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

Delphine Larivière

September 21st, 2023

VERTEBRATE GENOMES PROJECT

A PROJECT OF THE G10K CONSORTIUM

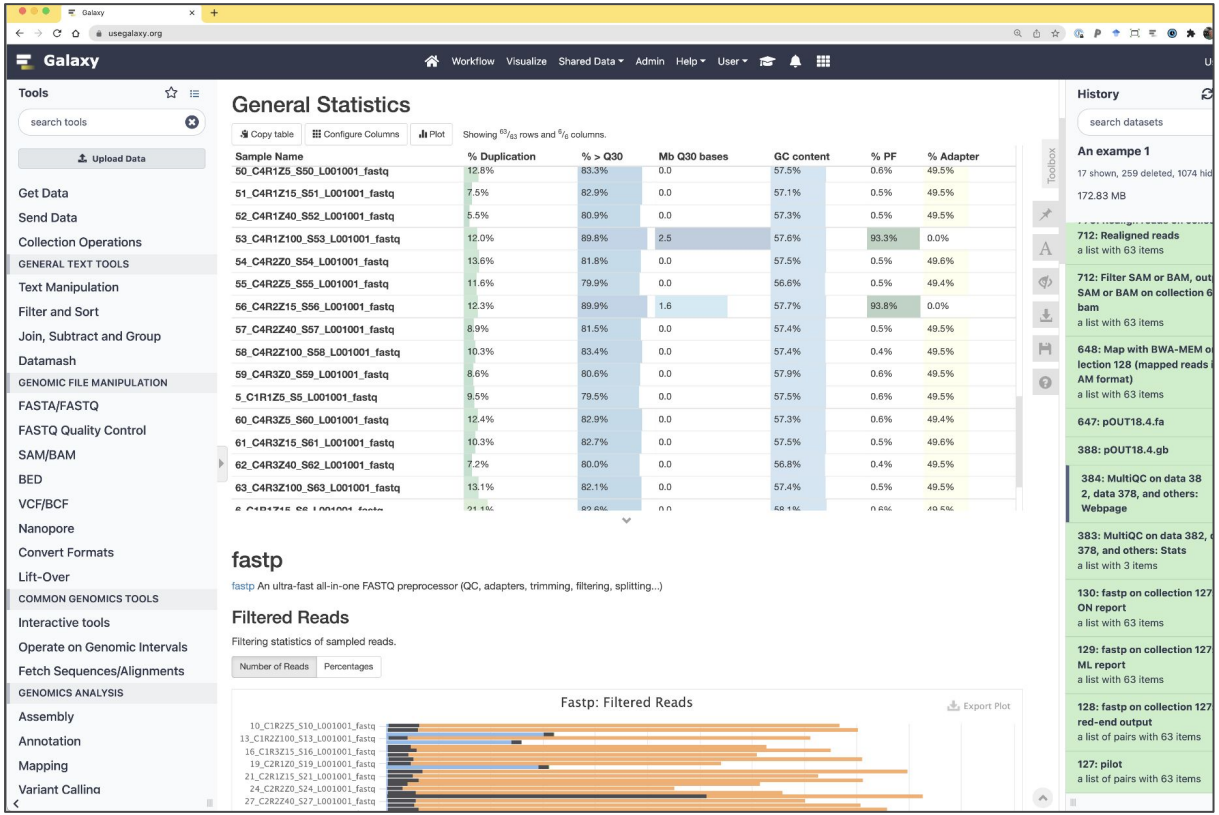
ERGA

EUROPEAN REFERENCE GENOME ATLAS

Galaxy PROJECT

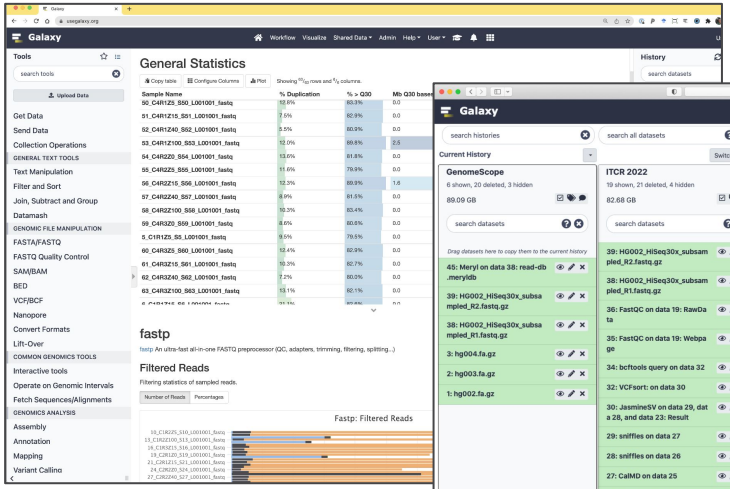
The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses

Introducing Galaxy!

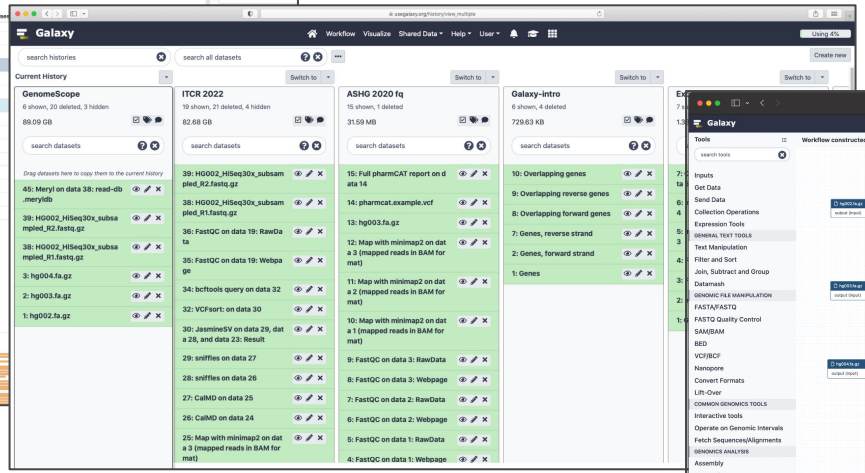


VERTEBRATE GENOMES PROJECT
A PROJECT OF THE G10K CONSORTIUM
ERGA EUROPEAN REFERENCE GENOME ATLAS
Galaxy PROJECT

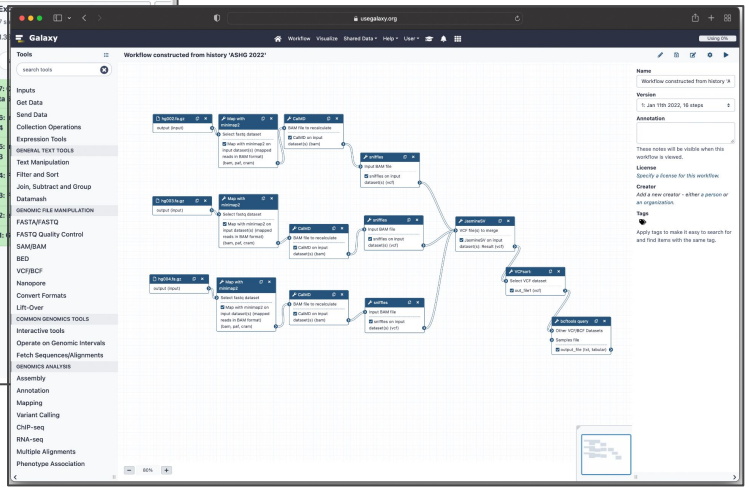
Accessible, Reproducible and Collaborative



One Analysis



Many Analyses



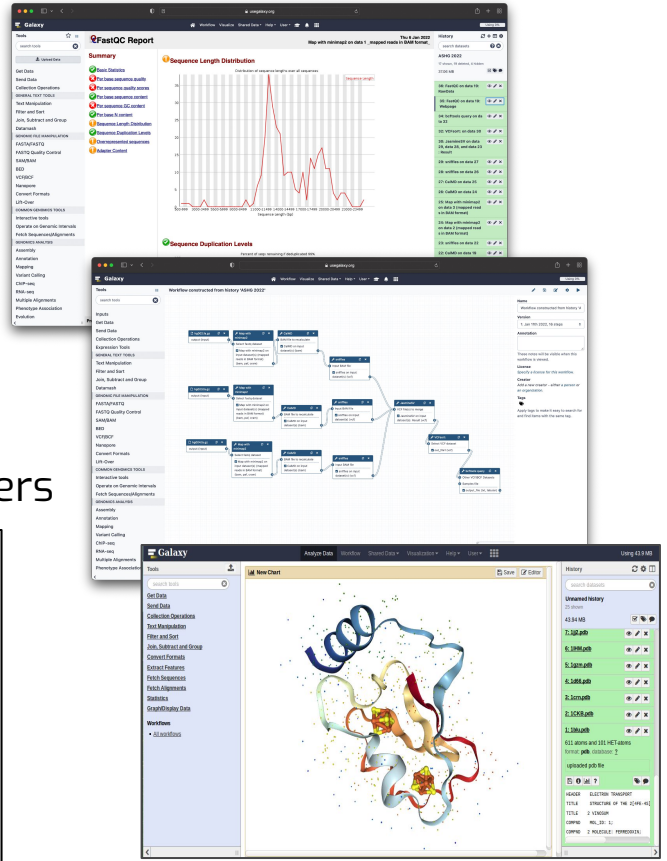
Collaborative
Data & Workflows

(The Galaxy Community, NAR, 2022)

VERTEBRATE GENOMES PROJECT
A PROJECT OF THE G10K CONSORTIUM
ERGA EUROPEAN REFERENCE GENOME ATLAS
Galaxy PROJECT

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

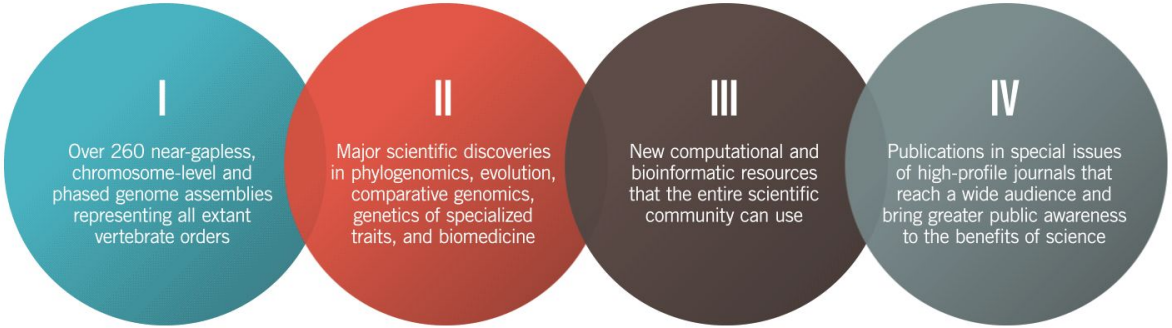
The Vertebrate Genomes Project

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

Shutterstock

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



GENOME 10K

B10K

BAT 1K

EARTH BIOGENOME PROJECT

A PROJECT OF THE G10K CONSORTIUM

ERGA EUROPEAN REFERENCE GENOME ATLAS

Galaxy PROJECT

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	256 species	173 species	20 species	63 species
VGP	250 species	172 species	15 species	63 species
T2T	6 species	none	6 species	none

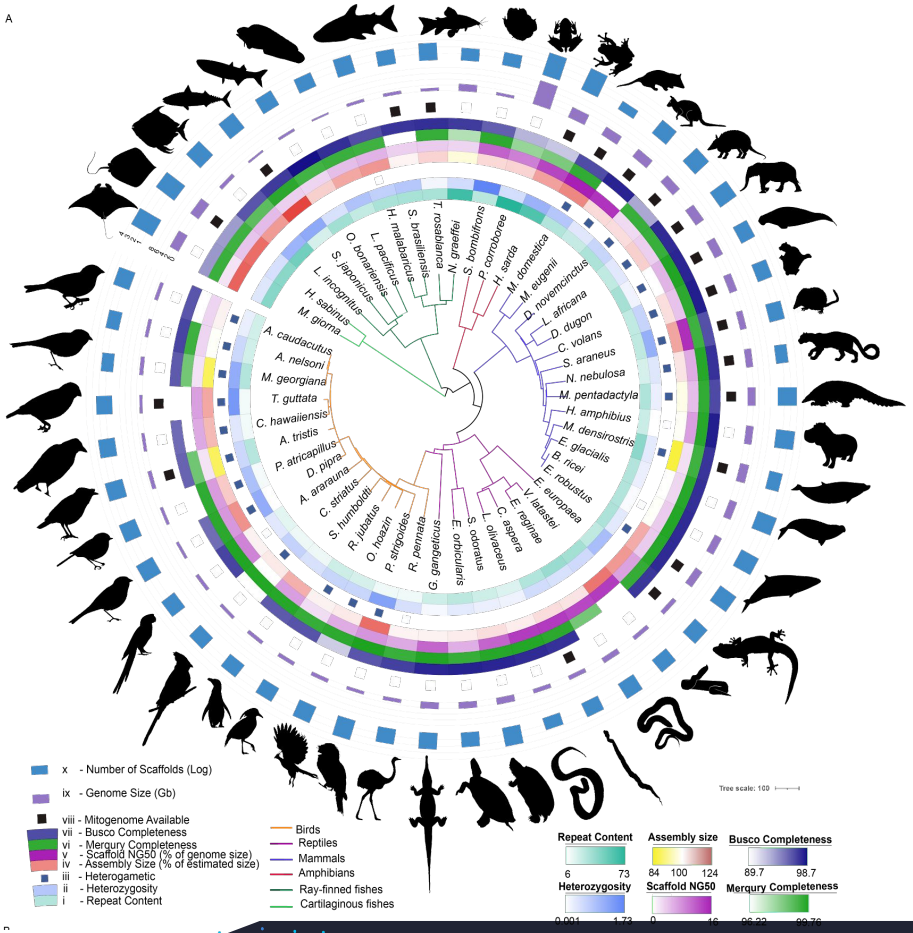
Number of species at each level of completion.

