

Genome assembly for everyone:

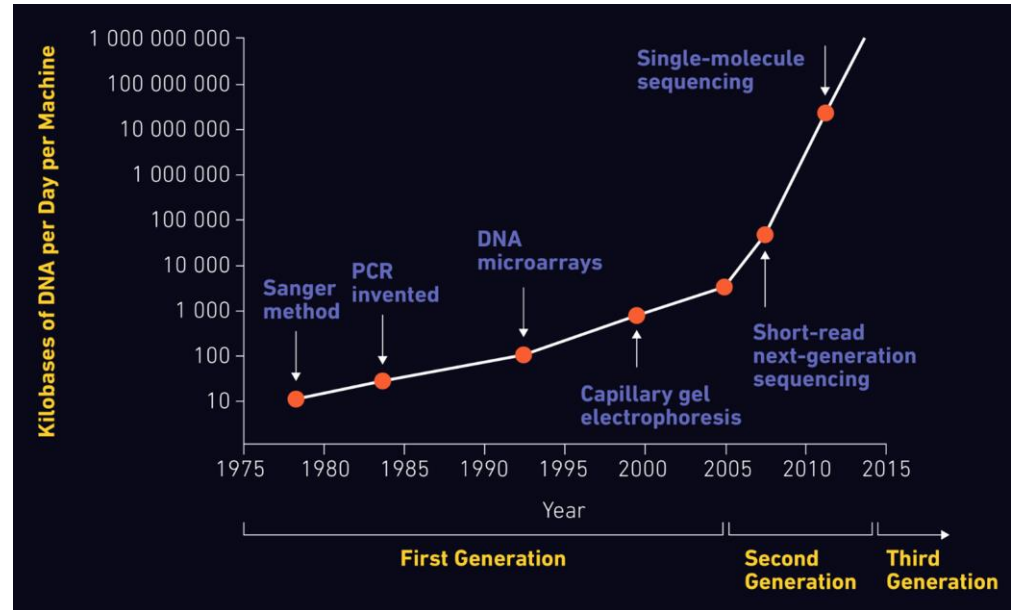
Galaxy Project-VGP collaboration

Cristóbal Gallardo
Galaxy Europe, University of Freiburg

Why do we need to make genome assembly accessible?

Fact:

Recent improvements in sequencing technologies and assembly tools promise to generate high-quality reference genomes for all species.



Why do we need to make genome assembly accessible?

Problem:

the genome assembly process is still laborious, costly, requires significant expertise.

Why do we need to make genome assembly accessible?

Solution:

make the pipeline freely accessible through the public computational infrastructure (**Galaxy**), and provide the required training (**Galaxy Training Network**).





VERTEBRATE GENOMES PROJECT

A PROJECT OF THE G10K CONSORTIUM

Phase I



54

BIRDS



58

MAMMALS



32

REPTILES



30

AMPHIBIANS



90

FISH



4

INVERTEBRATES

OVER 260 SPECIES



- **Open source platform** for accessible, reproducible, and transparent computational research
- **Public computational infrastructure** that provides a free analysis environment
 - European server: over 9000 CPU cores, 50TB of RAM, 4PB data storage
- The web-based graphical user interface allows interactive analyses

- **Training infrastructure Service (Tlaas)**
 - Private queue where only your training's jobs will run
 - See how your students are progressing
- **Galaxy Training Network (GTN) provides training material**



Tools



search tools



Upload Data

Get Data

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

FASTA/FASTQ

SAM/BAM

BED

COMMON GENOMICS TOOLS

Operate on Genomic Intervals

Annotation

...



Welcome to Galaxy for Genome Assembly

The **Genome Assembly Workbench** is a comprehensive set of analysis tools and consolidated workflows to assist in Genome Assembly. The workbench is based on the [Galaxy framework](#), which guarantees simple access, easy extension, flexible adaption to personal and security needs, and sophisticated analyses independent of command-line knowledge.

Vertebrate Genomes Project

The workbench is optimized to include all data, tools, and workflows associated with the [Vertebrate Genomes Project \(VGP\)](#). All raw data published by the VGP is available from the remote data repository **Genome Ark** in the data uploader. The VGP assembly workflows are available from the **Workflows** tab within **Shared Data**. As new assemblies are generated, they will appear in **Histories** in the **Shared Data** tab. Currently, we have assembled **23** genomes.

Human Pangenome Reference Project

The workbench has partnered with the [Human Pangenome Reference Consortium \(HPRC\)](#) to provide the latest genome assembly resources for the generation of high-quality diploid reference genomes. High-quality human datasets are available through the consortium, including multiple datatypes for the [HG002](#) benchmark and dozens of individuals from the [1000 Genomes Project](#). All data can directly be imported in Galaxy as input to the workflows.

