

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 shows the Galaxy web interface with the following details:

- General Statistics:** A table showing statistics for various FASTQ files. The columns include Sample Name, % Duplication, % > Q30, Mb Q30 bases, GC content, % PF, and % Adapter.
- Tools Sidebar:** Lists various tools categorized under GENERAL TEXT TOOLS, GENOMIC FILE MANIPULATION, and COMMON GENOMICS TOOLS.
- History Sidebar:** Shows a list of recent items, including "An example 1" (17 shown, 259 deleted, 1074 hid), "712: Realigned reads" (a list with 63 items), "712: Filter SAM or BAM, out", "648: Map with BWA-MEM on collection 128 (mapped reads in AM format)", "647: pOUT18.4.fa", "388: pOUT18.4.gb", "384: MultiQC on data 38", "383: MultiQC on data 382, 378, and others: Stats", "130: fastp on collection 127 ON report", "129: fastp on collection 127 ML report", "128: fastp on collection 127 red-end output", and "127: pilot" (a list of pairs with 63 items).
- Plot:** A horizontal bar chart titled "Fastp: Filtered Reads" showing the length distribution of filtered reads for various samples.

Accessible, Reproducible and Collaborative

The figure consists of three side-by-side screenshots of the Galaxy web interface:

- General Statistics:** Shows a table of sample names, their duplication percentages, and file sizes. It includes a search bar and a "History" tab.
- Many Analyses:** Shows a list of 30 analysis history items, each with a preview icon and a delete button. The items include various bioinformatics tools and datasets, such as ITCR 2022, ASHG 2020 fq, Galaxy-intro, and several FastQC and bcfTools analyses.
- Collaborative Data & Workflows:** Shows a detailed workflow history titled "Workflow constructed from history 'ASHG 2022'". The workflow consists of multiple steps: Input, Get Data, Send Data, Data�ip, Bioinformatics Tools (including FASTQ/FASTA, Quality Control, and Variant Calling), and Output. Annotations and notes are visible on the right side of the workflow steps.

One Analysis

Many Analyses

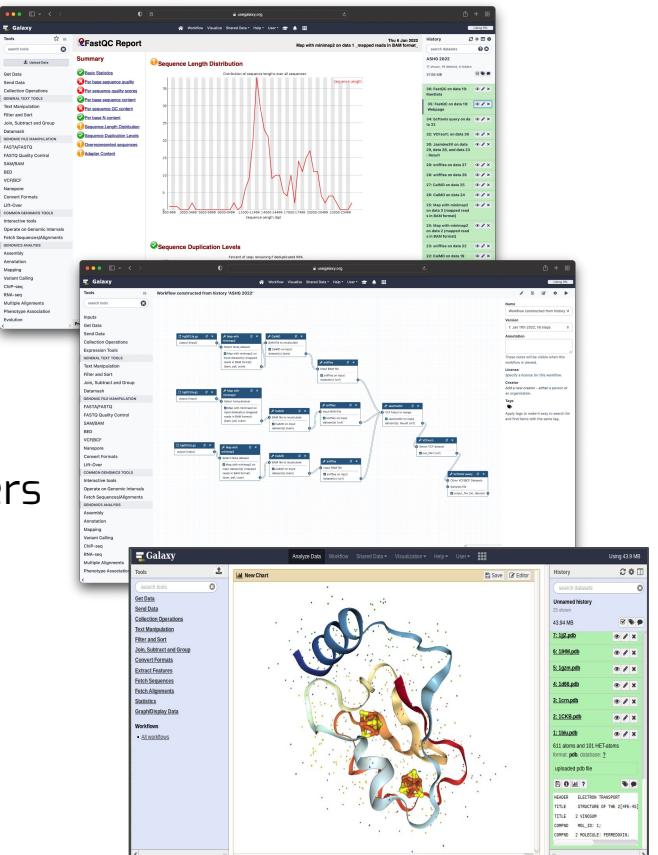
Collaborative
Data & Workflows

(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



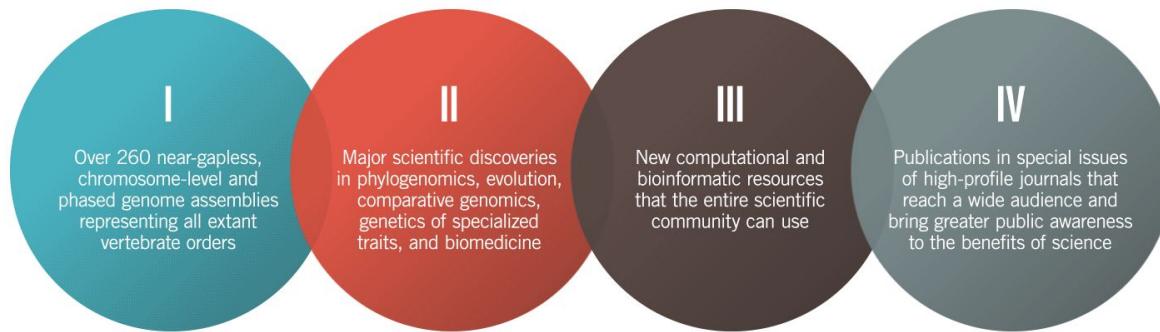
The Vertebrate Genomes Project

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

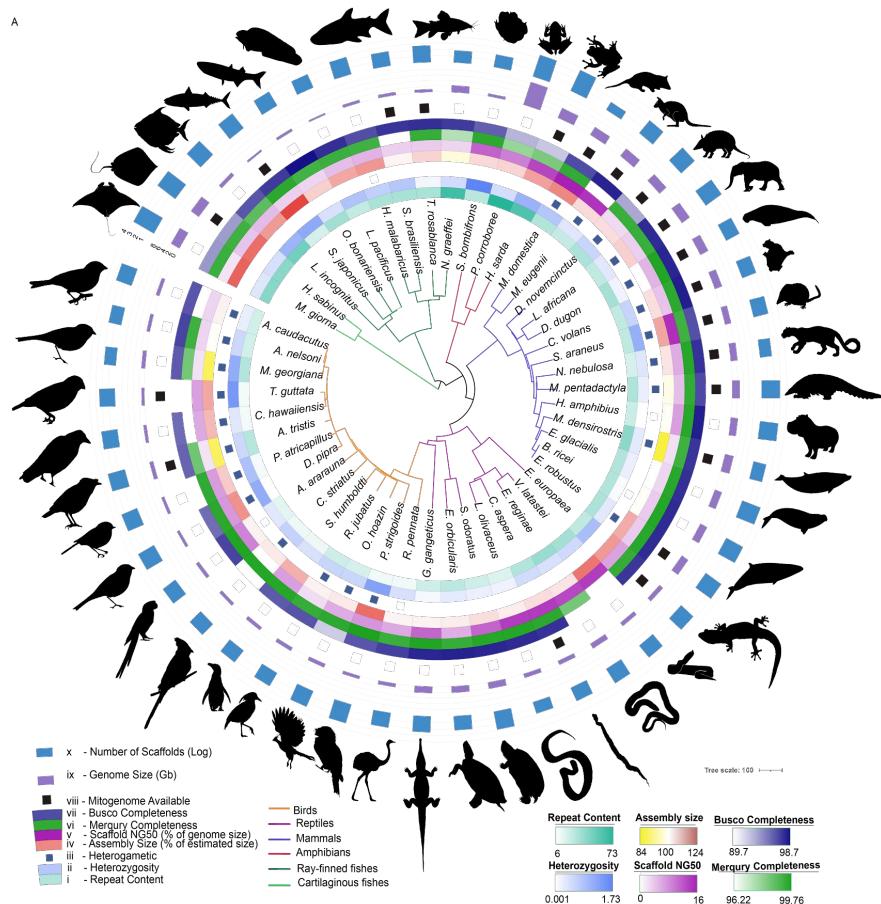
Number of species at each level of completion.

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



VGP assembly Pipeline v2

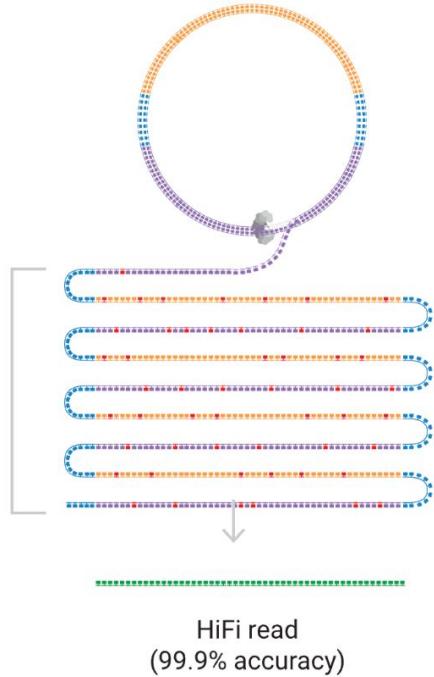
- Developed in Galaxy
- Available on global instances :
 - .eu
 - .org
 - .org.au (soon)
- Dozens of genomes assembled
 - Hundreds to thousands planned in the coming year
- Technologies :
 - PacBio HiFi
 - Bionano Cmap
 - Arima HiC



Data - PacBio HiFi



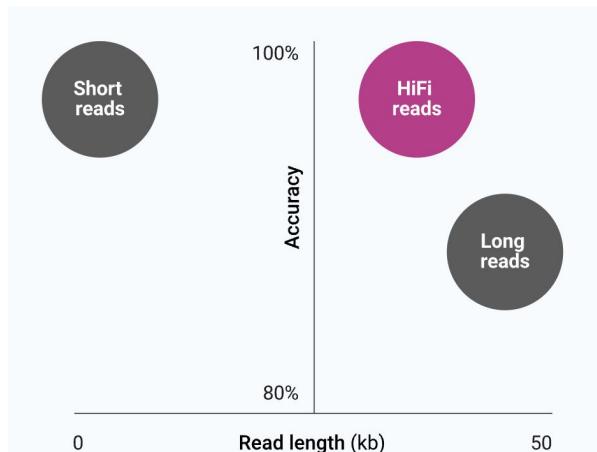
Circularized DNA
is sequenced in
repeated passes