https://genomeark.github.io/

VERTEBRATE GENOMES PROJECT
A PROJECT OF THE G10K CONSORTIUM
ERGA EUROPEAN REFERENCE GENOME ATLAS
Galaxy PROJECT

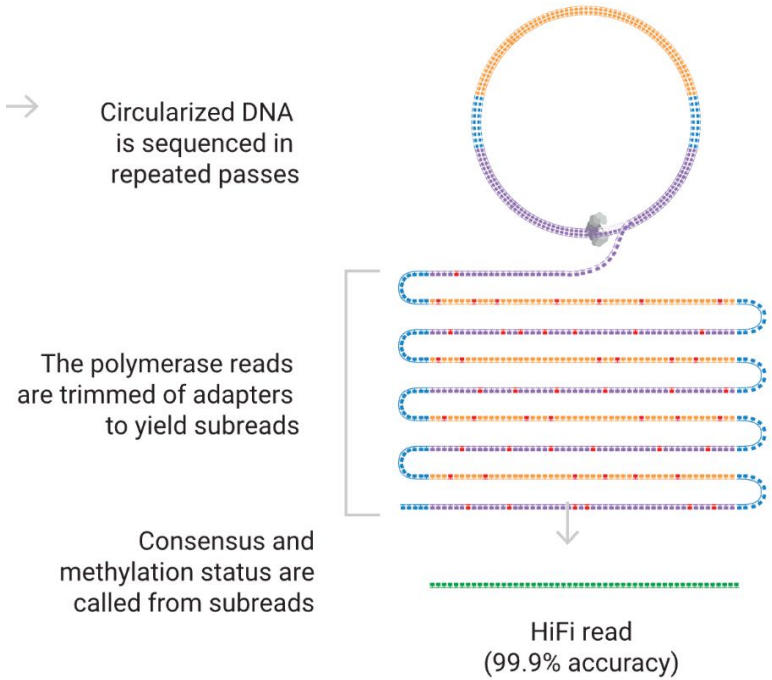
VGP assembly Pipeline v2

- 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

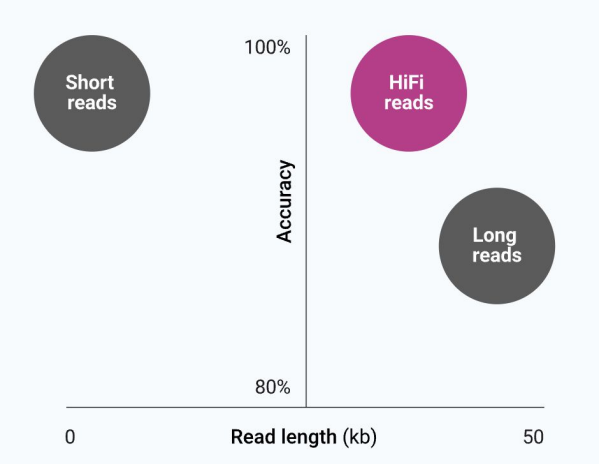


VERTEBRATE GENOMES PROJECT
A PROJECT OF THE G10K CONSORTIUM
ERGA EUROPEAN REFERENCE GENOME ATLAS
Galaxy PROJECT

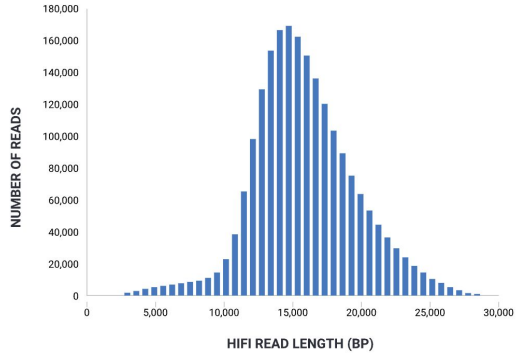
Data - PacBio HiFi



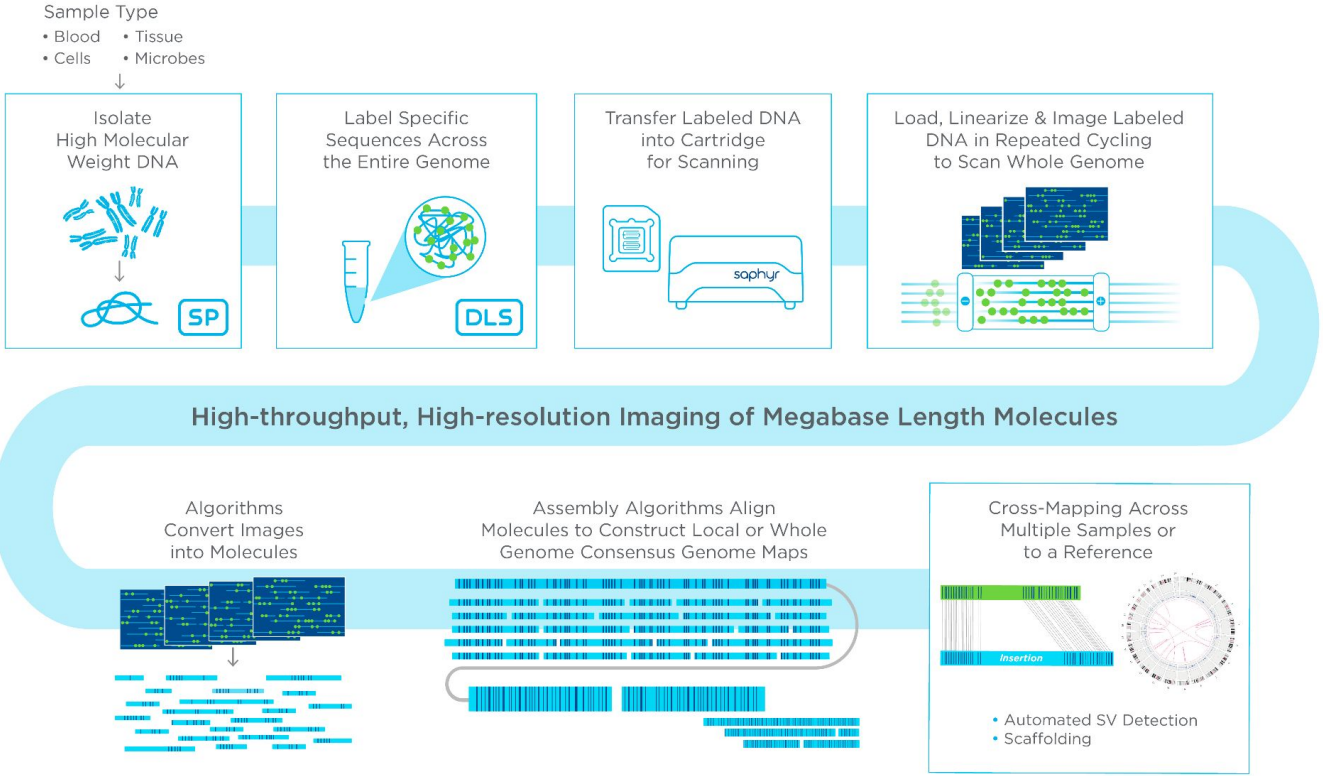
https://www.pacb.com/technology/hifi-sequencing/



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



Data - Bionano Optical Mapping



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

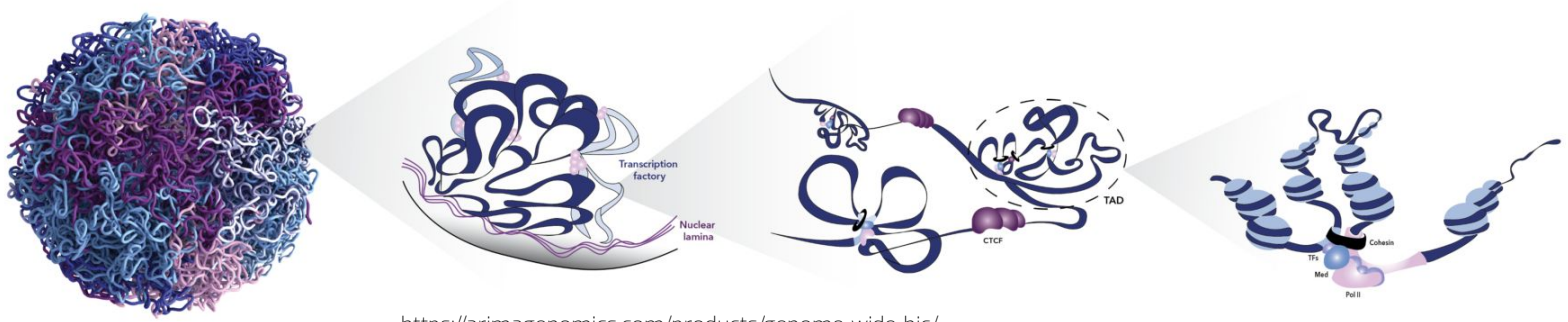
Data - Arima HiC

In the nucleus chromosomes are organized into chromosome territories

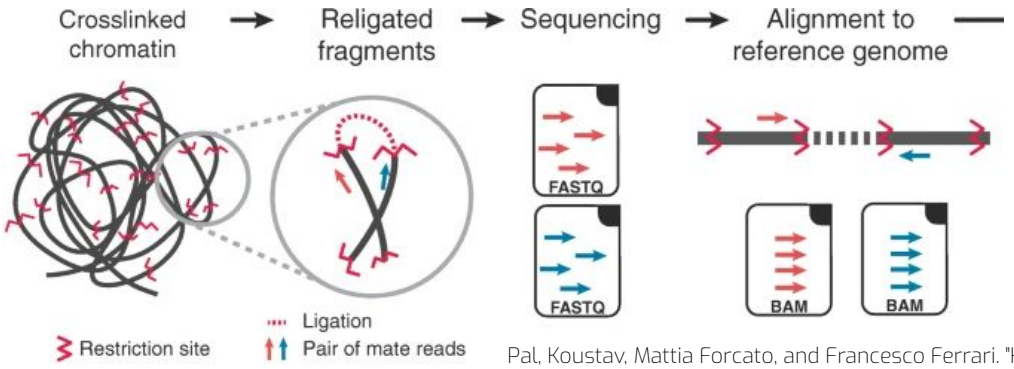
Chromosomes are divided into cell-specific A/B compartments

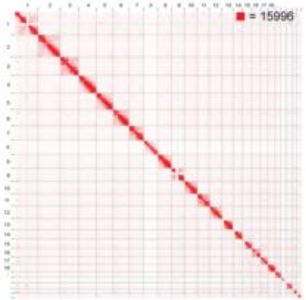
Compartments are organized into topologically associated domains (TADs)

