

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. On the left, a sidebar lists various tool categories: Tools, 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, Fetch Sequences/Alignments, GENOMICS ANALYSIS, Assembly, Annotation, Mapping, and Variant Calling.

The main content area displays "General Statistics" for a collection of FASTQ files. The table includes columns for Sample Name, % Duplication, % > Q30, Mb Q30 bases, GC content, % PF, and % Adapter. A "Plot" button is also present in the header of this section.

Below the statistics table, there is a section titled "fastp" which describes fastp as an ultra-fast all-in-one FASTQ preprocessor. It includes a "Filtered Reads" plot showing the distribution of read lengths for several samples.

The right side of the interface features a "History" panel listing various workflow steps and their details, such as "712: Realigned reads" and "648: Map with BWA-MEM on collection 128 (mapped reads in AM format)".

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)

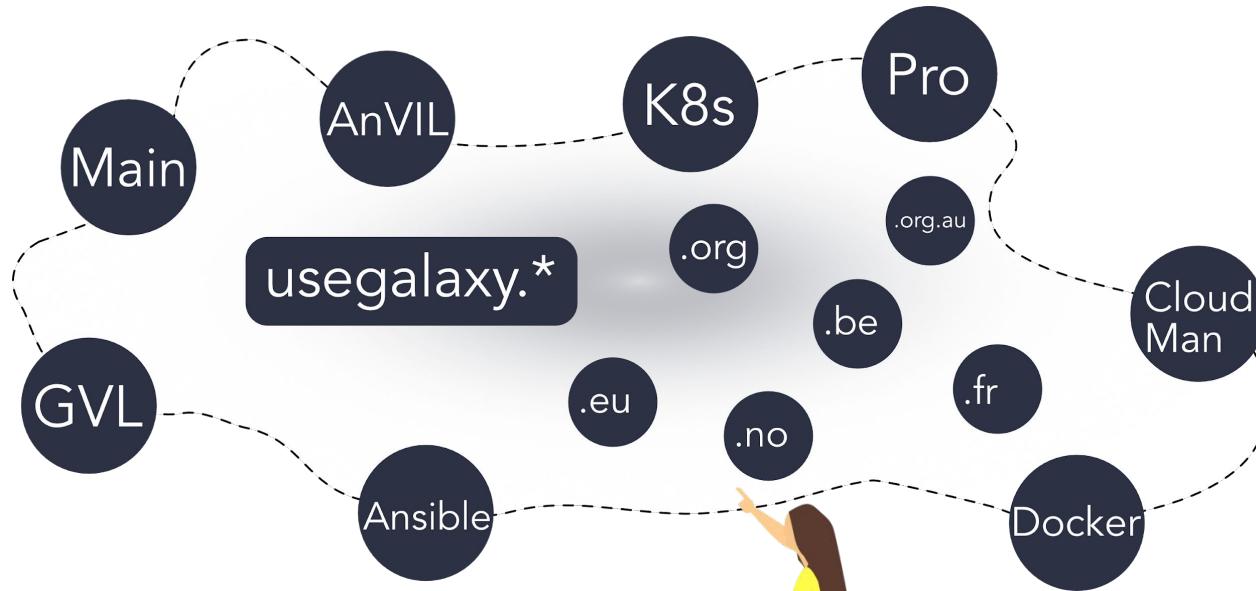
So ... what is Galaxy?

- 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



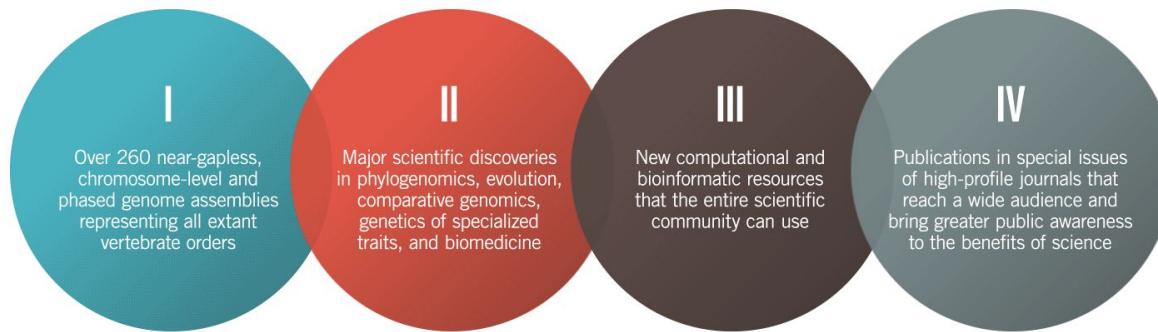
The Vertebrate Genomes Project

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

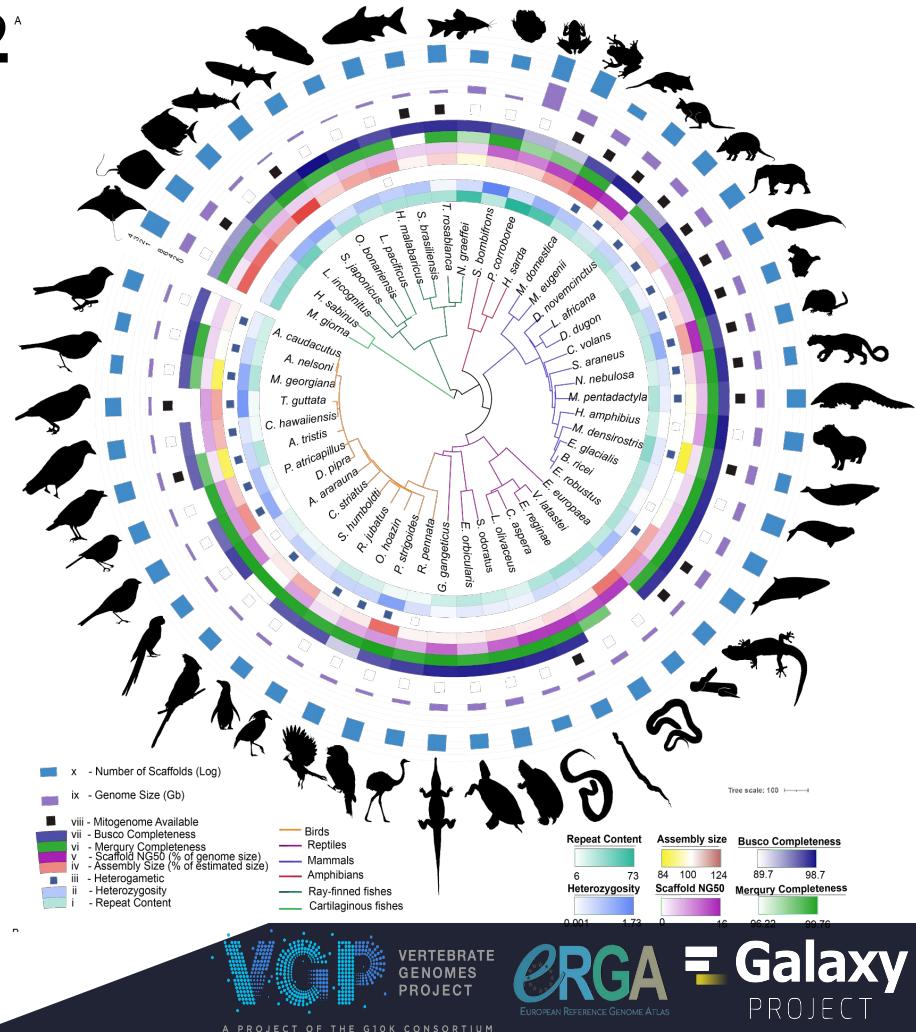
Number of species at each level of completion.

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



VGP assembly Pipeline v2^A

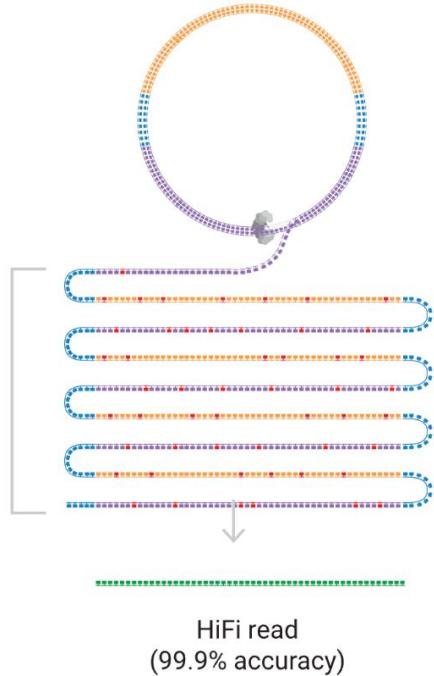
- Developed in Galaxy
- Available on global instances :
 - .eu
 - .org
 - .org.au (soon)
- Dozens of genomes assembled
 - Hundreds to thousands planned in the coming year
- Technologies :
 - PacBio HiFi
 - Bionano Cmap
 - Arima HiC



Data - PacBio HiFi



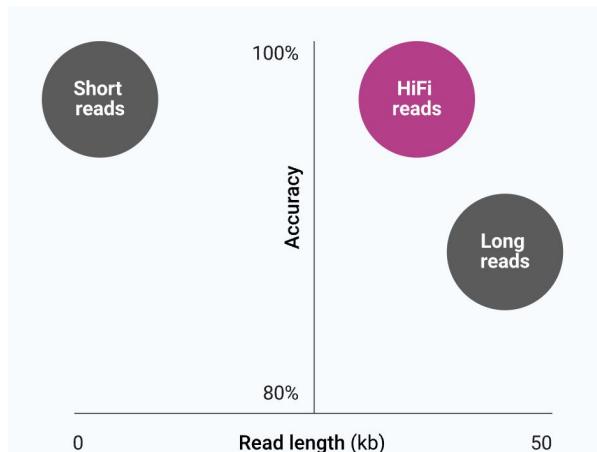
Circularized DNA
is sequenced in
repeated passes