Content

- [Vertebrate Genomes Project](#)
- [Human Pangenome Reference Project](#)
- [Get started](#)
- [VGP assembly training](#)

OPEN CHAT

History



search datasets



VGP assembly: training workflow

42.7 MB

83

71



92 : Pretext Snapshot on data 91



a list with 24 png datasets

#Bionano #Hi-C #hifi

91 : PretextMap on data 90



#Bionano #Hi-C #hifi

90 : Filter and merge on data 89 and data 88



#Bionano #Hi-C #hifi

89 : Map with BWA-MEM on data 6 and data 83 (mapped reads in BAM format)



Tool search panel

View panel

History panel



Tools



BWA-MEM



Upload Data

Show Sections

Map with BWA - map short reads (< 100 bp) against reference genome

Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome

BWA-MEM2 - map medium and long reads (> 100 bp) against reference genome

FilterSamReads include or exclude aligned and unaligned reads and read lists

Map with minimap2 A fast pairwise aligner for genomic and spliced nucleotide sequences

Cutadapt Remove adapter sequences from FASTQ/FASTA

Convert SOLiD output to fastq

Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome (Galaxy Version 0.7.17.2)

Will you select a reference genome from your history or use a built-in index?

Use a genome from history and build index

Built-ins were indexed using default options. See 'Indexes' section of help below

Use the following dataset as the reference sequence

83: SALSA on data 82 and data 70: FASTA assembly

You can upload a FASTA sequence to the history and use it as reference

Algorithm for constructing the BWT index

Auto. Let BWA decide the best algorithm to use

(-a)

Single or Paired-end reads

Single

Select between paired and single end data

Select fastq dataset

83: SALSA on data 82 and data 70: FASTA assembly

Specify dataset with single reads

Set read groups information?

History

search datasets

VGP assembly: training workflow

42.7 MB

83

71

89 : Map with BWA-MEM on data 6 and data 83 (mapped reads in BAM format)

#Bionano × #Hi-C × #hifi ×

14.6 GB

format qname_sorted.bam, database ?

[bwa_index] Pack FASTA... 0.24 sec
[bwa_index] Construct BWT for the packed sequence...

Binary bam alignments file

A web interface for each tool, so not **command line skills** are required for performing complex analysis



Queue

User	Created	Tool	State	Job Runner ID
9be9d8	2019-06-17 14:16:26	iuc/multiqc/multiqc/1.7	ok	859583
a81b3a	2019-06-17 14:14:38	devteam/samtool_filter2/samtool_filter2/1.8	ok	859579
a81b3a	2019-06-17 14:14:38	devteam/samtool_filter2/samtool_filter2/1.8	ok	859580
a81b3a	2019-06-17 14:14:38	devteam/samtool_filter2/samtool_filter2/1.8	ok	859578
a81b3a	2019-06-17 14:14:15	devteam/samtool_filter2/samtool_filter2/1.8	ok	859576
a81b3a	2019-06-17 14:14:15	devteam/samtool_filter2/samtool_filter2/1.8	ok	859575
a81b3a	2019-06-17 14:14:15	devteam/samtool_filter2/samtool_filter2/1.8	ok	859577
0c1d3c	2019-06-17 14:10:15	iuc/multiqc/multiqc/1.7	ok	859592

Jobs assigned to training groups preferentially run on a training machine with **dedicated resources**.



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	14
Climate	6
Computational chemistry	8
Ecology	8
Epigenetics	7
Genome Annotation	14

Welcome to the GTN!

Find out more about Galaxy Training Network



Video created by Geert Bonamie.

<https://training.galaxyproject.org/>

The GTN Materials in May 2022 has 260+ tutorials covering 23 topics, developed by over 260+ contributors!