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



https://arimagenomics.com/products/genome-wide-hic/

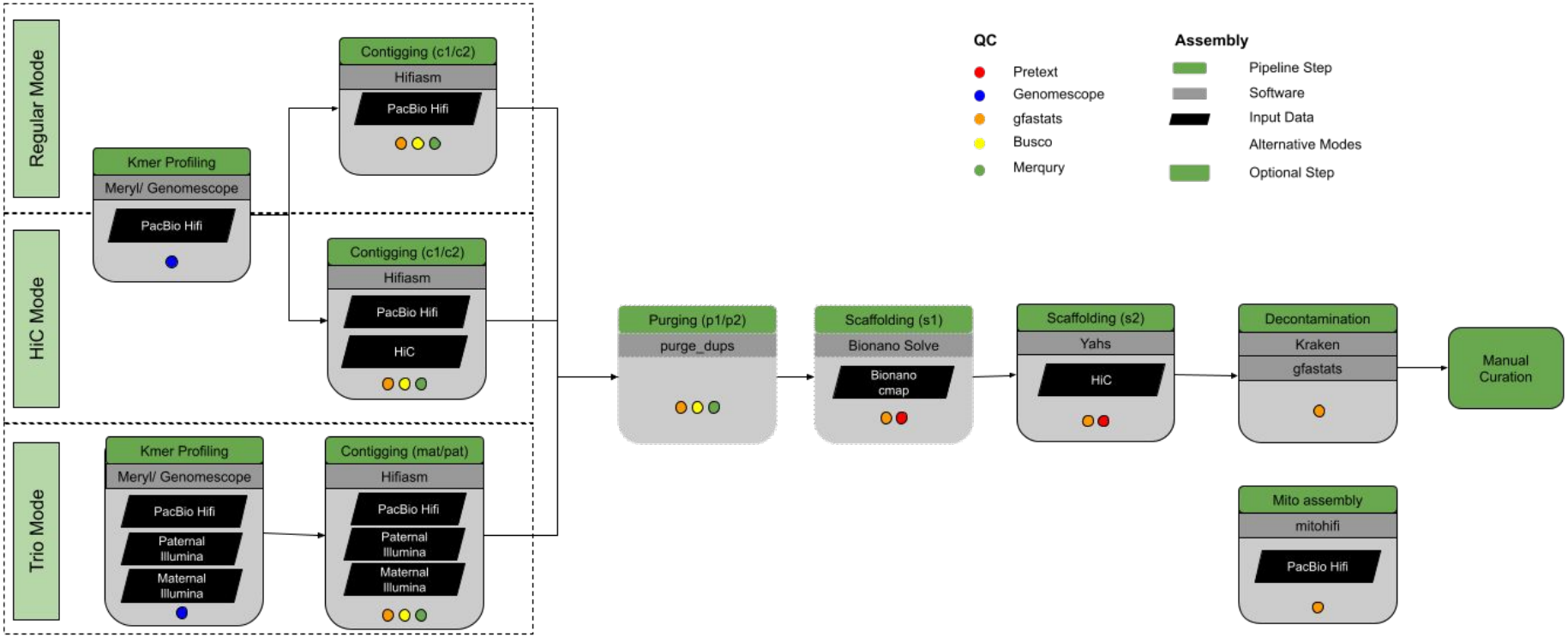




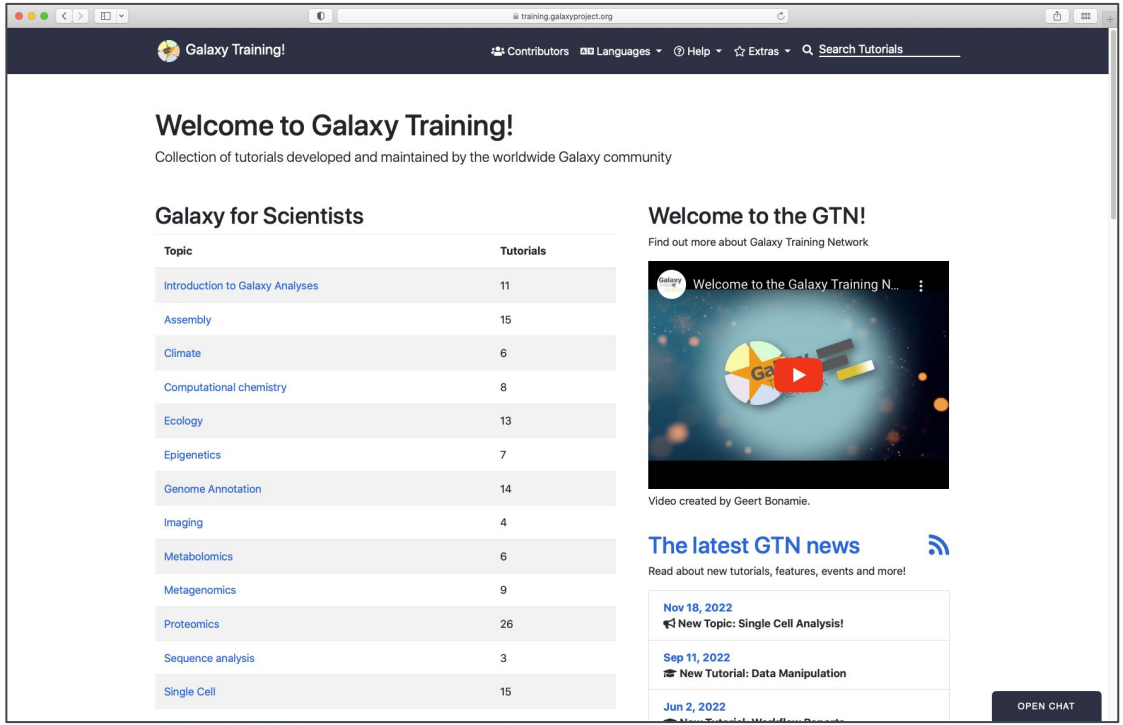
Pal, Koustav, Mattia Forcato, and Francesco Ferrari. "Hi-C analysis: from data generation to integration." Biophysical reviews 11.1 (2019): 67-78.

VERTEBRATE GENOMES PROJECT
A PROJECT OF THE G10K CONSORTIUM
ERGA European Reference Genome Atlas
Galaxy PROJECT

Assembly Pipeline Overview



training.galaxyproject.org



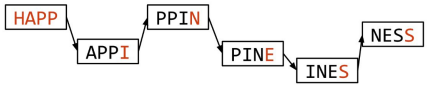
Contiging

TTGACACTTACCGA **Read**

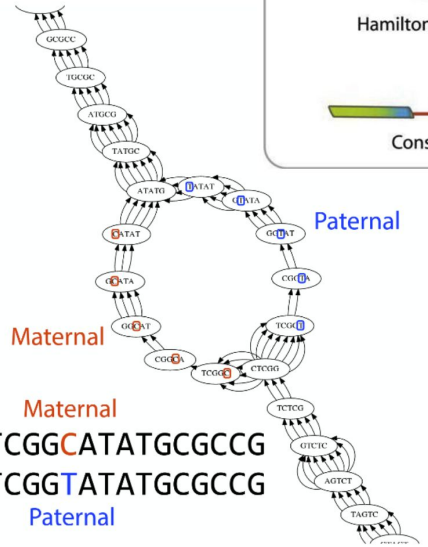
TTGACACTTACC
TGACACTTACCG
GACACTTACCGA
k-mers for k=12

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

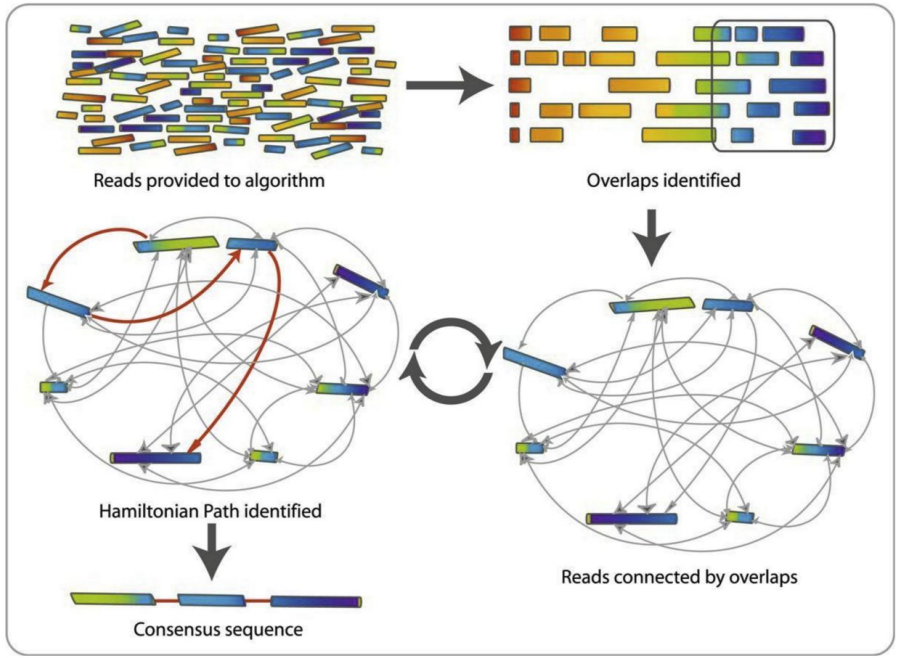
k = 4 k-mers:
HAPP APPI
PINE PPIN
INES NESS



HAPPINESS



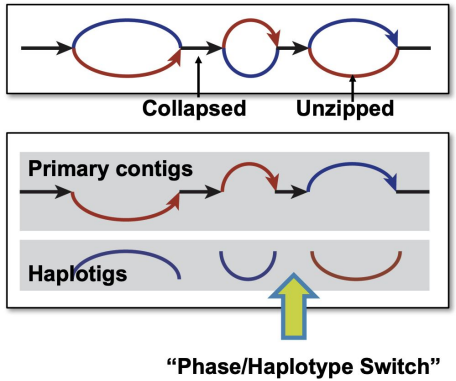
GTAGTCTCGGCATATGCGCCG
GTAGTCTCGGTATATGCGCCG



VERTEBRATE GENOMES PROJECT
A PROJECT OF THE G10K CONSORTIUM
ERGA EUROPEAN REFERENCE GENOME ATLAS
Galaxy PROJECT

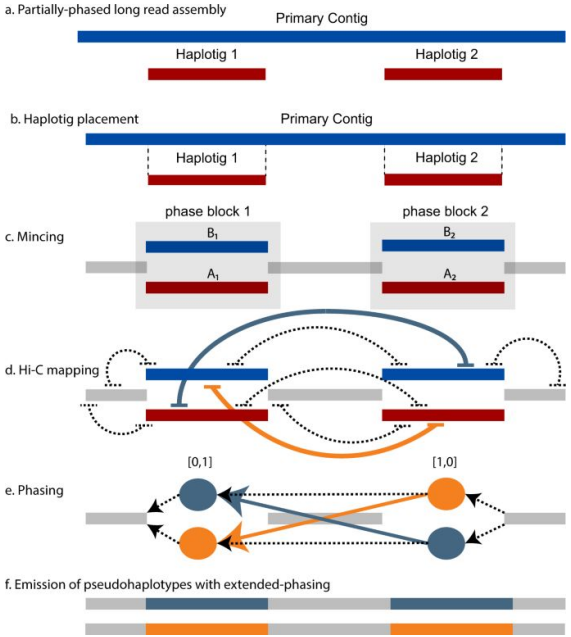
Phased Assemblies

Simple Phasing



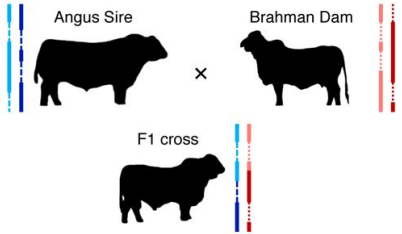
Chin, C.S. et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nature Methods. 13(12), 1050.

Hi-C-based Phasing



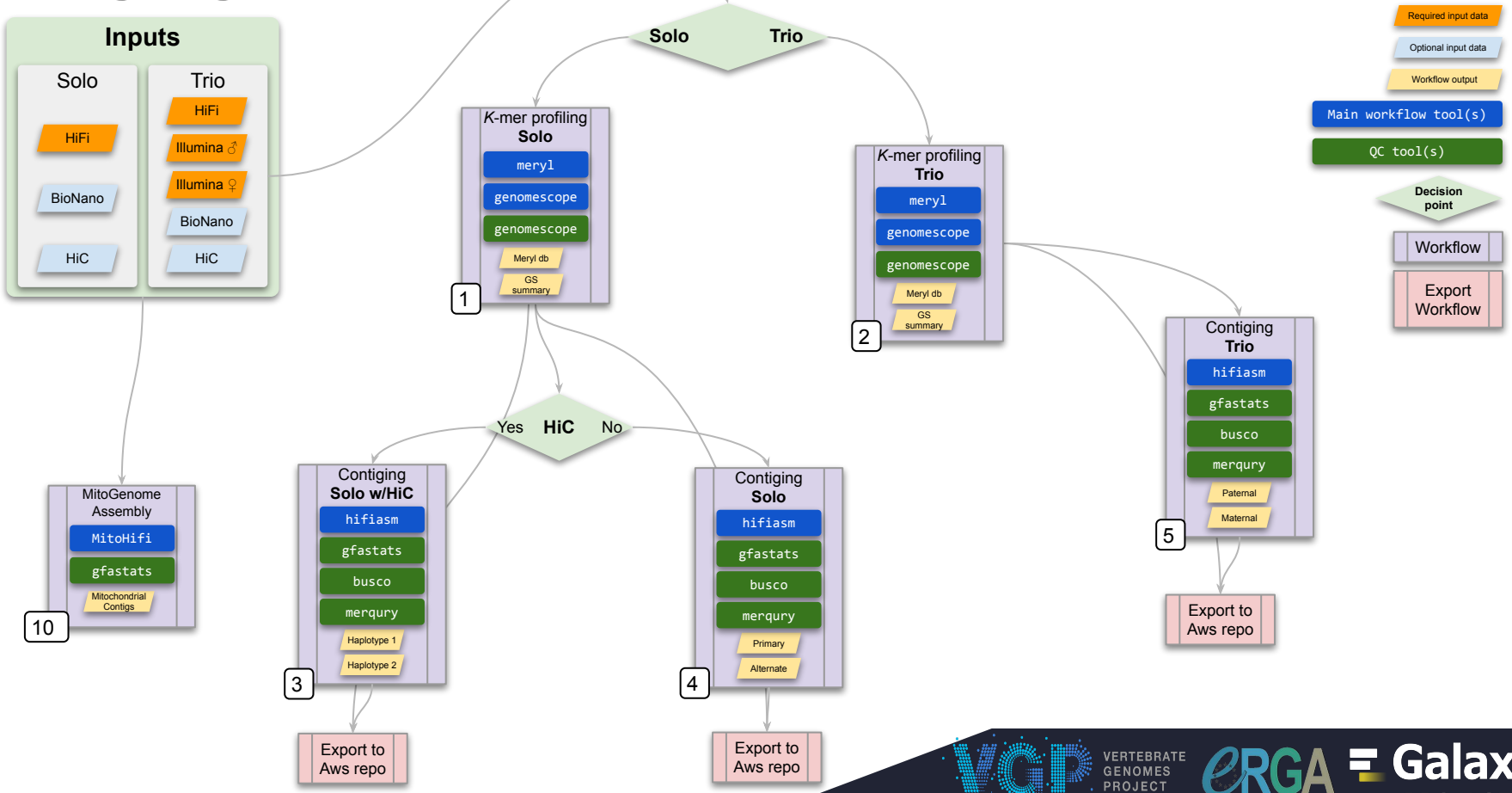
Kronenberg, Z. N., Rhie, A., Koren, S., Concepcion, G. T., Peluso, P., Munson, K. M., ... & Kingan, S. B. (2021). Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. Nature communications, 12(1), 1935.