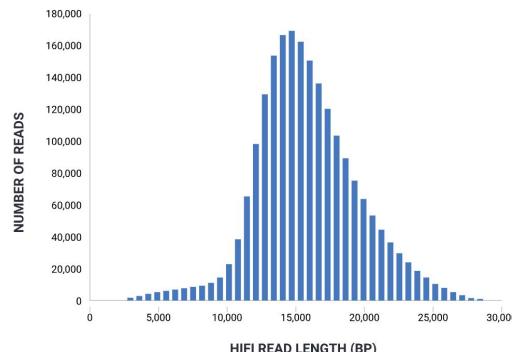
The polymerase reads
are trimmed of adapters
to yield subreads

Consensus and
methylation status are
called from subreads

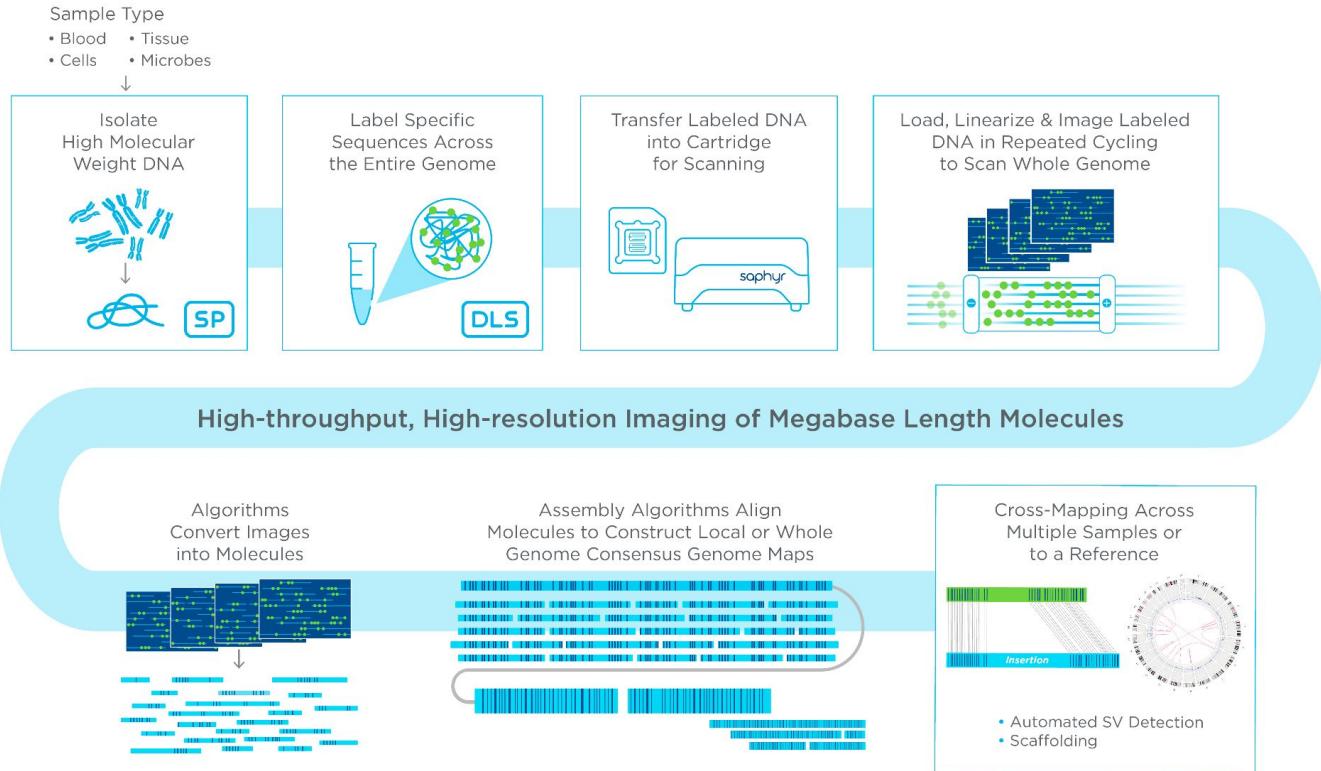
<https://www.pacb.com/technology/hifi-sequencing/>



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



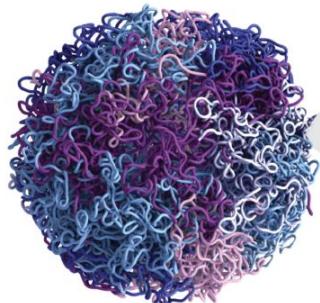
Data - Bionano Optical Mapping



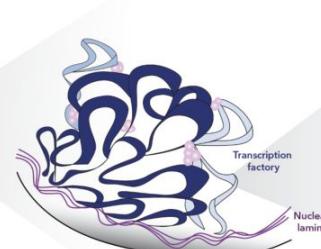
<https://bionanogenomics.com/technology/platform-technology/>

Data - Arima HiC

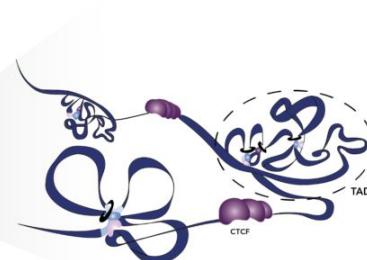
In the nucleus chromosomes are organized into chromosome territories



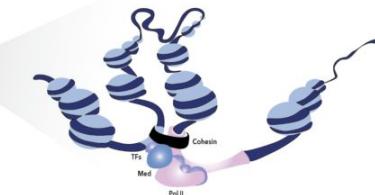
Chromosomes are divided into cell-specific A/B compartments



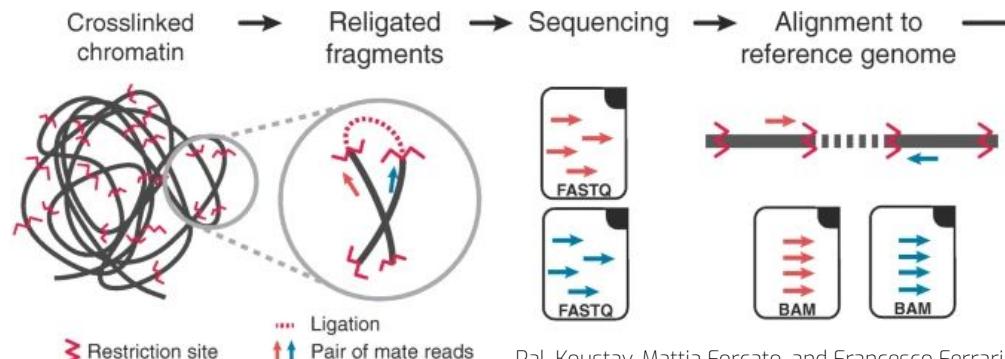
Compartments are organized into topologically associated domains (TADs)



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



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

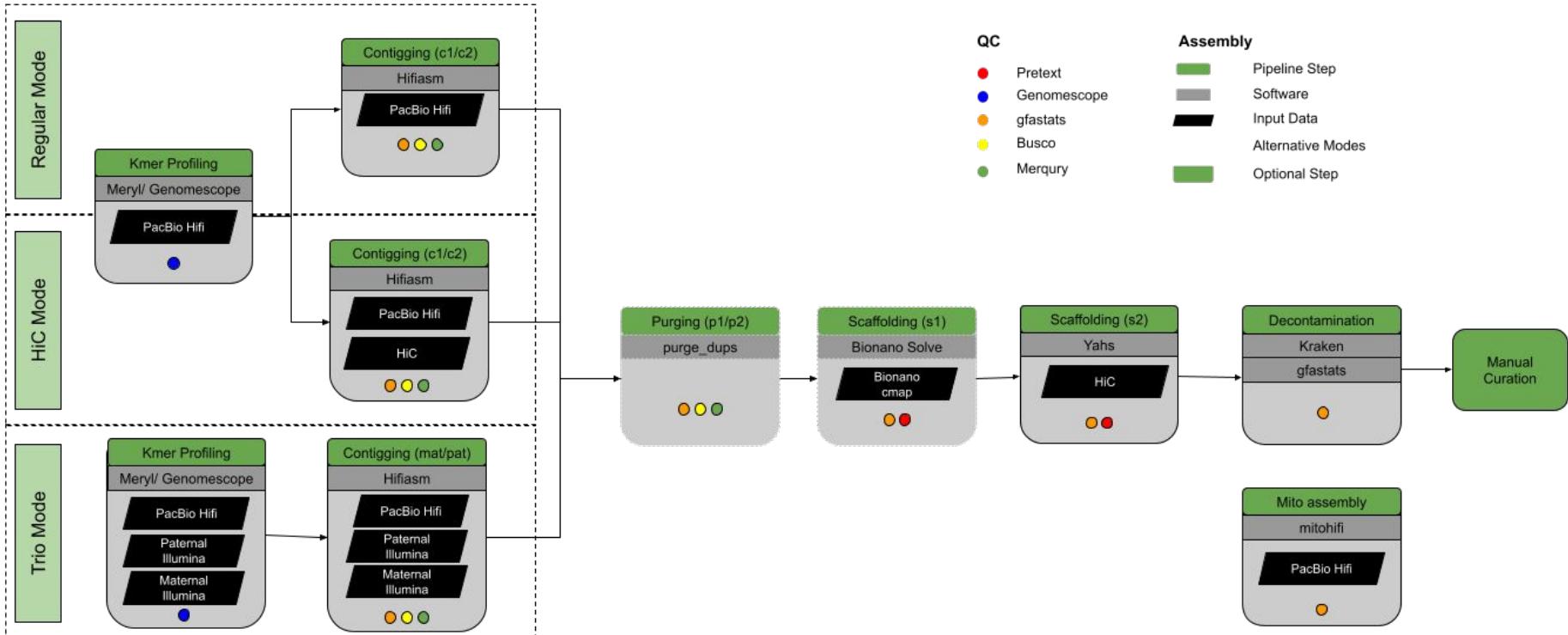


Pal, Koustav, Mattia Forcato, and Francesco Ferrari. "Hi-C analysis: from data to biological integration." *Biophysical reviews* 11.1 (2019): 67-78.