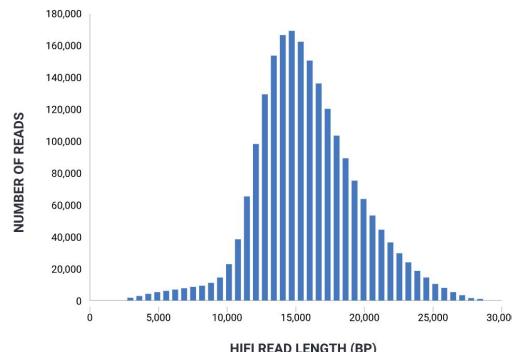
The polymerase reads
are trimmed of adapters
to yield subreads

Consensus and
methylation status are
called from subreads

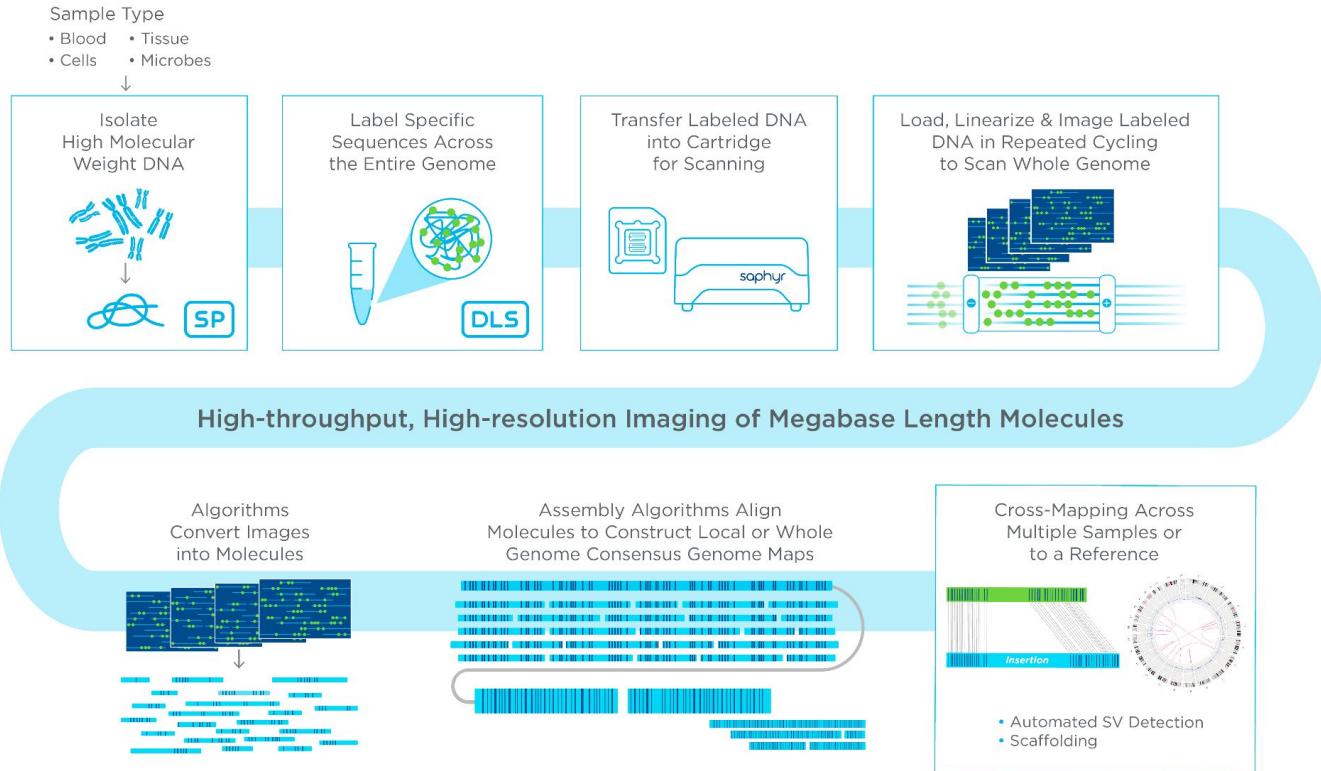
<https://www.pacb.com/technology/hifi-sequencing/>



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



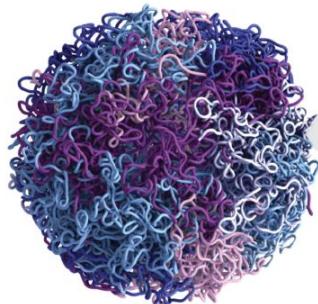
Data - Bionano Optical Mapping



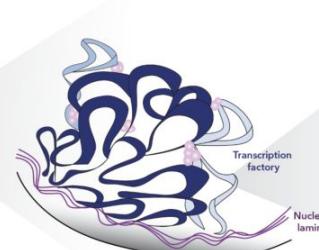
<https://bionanogenomics.com/technology/platform-technology/>

Data - Arima HiC

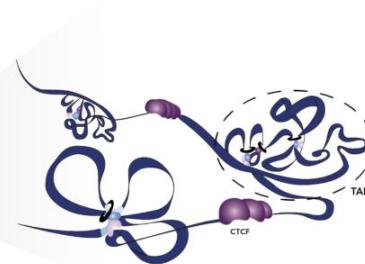
In the nucleus chromosomes are organized into chromosome territories



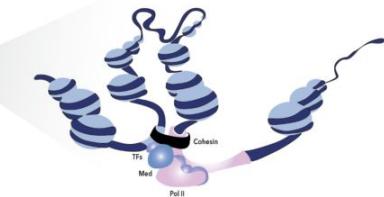
Chromosomes are divided into cell-specific A/B compartments



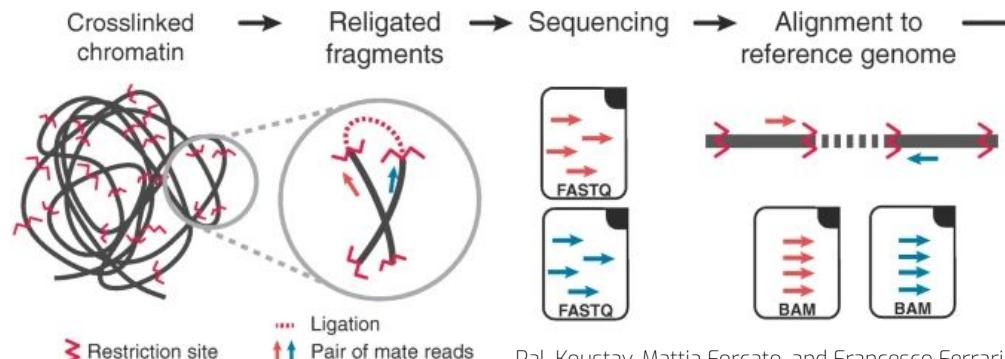
Compartments are organized into topologically associated domains (TADs)



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



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

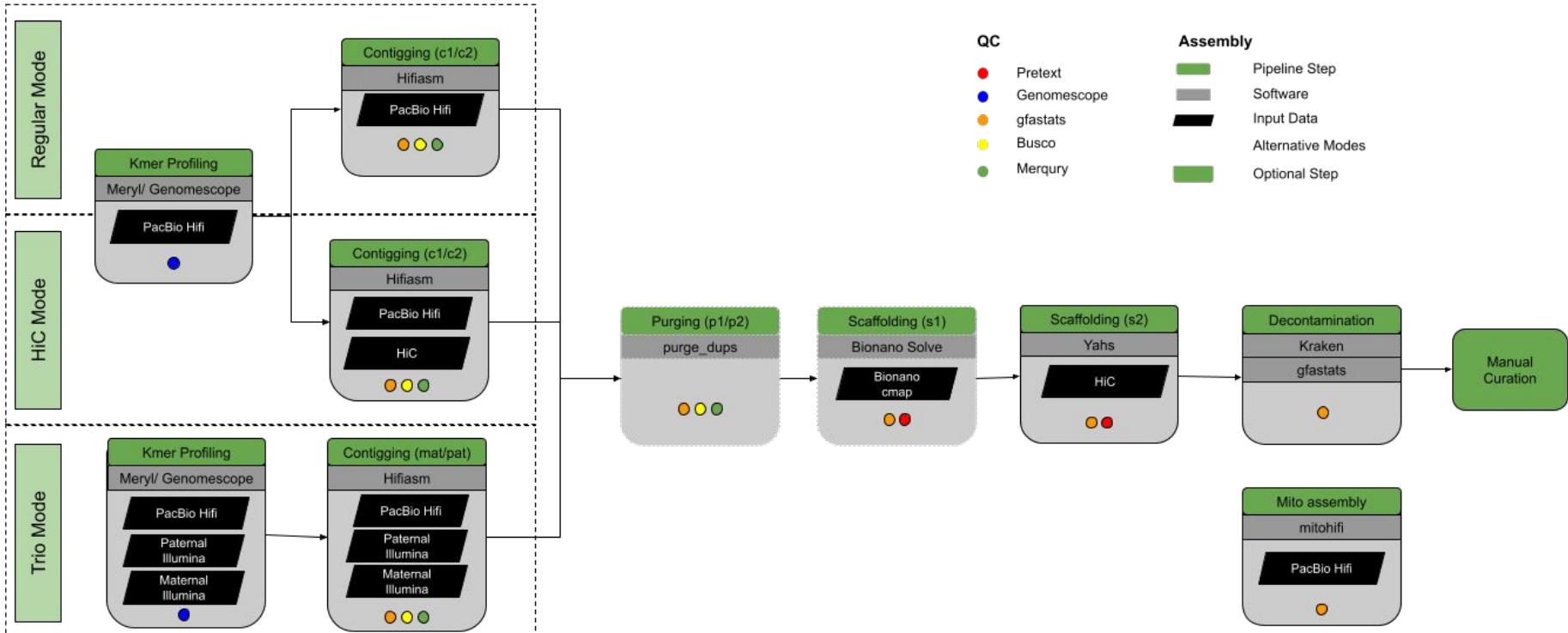


Pal, Koustav, Mattia Forcato, and Francesco Ferrari. "Hi-C analysis: from data to biological integration." *Biophysical reviews* 11.1 (2019): 67-78.