Trio-based Phasing

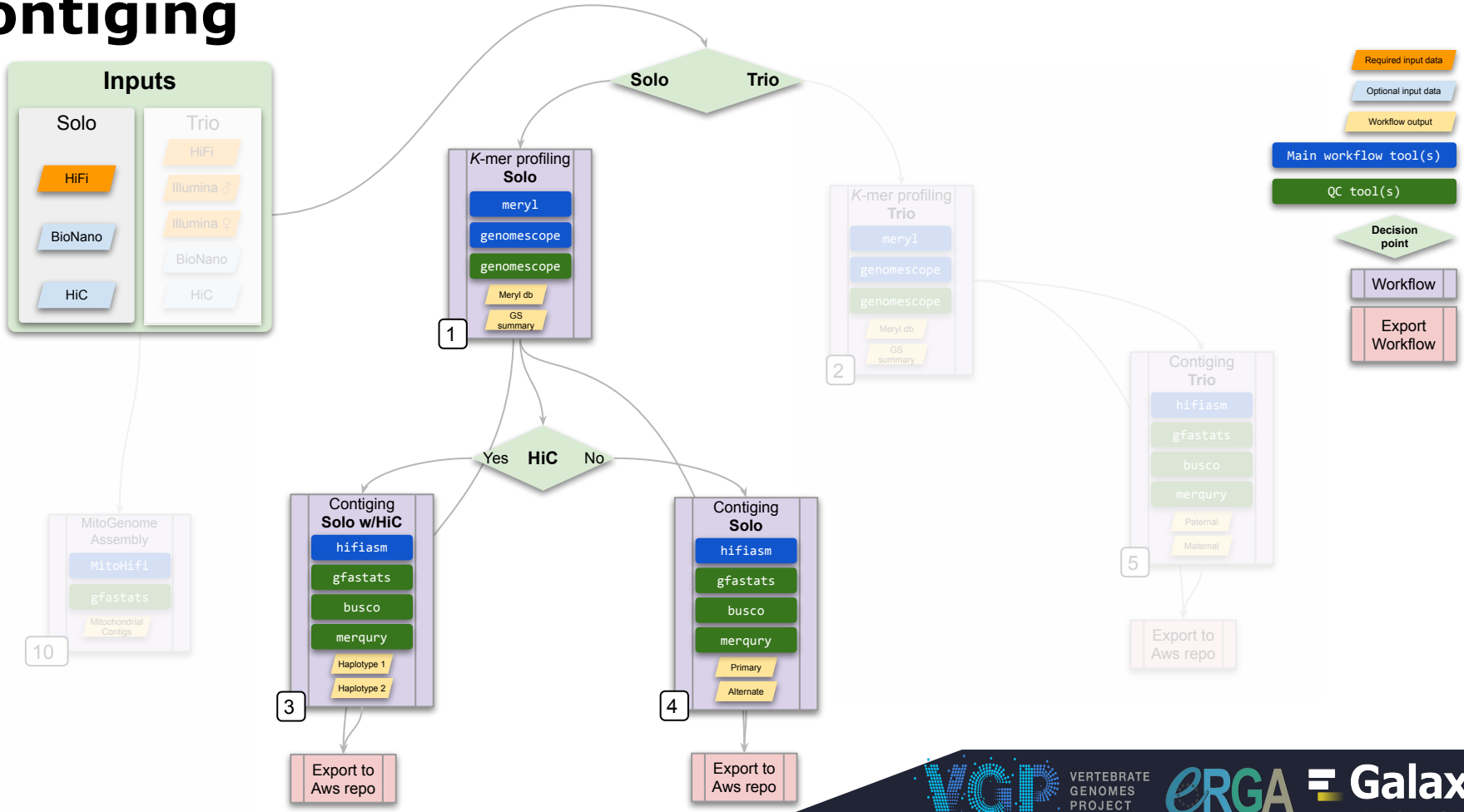


Koren et al. (2018). De novo assembly of haplotype-resolved genomes with trio binning. Nature Biotechnology. 10.1038/nbt.4277

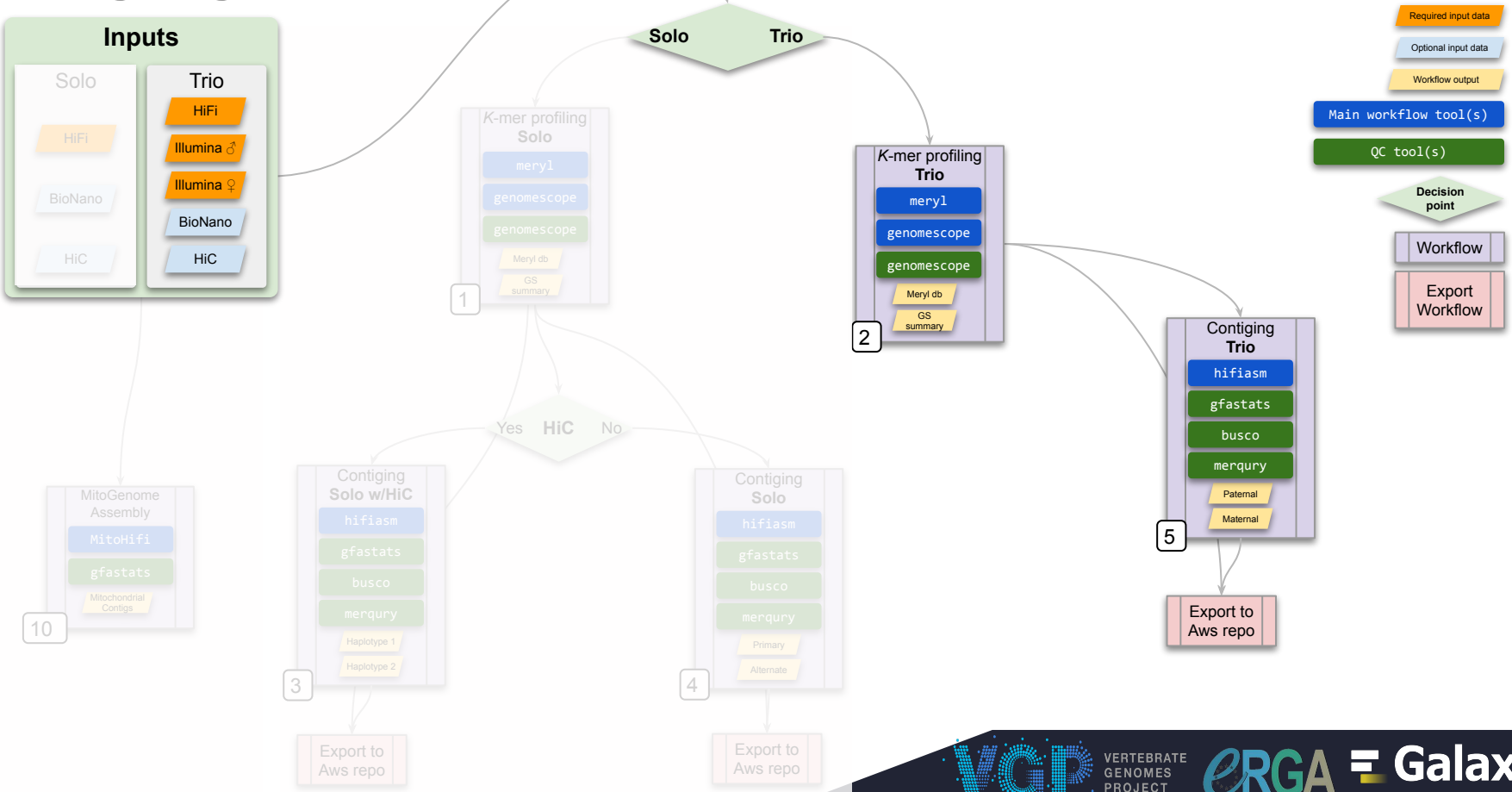
Contiging



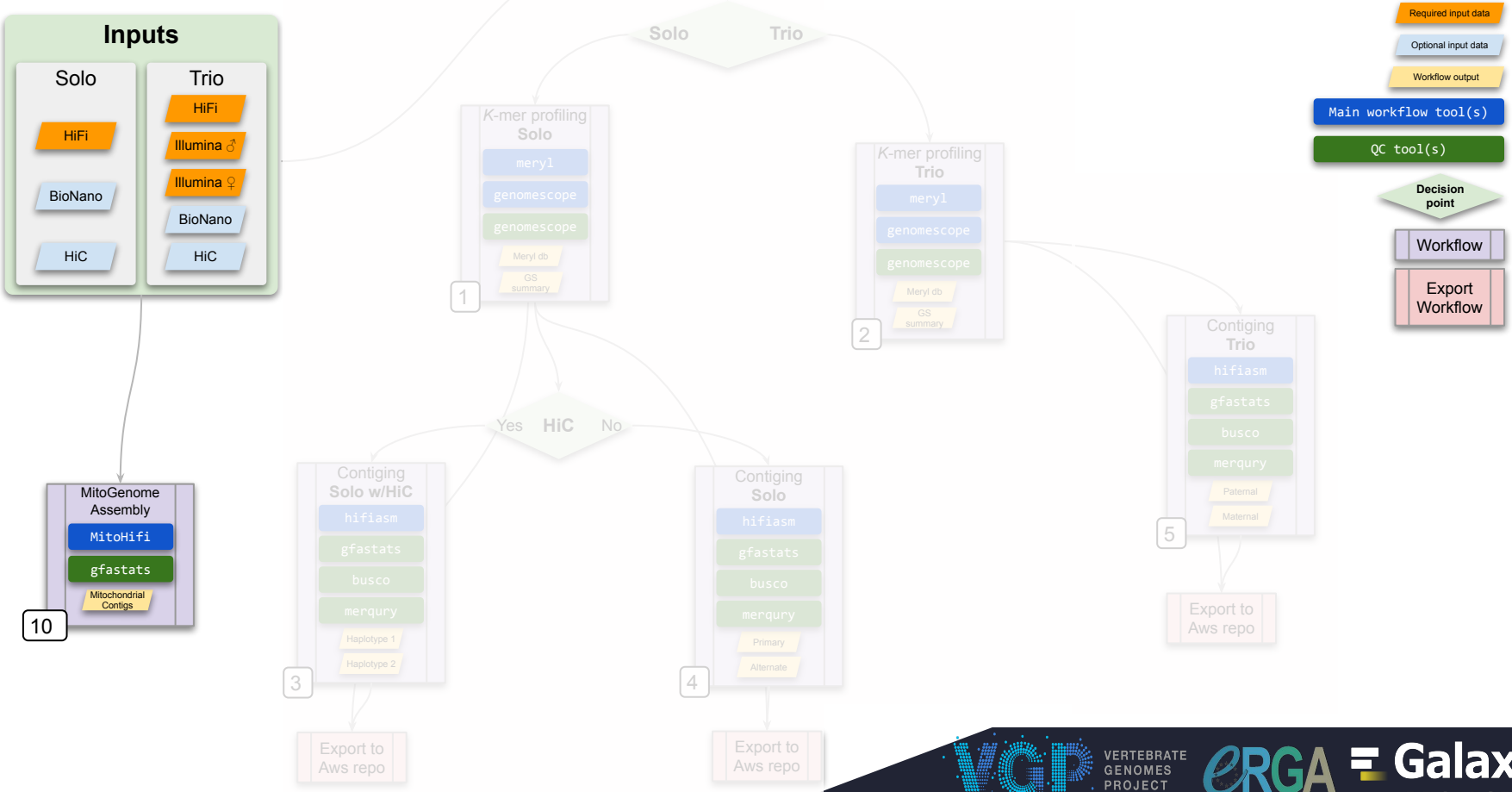
Contiging



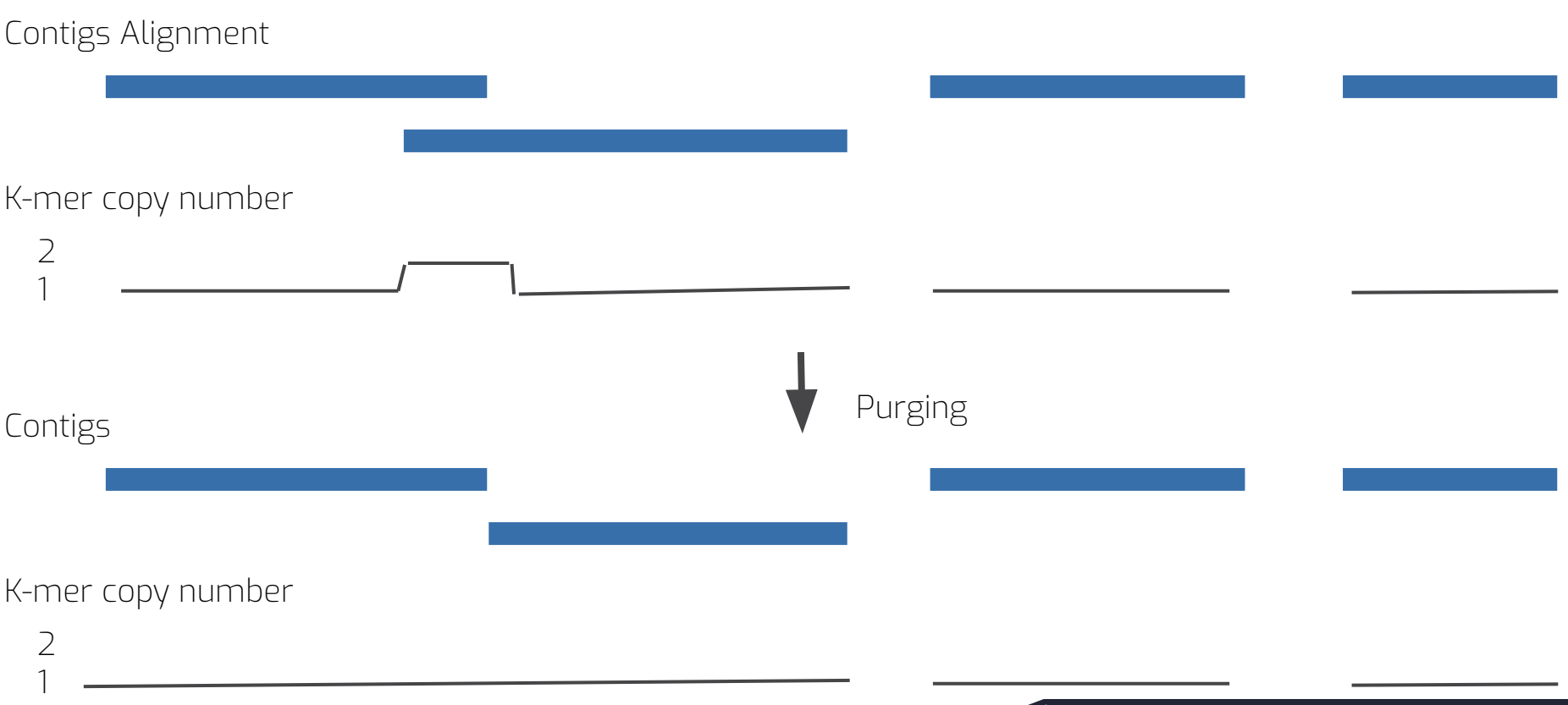
Contiging


Inputs
Solo
HiFi
BioNano
HiC
Trio
HiFi
Illumina ♂
Illumina ♀
BioNano
HiC
Solo
Trio
K-mer profiling Solo
meryl
genomescope
genomescope
Meryl db
GS summary
1
K-mer profiling Trio
meryl
genomescope
genomescope
Meryl db
GS summary
2
Contiging Trio
hifiasm
gfastats
busco
merqury
Paternal
Maternal
5
Export to Aws repo
Yes
HiC
No
Contiging Solo w/HiC
hifiasm
gfastats
busco
merqury
Haplotype 1
Haplotype 2
3
Export to Aws repo
Contiging Solo
hifiasm
gfastats
busco
merqury
Primary
Alternate
4
Export to Aws repo
MitoGenome Assembly
MitoHifi
gfastats
Mitochondrial Contigs
10
Required input data
Optional input data
Workflow output
Main workflow tool(s)
QC tool(s)
Decision point
Workflow
Export Workflow