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.

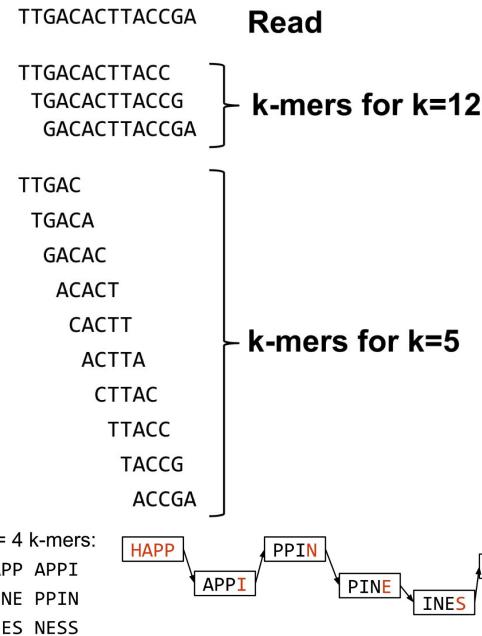
The latest GTN news

Read about new tutorials, features, events and more!

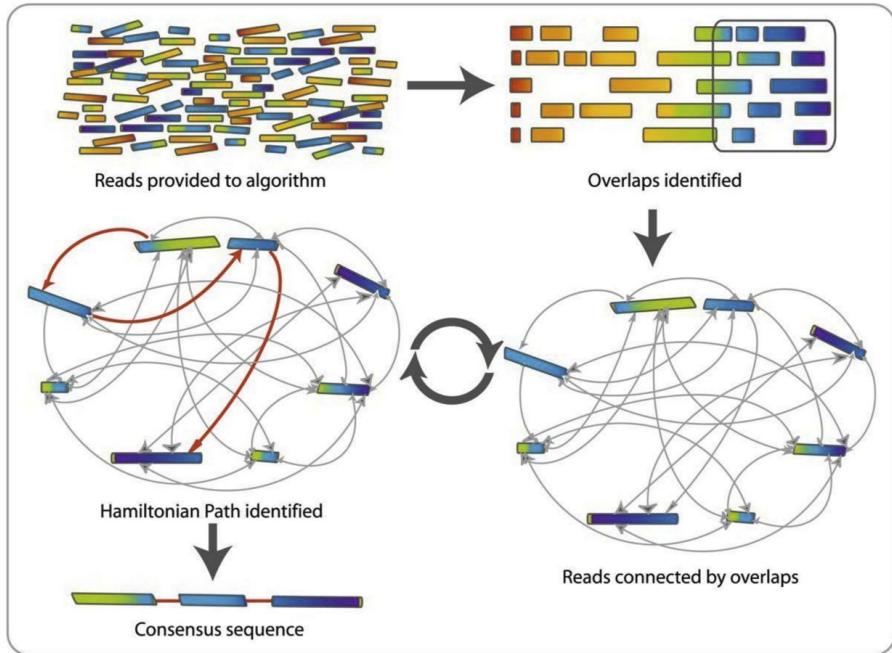
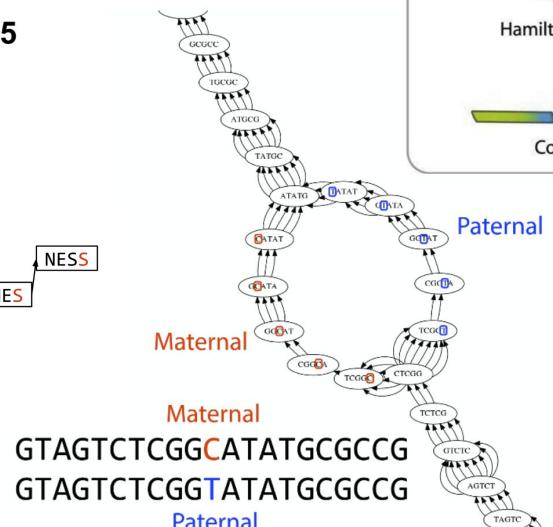
Nov 18, 2022	New Topic: Single Cell Analysis!
Sep 11, 2022	New Tutorial: Data Manipulation
Jun 2, 2022	New Tutorial: Workflow Examples

OPEN CHAT

Contiging

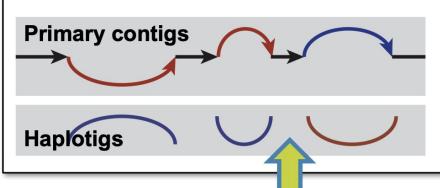
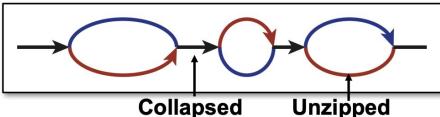


HAPPINESS



Phased Assemblies

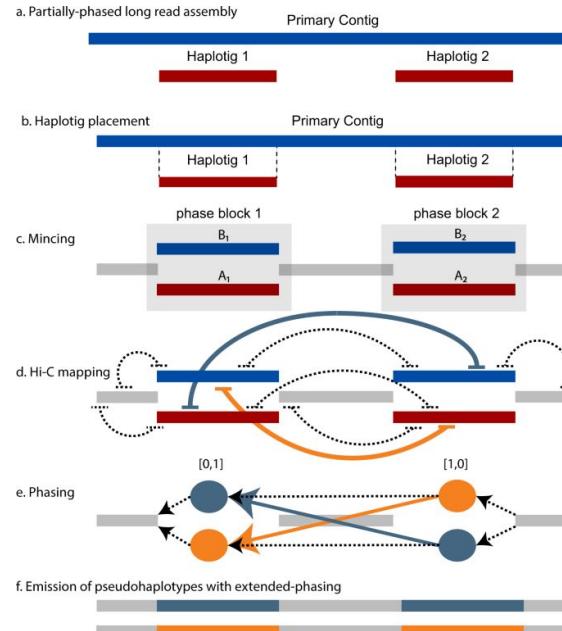
Simple Phasing



"Phase/Haplotype Switch"

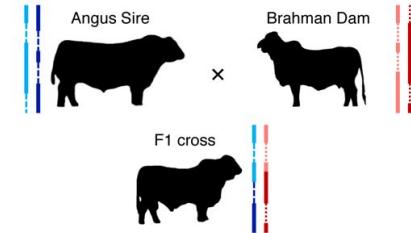
Chin, C.S. et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*. 13(12), 1050.

