

Assembling Genomes with the VGP-Galaxy Pipeline : A Hands-on Workshop

Delphine Larivière
September 21st, 2023



The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses



Introducing Galaxy!

The screenshot displays the Galaxy web interface. The top navigation bar includes 'Workflow', 'Visualize', 'Shared Data', 'Admin', 'Help', 'User', and a notification bell. The left sidebar contains a 'Tools' menu with categories like 'GENERAL TEXT TOOLS', 'GENOMIC FILE MANIPULATION', 'FASTQ Quality Control', 'SAM/BAM', 'BED', 'VCF/BCF', 'Nanopore', 'Convert Formats', 'Lift-Over', 'COMMON GENOMICS TOOLS', 'Interactive tools', 'Operate on Genomic Intervals', 'Fetch Sequences/Alignments', 'GENOMICS ANALYSIS', 'Assembly', 'Annotation', 'Mapping', and 'Variant Calling'.

General Statistics

Showing 63 rows and 6 columns.

Sample Name	% Duplication	% > Q30	Mb Q30 bases	GC content	% PF	% Adapter
50_C4R1Z5_S80_L001001_fastq	12.8%	83.3%	0.0	57.5%	0.6%	49.5%
51_C4R1Z15_S51_L001001_fastq	7.5%	82.9%	0.0	57.1%	0.5%	49.5%
52_C4R1Z40_S52_L001001_fastq	5.5%	80.9%	0.0	57.3%	0.5%	49.5%
53_C4R1Z100_S53_L001001_fastq	12.0%	89.8%	2.5	57.6%	93.3%	0.0%
54_C4R2Z0_S54_L001001_fastq	13.6%	81.8%	0.0	57.5%	0.5%	49.6%
55_C4R2Z5_S55_L001001_fastq	11.6%	79.9%	0.0	56.6%	0.5%	49.4%
56_C4R2Z15_S56_L001001_fastq	12.3%	89.9%	1.6	57.7%	93.8%	0.0%
57_C4R2Z40_S57_L001001_fastq	8.9%	81.5%	0.0	57.4%	0.5%	49.5%
58_C4R2Z100_S58_L001001_fastq	10.3%	83.4%	0.0	57.4%	0.4%	49.5%
59_C4R3Z0_S59_L001001_fastq	8.6%	80.6%	0.0	57.9%	0.6%	49.5%
5_C1R1Z5_S5_L001001_fastq	9.5%	79.5%	0.0	57.5%	0.6%	49.5%
60_C4R3Z5_S60_L001001_fastq	12.4%	82.9%	0.0	57.3%	0.6%	49.4%
61_C4R3Z15_S61_L001001_fastq	10.3%	82.7%	0.0	57.5%	0.5%	49.6%
62_C4R3Z40_S62_L001001_fastq	7.2%	80.0%	0.0	56.8%	0.4%	49.5%
63_C4R3Z100_S63_L001001_fastq	13.1%	82.1%	0.0	57.4%	0.5%	49.5%

fastp

fastp: An ultra-fast all-in-one FASTQ preprocessor (QC, adapters, trimming, filtering, splitting...)

Filtered Reads

Filtering statistics of sampled reads.

Number of Reads | Percentages

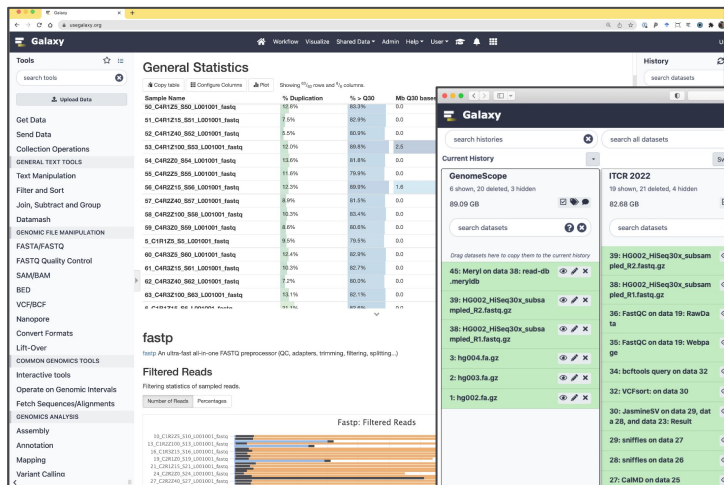
History

search datasets

An example 1
17 shown, 259 deleted, 1074 hidden
172.83 MB

- 712: Realigned reads a list with 63 items
- 712: Filter SAM or BAM, out SAM or BAM on collection 6 bam a list with 63 items
- 648: Map with BWA-MEM on collection 128 (mapped reads in BAM format) a list with 63 items
- 647: pOUT18.4.fa
- 388: pOUT18.4.gb
- 384: MultiQC on data 382, data 378, and others: Webpage
- 383: MultiQC on data 382, data 378, and others: Stats a list with 3 items
- 130: fastp on collection 127 ON report a list with 63 items
- 129: fastp on collection 127 ML report a list with 63 items
- 128: fastp on collection 127 red-end output a list of pairs with 63 items
- 127: pilot a list of pairs with 63 items

Accessible, Reproducible and Collaborative



Galaxy General Statistics

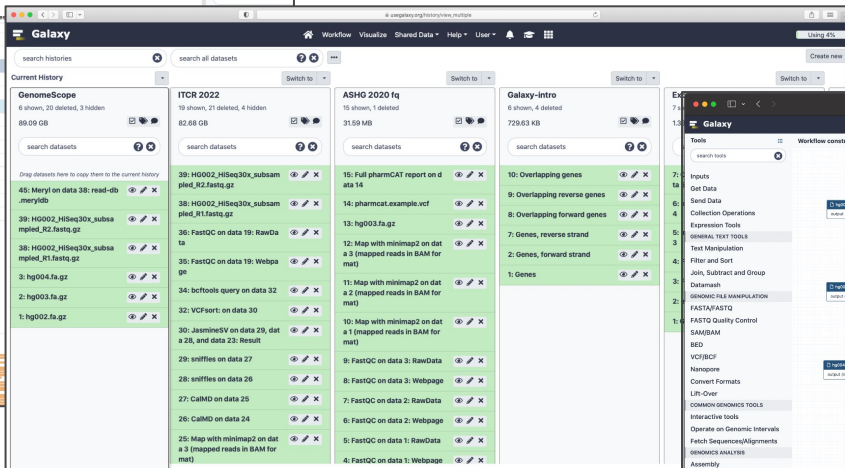
Sample Name	% Duplication	% > Q30	Meq G30 base
92_C4H1Z5_S56_L001001_fastq	15.9%	83.3%	0.0
91_C4H1Z5_S51_L001001_fastq	17.9%	82.9%	0.0
92_C4H1Z40_S52_L001001_fastq	1.5%	82.3%	0.0
93_C4H1Z100_S53_L001001_fastq	12.0%	86.8%	2.2
94_C4H1Z22_S54_L001001_fastq	13.6%	81.4%	0.0
95_C4H1Z215_S56_L001001_fastq	11.6%	79.9%	0.0
96_C4H1Z215_S56_L001001_fastq	12.3%	83.3%	1.6
97_C4H1Z246_S57_L001001_fastq	3.9%	81.5%	0.0
98_C4H1Z100_S58_L001001_fastq	19.3%	82.4%	0.0
99_C4H1Z22_S59_L001001_fastq	3.6%	80.6%	0.0
5_C1H1Z23_S5_L001001_fastq	3.6%	79.1%	0.0
60_C4H1Z22_S60_L001001_fastq	18.4%	82.3%	0.0
61_C4H1Z240_S62_L001001_fastq	19.3%	82.7%	0.0
62_C4H1Z100_S63_L001001_fastq	18.1%	82.1%	0.0
6_C4H1Z23_S64_L001001_fastq	16.1%	80.4%	0.0

fastp
An ultra-fast all-in-one FASTQ preprocessor (QC, adapters, trimming, rereading, splitting...)

Filtered Reads
Filtering statistics of sampled reads.

Number of Reads	Percentage
1: hg002.fa.gz	
2: hg003.fa.gz	
3: hg004.fa.gz	

One Analysis

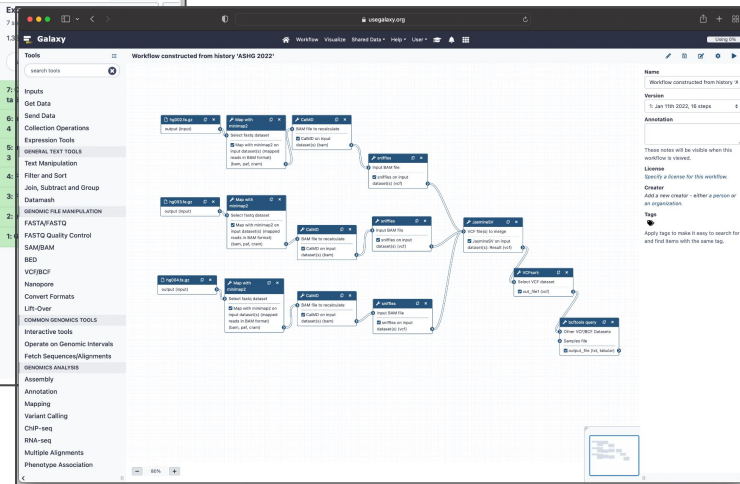


Galaxy Workflow

GenomeScope
ITCR 2022
ASHG 2020 fq
Galaxy-intro

- 39: HGO02_HiSeq30x_subsam pled_R2.fastq.gz
- 16: Full pharmacat report on ata 14
- 10: Overlapping genes
- 9: Overlapping reverse genes
- 8: Overlapping forward genes
- 7: Genes, reverse strand
- 2: Genes, forward strand
- 1: Genes

Many Analyses



Galaxy Workflow constructed from history 'ASHG 2022'

Tools: Get Data, Send Data, Collection Operations, Expression Tools, GENERAL TEXT TOOLS, Text Manipulation, Filter and Sort, Job, SubJob, and Group, Datamash, GENOMIC FILE MANIPULATION, FASTA/FASTQ, FASTQ Quality Control, SAM/BAM, BED, VCF/BCF, Nanopore, Convert Formats, COMMON GENOMICS TOOLS, LRF-Over, Interactive tools, Operate on Genomic Intervals, Fetch Sequences/Alignments, GENOMICS ANALYSIS, Assembly, Annotation, Mapping, Variant Calling, ChIP-seq, RNA-seq, Multiple Alignments, Phenotype Association.

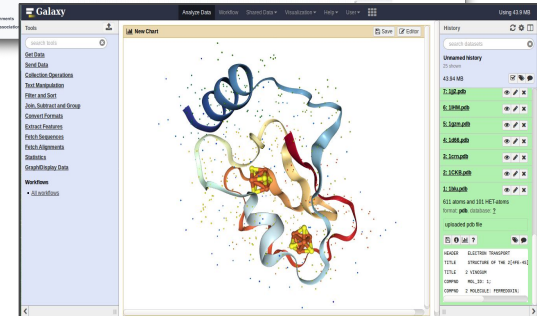
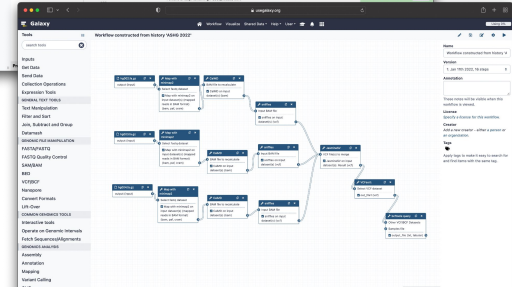
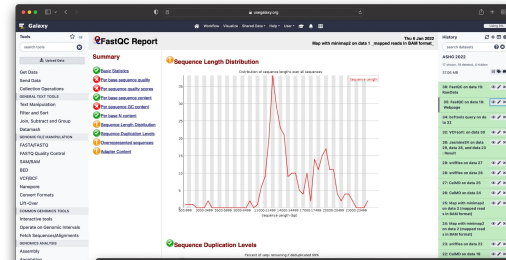
Collaborative Data & Workflows

(The Galaxy Community, NAR, 2022)

So ... what is Galaxy?