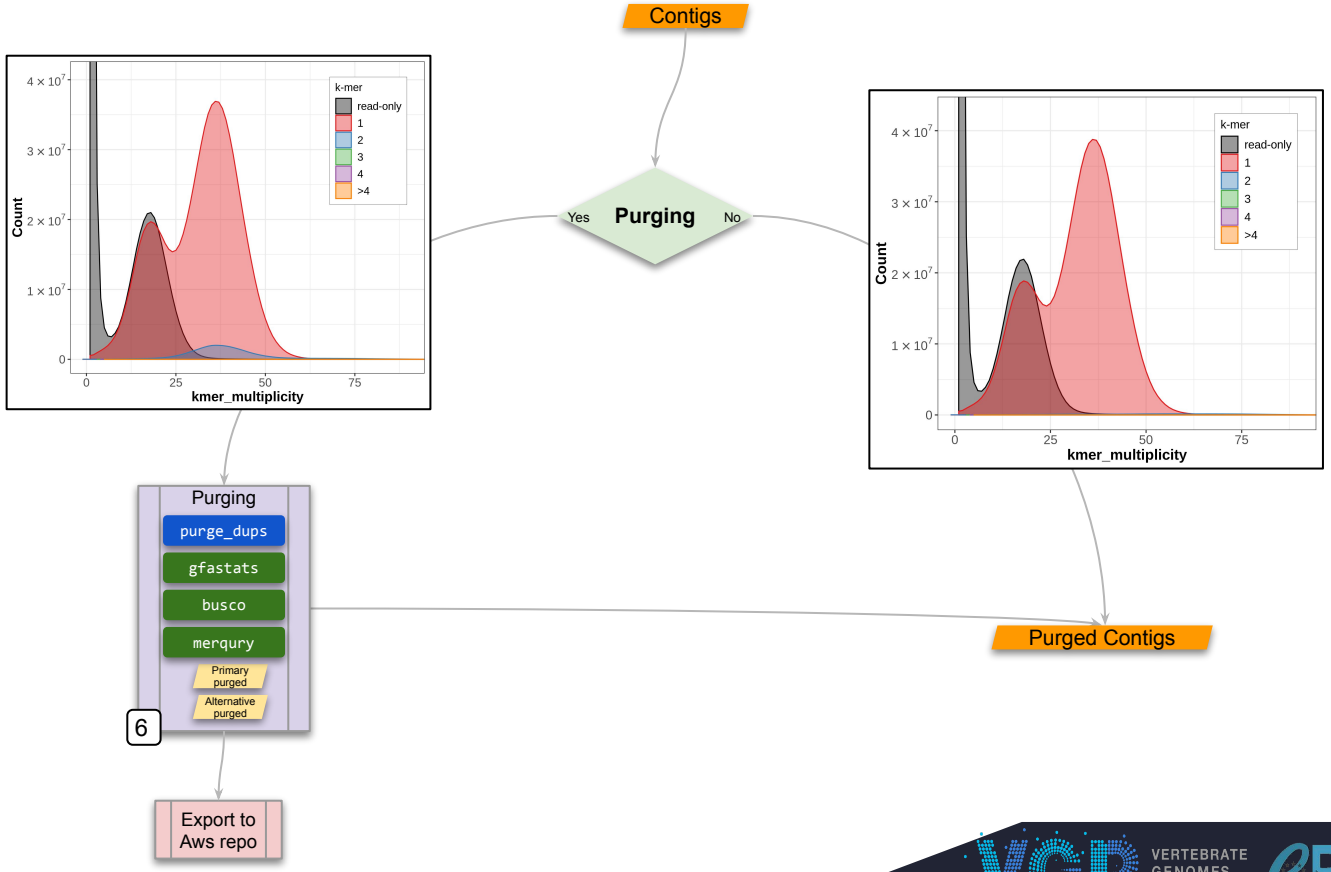
Contiging



Purging



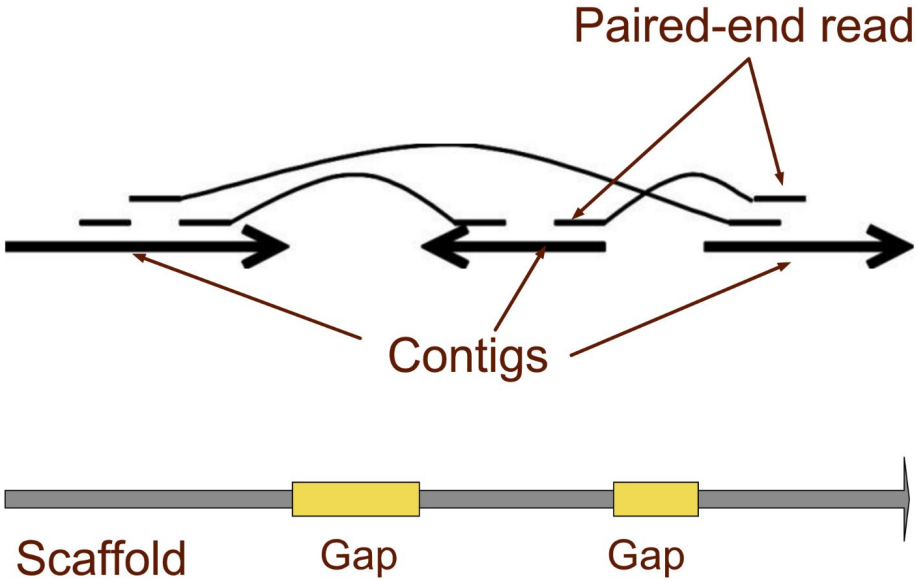
Contiging - purging



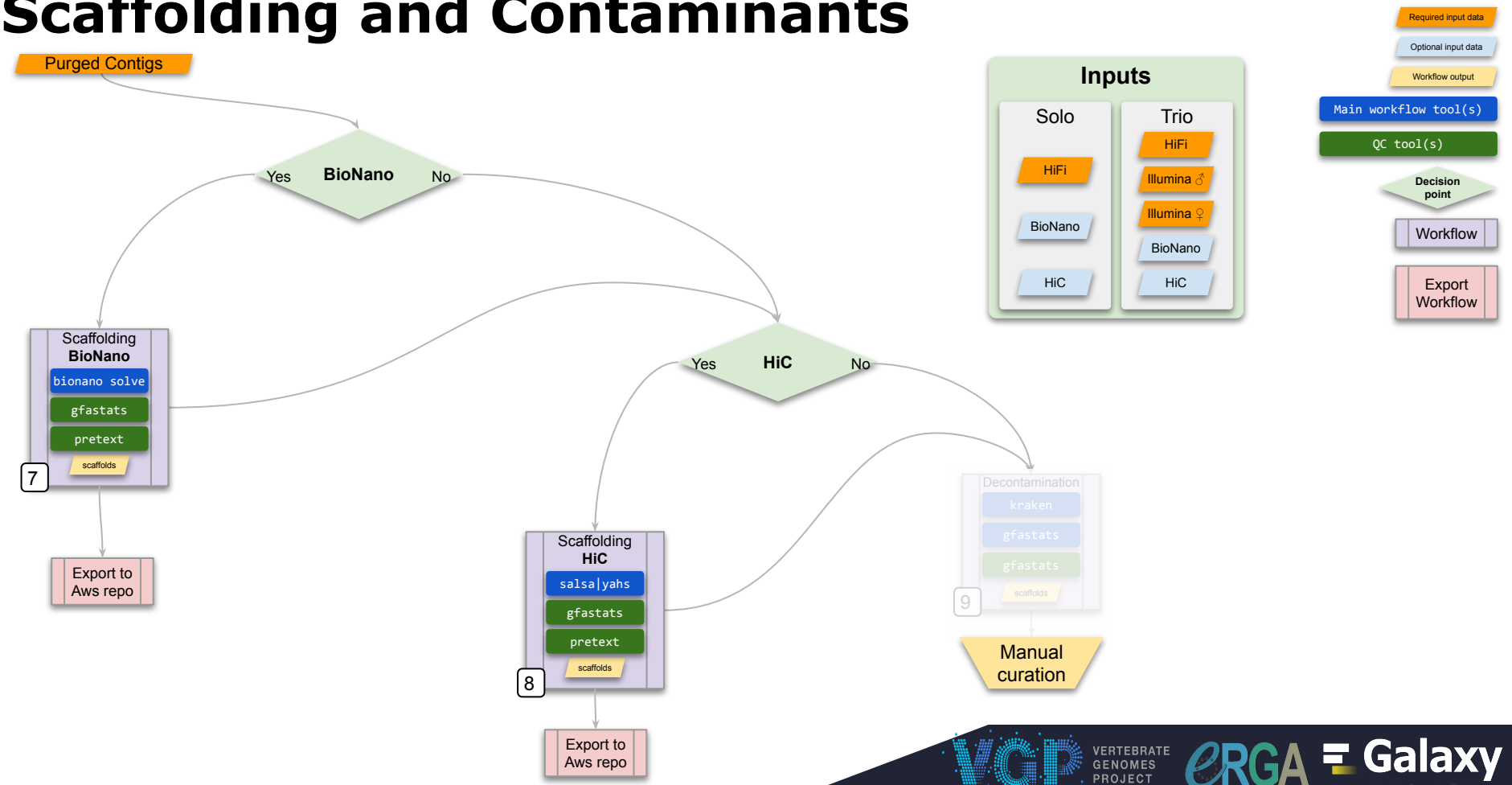
Required input data
Optional input data
Workflow output
Main workflow tool(s)
QC tool(s)
Decision point
Workflow
Export Workflow

VERTEBRATE GENOMES PROJECT
A PROJECT OF THE G10K CONSORTIUM
ERGA
European Reference Genome Atlas
Galaxy PROJECT