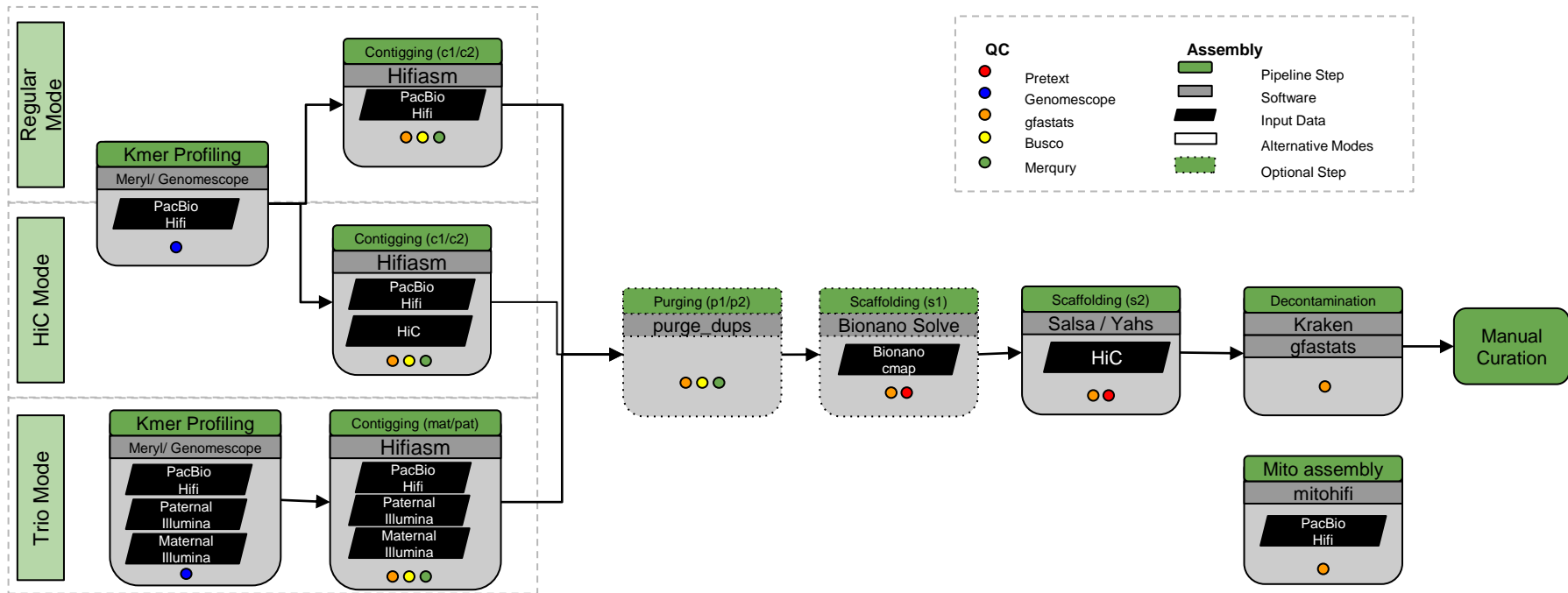


GTN tutorial are characterized by:

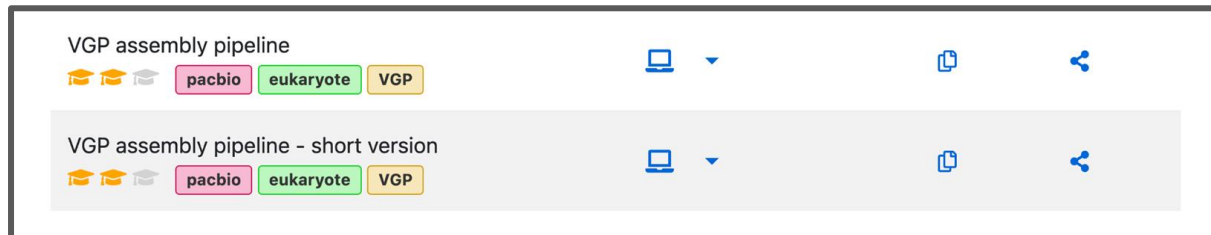
- Constructed around real world research problems
- No software installation is needed!



VERTEBRATE GENOMES PROJECT



Two training version available: extended and workflow-focused



Hands-on: Phased assembly with hifiasm

1. **Hifiasm** with the following parameters:
 - "Assembly mode": `Standard`
 - "Input reads": `HiFi_collection (trim)` (output of **Cutadapt**)
 - "Options for purging duplicates": `Specify`
 - "Purge level": `Light`
 - "Coverage upper bound": `114` (maximum depth previously obtained)
 - "Options for Hi-C partition": `Specify`
 - "Hi-C R1 reads": `Hi-C_dataset_F`
 - "Hi-C R2 reads": `Hi-C_dataset_R`
2. After the tool has finished running rename its outputs as follows:
 - Rename the `Hi-C hap1 balanced contig graph` as `Primary contigs graph` and add a `#primary` tag
 - Rename the `Hi-C hap2 balanced contig graph` as `Alternate contigs graph` and add a `#alternate` tag

[FAQs](#) | [Gitter Chat](#) | [Help Forum](#)

Hands-on: VGP purge assembly with purge_dups pipeline workflow

1. Click in the **Workflow** menu, located in the top bar
2. Click in the **Run workflow** button corresponding to `VGP purge assembly with purge_dups pipeline`
3. In the **Workflow: VGP purge assembly with purge_dups pipeline** menu:
 - `"Hifiasm Primary assembly": 39: Hifiasm HiC hap1`
 - `"Hifiasm Alternate assembly": 40: Hifiasm HiC hap2`
 - `"Pacbio Reads Collection - Trimmed": 22: Cutadapt`
 - `"Genomescope model parameters": 20: Genomescope on data 13 Model parameters`
4. Click in the **Run workflow** button