Hi-C-based Phasing



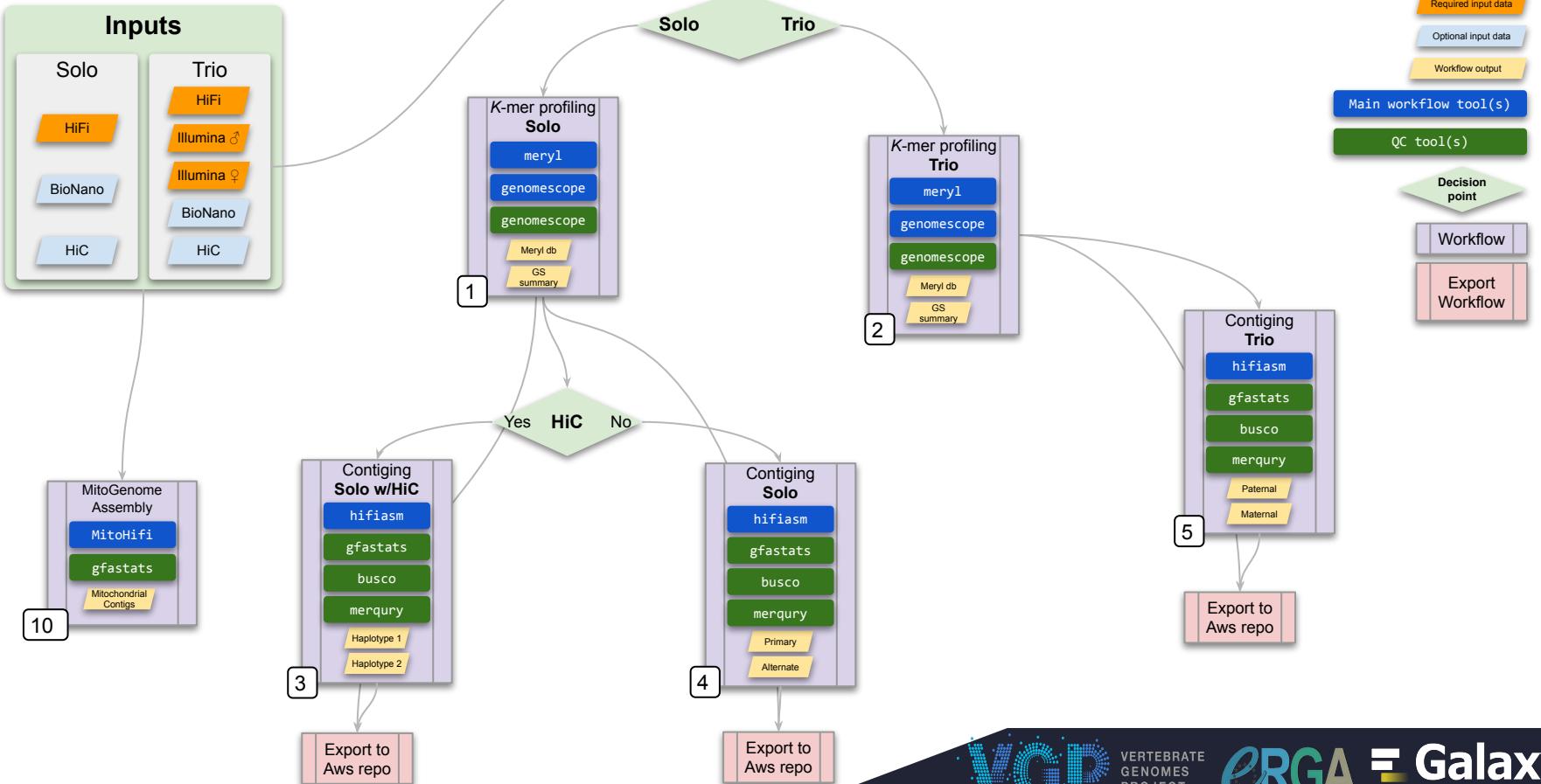
Kronenberg, Z. N., Rhee, A., Koren, S., Concepcion, G. T., Peluso, P., Munson, K. M., ... & Kingan, S. B. (2021). Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nature Communications*, 12(1), 1935.

Trio-based Phasing

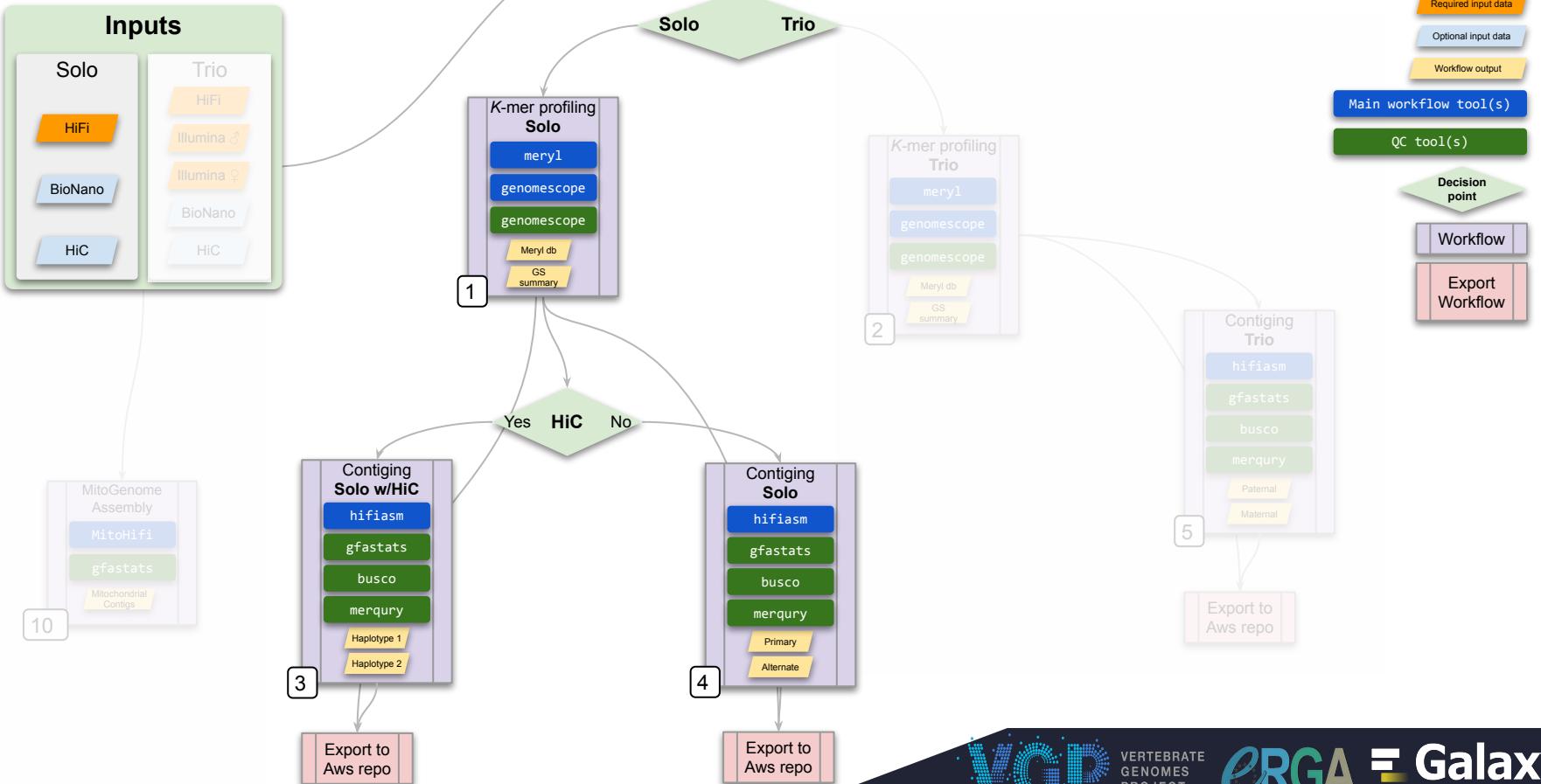


Koren et al. (2018). De novo assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology*. 10.1038/nbt.4277