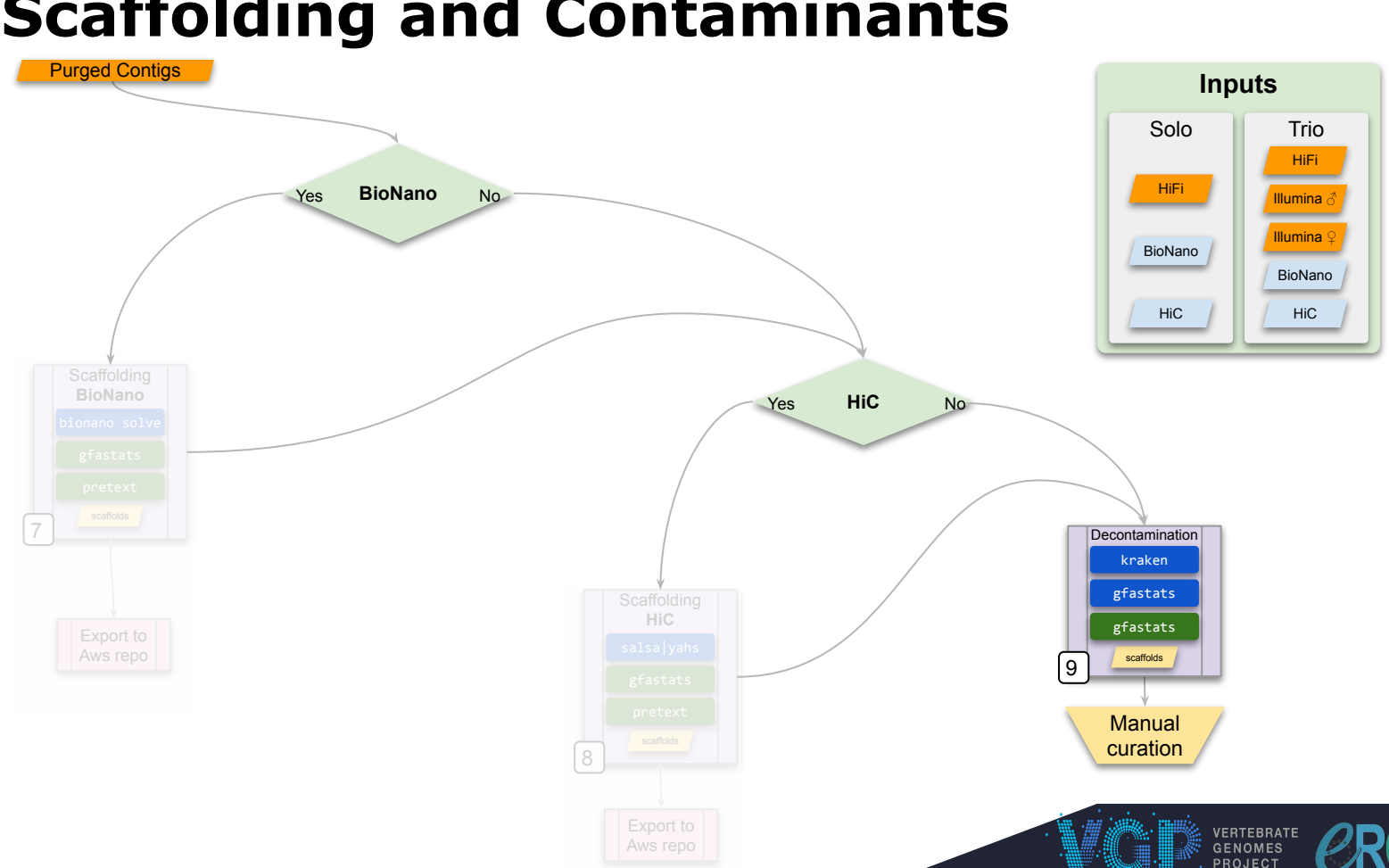
Scaffolding



Scaffolding and Contaminants

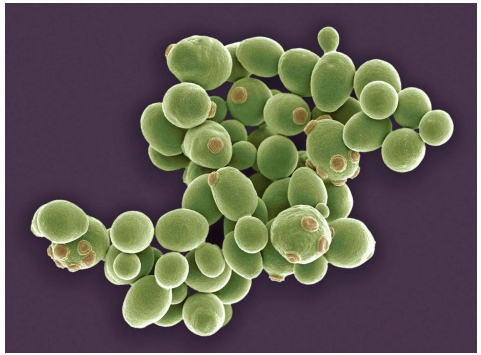


Scaffolding and Contaminants



Hands-on!

Species for Today



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

Model for eukaryotes

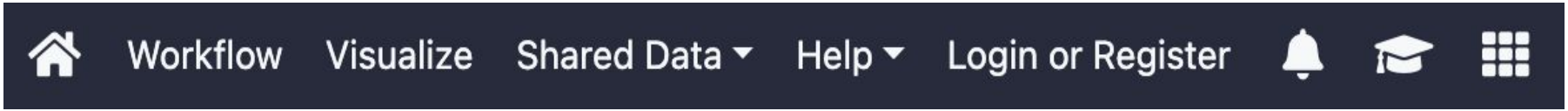
30x HiFi (SRR13577846)
10x HiC (SRR7126301)

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



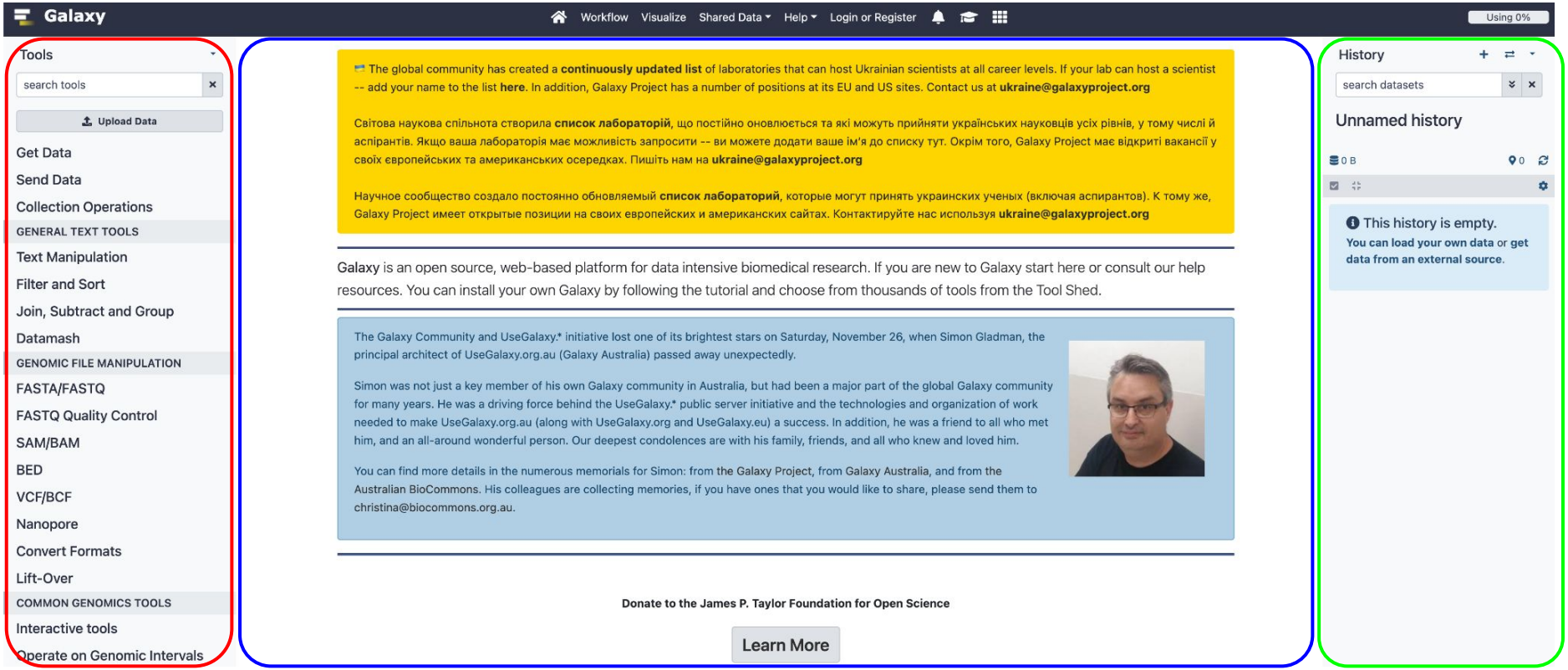
1. Create Account

https://usegalaxy.org

2. Join Training

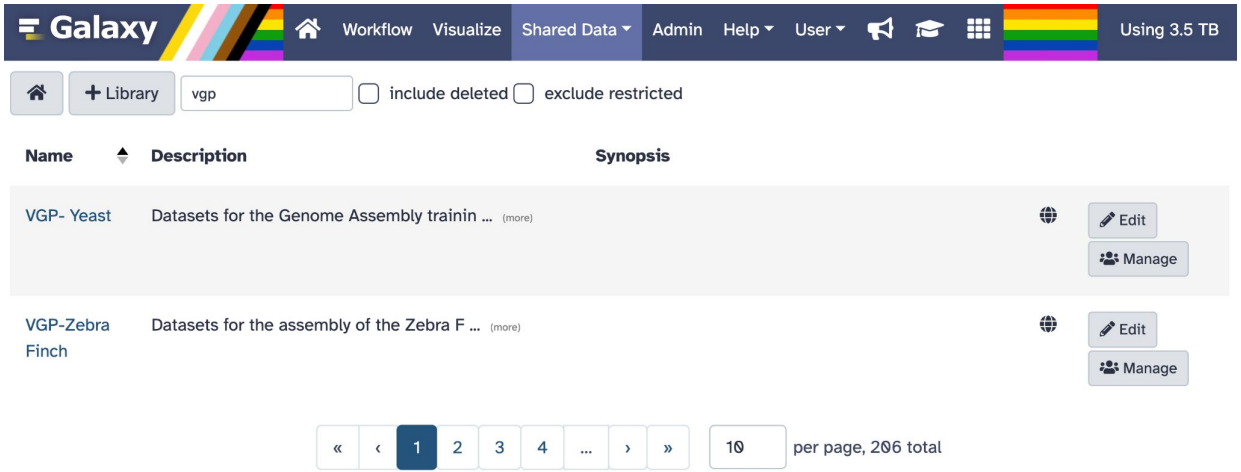
https://usegalaxy.org/join-training/vgpbga2023