- GUI for interactively running tools
- Toolshed with 1,000s of tools ready to run
- Terabytes of the latest, curated reference data
- Full featured workflow functionality
- Graphical interface for handling >1,000 samples
- Run Jupyter, RStudio, & Interactive Visualizations
- Extensive training tutorials and infrastructure
- Large international community of users and developers

All of this can be used on free and powerful public high performance computational infrastructure ... or on your institutional cluster ... or used on the cloud... or your own laptop... or a Raspberry Pi!



The universe has many Galaxies



usegalaxy.*: the big three

 **Galaxy**
Main
usegalaxy.org

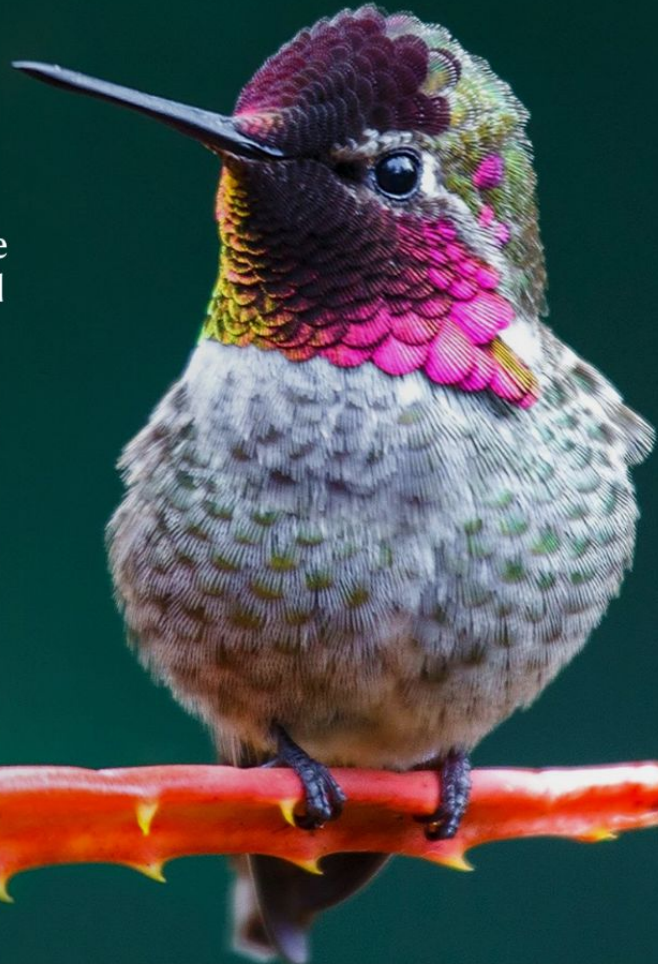
 **Galaxy**
EUROPE
usegalaxy.eu

 **Galaxy**
AUSTRALIA
usegalaxy.org.au



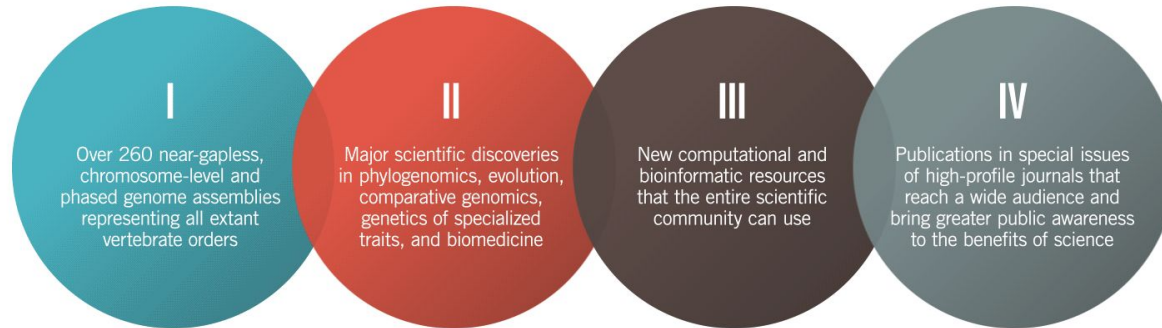
The Vertebrate Genomes Project

The Vertebrate Genomes Project (VGP) is a collaborative effort to generate high quality reference genomes for all vertebrate species.



The Vertebrate Genome Project

- Biodiversity under threat
 - Description of species in an effort to preserve species and ecosystems
- Generation of near error free reference genomes
 - Span across all vertebrate families



Genomeark

Data repository :

- Earth BioGenome Project
- Vertebrate Genomes Project
- Telomere-to-Telomere Consortium

by Project and
Completion

	All Species	Curated Assemblies	Draft Assemblies	Raw Data Only
All	520 species	297 species	60 species	163 species
VGP	388 species	257 species	30 species	101 species
T2T	7 species	none	6 species	1 species
ERGA	none	none	none	none
Bat1K	1 species	1 species	none	none

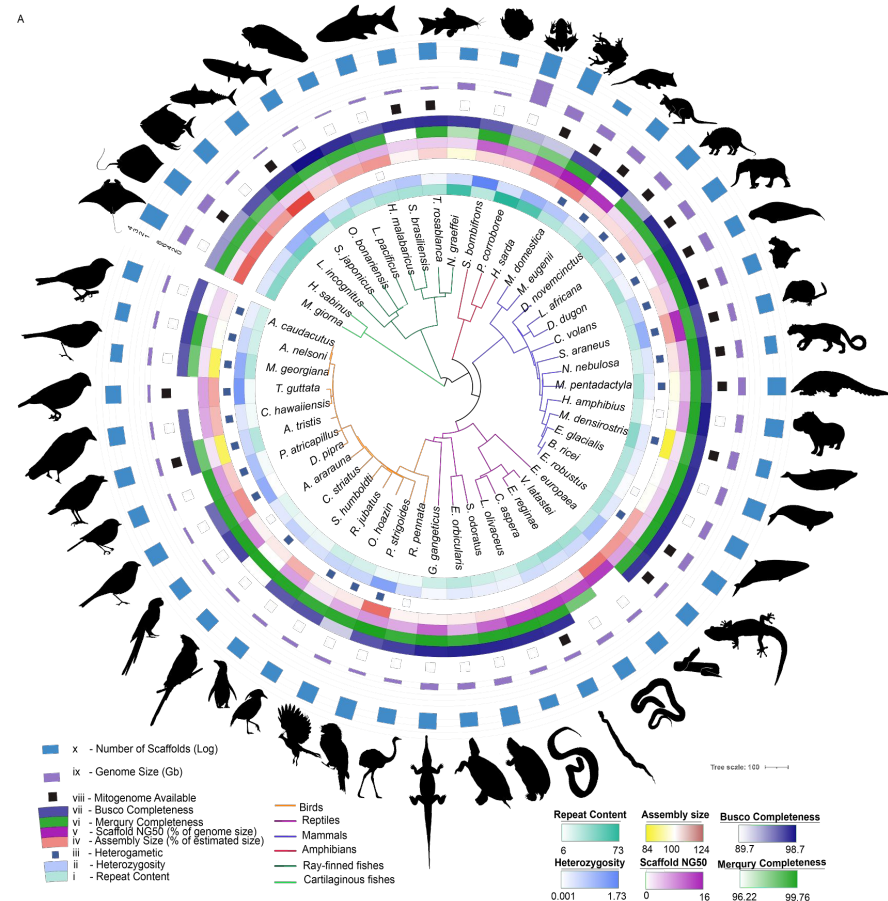
Number of species at each level of completion.

<https://genomeark.github.io/>

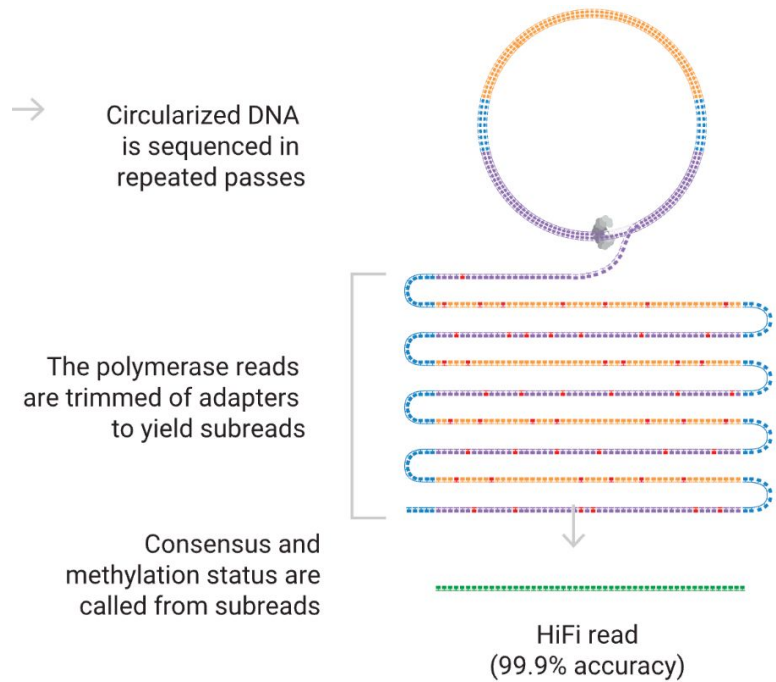


VGP assembly Pipeline v2

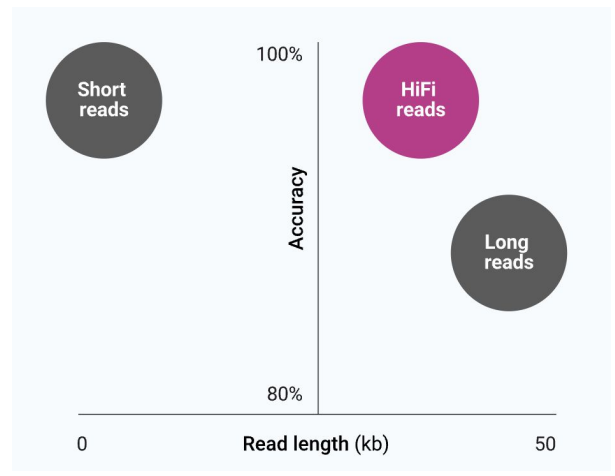
- Developed in Galaxy
- Available on global instances :
 - .eu
 - .org
 - .org.au (soon)
- Dozens of genomes assembled
 - Hundreds to thousands planned in the coming year
- Technologies :
 - PacBio HiFi
 - Bionano Cmap
 - Arima HiC