A PROJECT OF THE G10K CONSORTIUM

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.

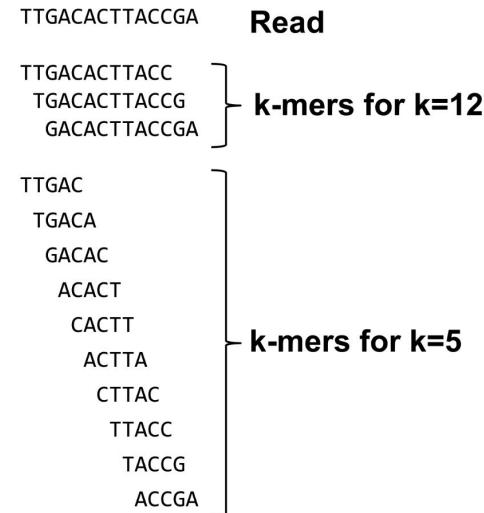
The latest GTN news

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: Workflow Examples

OPEN CHAT

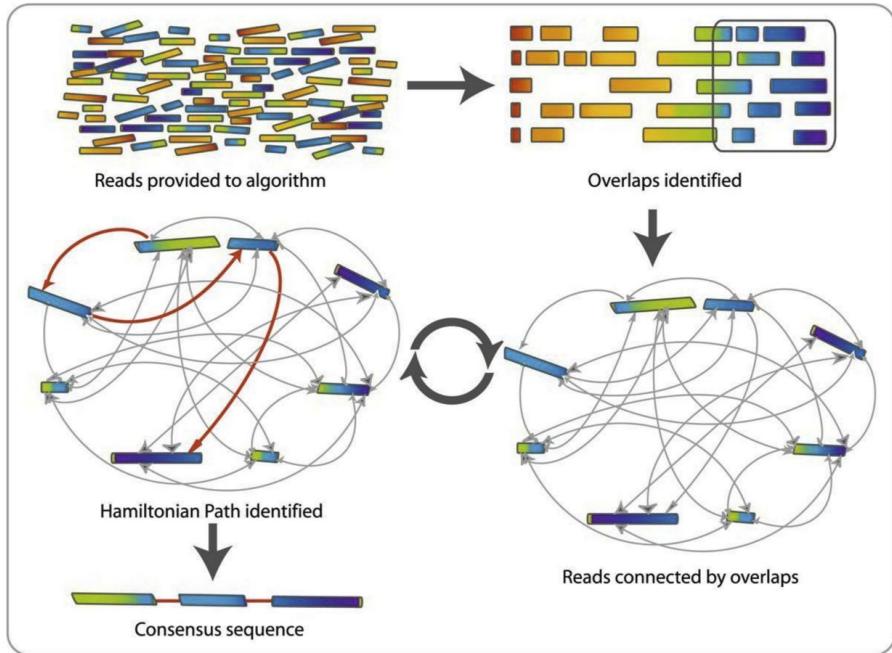
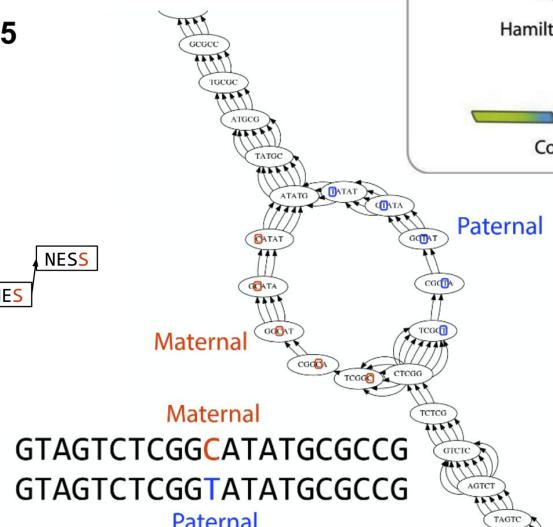
Contiging



$k = 4$ k-mers:

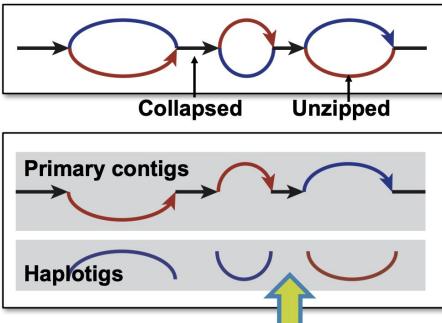
HAPP	PPIN
APP _I	PINE
PINE	PPIN
INES	NESS

HAPPINESS



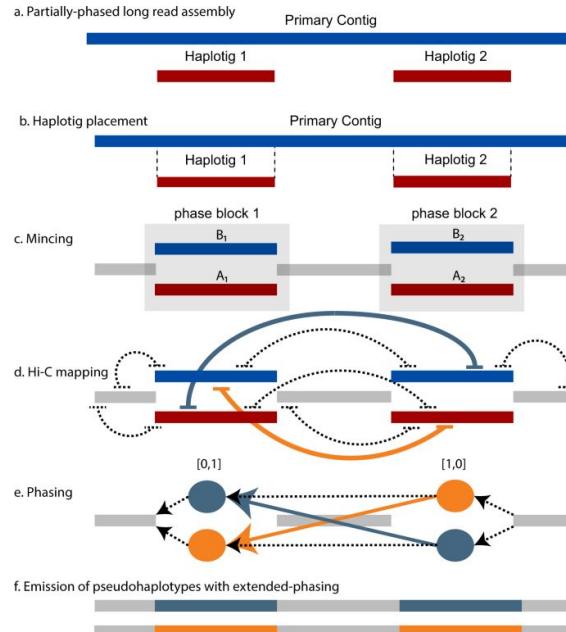
Phased Assemblies

Simple Phasing



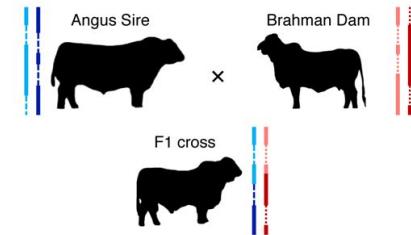
Chin, C.S. et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*. 13(12), 1050.

Hi-C-based Phasing



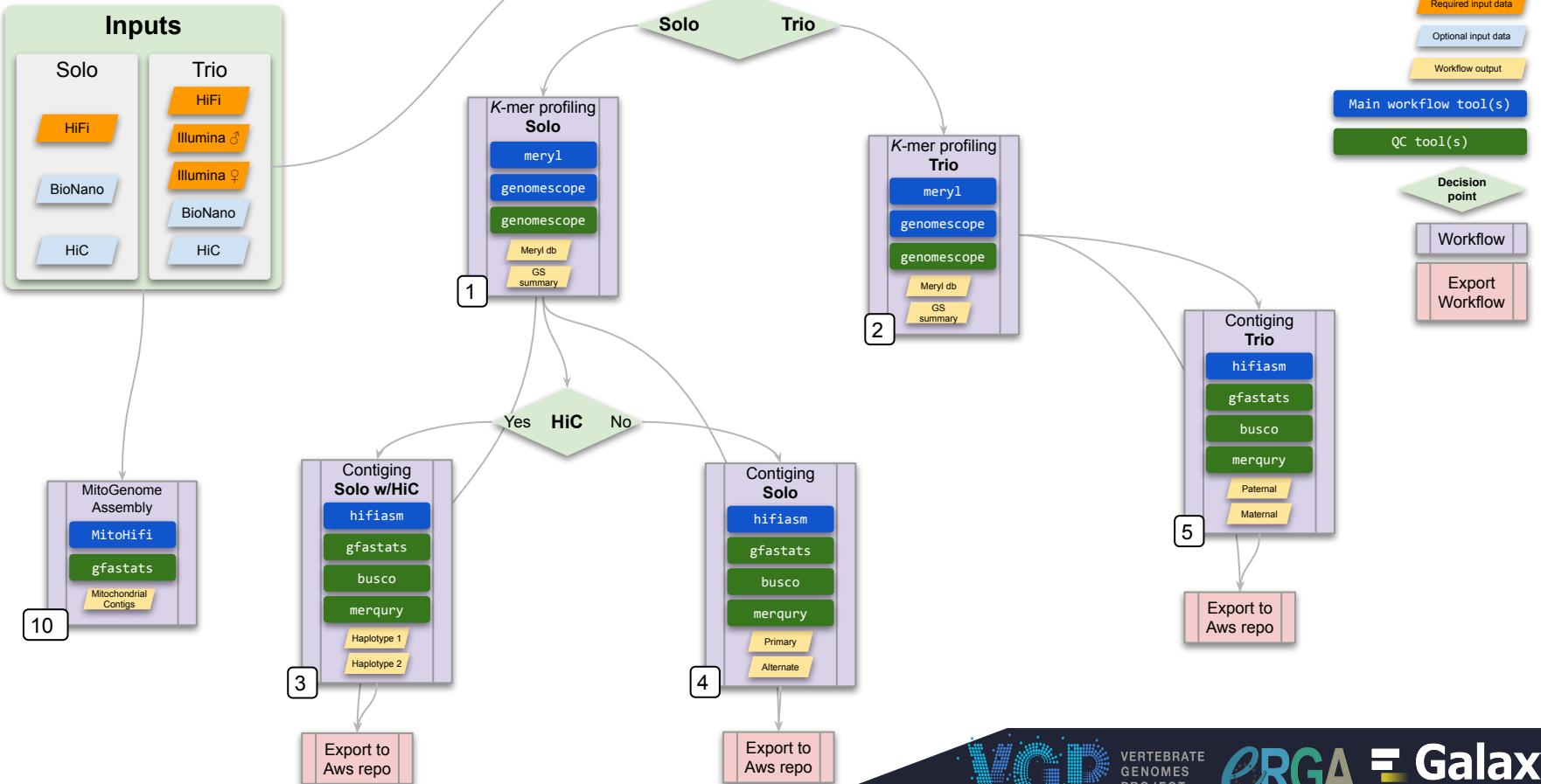
Kronenberg, Z. N., Rhee, A., Koren, S., Concepcion, G. T., Peluso, P., Munson, K. M., ... & Kingan, S. B. (2021). Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nature Communications*, 12(1), 1935.

