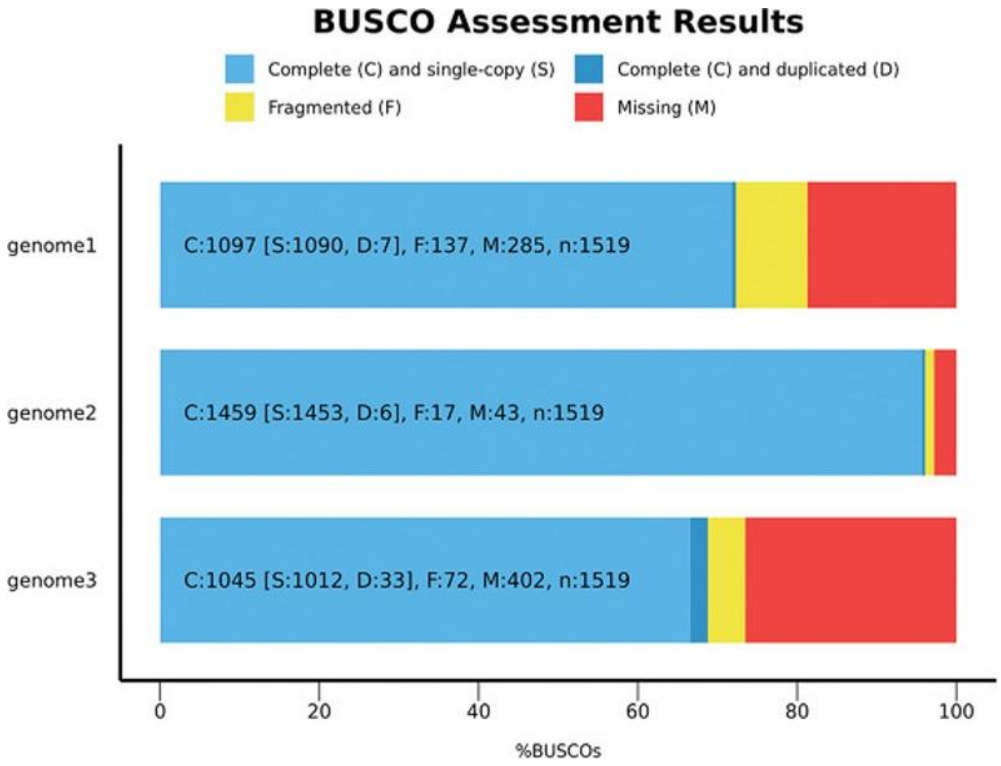


Sum of lengths: 56 kbp

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

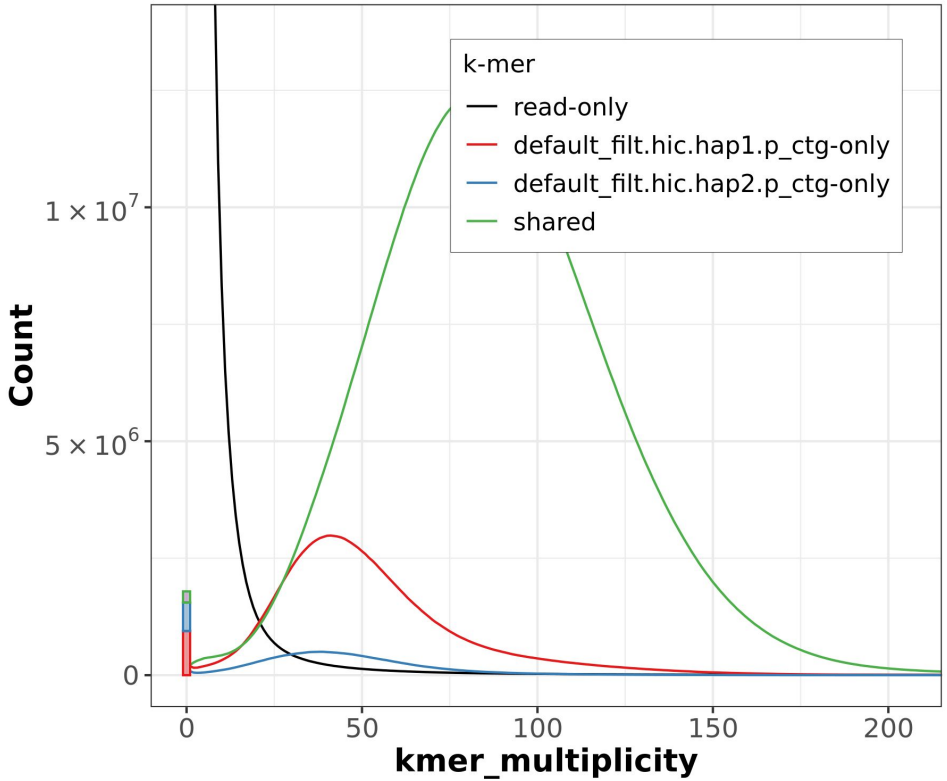
BUSCO

- 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



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 sawfly genome
    - GenomeScope2
    - Smudgeplot
    - HiFiAdapterFilt
    - hifiasm
    - YaHS
- 10:00-10:15 Coffee break
- 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

14:00-16:00 Decontamination and manual curation

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

Summary - Validation

Results from gfastats:

```
+++Assembly summary+++:
# scaffolds: 300
Total scaffold length: 187576100
Average scaffold length: 625253.67
Scaffold N50: 18558710
Scaffold auN: 19380832.84
Scaffold L50: 4
Largest scaffold: 32097451
Smallest scaffold: 1000
# contigs: 432
Total contig length: 187549700
Average contig length: 434142.82
Contig N50: 2100372
Contig auN: 2350525.83
Contig L50: 28
Largest contig: 8068699
Smallest contig: 1000
# gaps in scaffolds: 132
Total gap length in scaffolds: 26400
Average gap length in scaffolds: 200.00
Gap N50 in scaffolds: 200
Gap auN in scaffolds: 200.00
Gap L50 in scaffolds: 66
Largest gap in scaffolds: 200
Smallest gap in scaffolds: 200
Base composition (A:C:G:T): 55461815:38405655:38401625:55280605
GC content %: 40.95
# soft-masked bases: 0
# segments: 432
Total segment length: 187549700
Average segment length: 434142.82
# gaps: 132
# paths: 300
```

Summary - Validation

Results from gfastats:

```
***** Results: *****

C:95.3%[S:94.9%,D:0.4%],F:1.6%,M:3.1%,n:5991
5712    Complete BUSCOs (C)
5687    Complete and single-copy BUSCOs (S)
25      Complete and duplicated BUSCOs (D)
96      Fragmented BUSCOs (F)
183     Missing BUSCOs (M)
5991    Total BUSCO groups searched
```

Results from BUSCO:

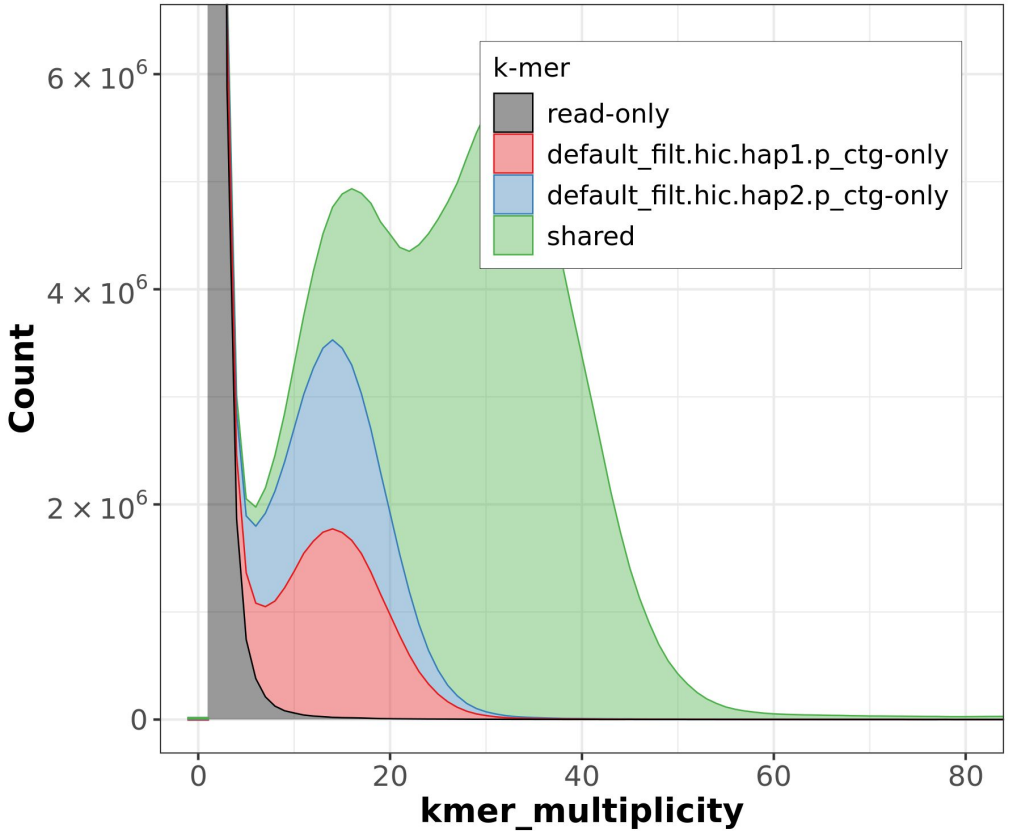
```
Assembly Statistics:
        423     Number of scaffolds
        423     Number of contigs
        187549700       Total length
        0.000%  Percent gaps
        2 MB    Scaffold N50
        2 MB    Contigs N50


Dependencies and versions:
        hmmsearch: 3.1
        bbtools: 39.01
        metaeuk: 6.a5d39d9
        busco: 5.4.5
```

Summary - Validation

Results from gfastats:

Results from BUSCO:

Results from merqury:

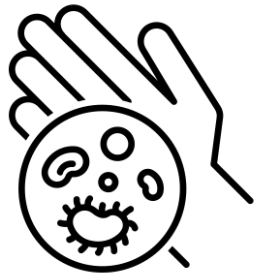


QV-values and K-mer completeness

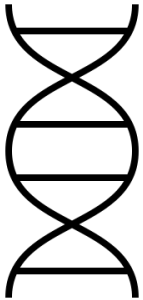
```
(base) 11:45:50-benedga@login-5:/cluster/projects/nn8013k/results/workshop_2023/
default_filt.hic.hap1.p_ctg     7565    183228249       57.0639 1.9661e-06
default_filt.hic.hap2.p_ctg     5999    178770625       57.9643 1.59798e-06
Both    13564   361998874       57.4853 1.7843e-06
(base) 11:45:57-benedga@login-5:/cluster/projects/nn8013k/results/workshop_20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

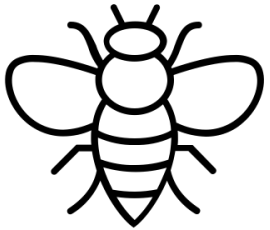
Why do we need to decontaminate our assemblies?



Contamination during handling of samples



Contamination during sequencing



Contamination from the sample itself

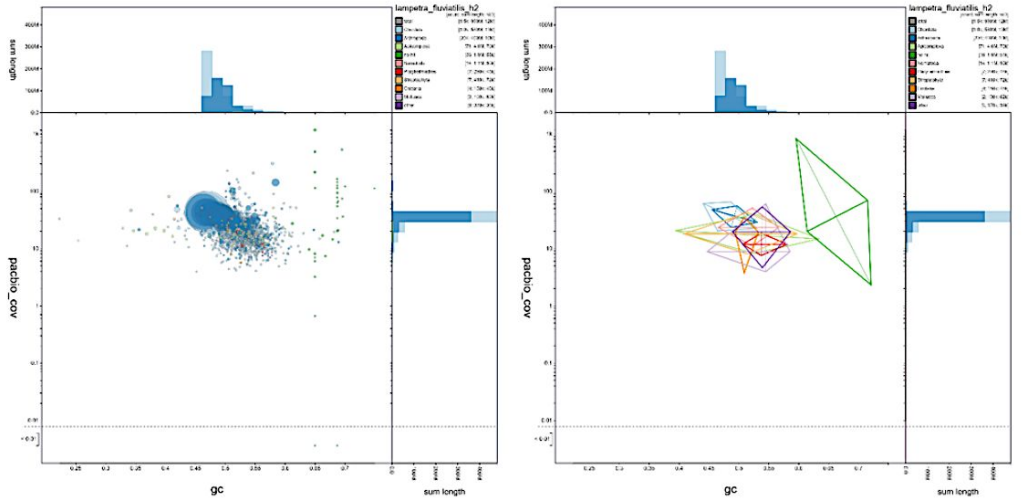
How do we decontaminate our assemblies?

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** ▾



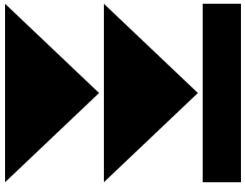
How do we decontaminate our assemblies?

FCS-adaptor and FCS-GX

Adaptor: searches for adaptor sequences

