Genome assembly for everyone:

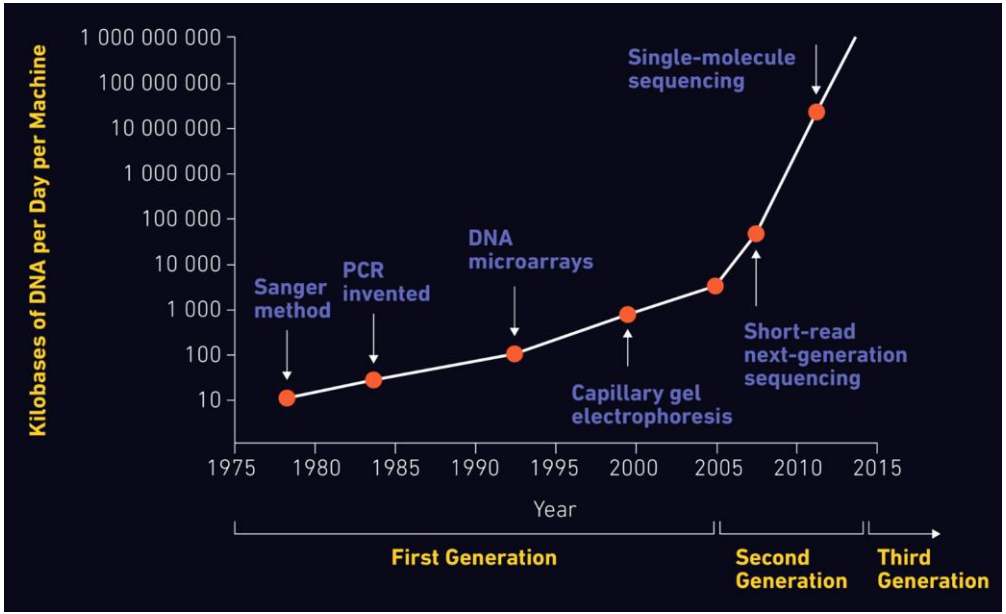
Galaxy Project-VGP collaboration

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.



https://www.technologynetworks.com/genomics/articles/an-overview-of-next-generation-sequencing-346532

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**).