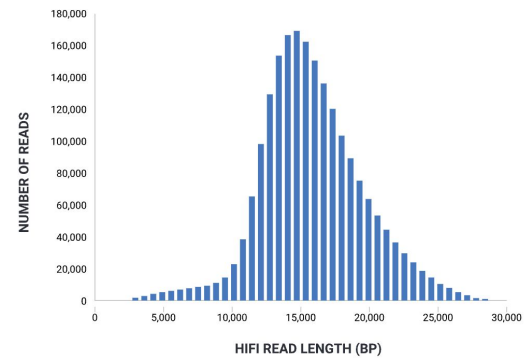
Data - PacBio HiFi



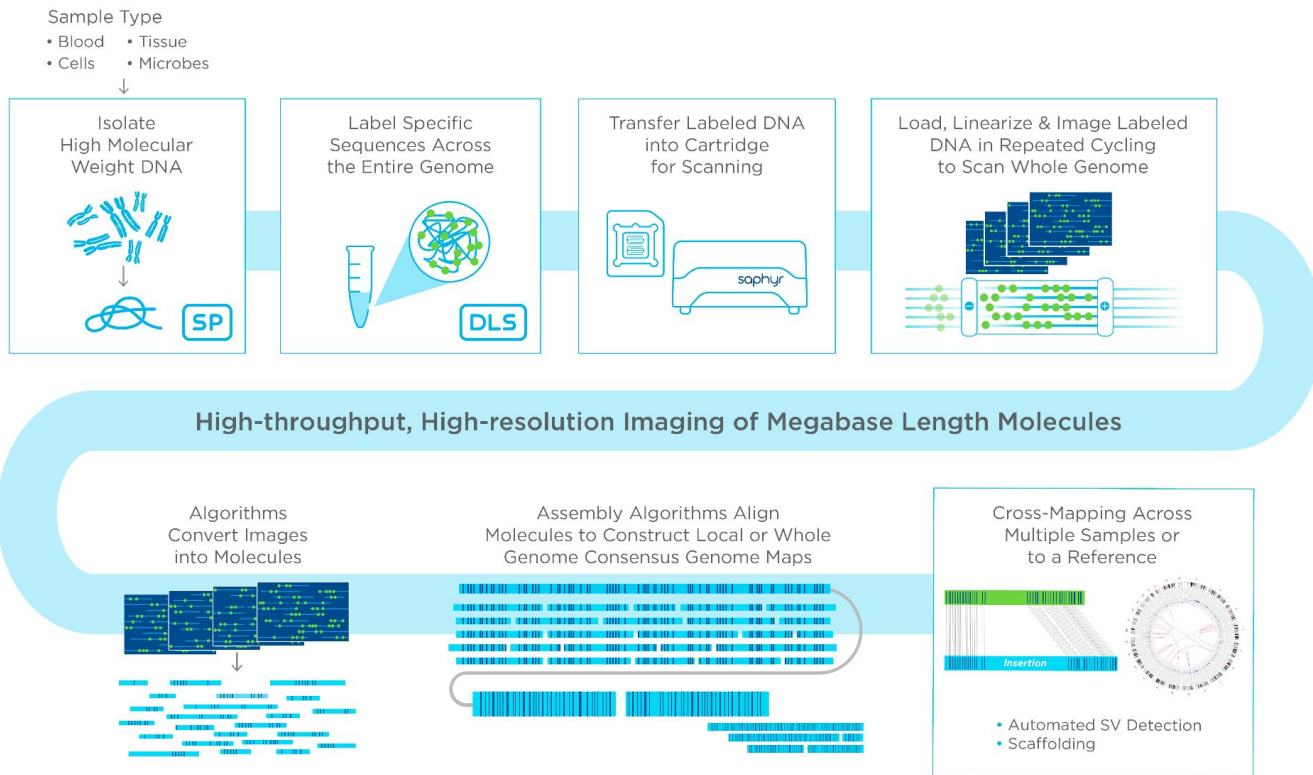
<https://www.pacb.com/technology/hifi-sequencing/>



<https://www.pacb.com/technology/hifi-sequencing/how-it-works/>



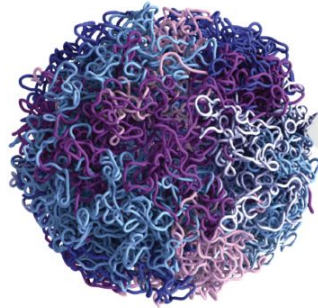
Data - Bionano Optical Mapping



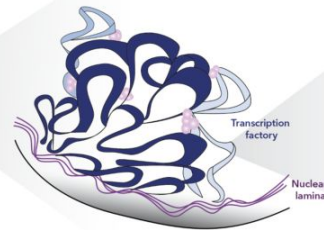
<https://bionanogenomics.com/technology/platform-technology/>

Data - Arima HiC

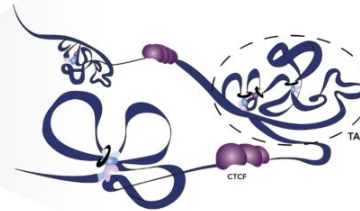
In the nucleus chromosomes are organized into chromosome territories



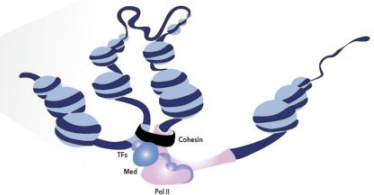
Chromosomes are divided into cell-specific A/B compartments



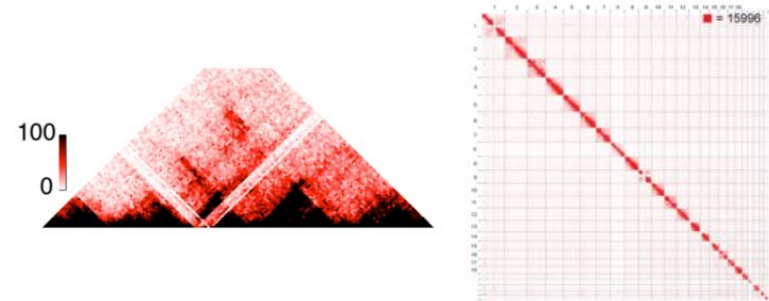
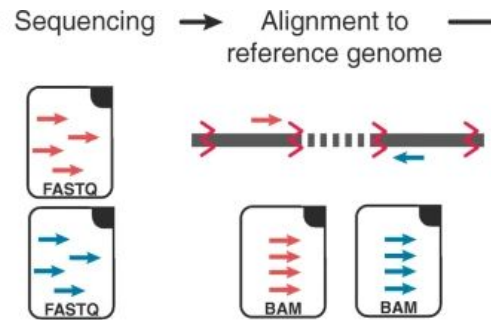
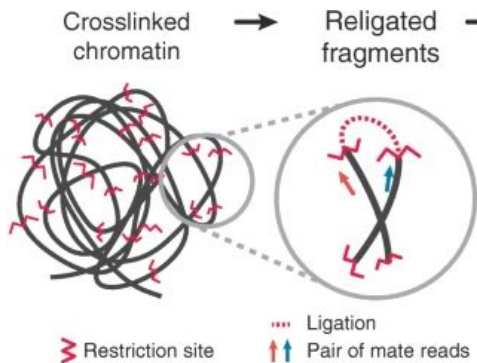
Compartments are organized into topologically associated domains (TADs)



Within TADs, DNA is looped together with the assistance of architectural proteins and histones

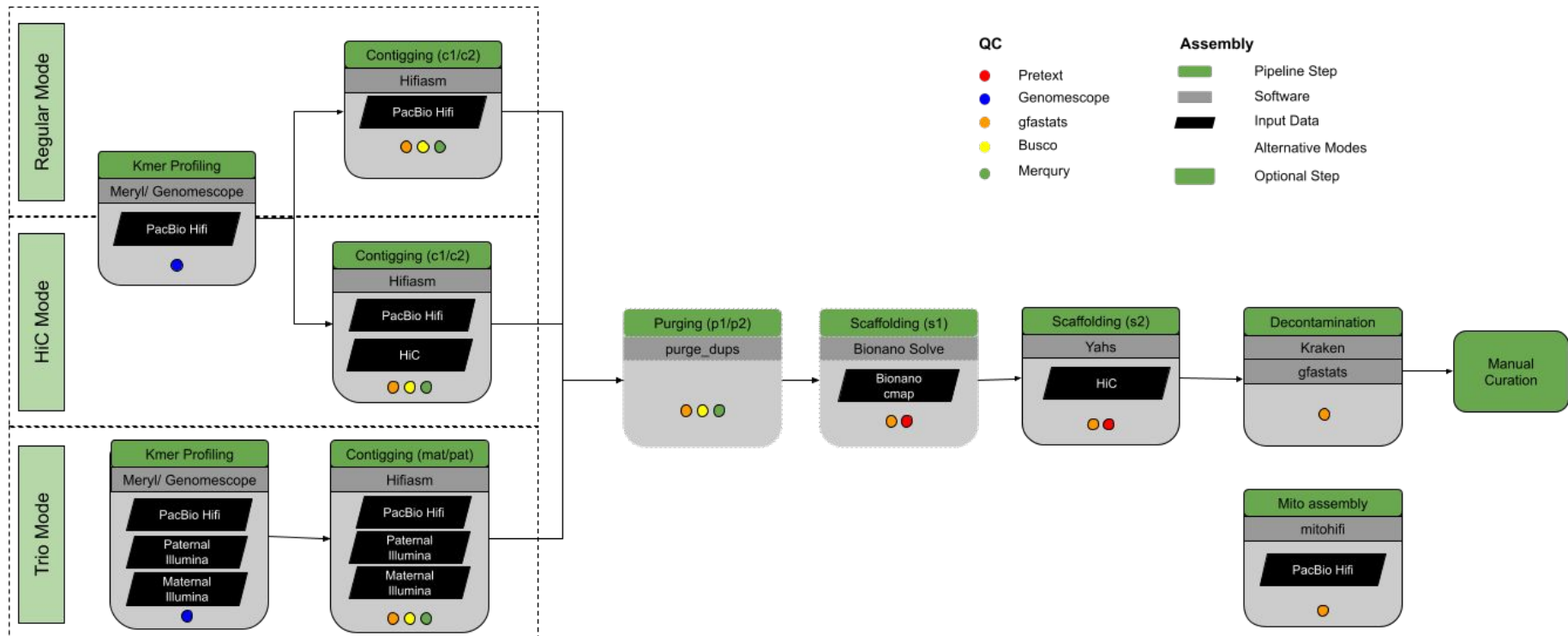


<https://arimagenomix.com/products/genome-wide-hic/>



Pal, Koustav, Mattia Forcato, and Francesco Ferrari. "Hi-C analysis: from data to biological integration." *Biophysical reviews* 11.1 (2019): 67-78.

Assembly Pipeline Overview



training.galaxyproject.org

Galaxy Training! Contributors Languages Help Extras Search Tutorials

Welcome to Galaxy Training!

Collection of tutorials developed and maintained by the worldwide Galaxy community

Galaxy for Scientists

Topic	Tutorials
Introduction to Galaxy Analyses	11
Assembly	15
Climate	6
Computational chemistry	8
Ecology	13
Epigenetics	7
Genome Annotation	14
Imaging	4
Metabolomics	6
Metagenomics	9
Proteomics	26
Sequence analysis	3
Single Cell	15

Welcome to the GTN!

Find out more about Galaxy Training Network

Video created by Geert Bonamie.

The latest GTN news

Read about new tutorials, features, events and more!