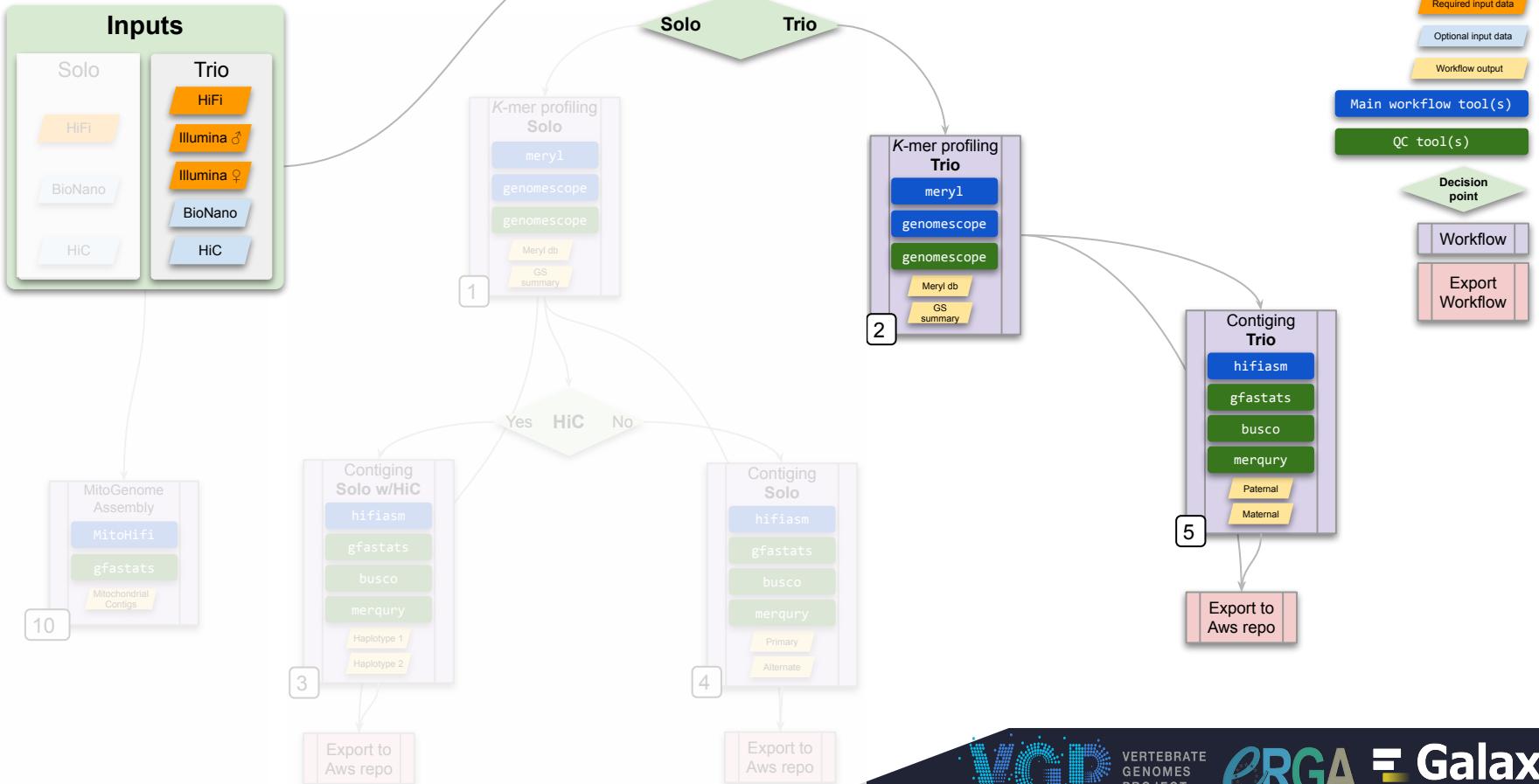
Contiging



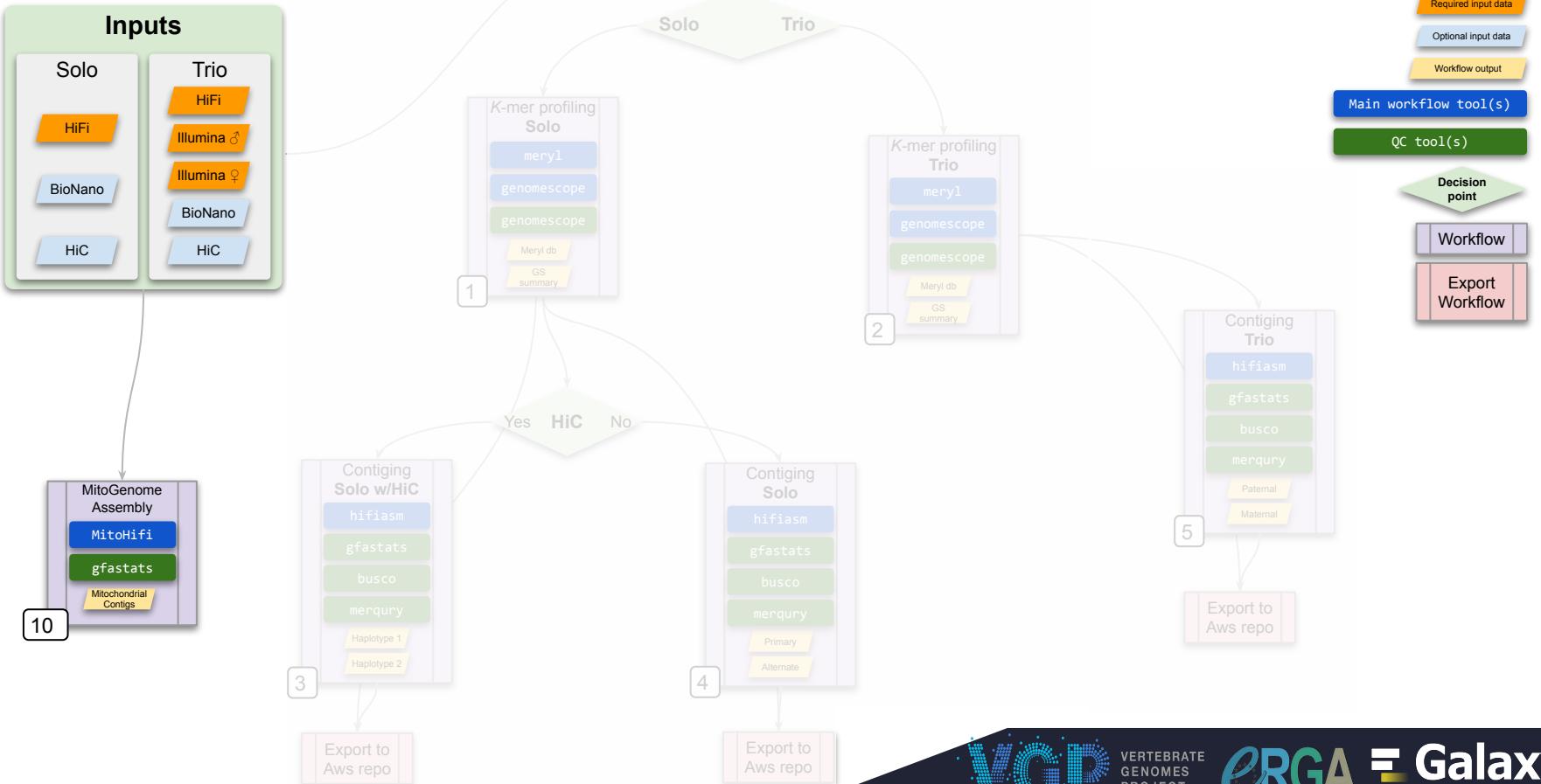
Contiging



Contiging



Contiging



Purging

Contigs Alignment



K-mer copy number

2
1



Contigs



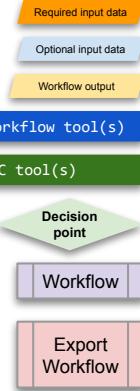
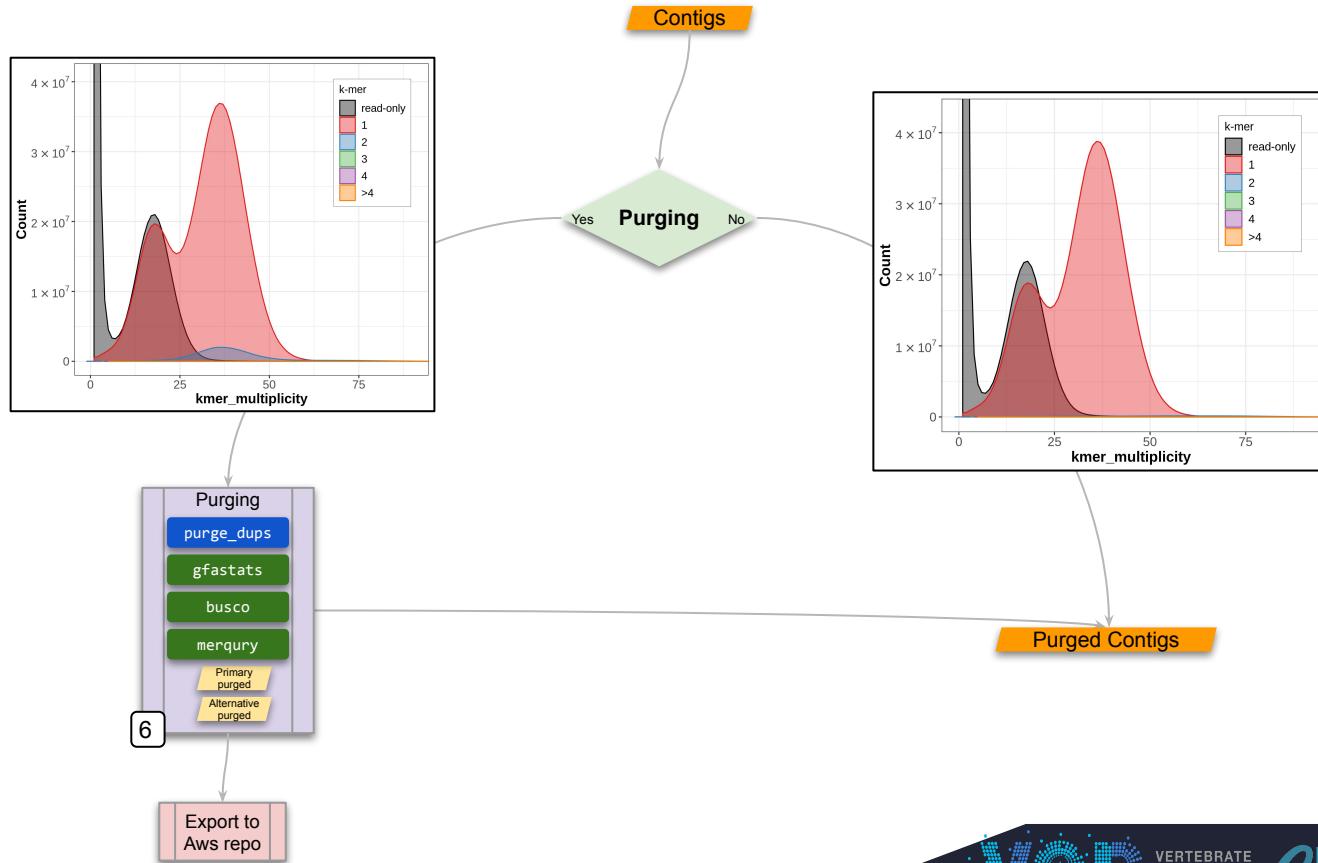
Purging

