INSTRUCTOR GUIDE · BIOINFORMATICS WORKSHOP

How to Teach This Workshop

Raw Data to Biology, in Two Days

A field manual for running the eight-module, two-day bioinformatics workshop — what to prep, how to pace each block, where learners get stuck, and how to keep a room of mixed skill levels moving together.

8 modules · 2 days · ~8 hrs | Beginner → confident | Hands-on / notebook-driven

01 The shape of the two days

Day 1 takes learners from a raw sequencer file to aligned reads. Day 2 turns those reads into biology — variants, expression, figures — and ends with an unassisted capstone. Each module is one notebook. Your job is less "lecture" and more "narrate the pipeline while they run it."

Day 1 — Raw Data to Aligned Reads

09:00	**M1** Linux CLI (45m)
09:45	**M2** Sequence formats (45m)
10:30	break
10:45	**M3** QC + trimming (60m)
11:45	**M4** Alignment (75m)

Day 2 — From Reads to Biology

09:00	**M5** Variant calling (70m)
10:10	**M6** RNA-seq + DESeq2 (70m)
11:20	break
11:35	**M7** Visualization (50m)
12:25	**M8** Capstone (35m)

The one principle to teach by

Every module is a station on one conveyor belt: FASTQ → QC → alignment → variants/expression → interpretation. Whenever a learner is lost, point back to the belt and ask "what's coming in, what's going out?" The tools change; that question never does.

02 Before the workshop

One week out

- Send participants the setup instructions (`environment/setup.sh`) and the checklist below.
- Test the environment on a **fresh** machine — install takes ~20 min.
- Confirm tool versions in `conda_env.yml` are still current.
- Run `download_reference.sh`; confirm chr22 files are intact.

The day before

- Run **every** pipeline script end-to-end on sample data.
- Pre-download the chr22 reference so learners copy it (saves 5 min Day 1).
- Load a shared folder / USB with: Miniconda installers (all platforms), a pre-built conda package cache, `chr22.fa` + `.fai`, `chr22.gtf`, and the BWA-MEM2 index.

Morning of Day 1

Arrive 30 minutes early. Test the projector and launch Jupyter Lab on your machine. Make sure WiFi can survive N participants downloading at once — if in doubt, serve files locally: `python3 -m http.server`.

03 Pacing each module

Module	Watch for	Move
M1 CLI	Runs short or long — people find the shell easy or hard in equal measure.	If fast, introduce `tmux` for parallel pipelines. If slow, skip exercise 8 (assign as homework).
M2 Formats	Eyes glaze on FLAGS / CIGAR.	Decode one real SAM line together on the board rather than reading the spec.
M3 QC	Easy win — keep the energy.	Have them compare a good vs bad FastQC report side by side.
M4 Align	Longest hands-on block; alignment takes minutes.	**Start it running, then explain while it runs.** Don't wait in silence.
M5 Variants	GATK is compute-heavy; may crawl on weak laptops.	Run on the instructor machine and project output; let learners follow along.
M6 RNA-seq	The conceptual jump (DNA vs RNA, normalization) is the hard part, not the code.	Anchor on genotype vs phenotype before touching DESeq2.
M7 Viz	Tempting to rabbit-hole on aesthetics.	Insist every figure answers one question. Move on once it does.
M8 Capstone	Mixed finish times.	Keep Tier 1 mandatory; Tier 2/3 optional for advanced learners.

04 Energy management

- The **Day 1 break at 10:30** is non-negotiable — people are absorbing a lot of new vocabulary.
- The **Day 2 break at 11:20** lands naturally after DESeq2 finishes its processing.
- After each module, run a 30-second "what did we just learn?" check-in. It surfaces confusion early and resets attention.
- Mixed-pace rooms drift apart fast. Pair a fast learner with a stuck one — teaching cements both.

05 Common issues & fixes

"conda activate isn't working"

```
conda init bash && source ~/.bashrc
conda activate bioworkshop
```

"Java not found" (Picard / GATK)

```
conda install -c conda-forge \
  openjdk -n bioworkshop
```

"BWA-MEM2 index not found"

The index must sit in the same directory as the `.fa` file. Files:

.bwt.2bit.64 .pac .sa .ann .amb

Low-RAM machines (< 8 GB)

- Drop `--threads` to 2 everywhere.
- `samtools sort -m 1G` to cap sort memory.
- M5 GATK may stall — demo on the projector while learners follow.

Windows participants

- Must use **WSL2** — not Git Bash or Cygwin.
- Docker fallback: `docker run -p 8888:8888 bioworkshop-image`.

06 Questions to anticipate

Q What's the difference between BWA-MEM and BWA-MEM2?
A MEM2 is 3–4× faster, same algorithm, uses SIMD/AVX instructions. Always use MEM2.

Q Why not just use STAR instead of HISAT2?
A Both are good. HISAT2 uses far less RAM (~8 GB vs ~30 GB genome-wide), which matters in a classroom. STAR is faster on large genomes when RAM is plentiful.

Q featureCounts or Salmon?
A Salmon (quasi-mapping) is faster and great for well-annotated genomes. featureCounts gives a BAM you can inspect visually — better for teaching.

Q When do I use hard filtering vs VQSR?
A VQSR needs many samples (30+ for SNPs, 50+ for indels). Single-sample or small cohorts: hard filtering or DeepVariant.

07 Pre-flight checklists

Send to participants (1 week out)

- [ ] Install Miniconda
- [ ] Clone / download the workshop repo
- [ ] `bash environment/setup.sh` (~20 min)
- [ ] `bash data/reference/download_reference.sh` (~5 min)
- [ ] Verify: `conda activate bioworkshop && fastqc --version`
- [ ] Install IGV and RStudio (M6 R section)
- [ ] Free up 10 GB of disk

Instructor pre-workshop

- [ ] conda env builds from `conda_env.yml`
- [ ] All tools answer `--version`
- [ ] Jupyter Lab launches; notebooks open
- [ ] R packages install (DESeq2, clusterProfiler)
- [ ] chr22 indexed (BWA-MEM2 + HISAT2)
- [ ] All four pipeline scripts run to completion
- [ ] Projector + backup USB + power outlets confirmed