- Nov 18, 2022
New Topic: Single Cell Analysis!
- Sep 11, 2022
New Tutorial: Data Manipulation
- Jun 2, 2022
New Tutorial: Workflows

OPEN CHAT

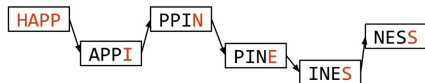
Contigging

TTGACTTACCGA **Read**

TTGACTTACC
TGACTTACCG
GACTTACCGA } **k-mers for k=12**

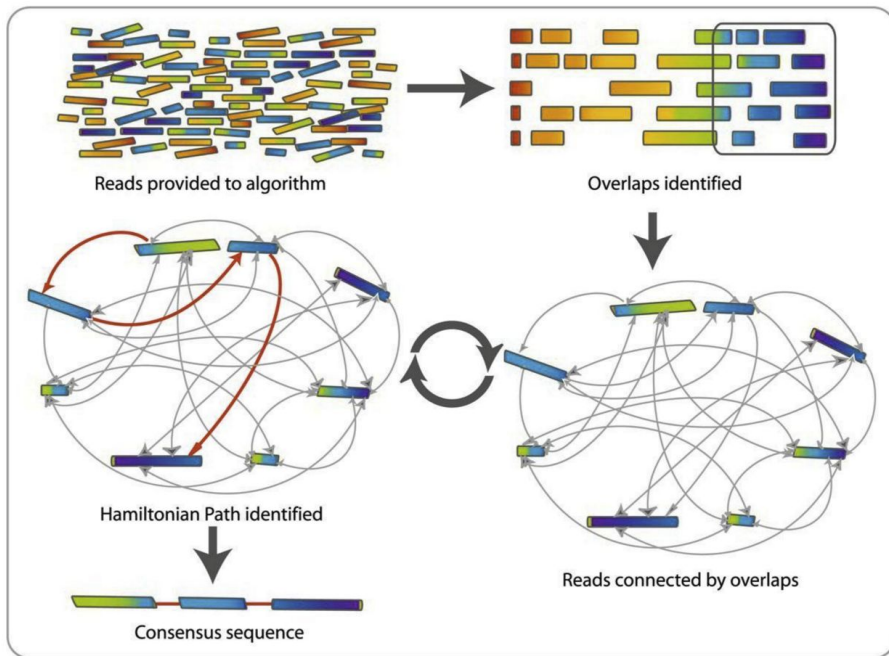
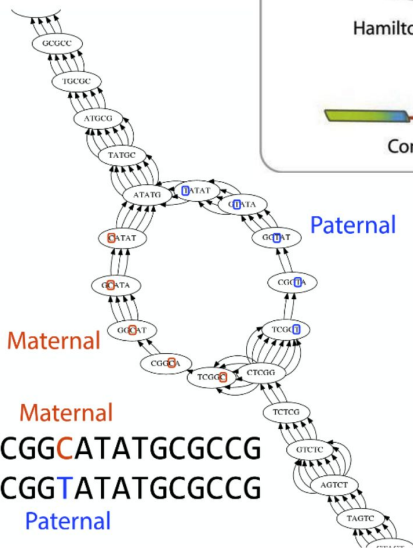
TTGAC
TGACA
GACAC
ACACT
CACTT
ACTTA
CTTAC
TTACC
TACCG
ACCGA } **k-mers for k=5**

k = 4 k-mers:
HAPP APPI
PINE PPIN
INES NESS



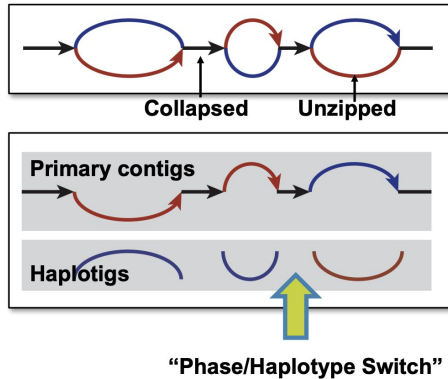
HAPPINESS

Maternal
GTAGTCTCGGCATATGCGCCG
Paternal
GTAGTCTCGGTATATGCGCCG



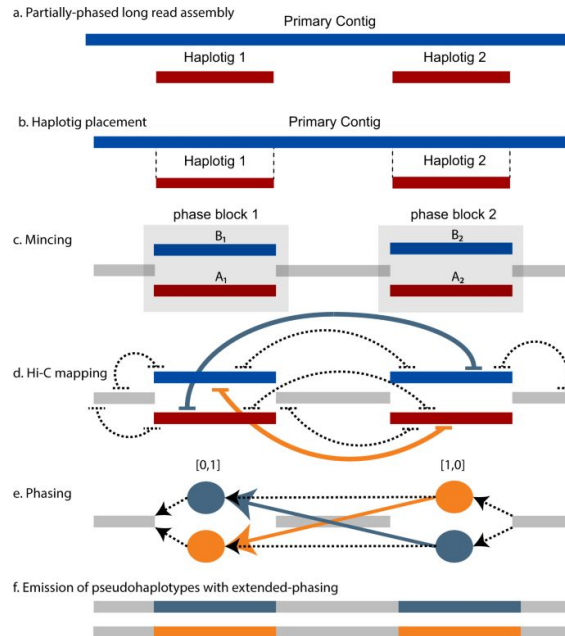
Phased Assemblies

Simple Phasing



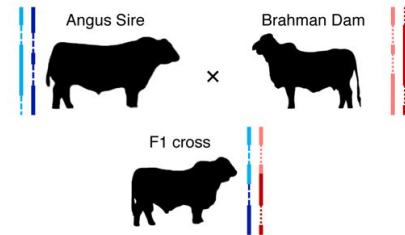
Chin, C.S. et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*. 13(12), 1050.

Hi-C-based Phasing



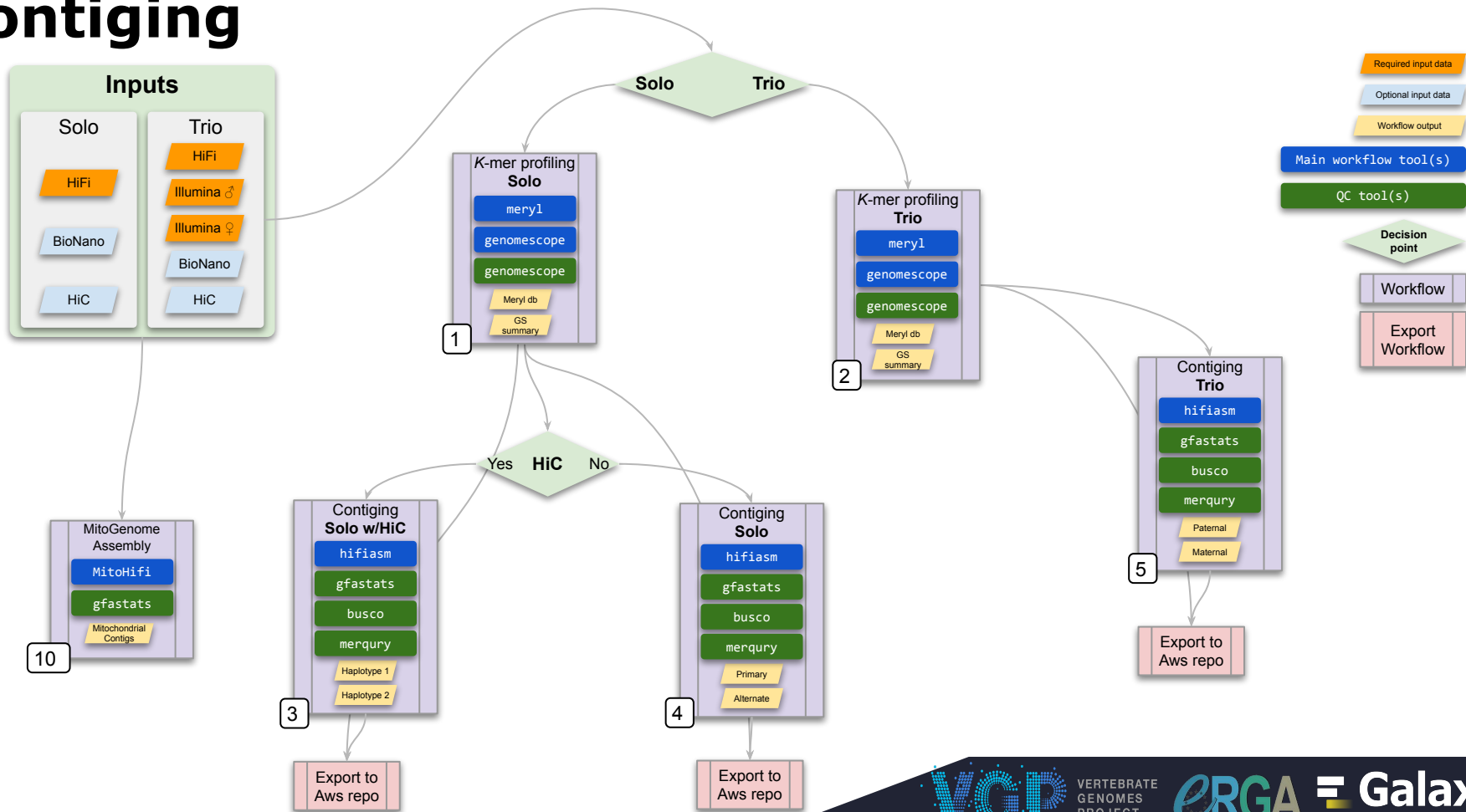
Kronenberg, Z. N., Rhie, A., Koren, S., Concepcion, G. T., Peluso, P., Munson, K. M., ... & Kingan, S. B. (2021). Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nature communications*, 12(1), 1935.

Trio-based Phasing

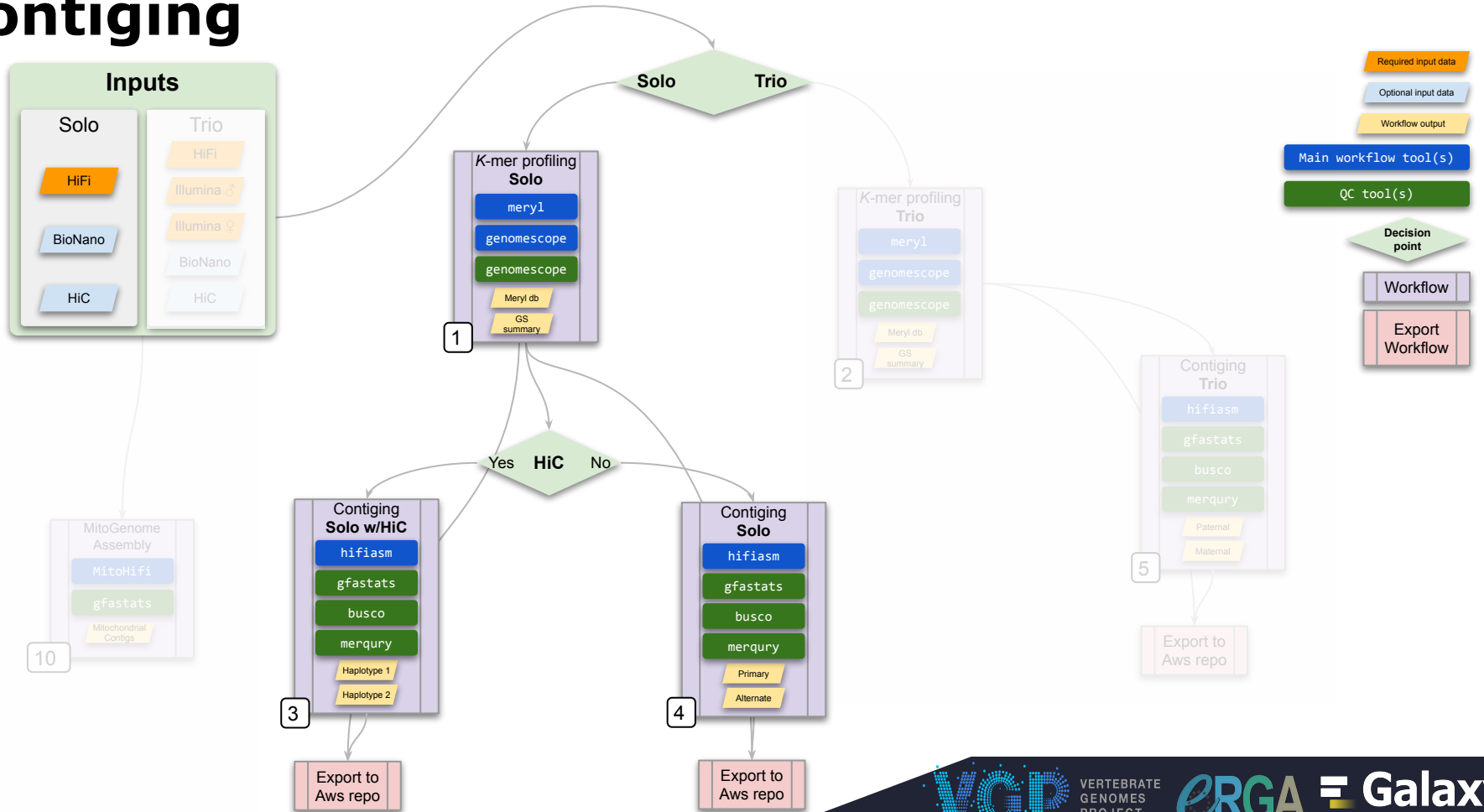


Koren et al. (2018). De novo assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology*. 10.1038/nbt.4277