K-mer copy number

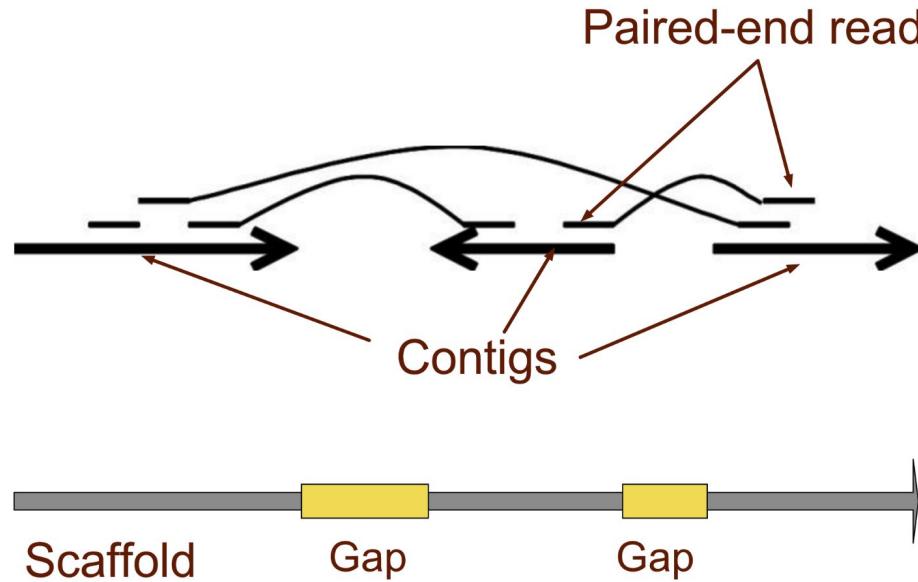
2
1



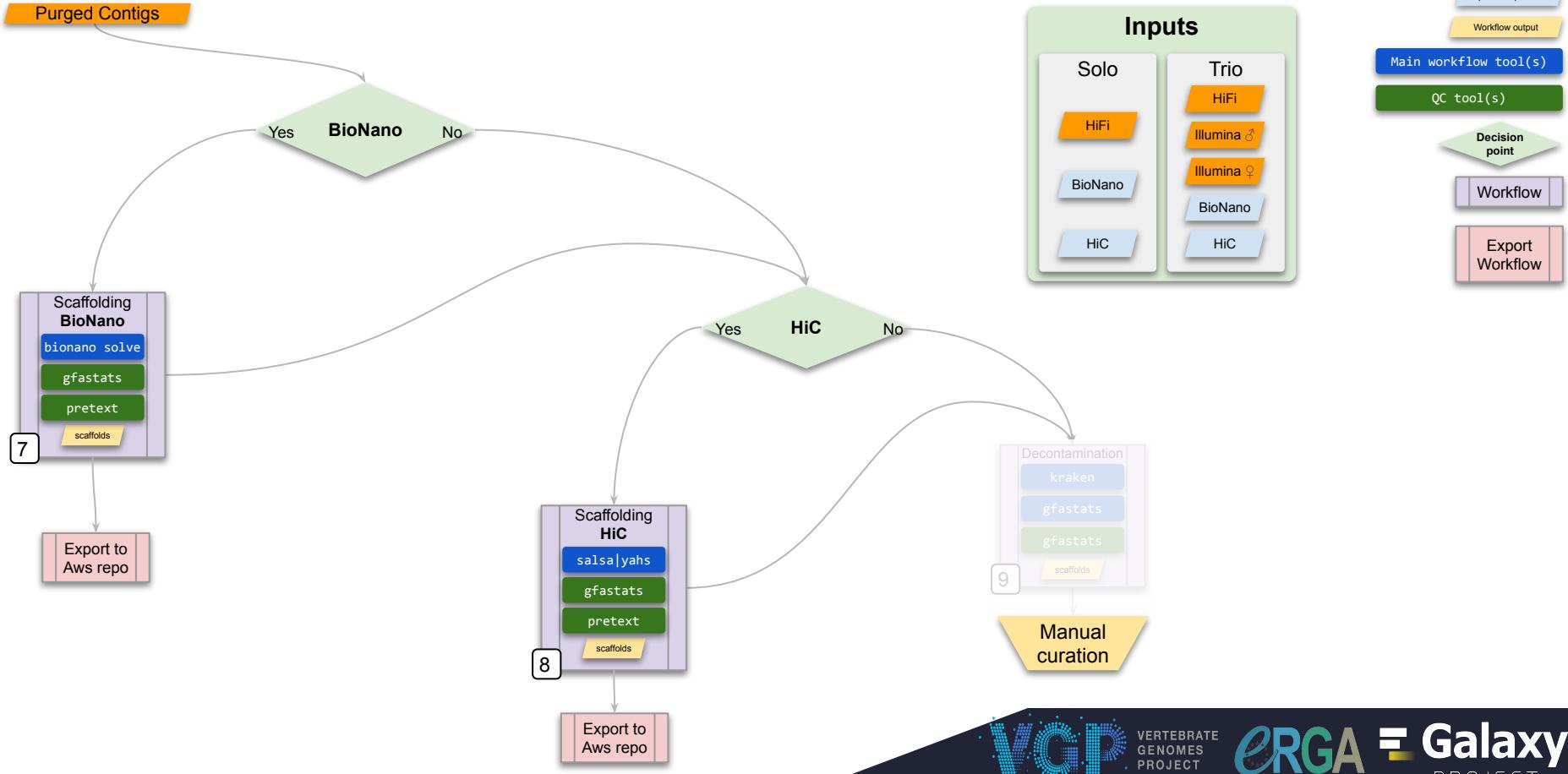
Contiging - purging



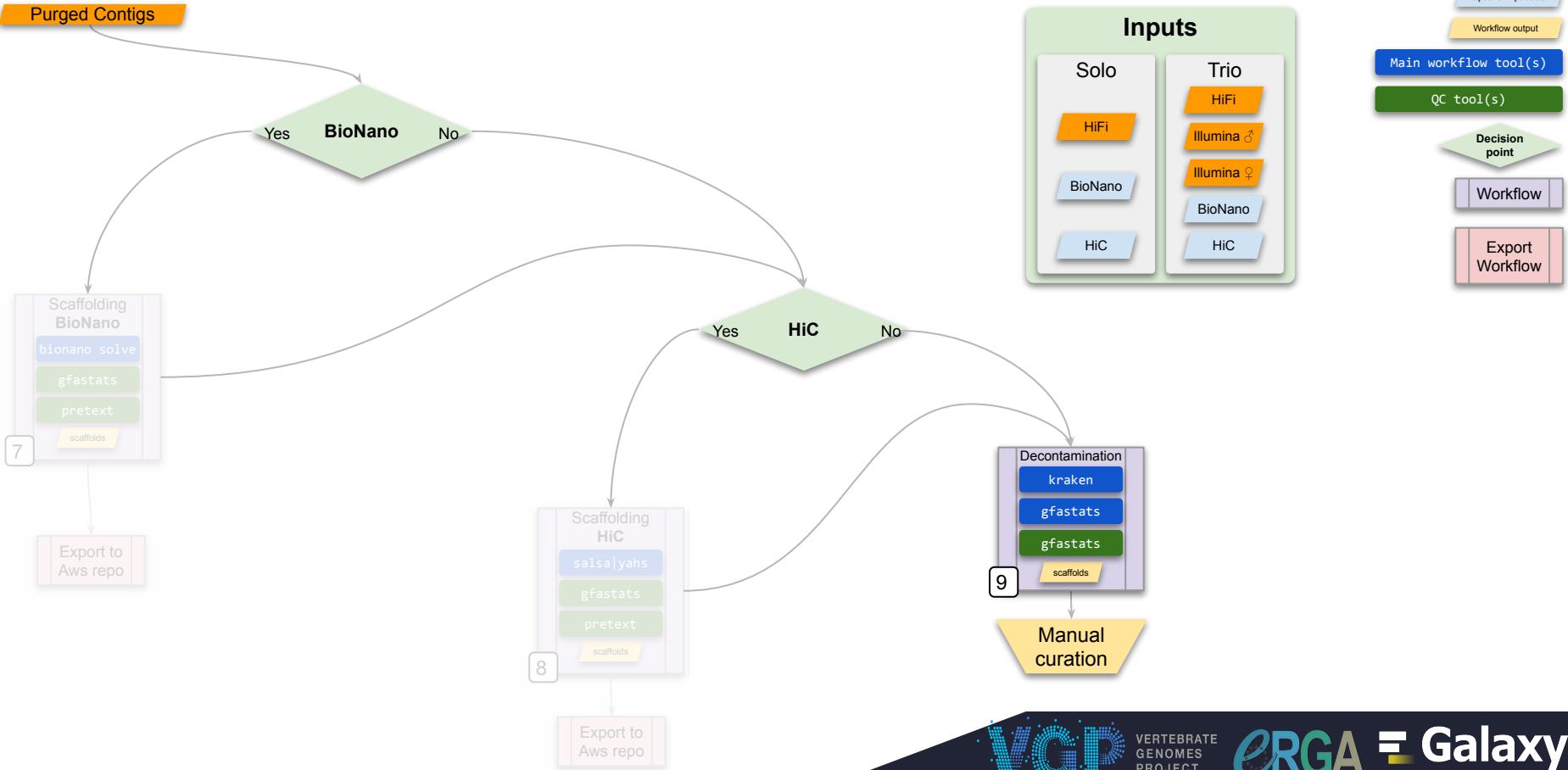
Scaffolding



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)



A PROJECT OF THE G10K CONSORTIUM

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 buttons for '+ Create' and 'Import'. The main content area is titled 'Published Workflows'. A table displays the following data:

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

WF1: HiFi reads-based Kmer-counting

Workflow Visualize Shared Data Admin Help User ? + Create Import

Search Workflows

Name	Tags	Updated	Sharing	Bookmarked
▼ K-mer profiling and QC (WF1) ✓	Reviewed X VGP X	less than a minute ago	★	 Run workflow

Workflow Visualize Shared Data Admin Help User ?