Trio-based Phasing

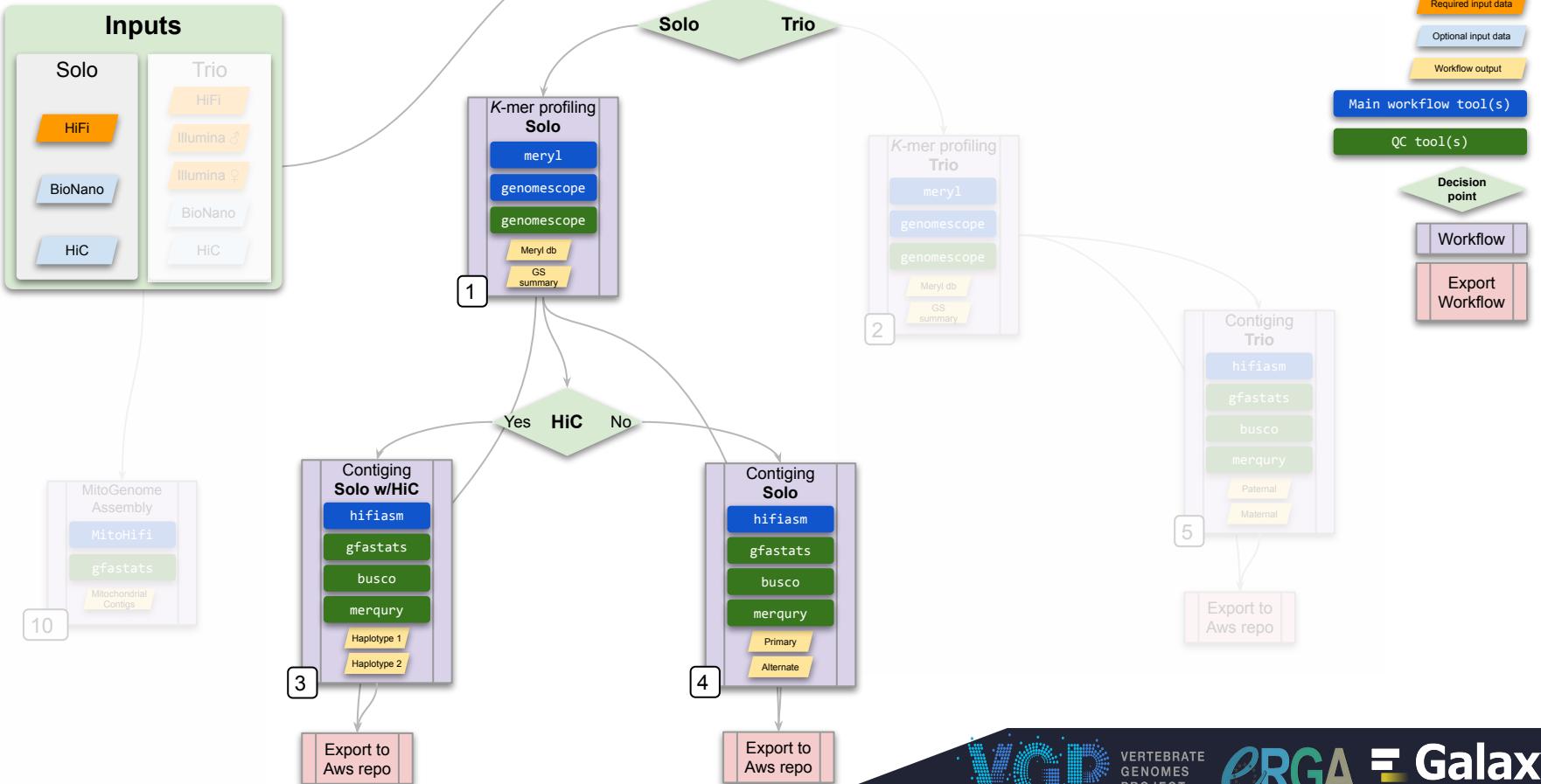


Koren et al. (2018). De novo assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology*. 10.1038/nbt.4277