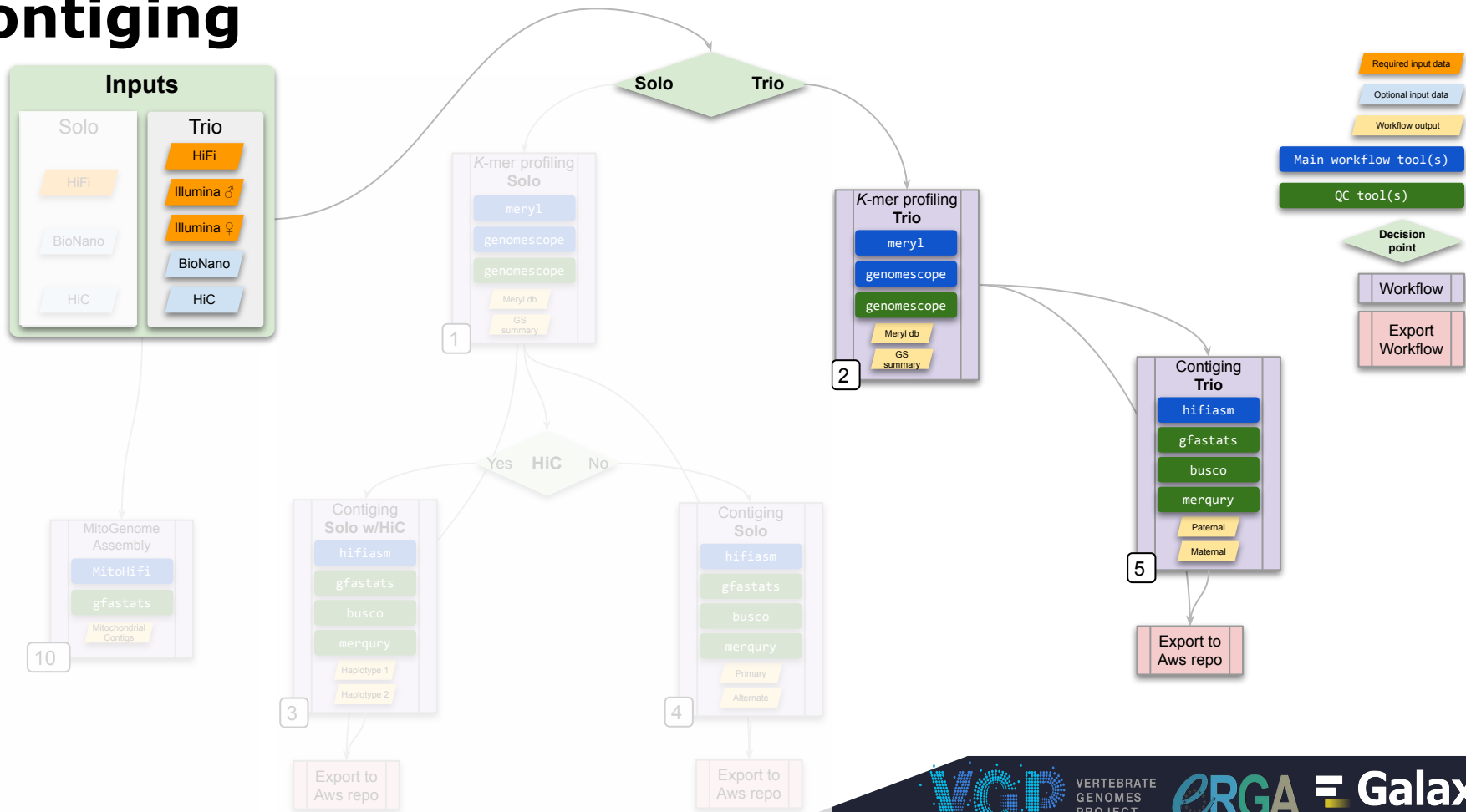
Contigging



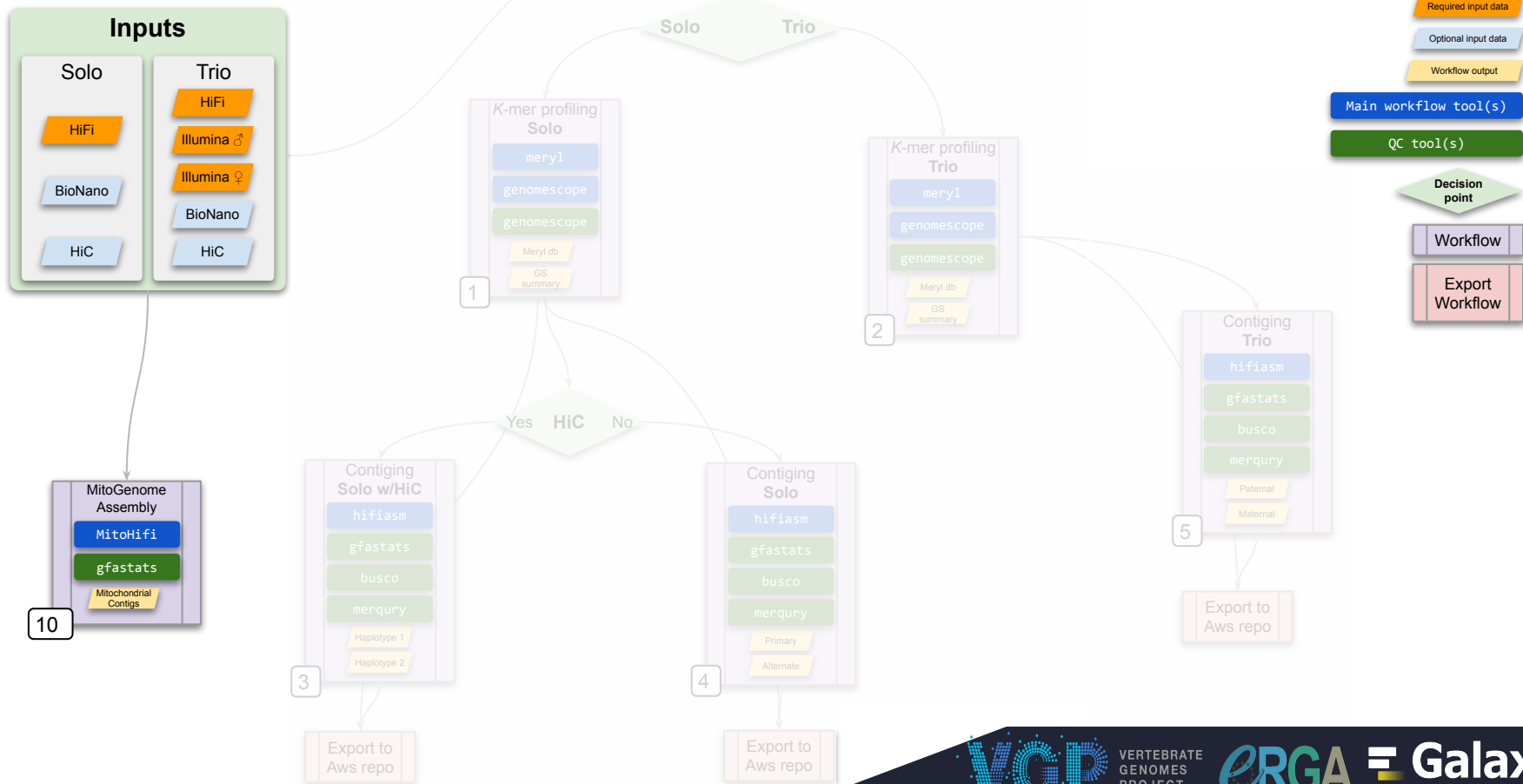
Contigging



Contigging



Contigging

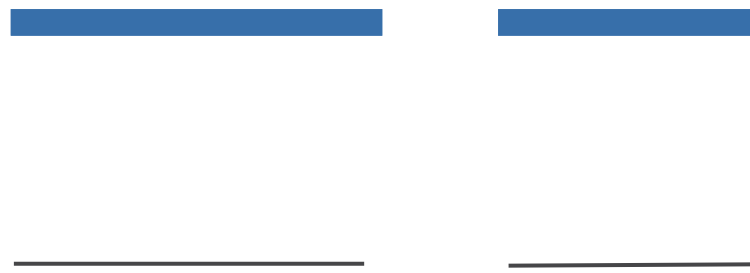


Purging

Contigs Alignment



K-mer copy number



↓ Purging

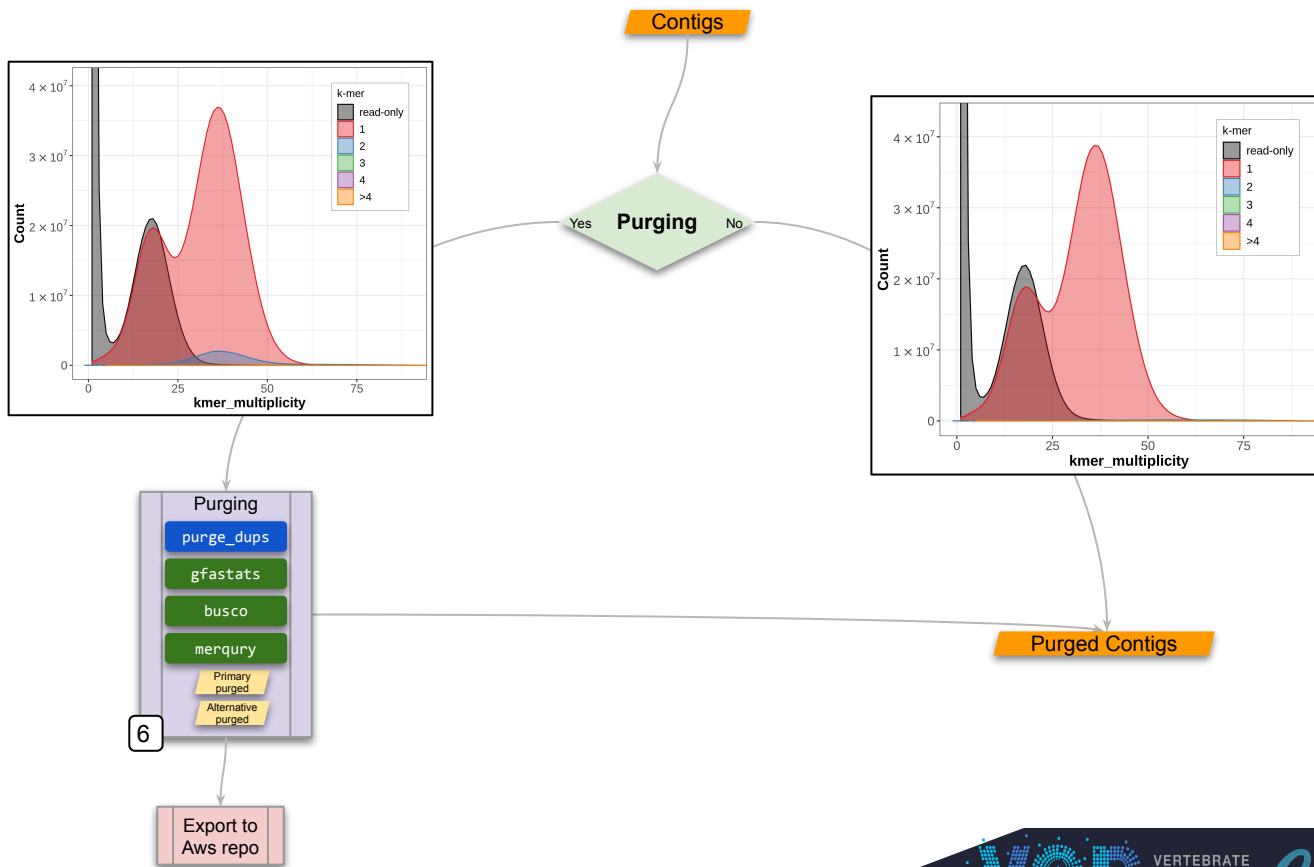
Contigs



K-mer copy number



Contigging - purging



Required input data

Optional input data

Workflow output

Main workflow tool(s)