Workflow: K-mer profiling and QC (WF1)

Run Workflow

Collection of Pacbio Data

2: PacBio reads

K-mer length

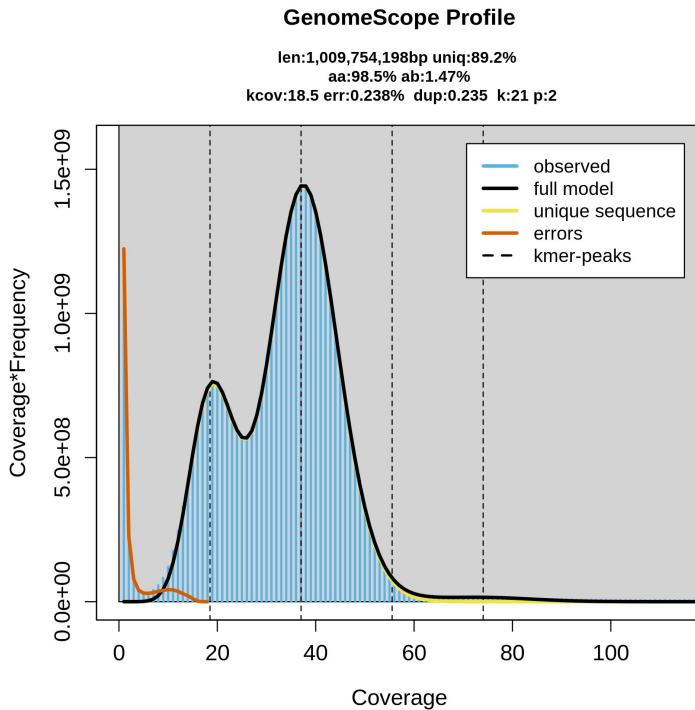
21

Ploidy

2

Expand to full workflow form.

WF1: Outputs

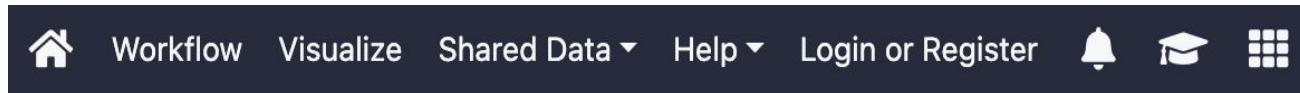


Kmer profiling Zebra Finch

GenomeScope 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



[Create](#) [Import](#)

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:

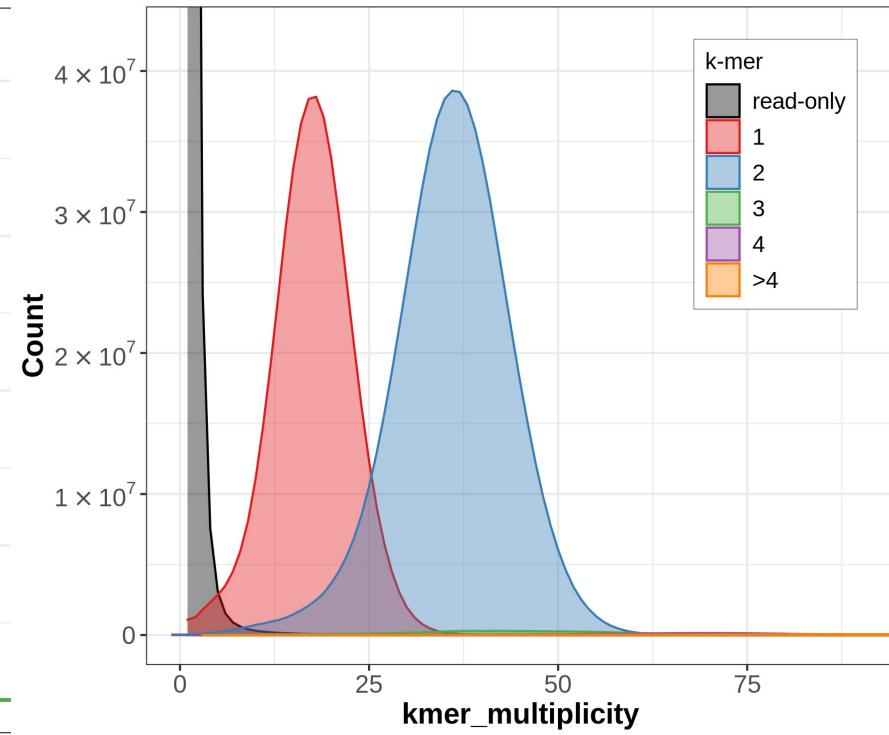
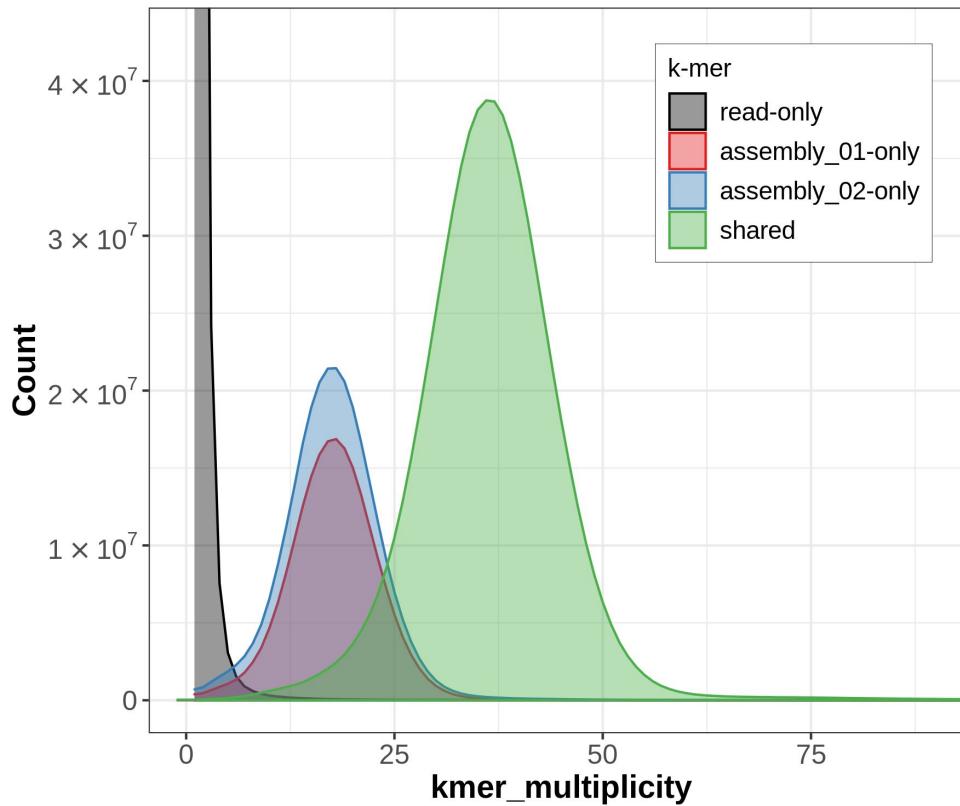
#workflow/github.com/Delphine-L/iwc/WF3-Assembly

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 reads), "Meryl Database" (with 8 entries), "Genomescope Summary" (with 16 entries, one of which is highlighted in blue), and "SAK input file" (with 16 entries). On the right side of the workflow configuration, there are "Run Workflow" and "Run Workflow" buttons.

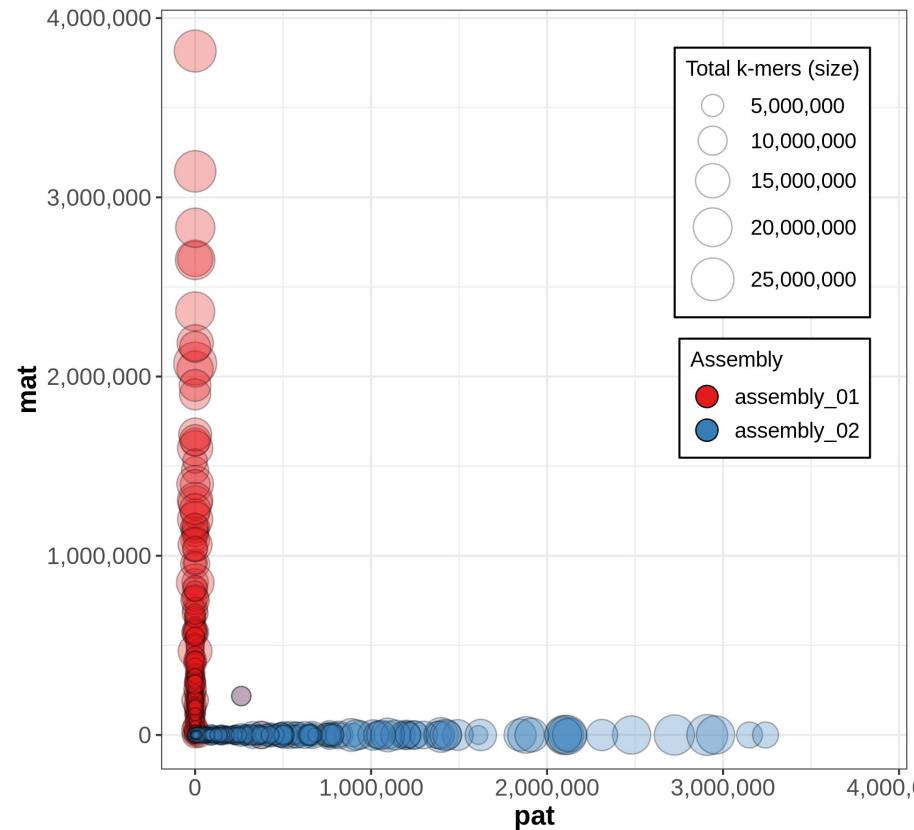
WF3-5: Outputs

Assembly with HiC phasing Zebra Finch



WF5: Outputs

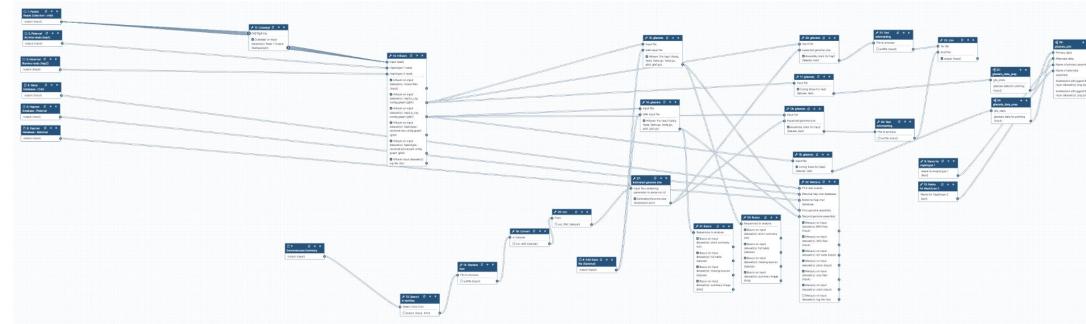
Trio Assembly Zebra Finch



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

Workflow: VGP HiC (Yahs)



Run Workflow

Scaffolded Assembly

31: Hifiasm Primary assembly

Scaffold generated by "VGP Bionano" workflows or contigs Generated by the "VGP Hifiasm" workflow.