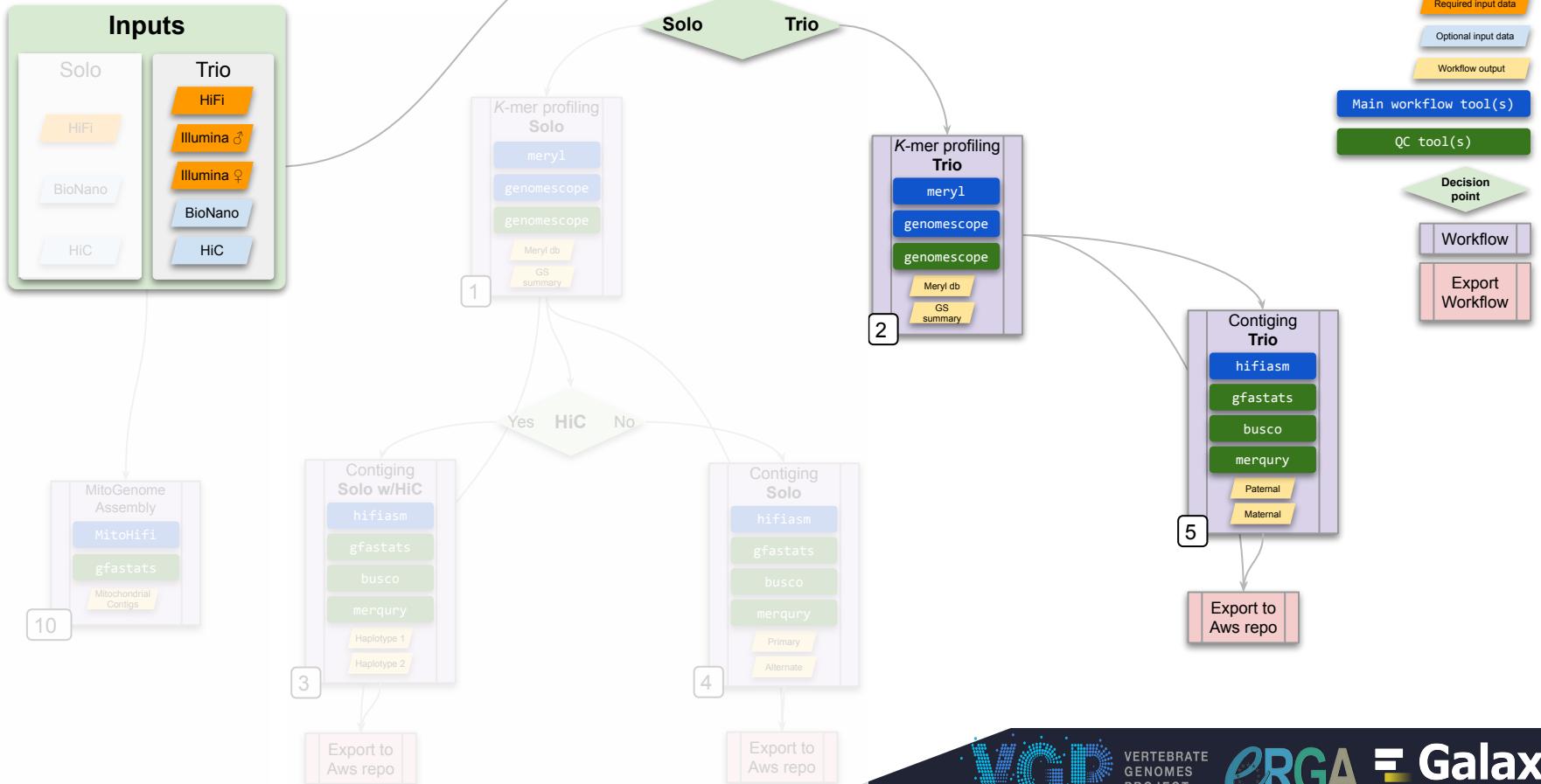
Contiging



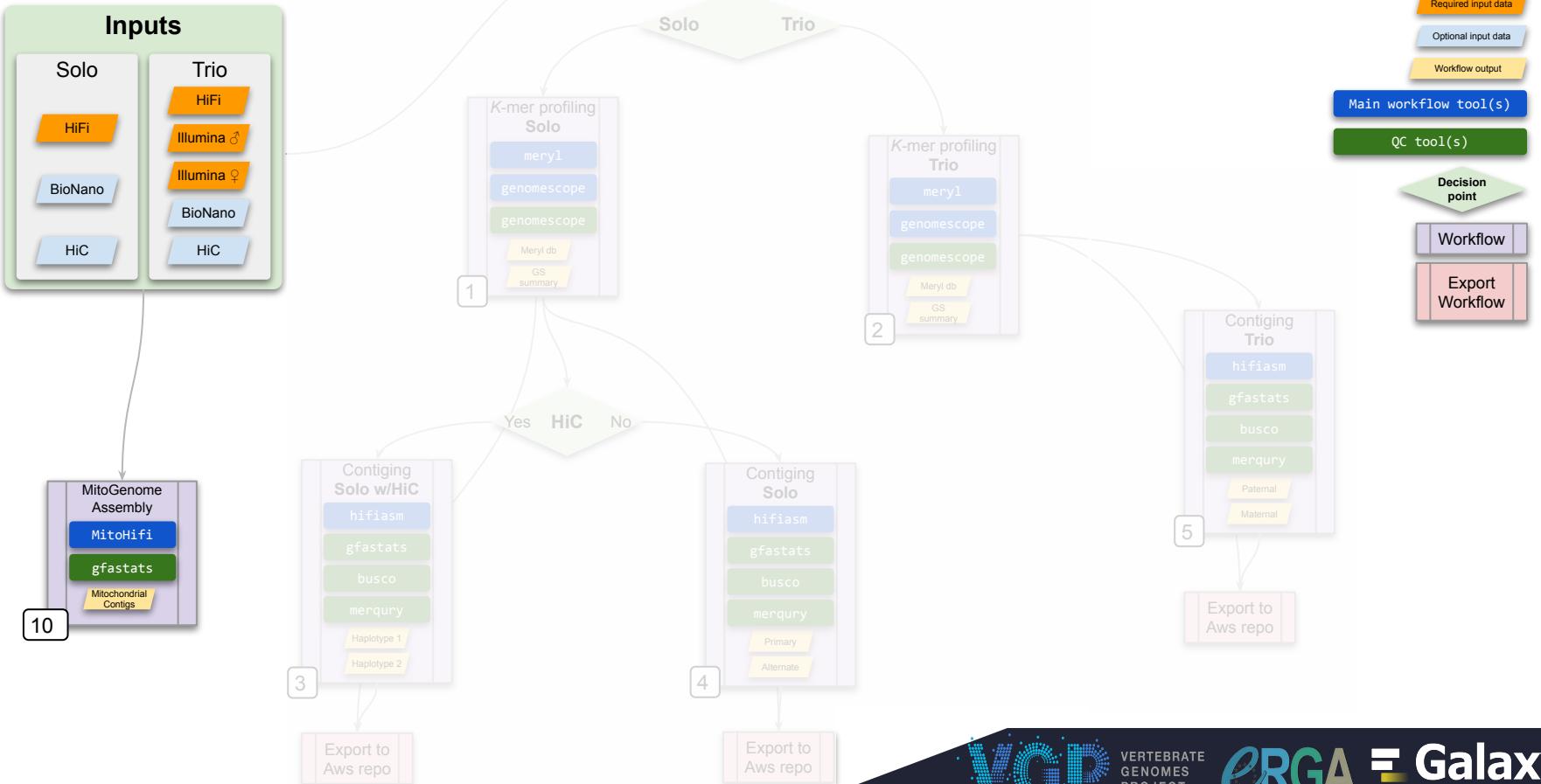
Contiging



Contiging



Contiging



Purging

Contigs Alignment



K-mer copy number

2
1



Purging

Contigs

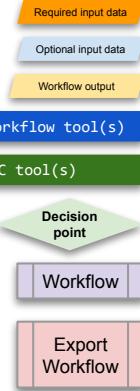
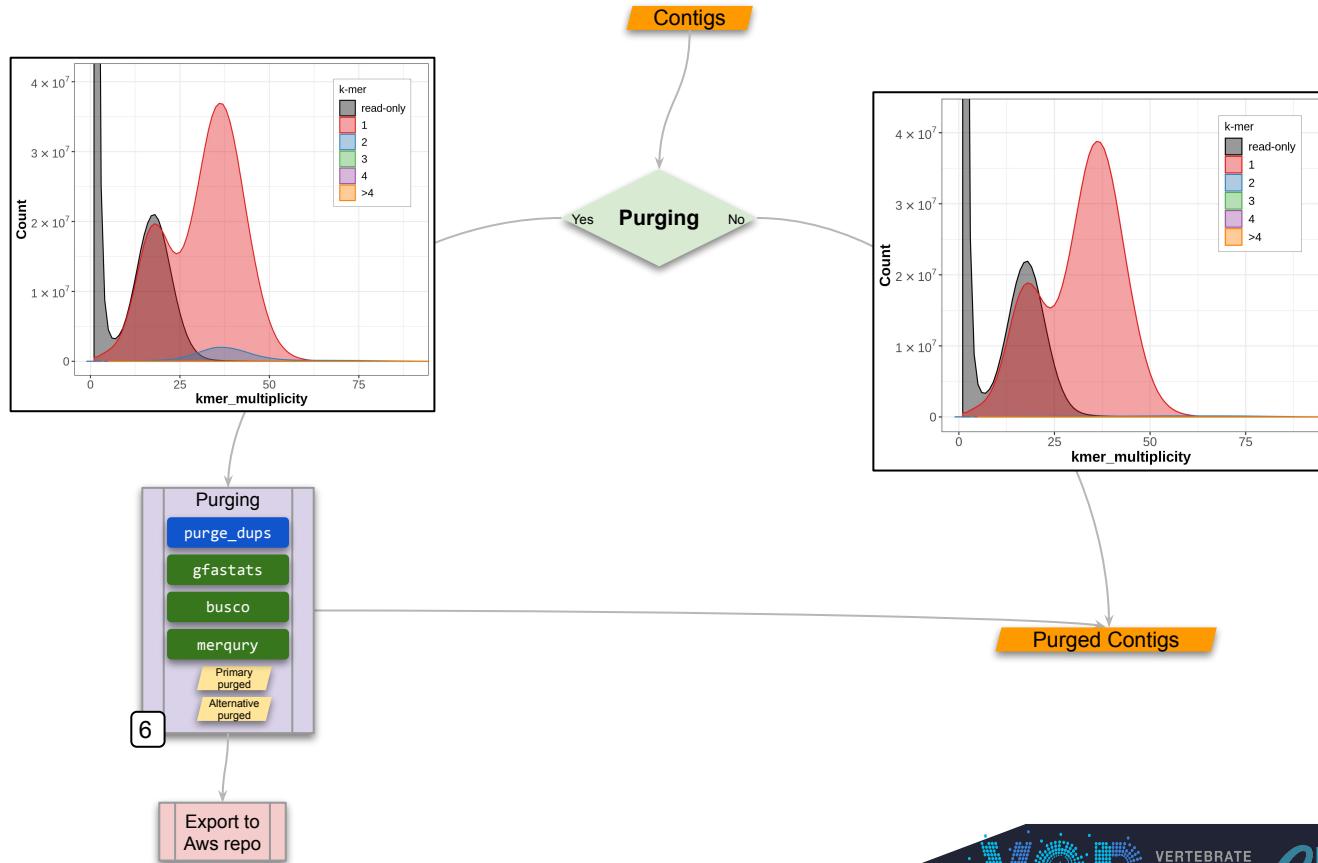