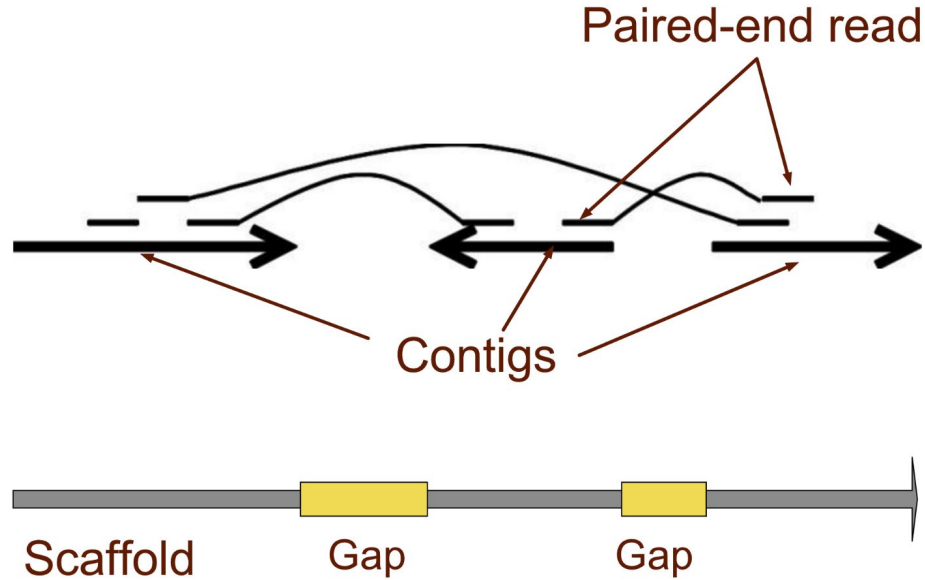
QC tool(s)

Decision point

Workflow

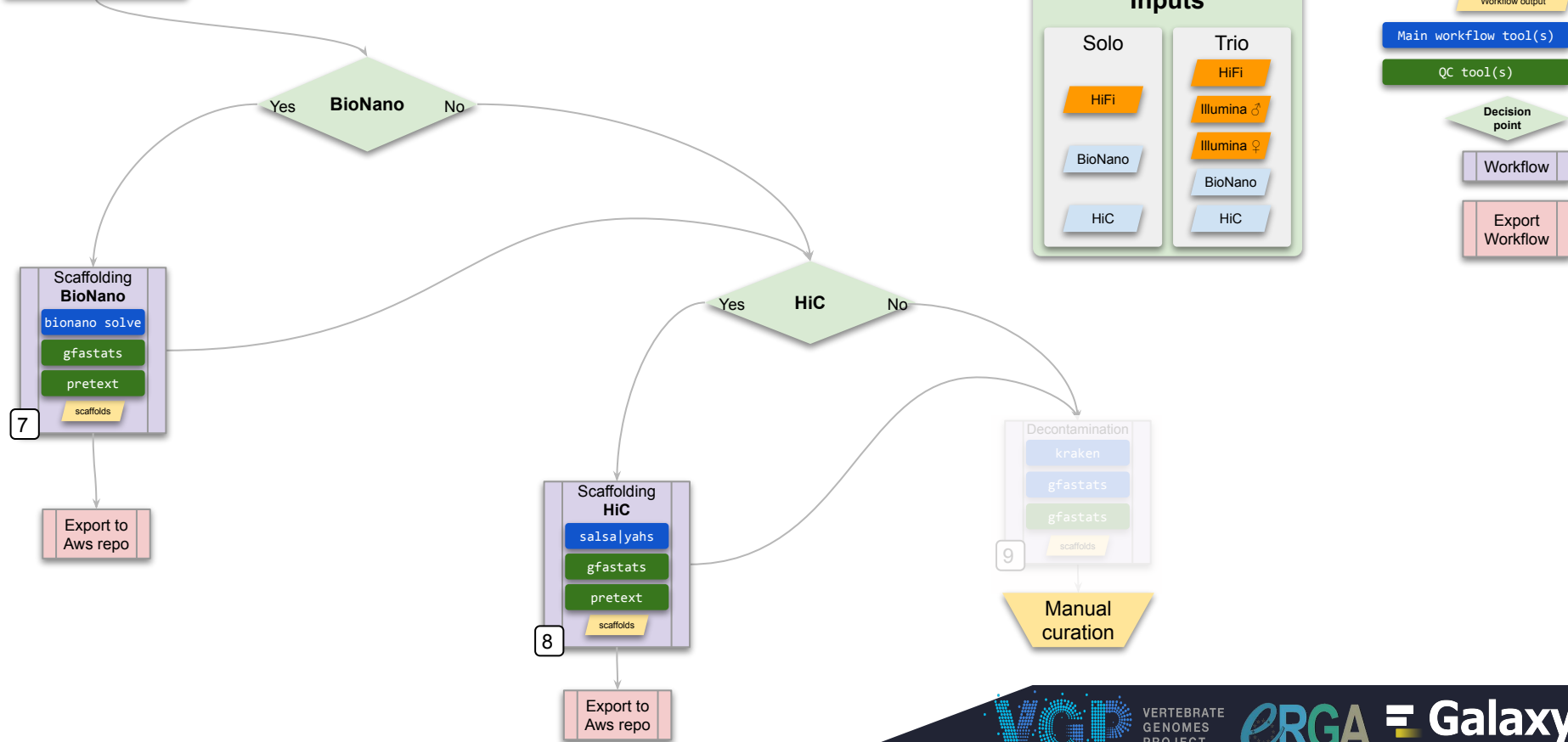
Export Workflow

Scaffolding



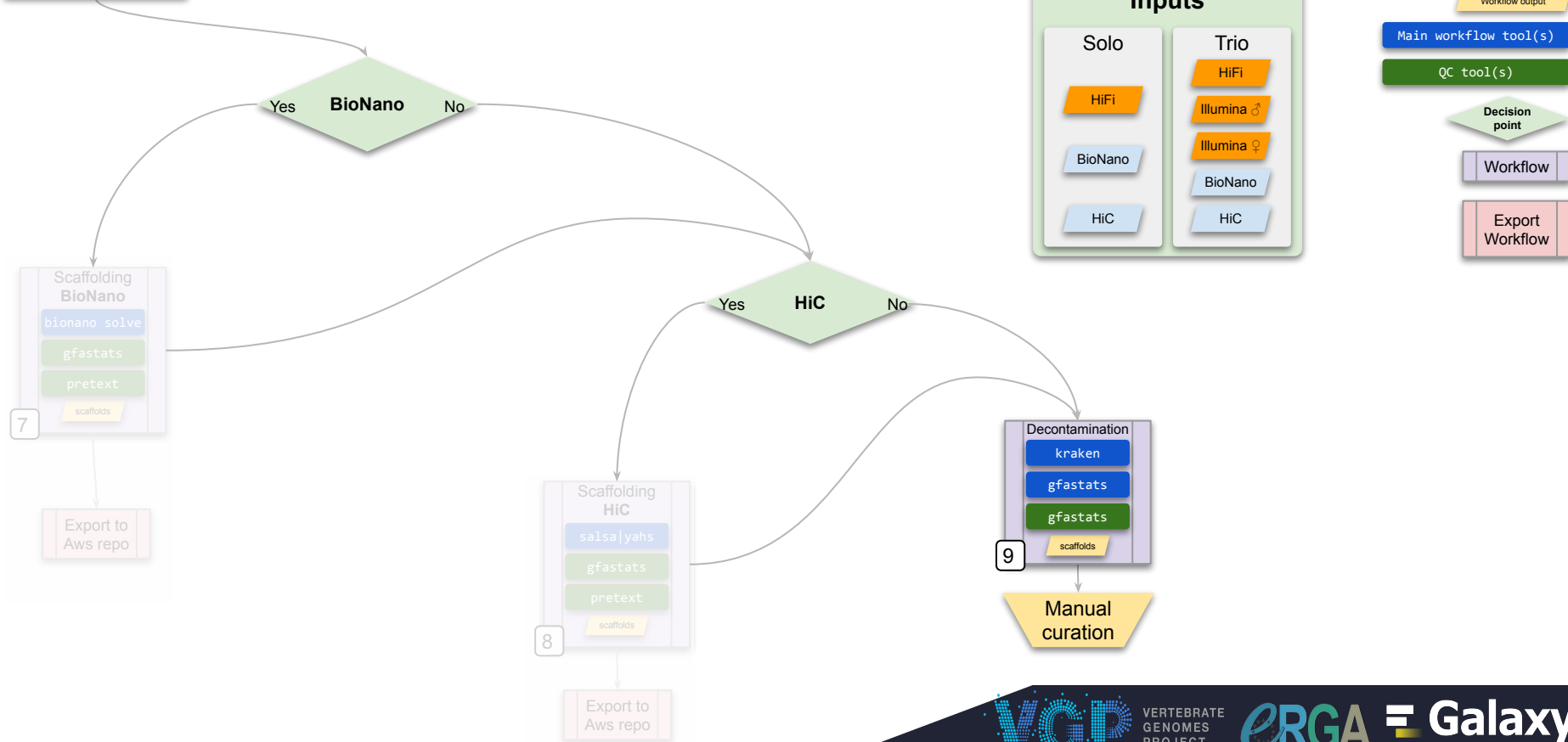
Scaffolding and Contaminants

Purged Contigs



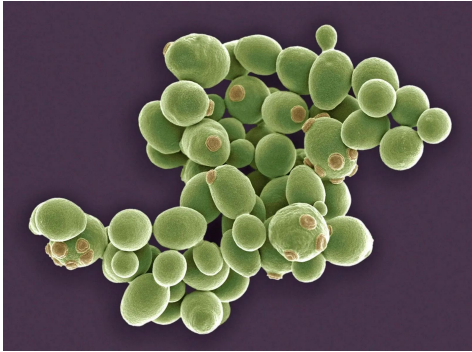
Scaffolding and Contaminants

Purged Contigs



Hands-on!