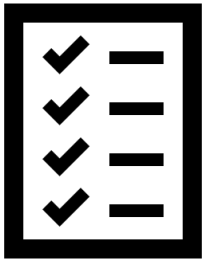
GX: searches for sequences from a wide range of organisms from the RefSeq prokaryotes, RefSeq eukaryotes, ReSeq viruses, RefSeq plasmids and GenBank fungi, nematodes, protists and algae databases

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



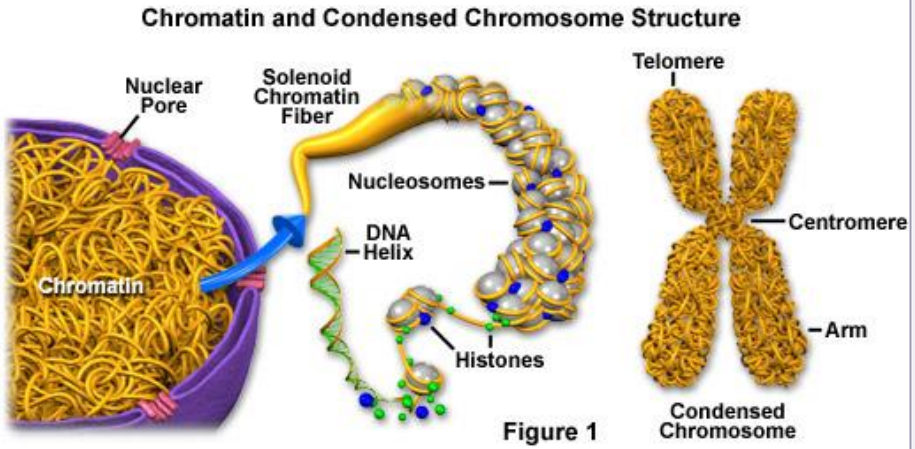
Great results: .txt reports, actionable

What is the end result after running FCS-GX?

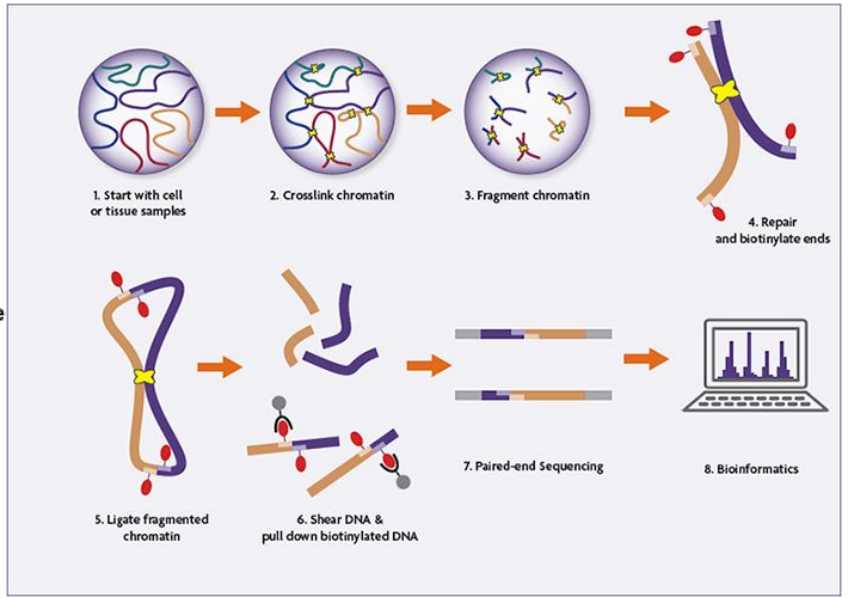
```
##[["FCS genome report", 2, 1], {"git-rev": "0.2.1-14-ga7e5602", "run-date": "Sat Jan 28 10:40:28 2023", "d
vg": 0.948, "asserted-div": "anml:insects", "primary-divs": ["anml:insects"]}}]
#seq_id start_pos       end_pos seq_len action  div     agg_cont_cov    top_tax_name
scaffold_10     1       3321559 3321559 EXCLUDE prok:g-proteobacteria   89      Entomomonas 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       48096   48096   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," *Klimik Dergisi/Klimik Journal*, 26(3), pp. 119–121. Available at: https://doi.org/10.5152/kd.2013.34.