K-mer copy number

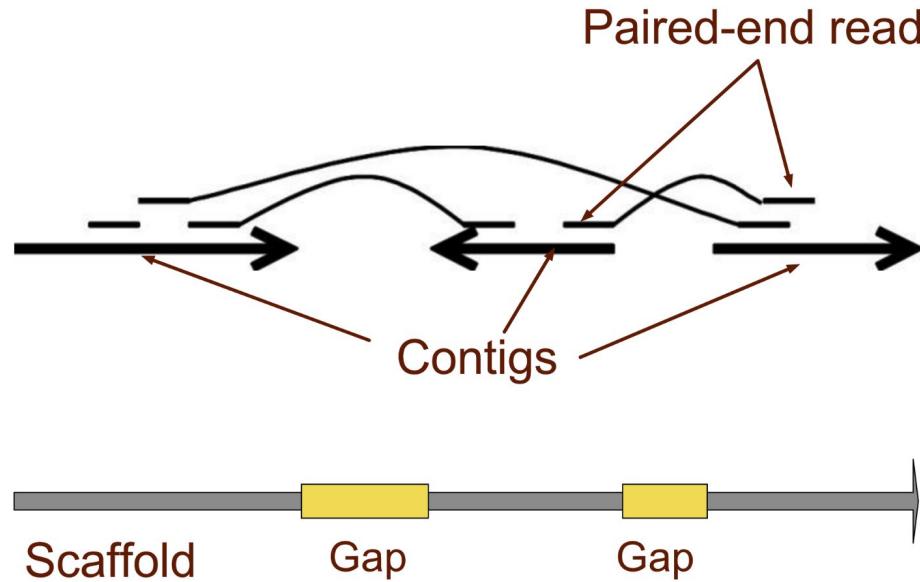
2
1



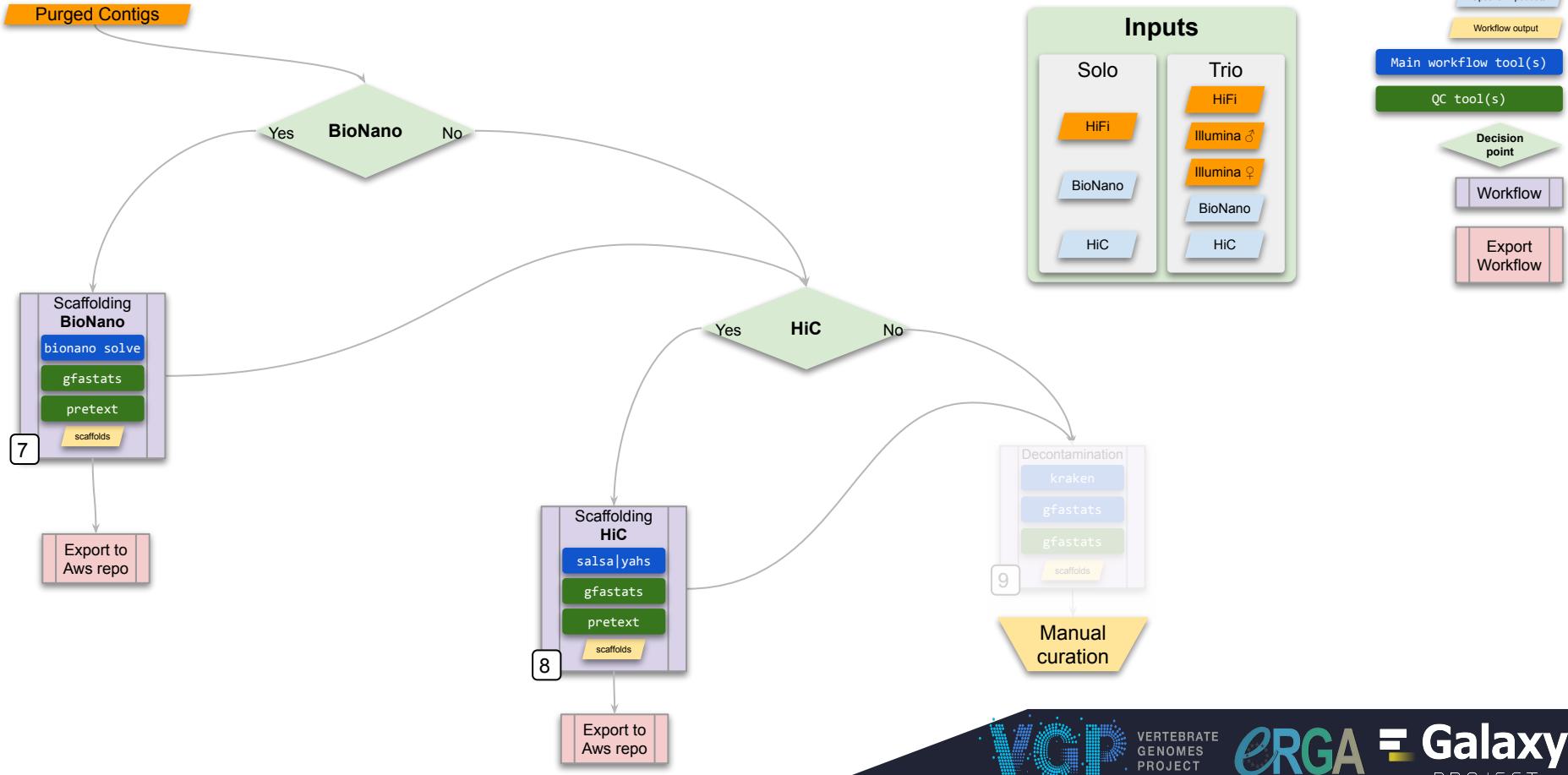
Contiging - purging



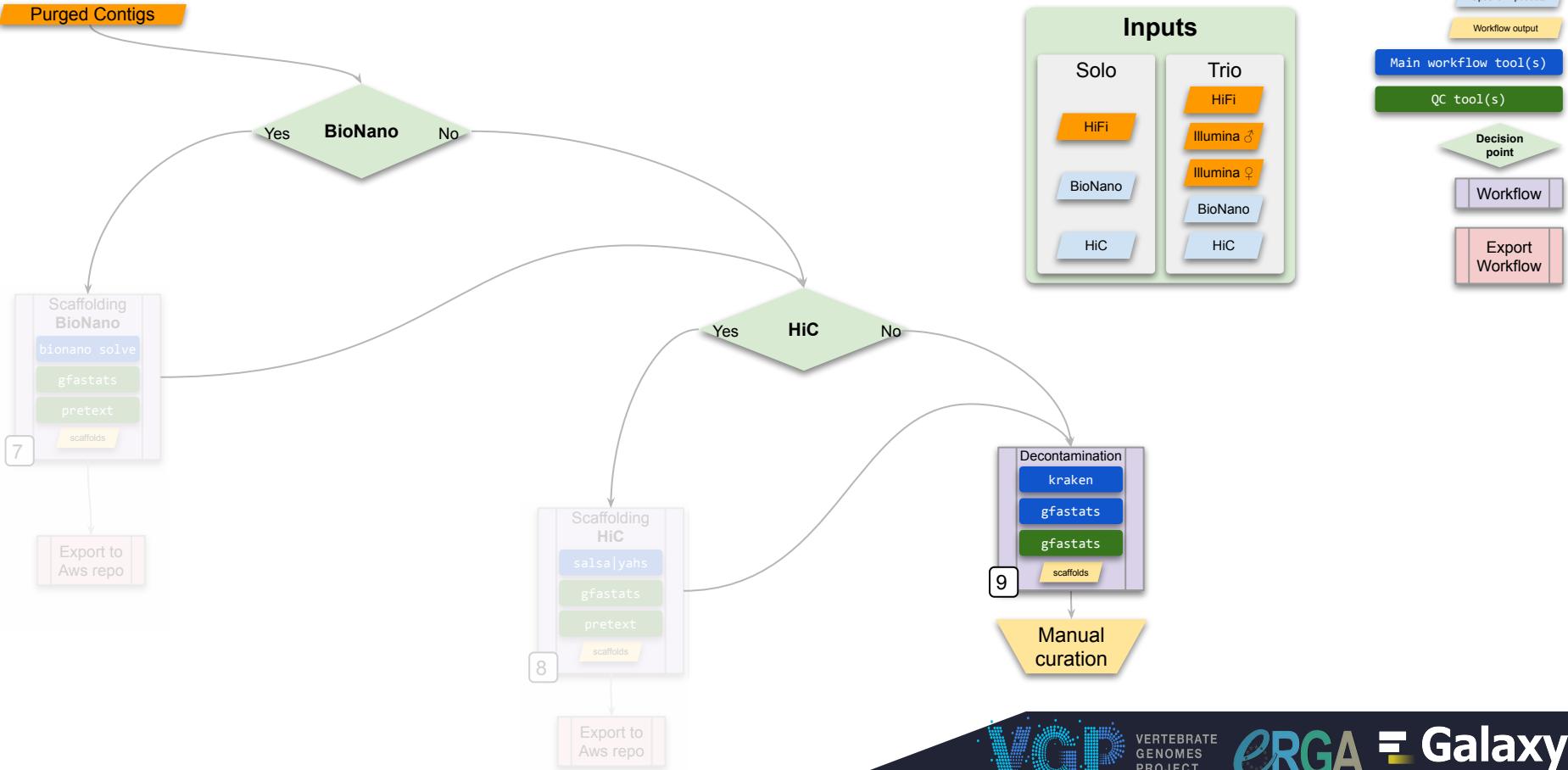
Scaffolding



Scaffolding and Contaminants

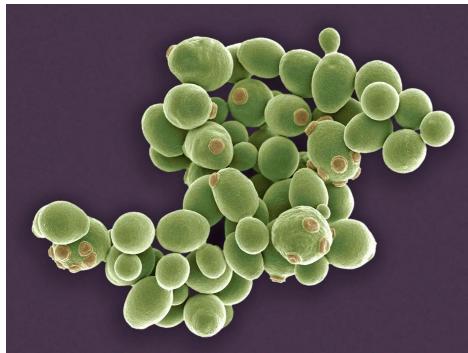


Scaffolding and Contaminants



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



Workflow

Visualize

Shared Data ▾

Help ▾

Login or Register



1. Create Account

<https://usegalaxy.org>

2. Join Training

<https://usegalaxy.org/join-training/vgpbg2023>

Using Galaxy

Galaxy

Workflow Visualize Shared Data Help Login or Register

Using 0%

Tools

search tools

Upload Data

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

Operate on Genomic Intervals

The global community has created a **continuously updated list** of laboratories that can host Ukrainian scientists at all career levels. If your lab can host a scientist -- add your name to the list [here](#). In addition, Galaxy Project has a number of positions at its EU and US sites. Contact us at ukraine@galaxyproject.org

Світова наукова спільнота створила **спісок лабораторій**, що постійно оновлюється та які можуть прийняти українських науковців усіх рівнів, у тому числі й аспірантів. Якщо ваша лабораторія має можливість запросити -- ви можете додати ваше ім'я до списку тут. Окрім того, Galaxy Project має відкріті вакансії у своїх європейських та американських осередках. Пишіть нам на ukraine@galaxyproject.org

Научное сообщество создало постоянно обновляемый **список лабораторий**, которые могут принять украинских ученых (включая аспирантов). К тому же, Galaxy Project имеет открытые позиции на своих европейских и американских сайтах. Контактируйте нас используя ukraine@galaxyproject.org

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

The Galaxy Community and UseGalaxy* initiative lost one of its brightest stars on Saturday, November 26, when Simon Gladman, the principal architect of UseGalaxy.org.au (Galaxy Australia) passed away unexpectedly.

Simon was not just a key member of his own Galaxy community in Australia, but had been a major part of the global Galaxy community for many years. He was a driving force behind the UseGalaxy* public server initiative and the technologies and organization of work needed to make UseGalaxy.org.au (along with UseGalaxy.org and UseGalaxy.eu) a success. In addition, he was a friend to all who met him, and an all-around wonderful person. Our deepest condolences are with his family, friends, and all who knew and loved him.

You can find more details in the numerous memorials for Simon: from the Galaxy Project, from Galaxy Australia, and from the Australian BioCommons. His colleagues are collecting memories, if you have ones that you would like to share, please send them to christina@biocommons.org.au.

Donate to the James P. Taylor Foundation for Open Science

Learn More

History

search datasets

Unnamed history

0 B 0

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 dark header bar with the Galaxy logo, navigation links (Workflow, Visualize, Shared Data, Admin, Help, User), and a status message "Using 3.5 TB". Below the header is a search bar with a library filter set to "vgp". Underneath is a table listing datasets:

Name	Description	Synopsis	Actions
VGP- Yeast	Datasets for the Genome Assembly trainin ... (more)		Edit Manage
VGP-Zebra Finch	Datasets for the assembly of the Zebra F ... (more)		Edit Manage

At the bottom, there is a pagination control with buttons for navigating through pages (1, 2, 3, 4, ..., 10) and a message indicating "per page, 206 total".

Importing Workflows : Public workflows

The screenshot shows the Galaxy web interface with a dark blue header bar. The header includes the Galaxy logo, a navigation menu with 'Workflow', 'Visualize', 'Shared Data', 'Admin', 'Help', 'User', and several icons for notifications, help, and user management. Below the header is a search bar containing the text 'bga'. To the right of the search bar are two buttons: '+ Create' and 'Import'. The main content area is titled 'Published Workflows'. A table displays a single workflow entry:

Name	Tags	Updated	Owner
▼ K-mer profiling and QC (WF1)	VGP BGA23	less than a minute ago	delphinel

WF1: HiFi reads-based Kmer-counting

The screenshot shows the Galaxy web interface. At the top, there is a dark header bar with navigation links: Home, Workflow, Visualize, Shared Data, Admin, Help, User, a bell icon, a graduation cap icon, and a grid icon. Below the header is a search bar labeled "Search Workflows" and a help icon. To the right are buttons for "+ Create" and "Import".