No agp dataset available.

Sequence graphs encoded in agp format for re-scaffolding. (-g)

HIC Forward reads

4: SRR7126301_1

HiC forward reads in Fastq Format. The reads must be in the same order as reverse reads.

Most recent gfa1

27: Hifiasm on data 19: primary assembly contig graph

Scaffold generated by "VGP Bionano" workflows or contigs Generated by the "VGP Hifiasm" workflow.

HIC reverse reads

3: SRR7126301_2

HiC reverse reads in Fastq Format. The reads must be in the same order as forward reads.

Restriction enzyme sequences

2: re_bases.txt

File containing the list of restriction enzymes used by Arima technology, the sequences are separated by a comma. In the VGP project AWS, the restriction enzyme sequences can be found in the repository containing HiC data.

Estimated genome size - Parameter File

55: Estimated Genome size

Parameter file generated in the VGP Hifiasm workflow. Estimated reference genome size (in bp) for computing NGx statistics

SAK input file (Optional)

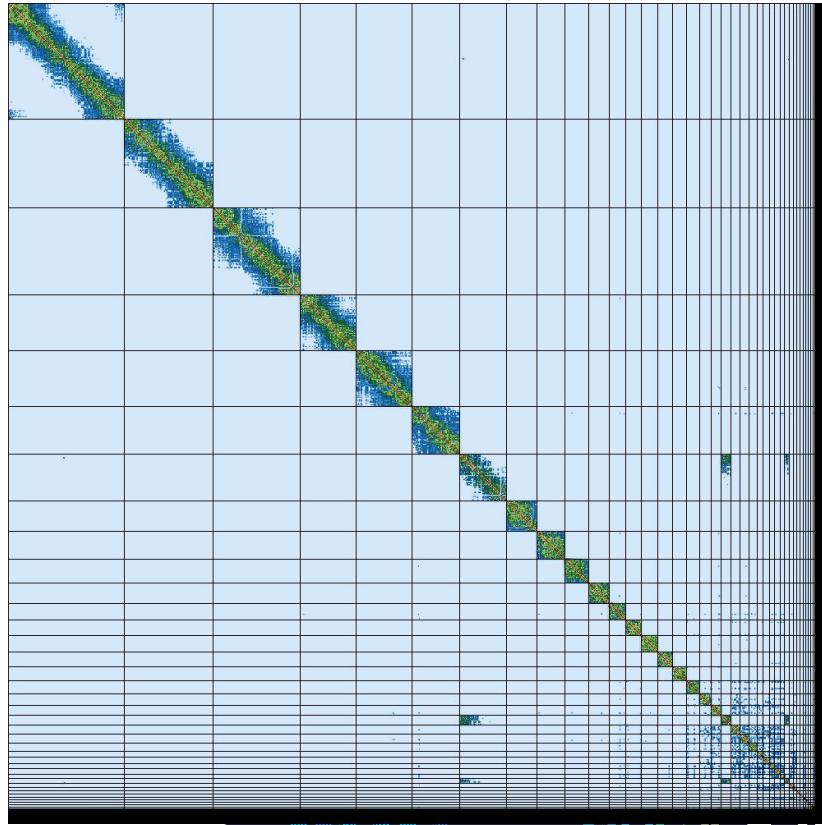
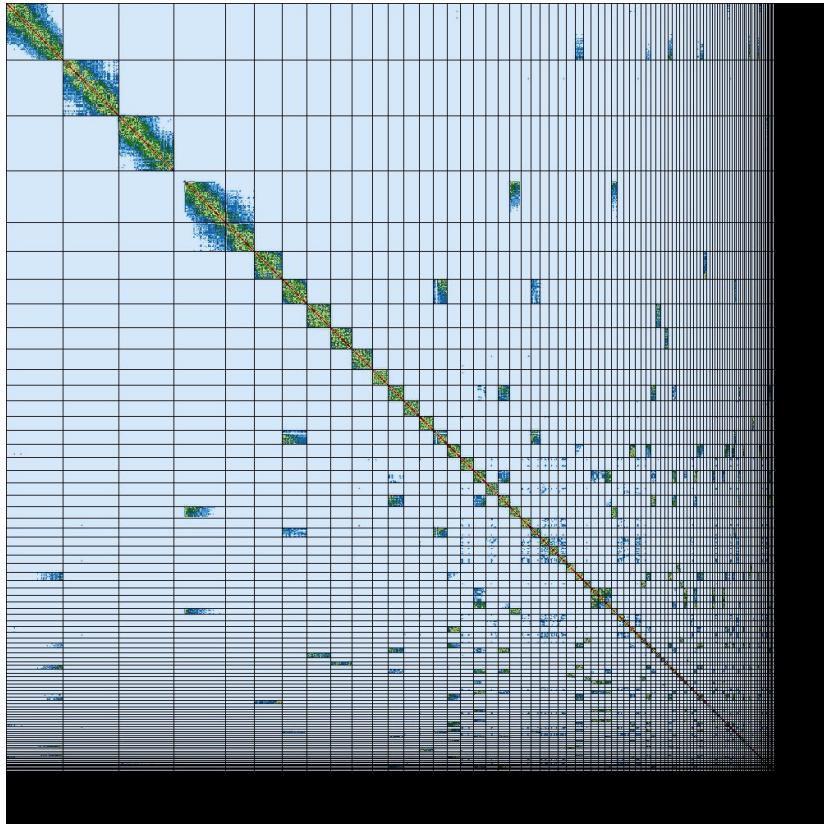
No text dataset available.

Data input 'swiss_army_knife' for conversion gfa to gasta of s1 and s2 assemblies (text) Set of instructions provided as an ordered list (--swiss-army-knife)

Expand to full workflow form.

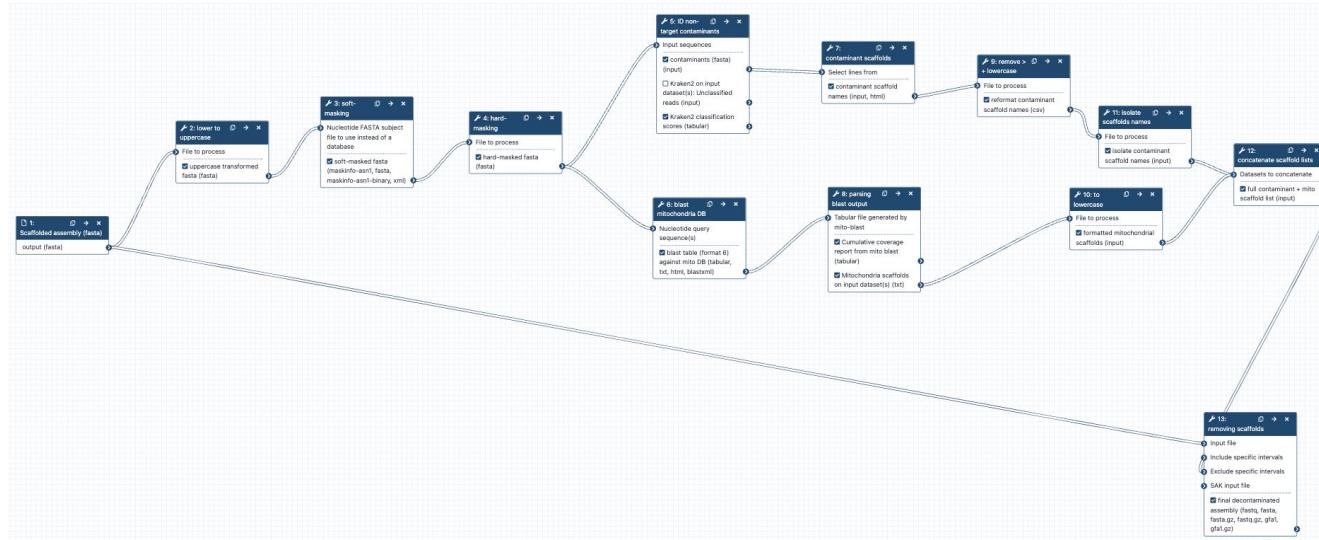
WF7-8: Outputs

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



Thank you!