Species for Today



Yeast (*S. cerevisiae*)
12Mbp - 16 chromosomes
Highly inbred

Model for eukaryotes

30x HiFi (SRR13577846)
10x HiC (SRR7126301)



Zebra Finch (*T. guttata*)
1.2 Gbp - 32 chromosomes
Highly Heterozygous

Model for vocal learning

36x HiFi (Genomeark)
40x HiC (Genomeark)

Create a Galaxy account



Workflow

Visualize

Shared Data ▾

Help ▾

Login or Register



1. Create Account

<https://usegalaxy.org>

2. Join Training

<https://usegalaxy.org/join-training/vgpbga2023>

Using Galaxy

The screenshot shows the Galaxy web interface with three distinct sections highlighted by colored borders:

- Tools (Red border):** A sidebar menu on the left containing categories like 'Get Data', 'Send Data', 'Collection Operations', 'GENERAL TEXT TOOLS', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Datamash', 'GENOMIC FILE MANIPULATION', 'FASTA/FASTQ', 'FASTQ Quality Control', 'SAM/BAM', 'BED', 'VCF/BCF', 'Nanopore', 'Convert Formats', 'Lift-Over', 'COMMON GENOMICS TOOLS', 'Interactive tools', and 'Operate on Genomic Intervals'.
- News Article (Blue border):** A central content area with a yellow header. The text discusses a 'continuously updated list of laboratories that can host Ukrainian scientists' and includes a Ukrainian version of the same text. It also mentions a memorial for Simon Gladman, the principal architect of UseGalaxy.org.au, and includes a 'Learn More' button.
- History (Green border):** A panel on the right titled 'History' with a search bar and a message stating 'This history is empty. You can load your own data or get data from an external source.'

Importing Data from data libraries


- Hands-on Data : Yeast

The screenshot shows the Galaxy web interface. At the top, there is a navigation bar with the Galaxy logo, a home icon, and menu items for Workflow, Visualize, Shared Data, Admin, Help, and User. On the right of the navigation bar, it says "Using 3.5 TB". Below the navigation bar, there is a search bar with the text "vgp" and two checkboxes: "include deleted" and "exclude restricted". Below the search bar, there is a table with three columns: "Name", "Description", and "Synopsis". The table contains two rows of data. The first row is for "VGP- Yeast" with the description "Datasets for the Genome Assembly trainin ..." and a "Synopsis" column. The second row is for "VGP-Zebra Finch" with the description "Datasets for the assembly of the Zebra F ..." and a "Synopsis" column. To the right of each row, there are "Edit" and "Manage" buttons. At the bottom of the table, there is a pagination control showing "1" selected, "10" per page, and "206 total".




Name	Description	Synopsis
VGP- Yeast	Datasets for the Genome Assembly trainin ... <small>(more)</small>	
VGP-Zebra Finch	Datasets for the assembly of the Zebra F ... <small>(more)</small>	

Importing Workflows : Public workflows

Published Workflows

Name	Tags	Updated	Owner
 K-mer profiling and QC (WF1)	 	less than a minute ago	delphinel

WF1: HiFi reads-based Kmer-counting

Search Workflows



+ Create

Import

Name

Tags

Updated

Sharing

Bookmarked

K-mer profiling and QC (WF1) ✓

Reviewed ×

VGP ×

less than a minute ago



Run workflow

Workflow: K-mer profiling and QC (WF1)



Run Workflow

Collection of Pacbio Data



2: PacBio reads

K-mer length

21

Ploidy

2

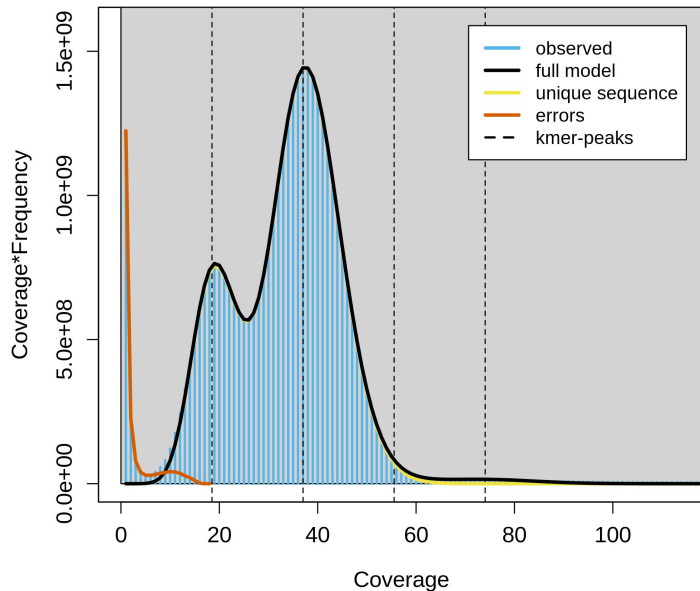
Expand to full workflow form.

WF1: Outputs

Kmer profiling Zebra Finch

GenomeScope Profile

len:1,009,754,198bp uniq:89.2%
aa:98.5% ab:1.47%
kcov:18.5 err:0.238% dup:0.235 k:21 p:2



GenomeScope version 2.0

input file =

/data/dnb07/galaxy_db/files/a/8/5/dataset_a855ad9d-4075-4de9-b552-dbe07c12f57c.dat

output directory = .

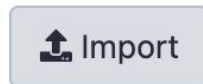
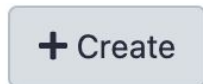
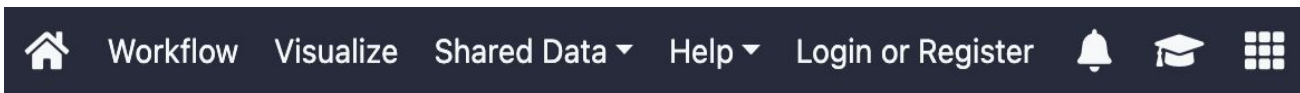
p = 2

k = 21

TESTING set to TRUE

property	min	max
Homozygous (aa)	98.5256%	98.5313%
Heterozygous (ab)	1.4687%	1.47441%
Genome Haploid Length	1,009,402,687 bp	1,009,754,198 bp
Genome Repeat Length	109,282,859 bp	109,320,915 bp
Genome Unique Length	900,119,828 bp	900,433,283 bp
Model Fit	89.557%	99.2734%
Read Error Rate	0.237557%	0.237557%

Importing Workflows : From Dockstore



Import a Workflow from Configured GA4GH Tool Registry Servers (e.g. Dockstore)

Use either the Galaxy search form or import from a TRS ID.

TRS Server: **Dockstore** ▼

TRS ID:

#workflow/github.com/Delphine-L/iwc/WF3-Assembly

WF3-5: Contig assembly

Workflow management interface showing a list of workflows. The selected workflow is "Assembly (WF3)".

Name	Tags	Updated	Sharing	Bookmarked
Assembly (WF3) ✓	VGP x Reviewed x	less than a minute ago		☆

Buttons: + Create, Import, Run workflow

Workflow: Assembly (WF3)

Run Workflow

Pacbio Reads Collection

5: PacBio reads

Meryl Database

8: Meryl on data 7: read-db.meryldb

Genomescope Summary

16: GenomeScope on data 9 Model parameters

SAK input file

16: GenomeScope on data 9 Model parameters

Name of primary

15: GenomeScope on data 9 Summary

Name of alternate

14: GenomeScope on data 9 Model

Alternate

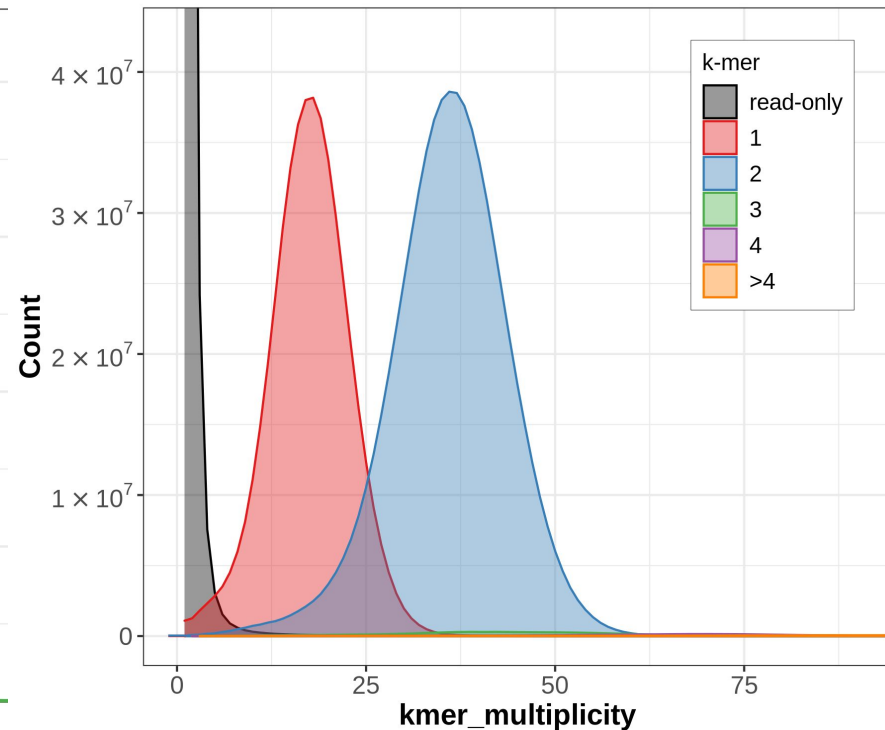
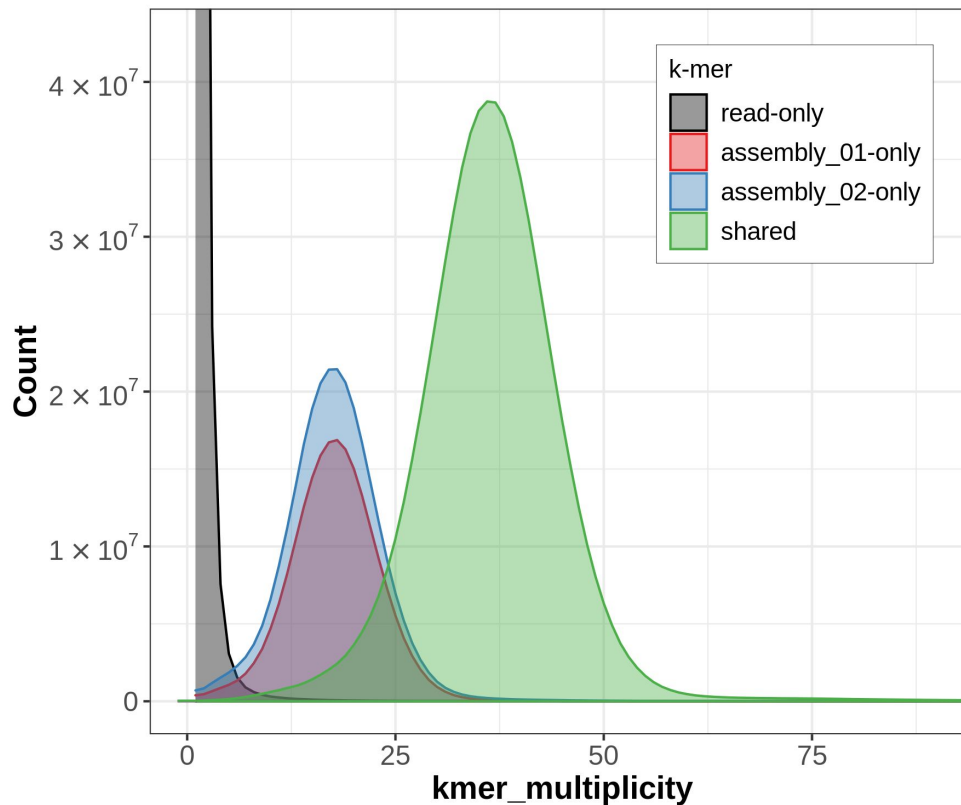
4: SRR7126301_1