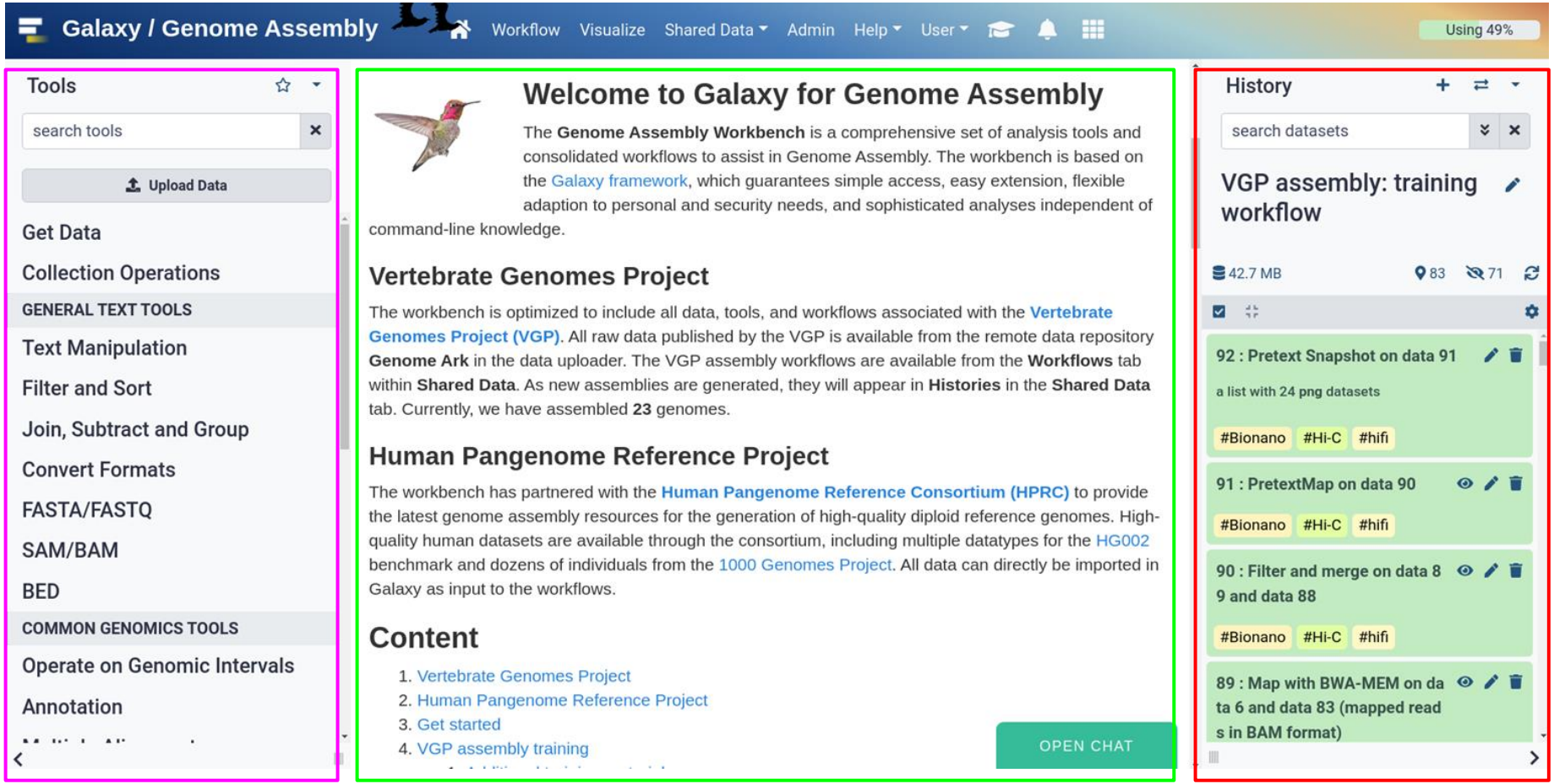
VERTEBRATE GENOMES PROJECT

A PROJECT OF THE G10K CONSORTIUM

Phase I

- **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

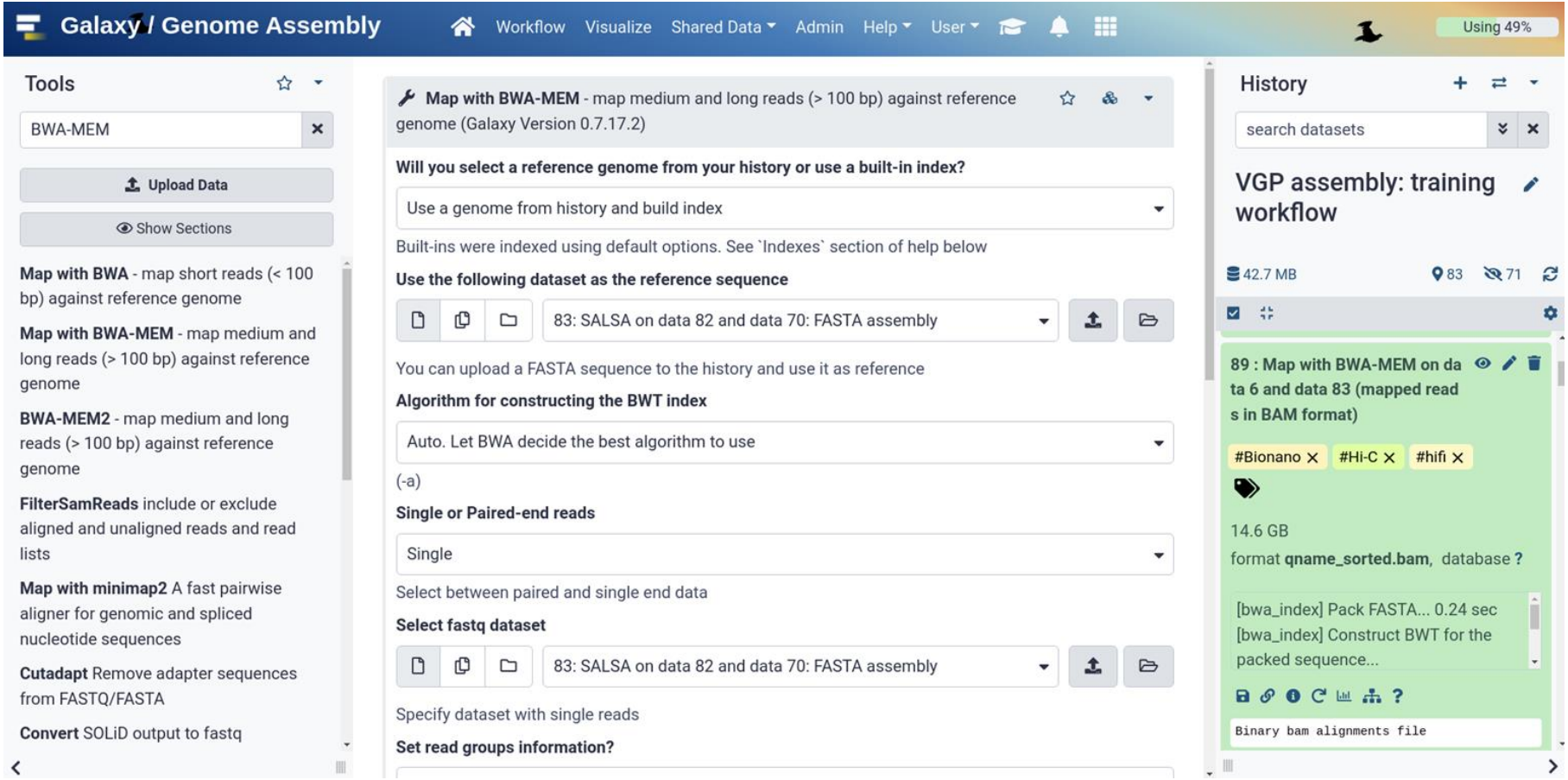
- **Training infrastructure Service (Tlaas)**
  - Private queue where only your training's jobs will run
  - See how your students are progressing
- **Galaxy Training Network (GTN) provides training material**