Using Galaxy

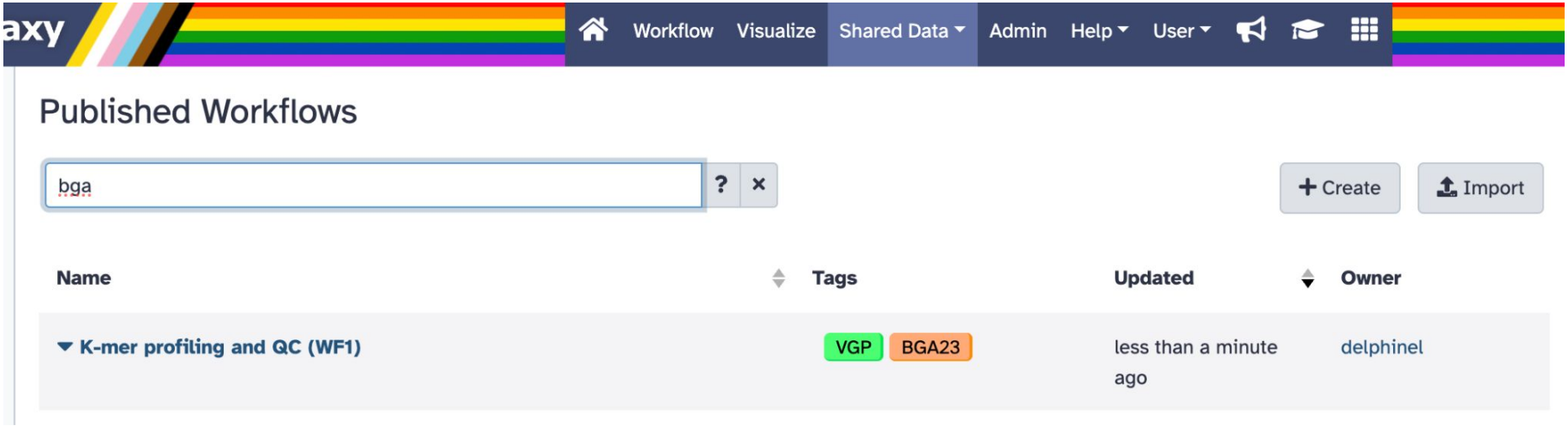


Importing Data from data libraries

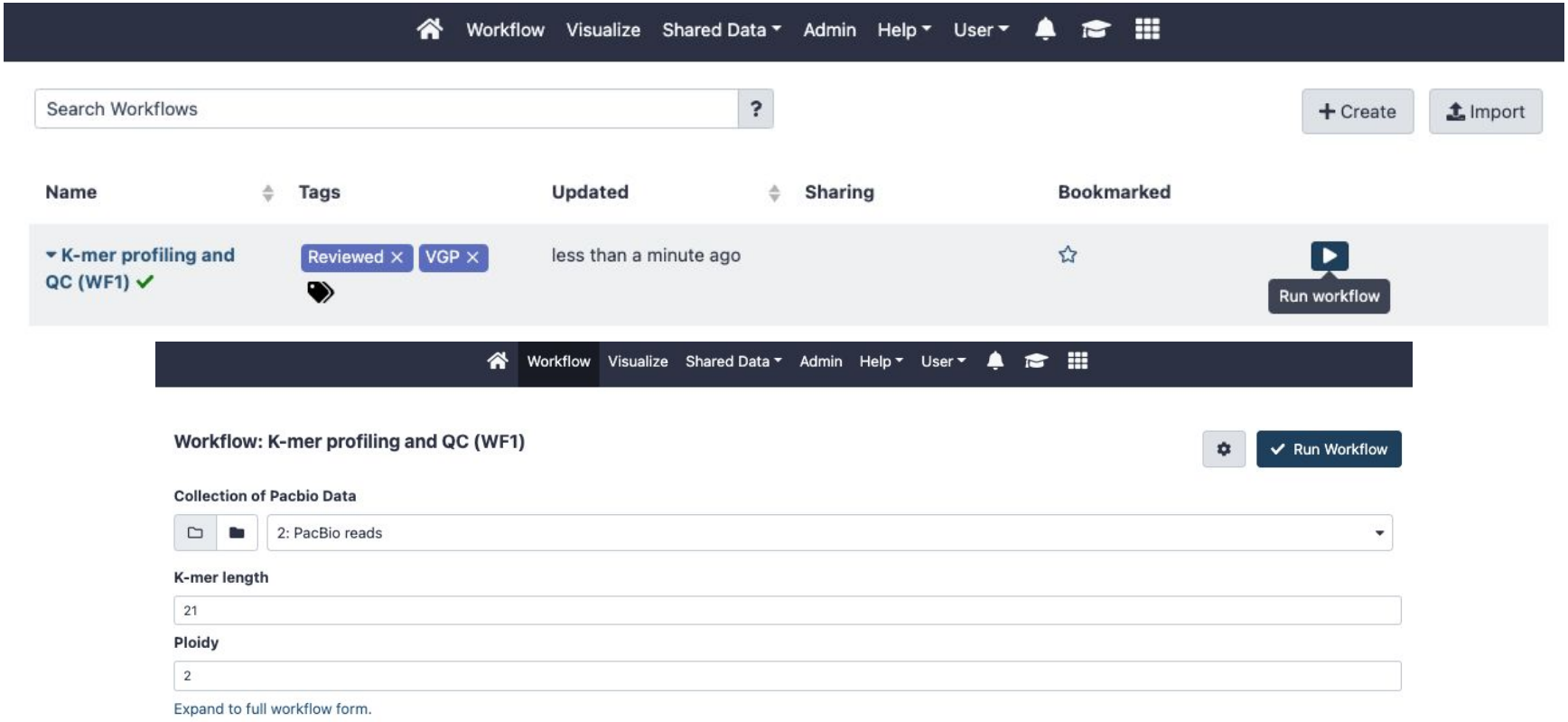
- Hands-on Data : Yeast



Importing Workflows : Public workflows



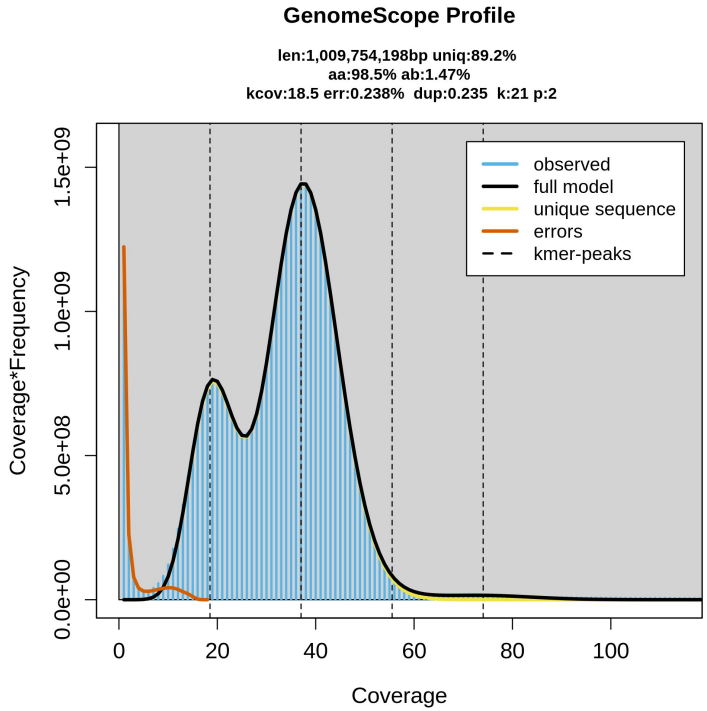
WF1: HiFi reads-based Kmer-counting



VERTEBRATE GENOMES PROJECT
A PROJECT OF THE G10K CONSORTIUM
ERGA
European Reference Genome Atlas
Galaxy PROJECT

WF1: Outputs

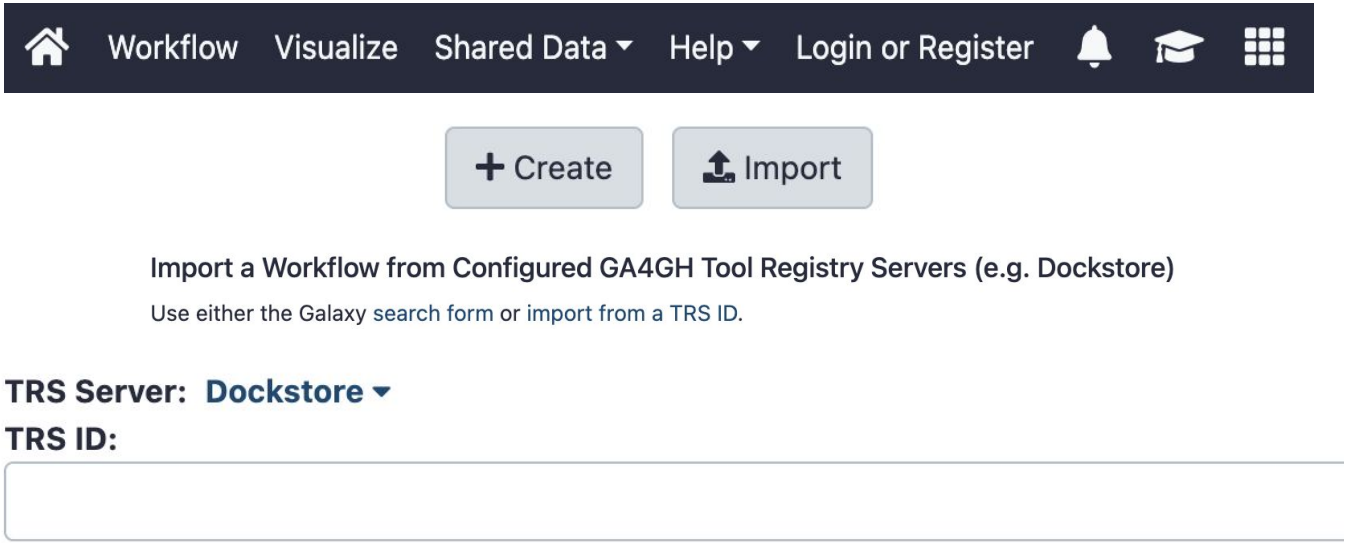
- Kmer profiling Zebra Finch



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

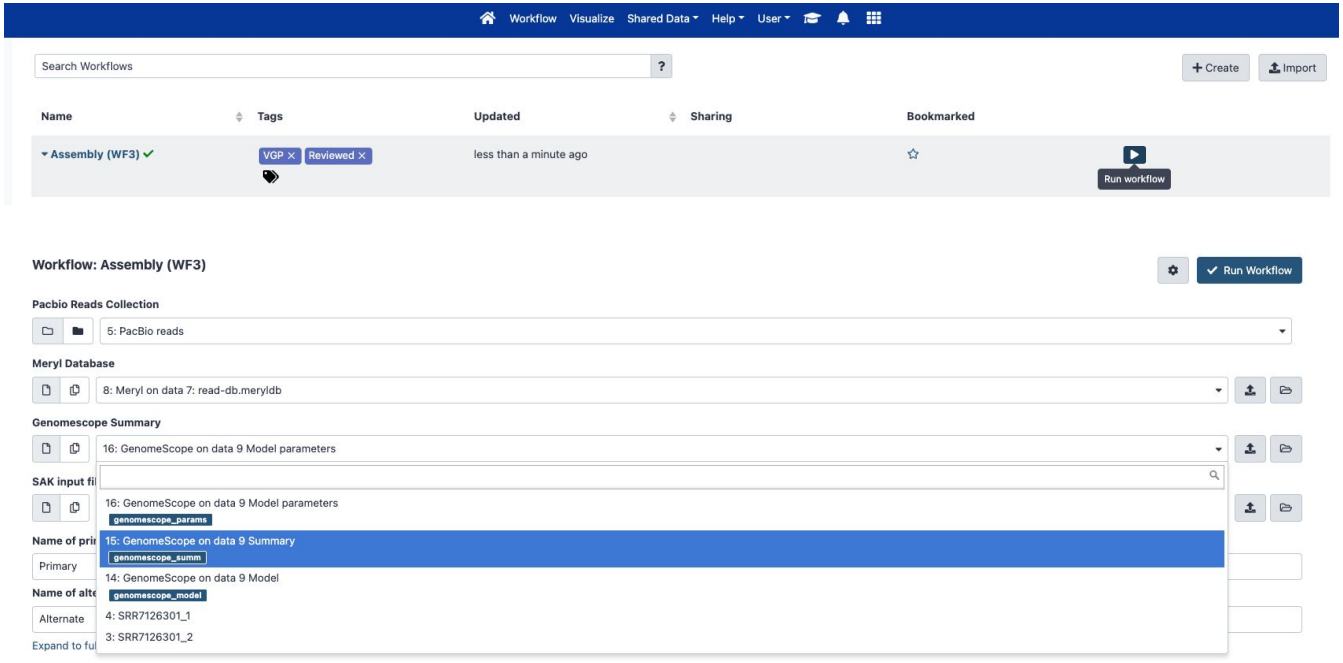
property	min	max
Homozygous (aa)	98.5256%	98.5313%
Heterozygous (ab)	1.4687%	1.47441%
Genome Haploid Length	1,009,402,687 bp	1,009,754,198 bp
Genome Repeat Length	109,282,859 bp	109,320,915 bp
Genome Unique Length	900,119,828 bp	900,433,283 bp
Model Fit	89.557%	99.2734%
Read Error Rate	0.237557%	0.237557%

Importing Workflows : From Dockstore



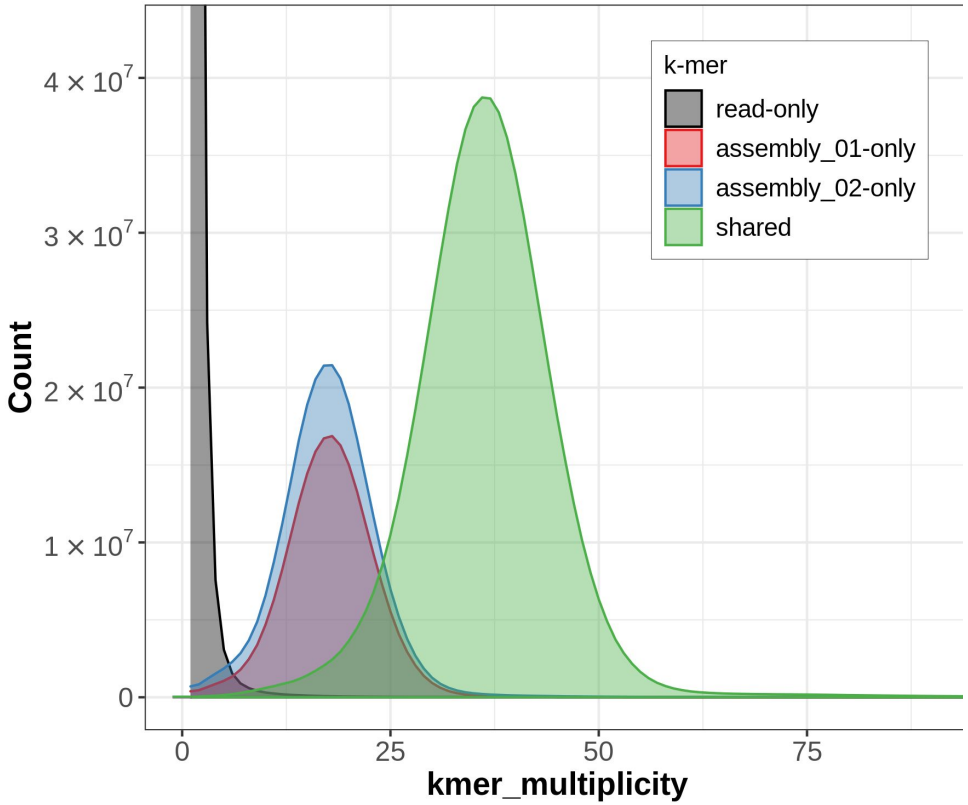
#workflow/github.com/Delphine-L/iwc/WF3-Assembly