The main area displays a table of workflows. The columns are: Name, Tags, Updated, Sharing, and Bookmarked. A row for "K-mer profiling and QC (WF1)" is selected, indicated by a blue background. The workflow details are shown below:

- Name:** K-mer profiling and QC (WF1)
- Tags:** Reviewed, VGP
- Updated:** less than a minute ago
- Sharing:** (indicated by a star icon)
- Bookmarked:** (indicated by a play button icon)
- Run workflow** button

This screenshot shows the configuration page for the selected workflow. The top bar is identical to the main interface.

The main content area is titled "Workflow: K-mer profiling and QC (WF1)". It contains several input fields:

- Collection of Pacbio Data:** A dropdown menu showing "2: PacBio reads".
- K-mer length:** A text input field containing "21".
- Ploidy:** A text input field containing "2".

At the bottom, there is a link "Expand to full workflow form." and a "Run Workflow" button with a checkmark icon.

Expand to full workflow form.

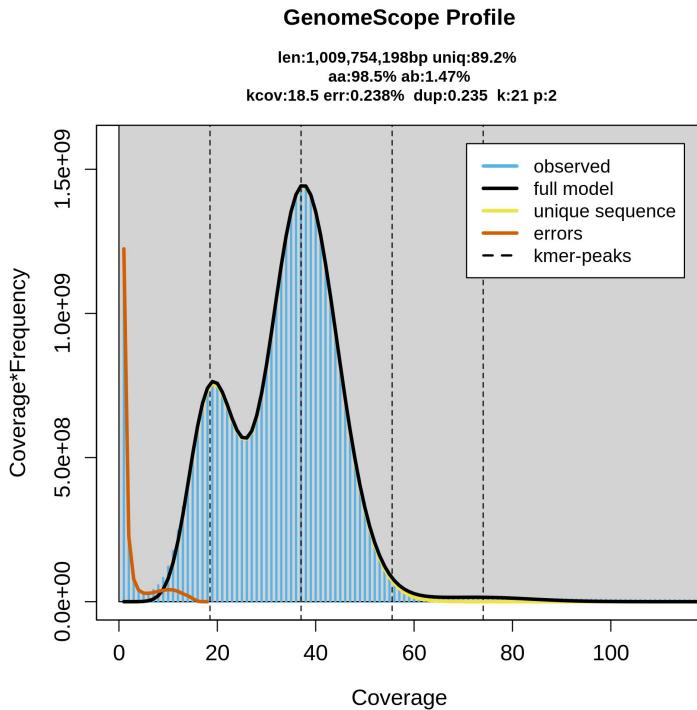


A PROJECT OF THE G10K CONSORTIUM



GALAXY
PROJECT

WF1: Outputs

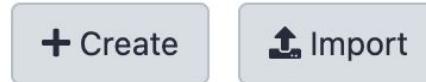
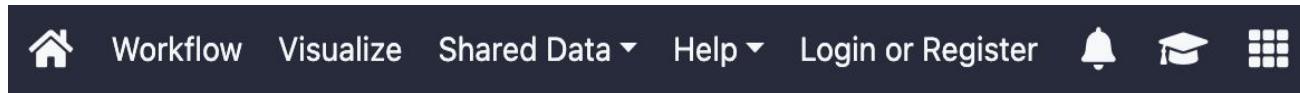


Kmer profiling Zebra Finch

GenomeScope version 2.0
input file =
/data/dnb07/galaxy_db/files/a/8/5/dataset_a855ad9d-4075-4de9-b552-dbe07c12f5
7c.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:

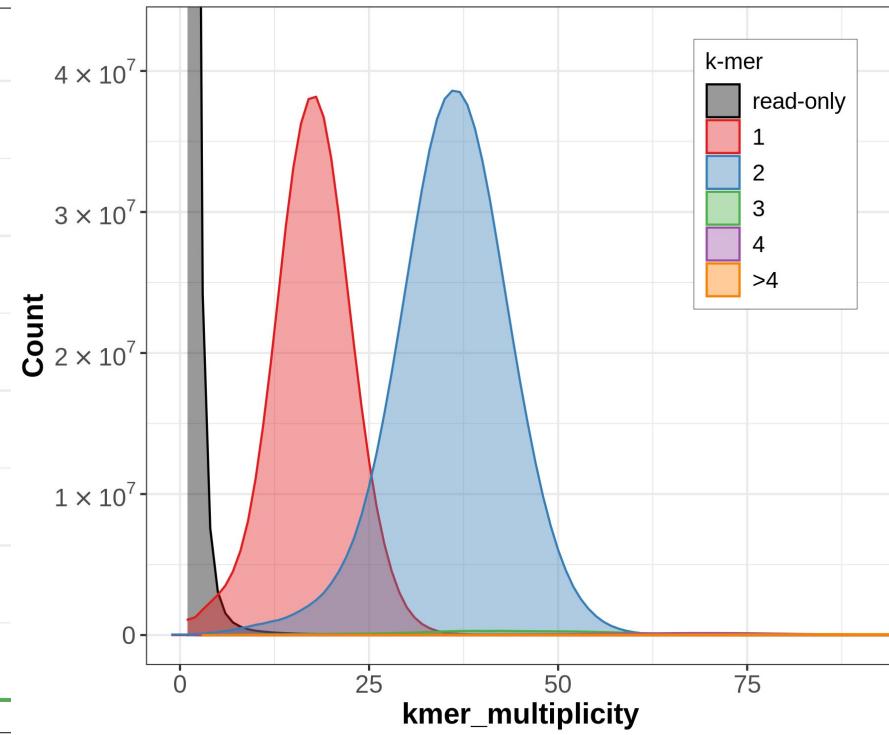
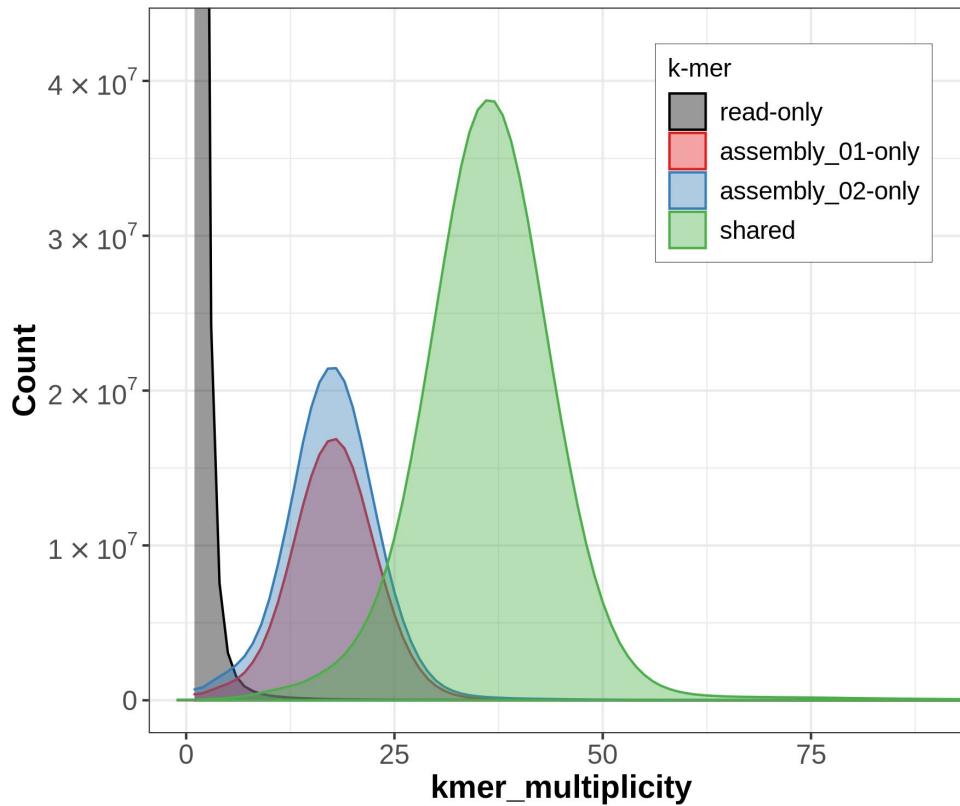
#workflow/github.com/Delphine-L/iwc/WF3-Assembly

WF3-5: Contig assembly

The screenshot shows the Galaxy web interface for the Assembly (WF3) workflow. At the top, there is a navigation bar with links for Home, Workflow, Visualize, Shared Data, Help, User, and a grid icon. Below the navigation bar is a search bar labeled "Search Workflows" and buttons for "+ Create" and "Import". The main area displays the "Assembly (WF3)" workflow, which is currently bookmarked and has a status of "less than a minute ago". The workflow configuration section includes fields for "Pacbio Reads Collection" (with 5: PacBio reads), "Meryl Database" (with 8: Meryl on data 7: read-db.meryldb), "Genomescope Summary" (with 16: GenomeScope on data 9 Model parameters, 15: GenomeScope on data 9 Summary, 14: GenomeScope on data 9 Model, 4: SRR7126301_1, and 3: SRR7126301_2), and "SAK input file" (with 16: GenomeScope on data 9 Model parameters). A "Run Workflow" button is located at the top right of the configuration area.