[FAQs](#) | [Gitter Chat](#) | [Help Forum](#)

Workflow Availability: IWC



README.md

IWC - Intergalactic Workflow Commission

Galaxy Workflow Tests for push and PR passing chat on gitter

The IWC maintains high-quality Galaxy Workflows

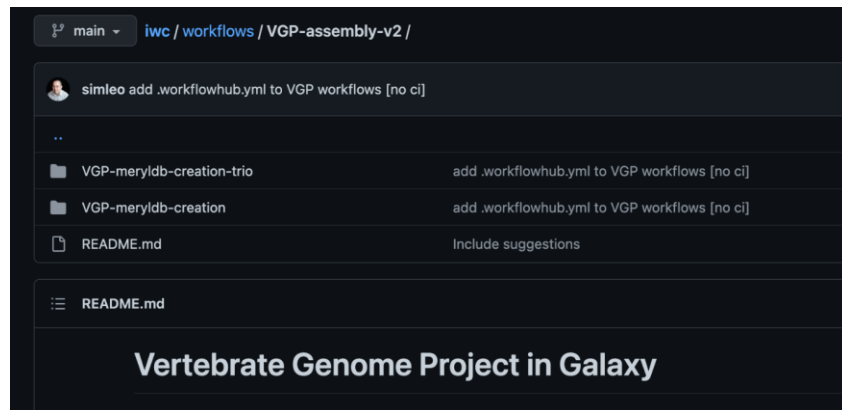
Workflows are categorized in the [workflows](#) directory, and listed in [Dockstore](#) and [WorkflowHub](#).

All workflows are reviewed and tested before publication and with every new Galaxy release. Deposited workflows follow [best practices](#) and are versioned using github releases. Workflows also contain important metadata, such as:

- License
- Author
- Institutes

Additionally the IWC will collect further best practices, tips and tricks, FAQs and assist the community in designing high-quality Galaxy workflows.

<https://github.com/galaxyproject/iwc>



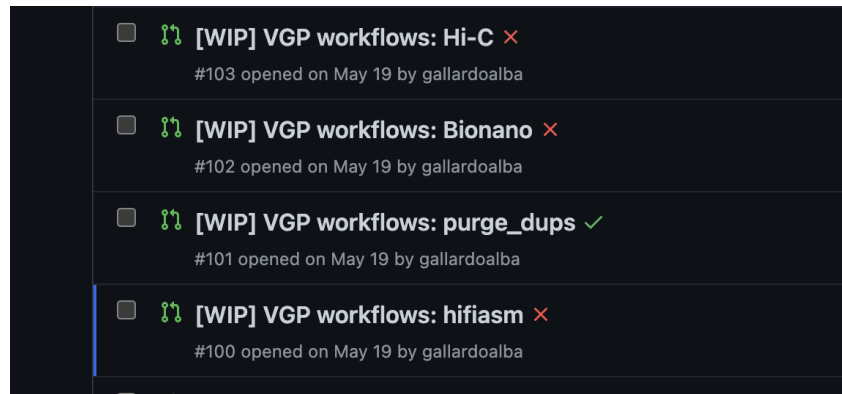
main iwc / workflows / VGP-assembly-v2 /

simleo add .workflowhub.yml to VGP workflows [no ci]

- VGP-meryldb-creation-trio add .workflowhub.yml to VGP workflows [no ci]
- VGP-meryldb-creation add .workflowhub.yml to VGP workflows [no ci]
- README.md Include suggestions

README.md

Vertebrate Genome Project in Galaxy



- 🚧 [WIP] VGP workflows: Hi-C ✗
#103 opened on May 19 by gallardoalba
- 🚧 [WIP] VGP workflows: Bionano ✗
#102 opened on May 19 by gallardoalba
- 🚧 [WIP] VGP workflows: purge_dups ✓
#101 opened on May 19 by gallardoalba
- 🚧 [WIP] VGP workflows: hifiasm ✗
#100 opened on May 19 by gallardoalba

Genomes Assembled on public Galaxy instances

- 21 Genomes in 6 months
 - 10 birds, 2 amphibians, 2 fish, 6 mammals, 1 reptile
 - 5 more in the works



Florida Museum photos by
Kenneth Krysko



Jacob Drucker
Hawaii, United States
[Macaulay Library ML 141519531](#)

Acknowledgments

VGP team:

- Giulio Formenti
- Linelle Abueg
- Nadolina Brajuka
- Marc Palmada Flores

Galaxy team:

- Alex Ostrovsky
- Delphine Lariviere
- Anton Nekrutenko
- Bjorn Grüning
- Michael Schatz
- And everyone else