Tool search panel

View panel

History panel

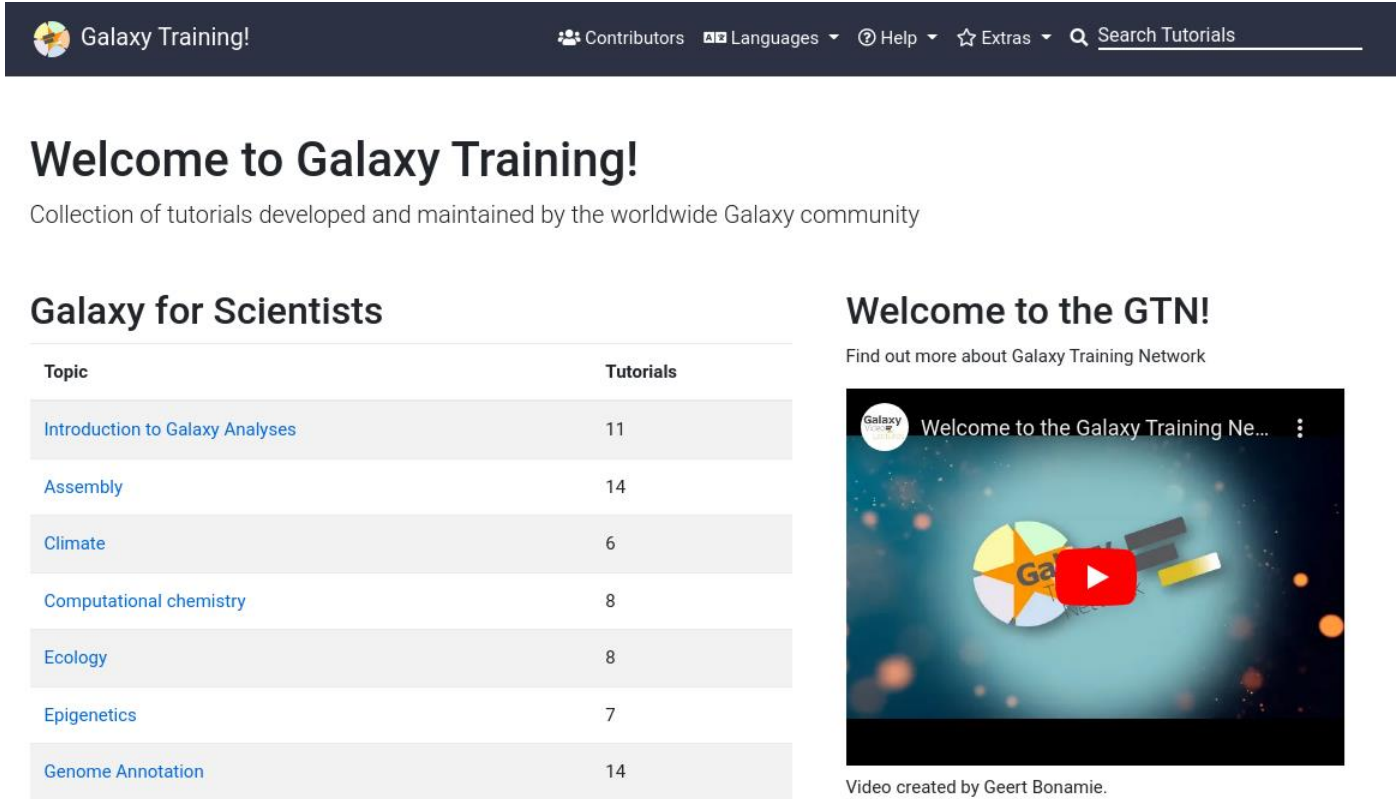


A web interface for each tool, so not **command line skills are required** for performing complex analysis

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
0c1dac	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**.



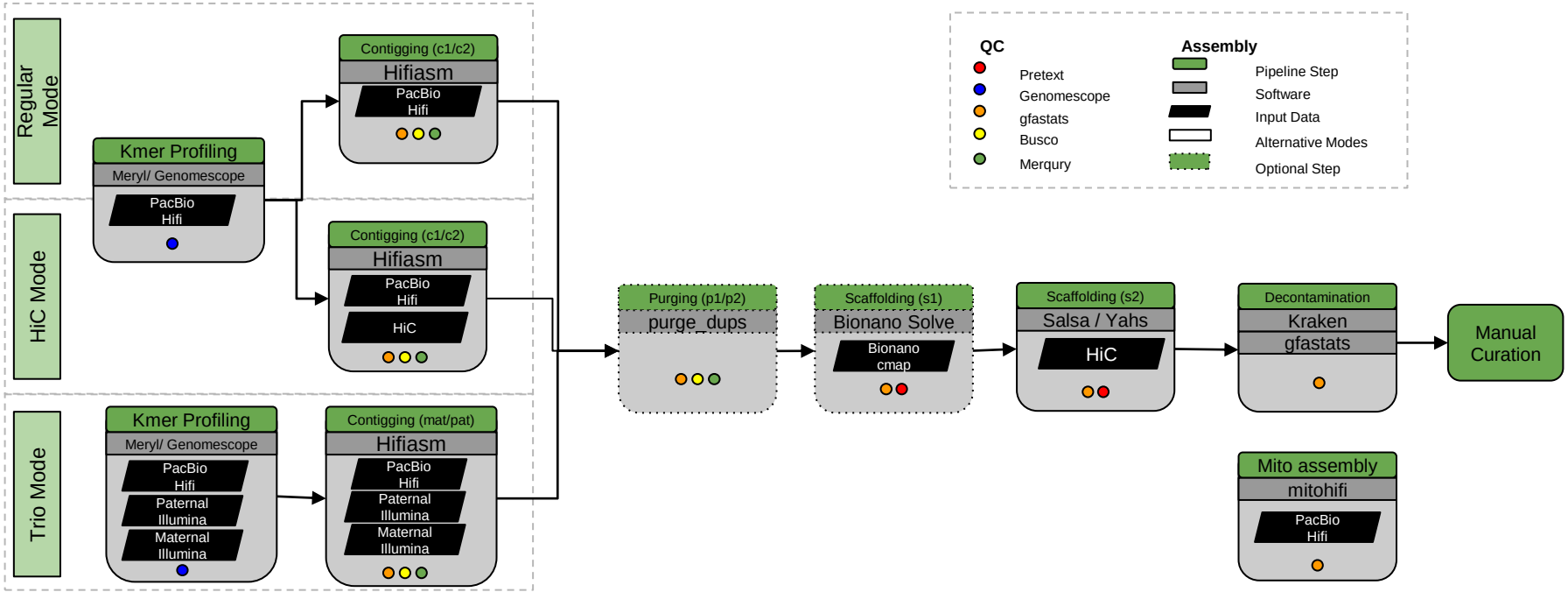
https://training.galaxyproject.org/

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

GTN tutorial are characterized by:

- Constructed around a **real world research problems**
- ... to complete them is bundled with the tutorial
- 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** buttom 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` buttom

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

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