Re-introducing Hi-C sequencing

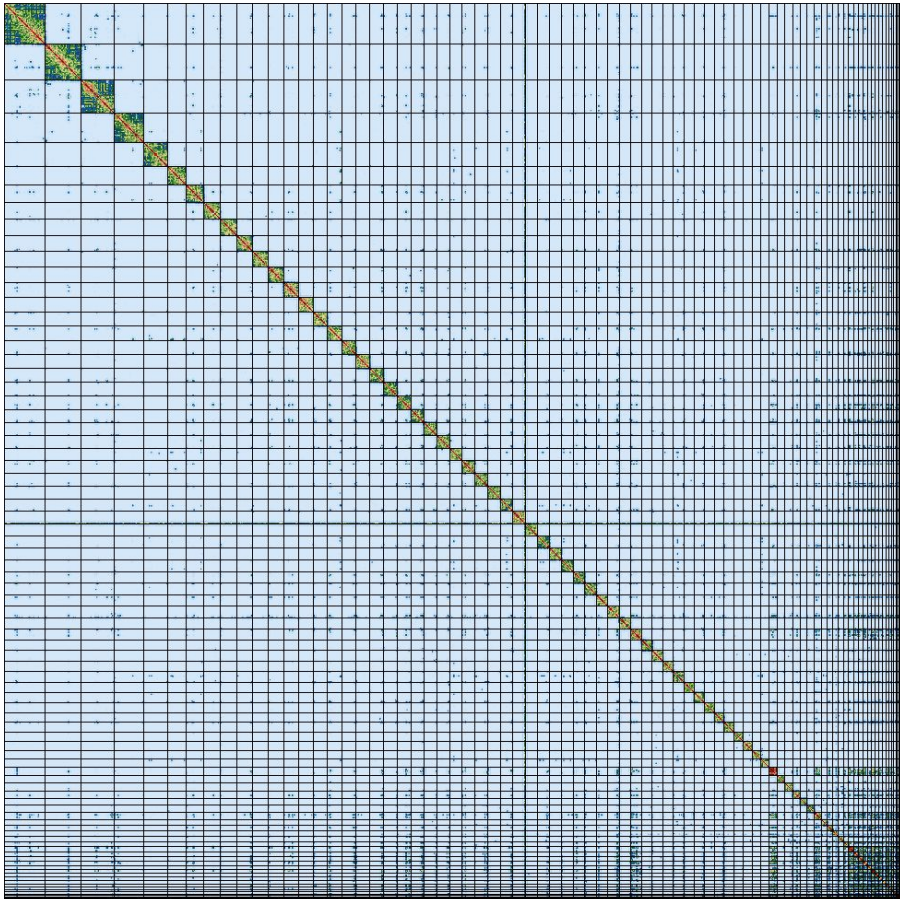


https://micro.magnet.fsu.edu/cells/nucleus/chromatin.html

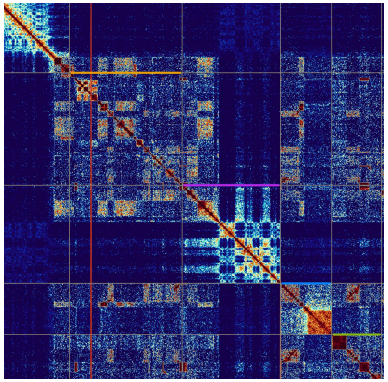


https://www.activemotif.com/blog-hi-c

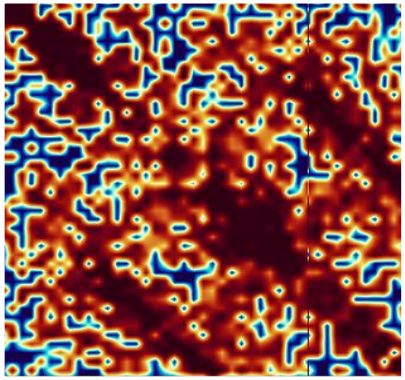
What is a Hi-C contact map?



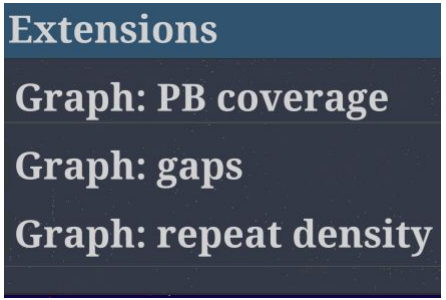
Some situations you may encounter



Ambiguous contact signals



Haplotypic duplications



Lack of data

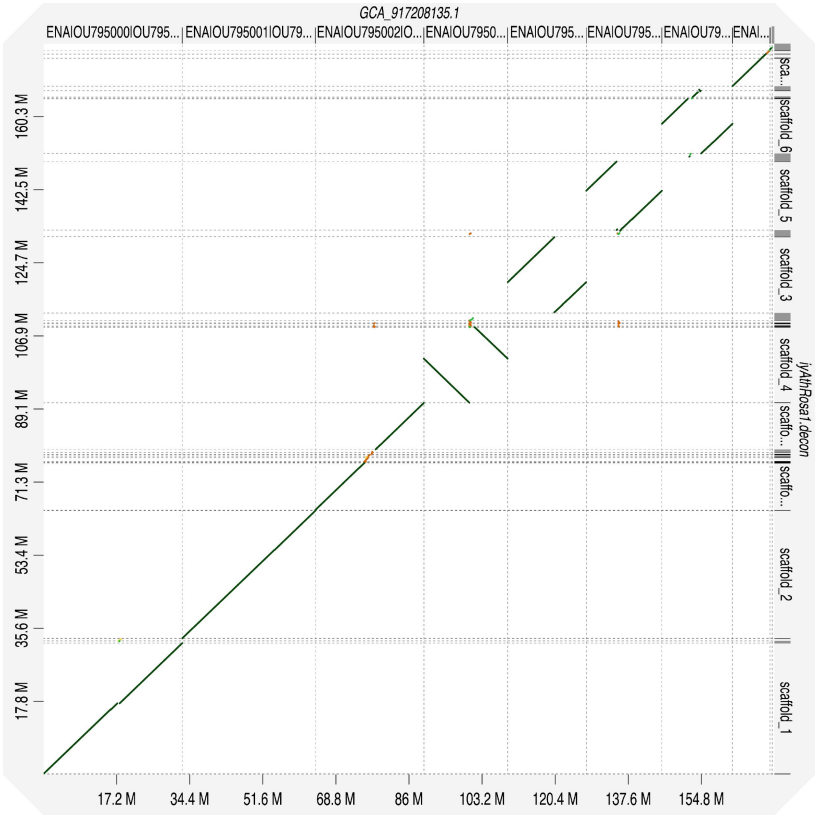
Now it’s time for you to do this!

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**