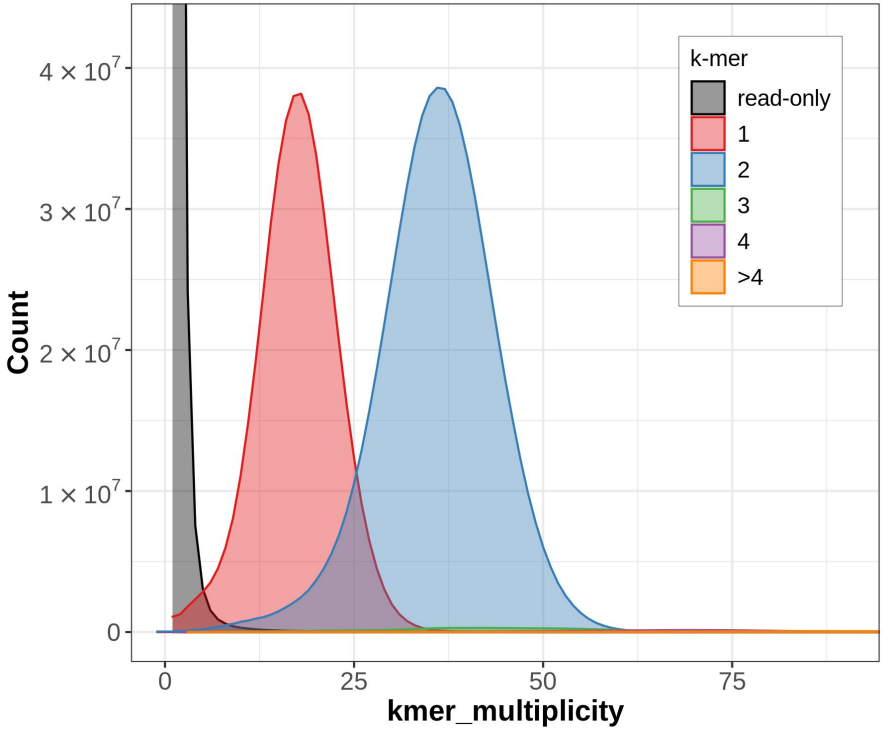
WF3-5: Contig assembly



WF3-5: Outputs

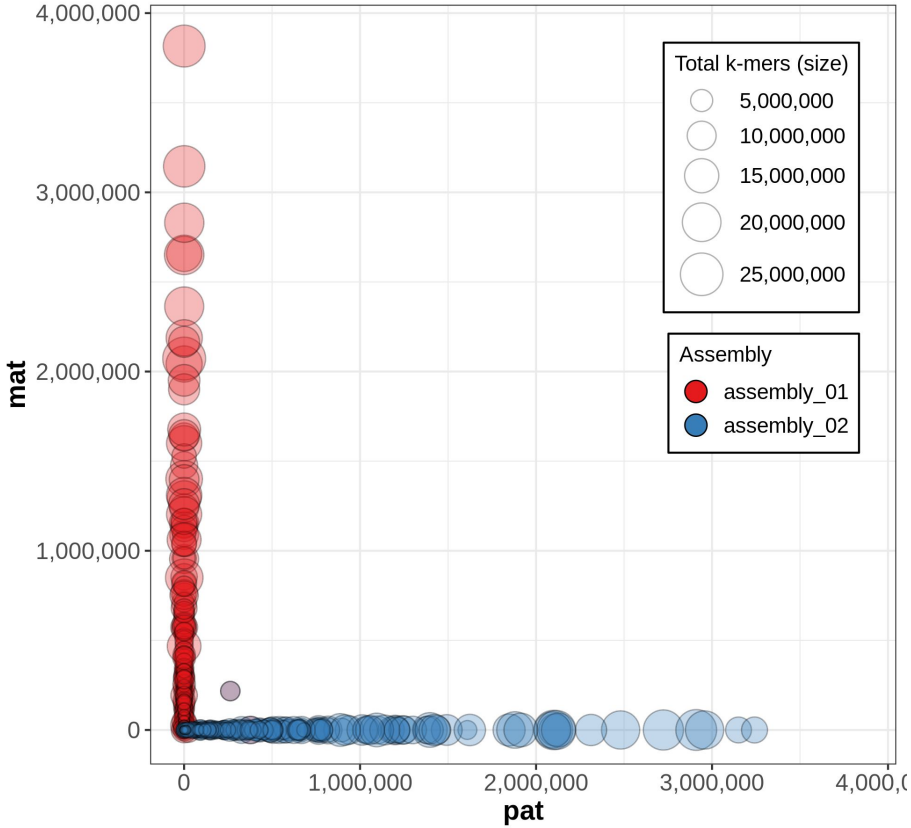
Assembly with HiC phasing Zebra Finch





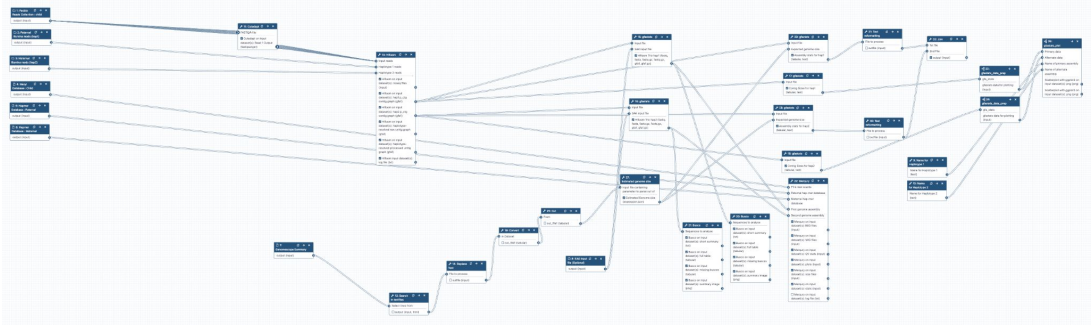
WF5: Outputs

Trio Assembly Zebra Finch



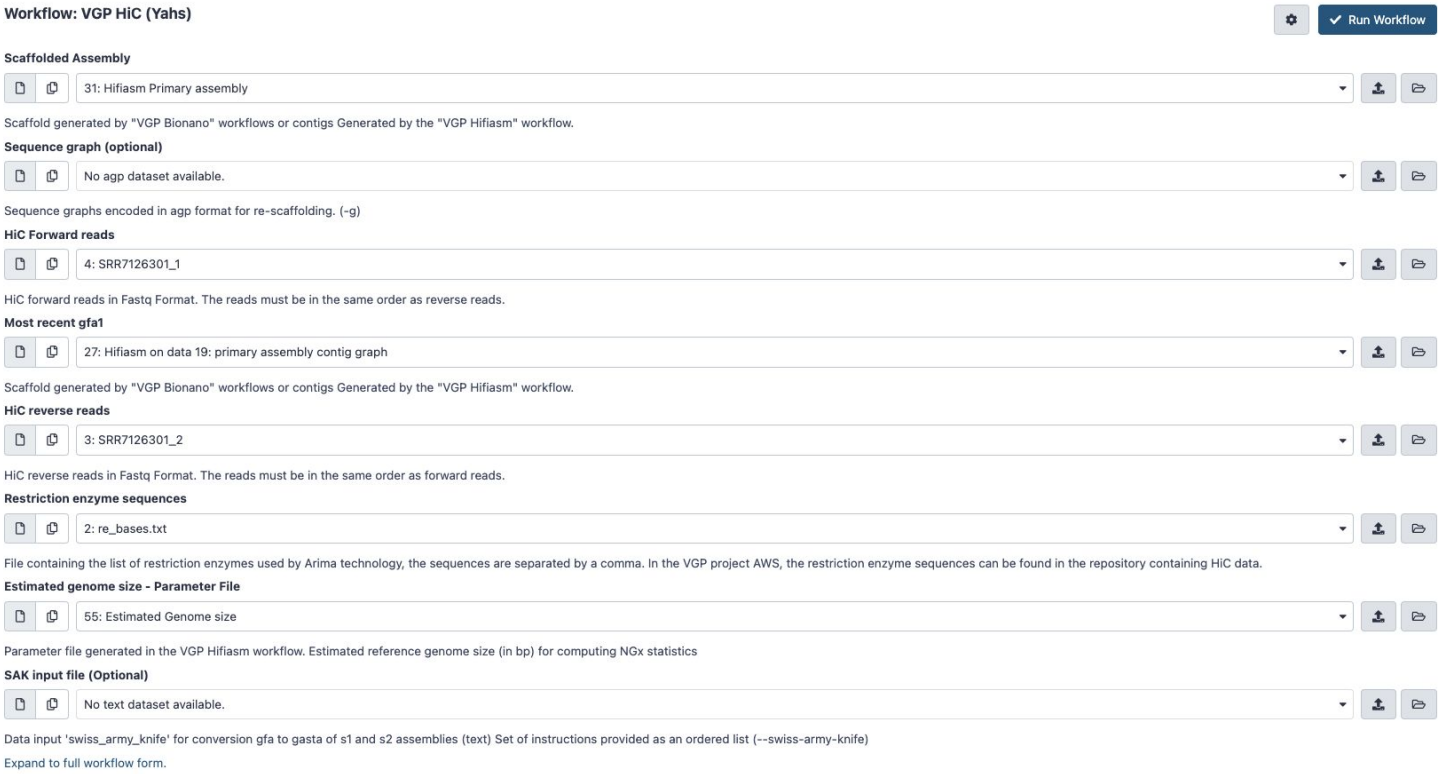
WF7-8: Scaffolding

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



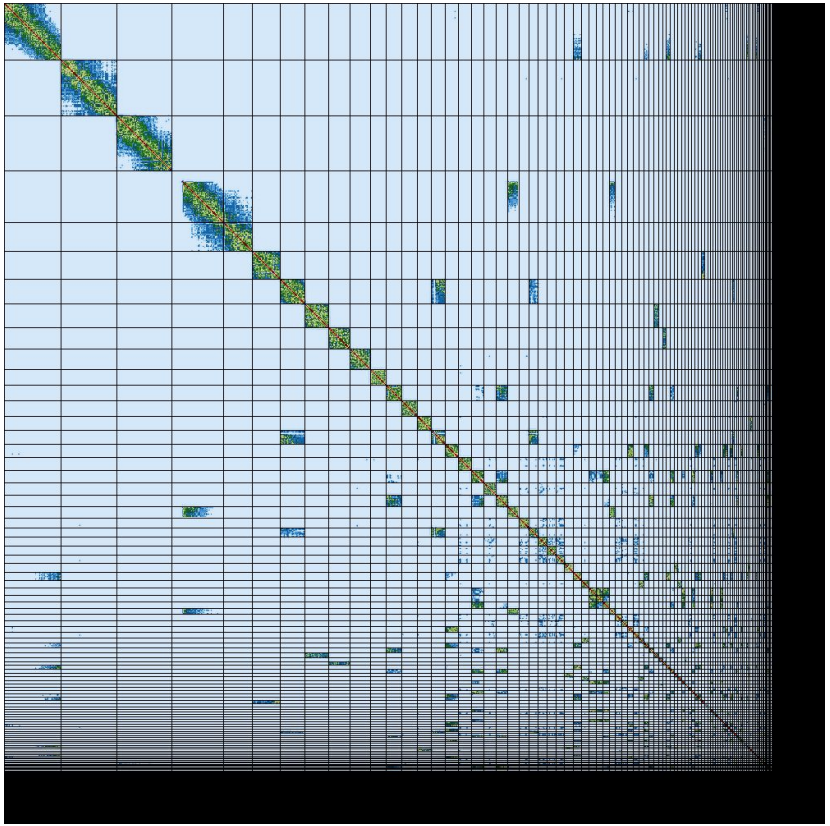
- If Bionano data:
  - #workflow/github.com/Delphine-L/iwc/WF7-Scaffolding_Bionano
  - Tools
    - Bionano solve
    - gfastats
    - Pretext

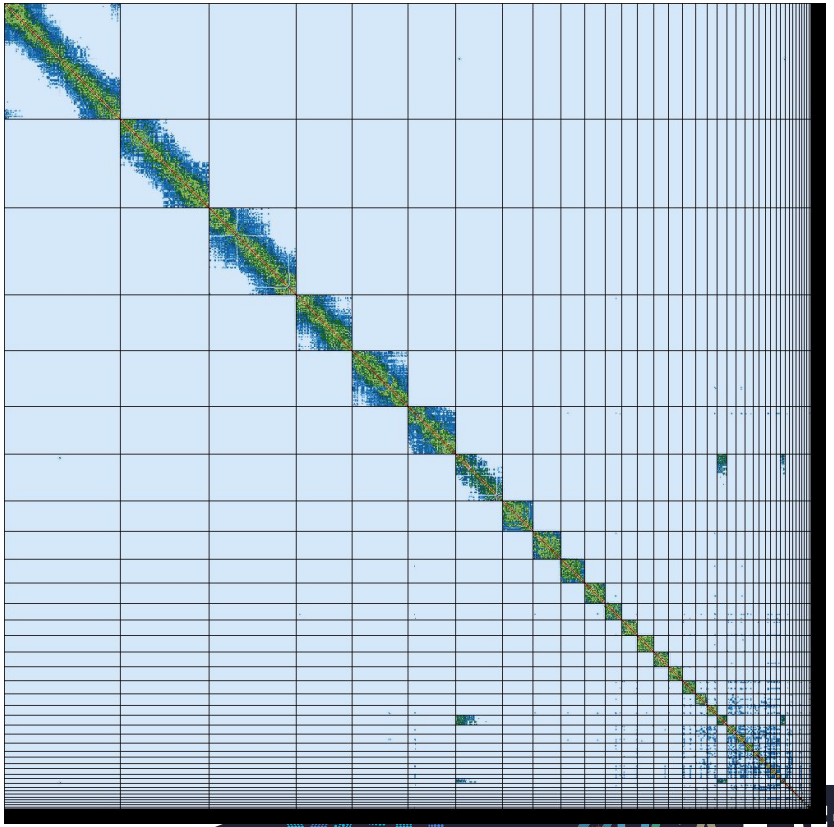
WF7-8: Scaffolding



WF7-8: Outputs

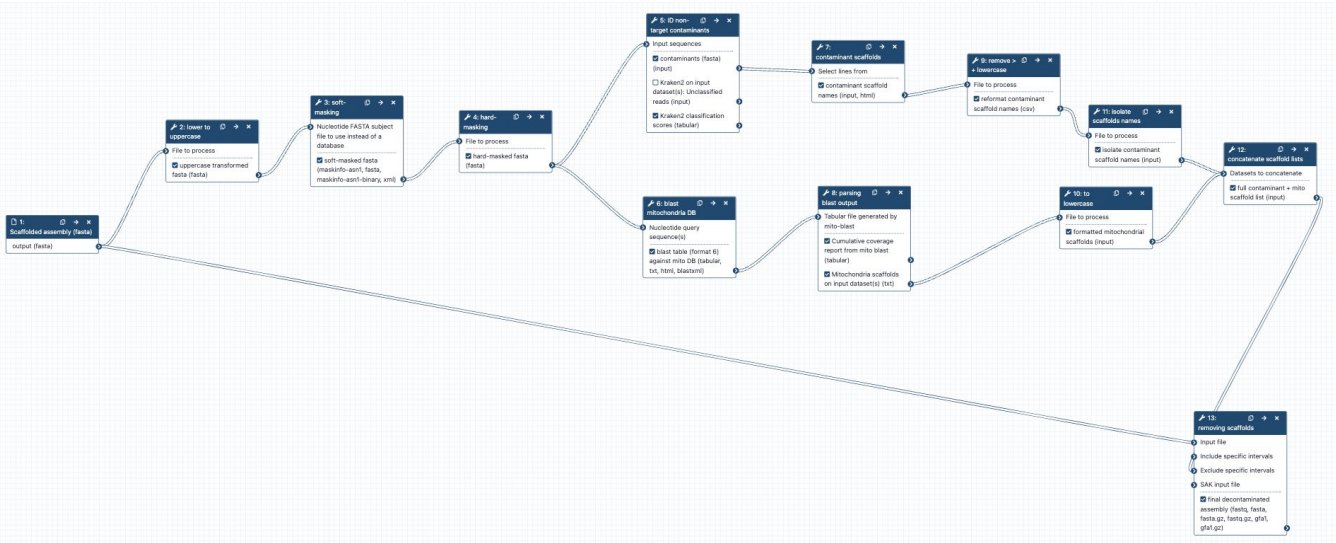
HiC Scaffolding Zebra Finch





WF9: Decontamination (optional)

- TRS ID: #workflow/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

https://galaxyproject.org/

Thank you!


VGP
VERTEBRATE
GENOMES
PROJECT
A PROJECT OF THE G10K CONSORTIUM
eRGA
European Reference Genome Atlas
Galaxy
PROJECT