Expand to full

3: SRR7126301_2

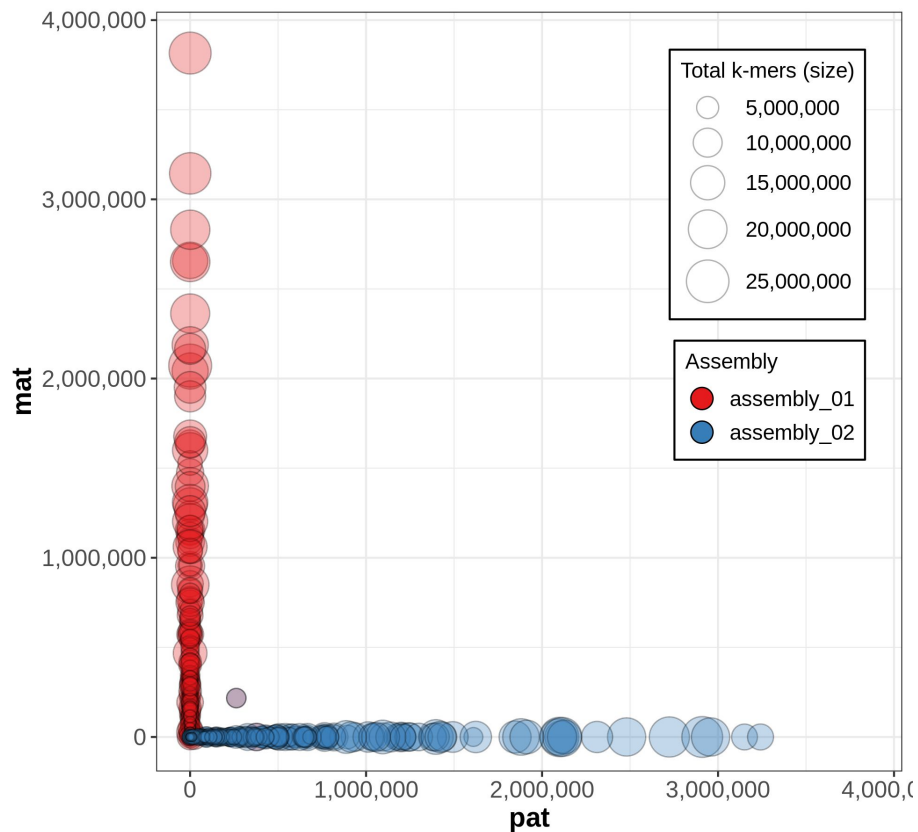
WF3-5: Outputs

Assembly with HiC phasing Zebra Finch



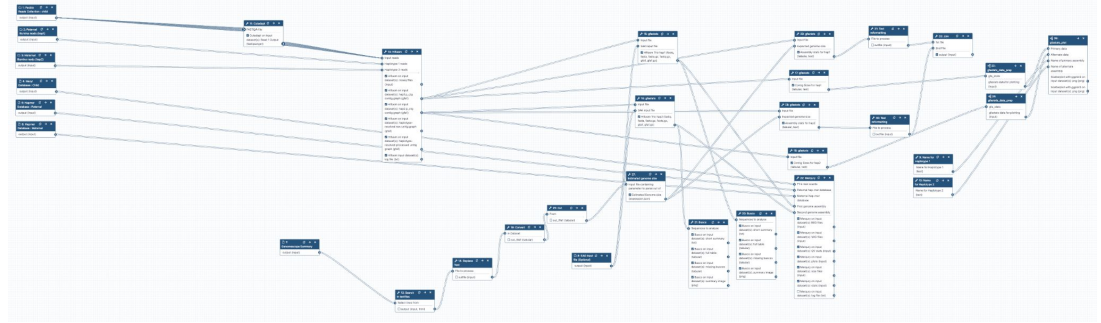
WF5: Outputs

Trio Assembly Zebra Finch



WF7-8: Scaffolding

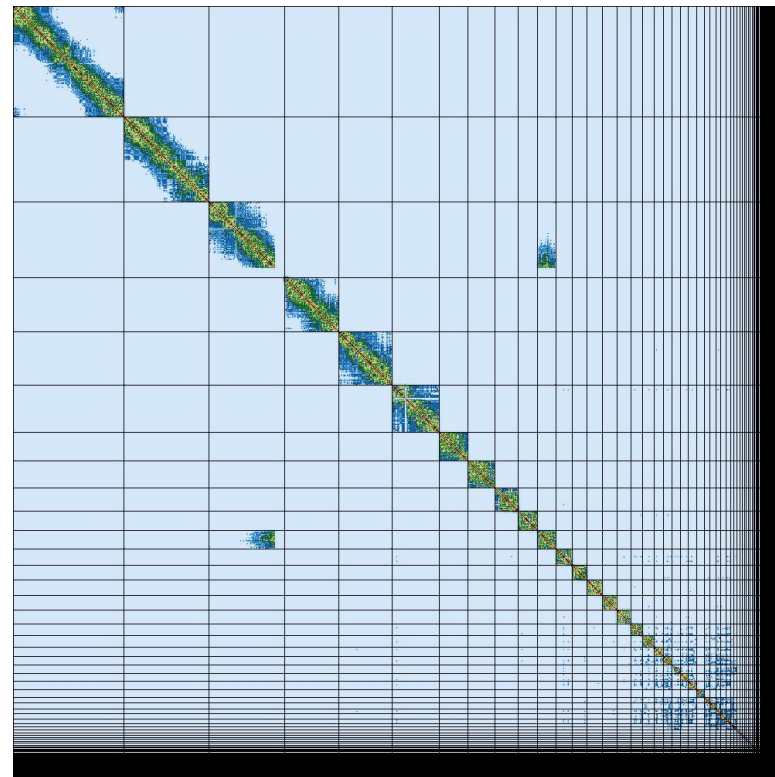
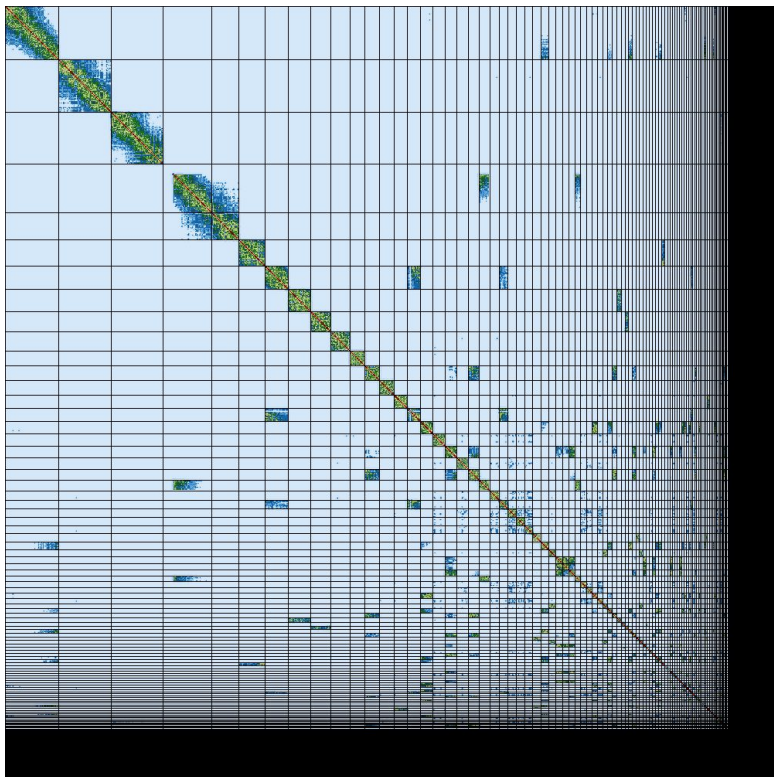
- If HiC data
 - [#workflow/github.com/Delphine-L/iwc/WF8a-Scaffolding_HiC_Yahs](#)
 - Tools
 - Yahs
 - Gfastats
 - Pretext



- If Bionano data:
 - [#workflow/github.com/Delphine-L/iwc/WF7-Scaffolding_Bionano](#)
 - Tools
 - Bionano solve
 - gfastats
 - Pretext

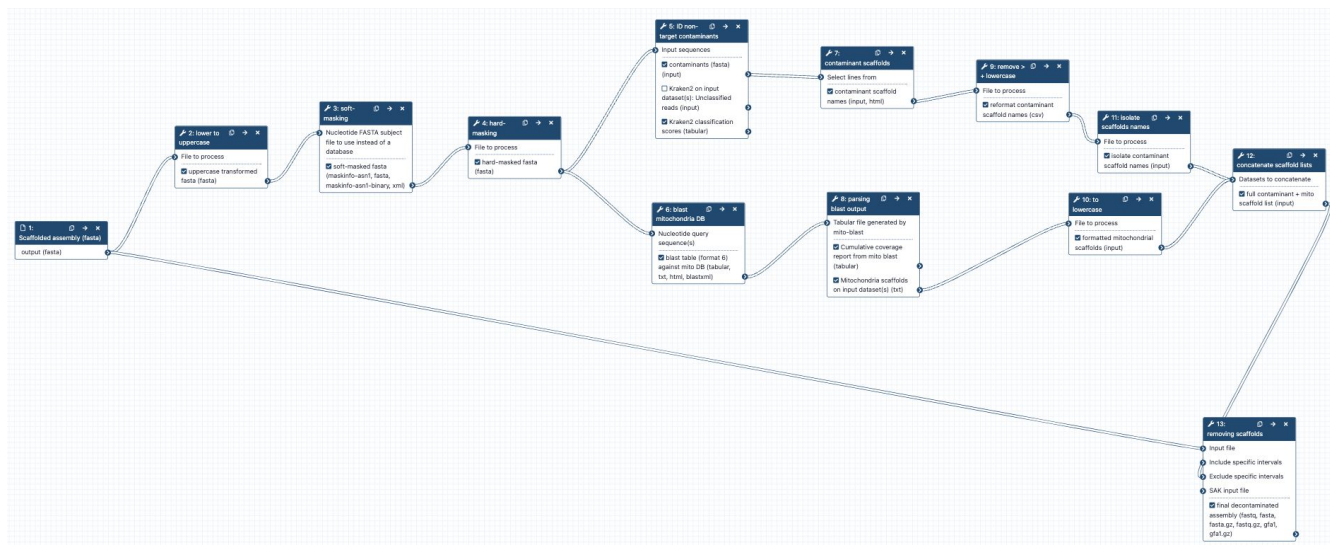
WF7-8: Outputs

HiC Scaffolding Zebra Finch



WF9: Decontamination (optional)

- TRS ID: [#workflow/github.com/Delphine-L/iwc/WF9-Decontamination](https://github.com/Delphine-L/iwc/WF9-Decontamination)
- Tools:
 - Kraken
 - gfastats



Run your own assembly!

Data Requirements:

- >30x coverage for HiFi data
- >60x coverage for HiC data

Find more information:

Galaxy Project Hub :

<https://galaxyproject.org/projects/vgp/>

Galaxy Training Network:

https://training.galaxyproject.org/training-material/topics/assembly/tutorials/vgp_genome_assembly/tutorial.html

Acknowledgments

VGP team:

- Giulio Formenti
- Linelle Abueg
- Nadolina Brajuka
- Marc Palmada Flores
- Byung June Ko
- Arang Rhie
- Mark Chaisson
- Jo Collins
- Erich Jarvis
- Adam Phillippy
- And everyone else

Galaxy team:

- Delphine Lariviere
- Cristobal Gallardo
- Alex Ostrovsky
- Bjorn Grüning
- Anton Nekrutenko
- Michael Schatz
- Marius van den Beek
- And everyone else



JOHNS HOPKINS
UNIVERSITY



<https://galaxyproject.org/>



Thank you!