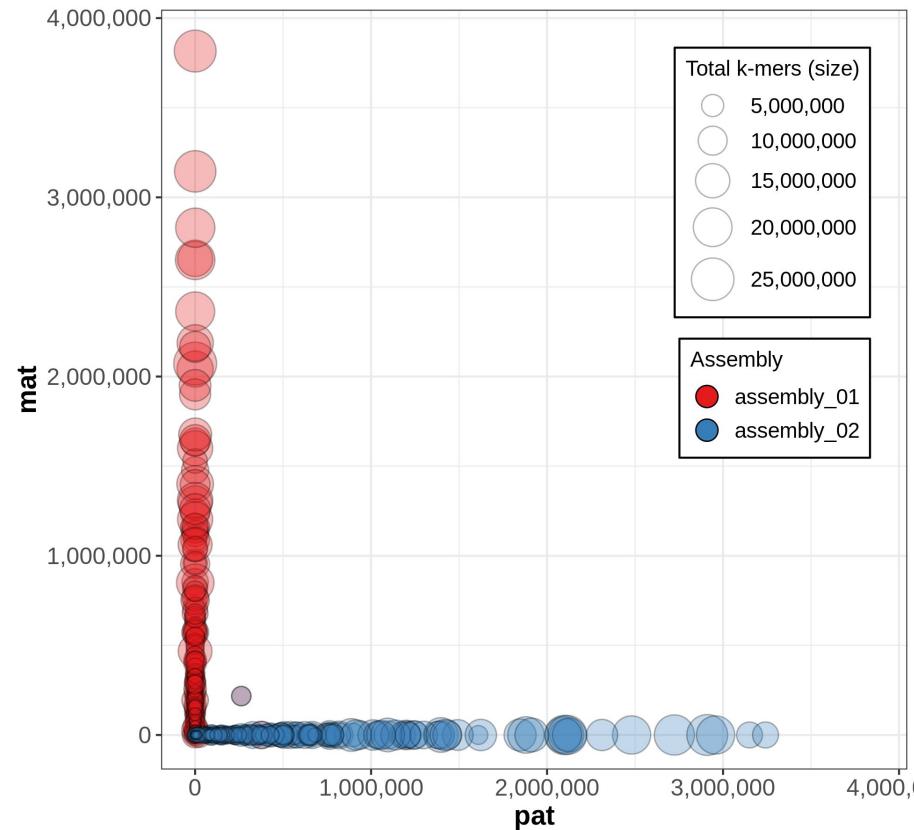
WF3-5: Outputs

Assembly with HiC phasing Zebra Finch



WF5: Outputs

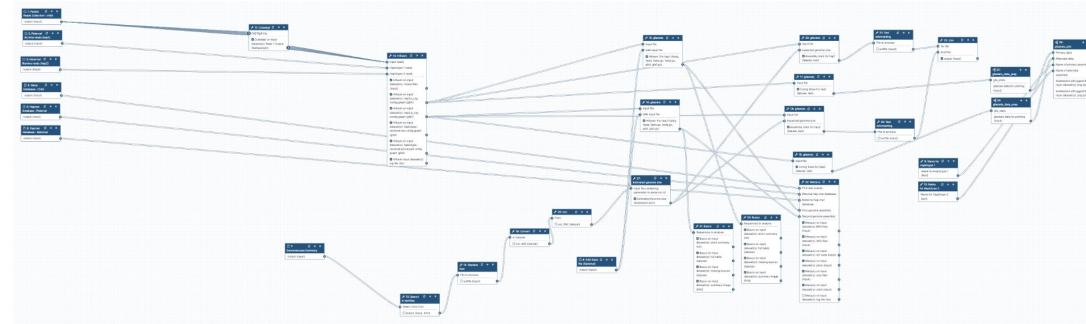
Trio Assembly Zebra Finch



WF7-8: Scaffolding

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

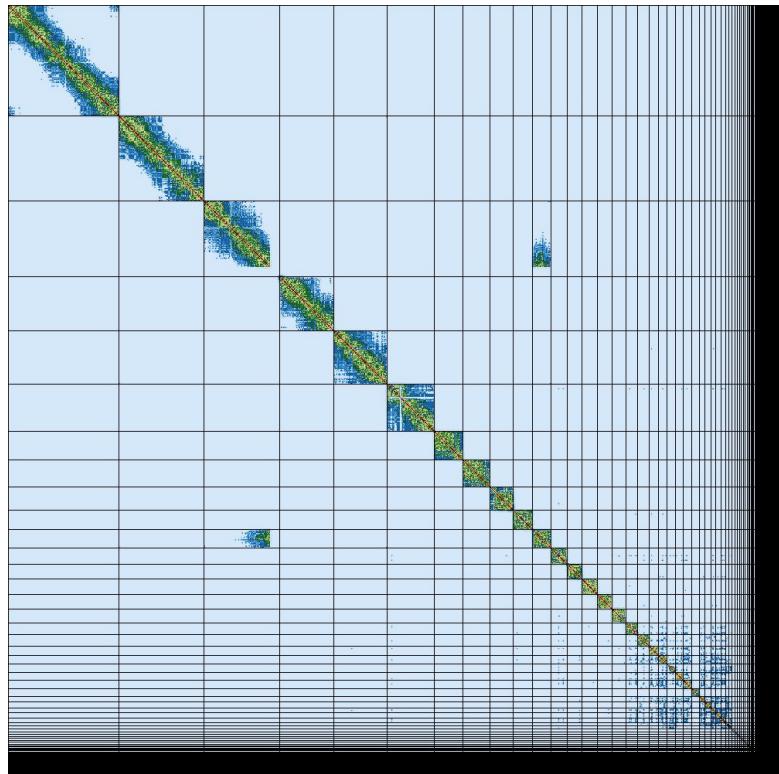
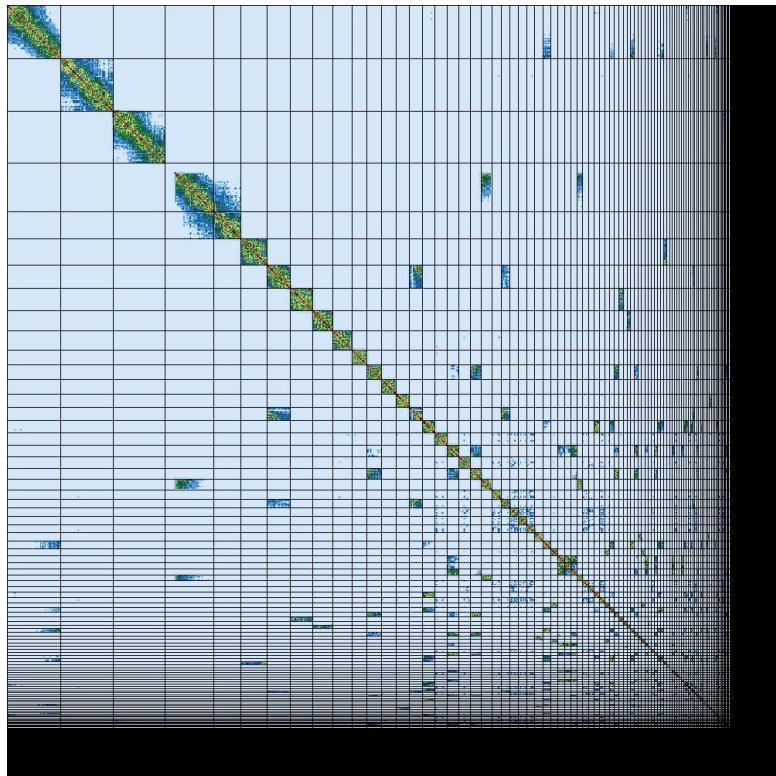
- Tools
 - Yahs
 - Gfastats
 - Pretext



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

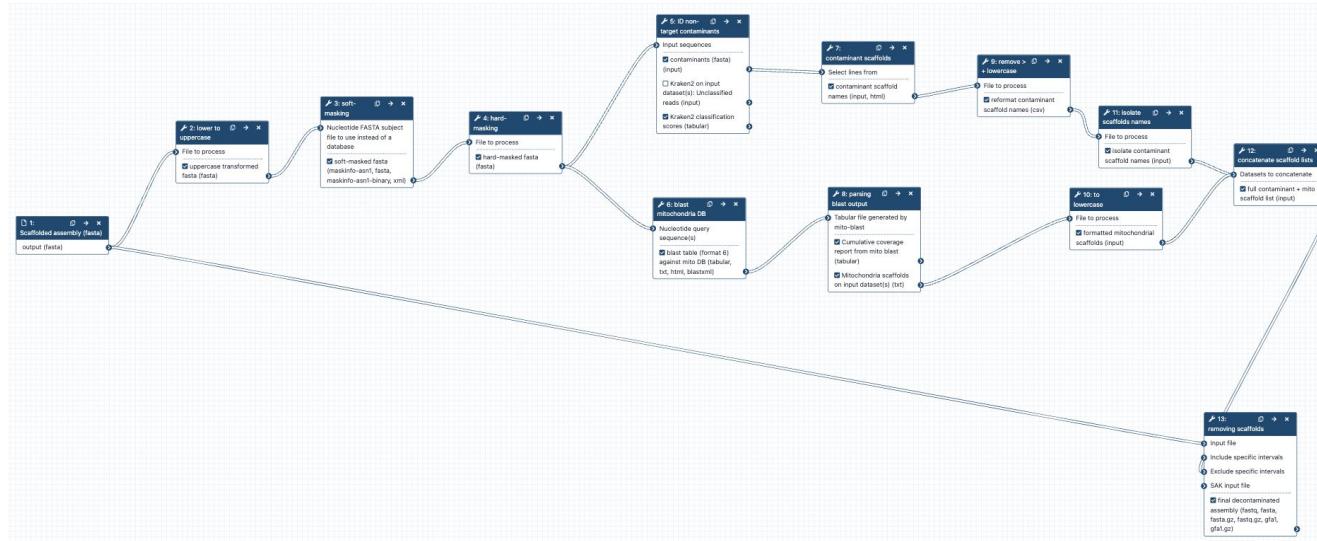
WF7-8: Outputs

HiC Scaffolding Zebra Finch



WF9: Decontamination (optional)

- TRS ID: #workflow/github.com/Delphine-L/iwc/WF9-Decontamination
- 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



National Institutes
of Health



NSF



JOHNS HOPKINS
UNIVERSITY



A PROJECT OF THE G10K CONSORTIUM



<https://galaxyproject.org/>



A PROJECT OF THE G10K CONSORTIUM



EUROPEAN REFERENCE GENOME ATLAS

Thank you!