Bioinformatics Containerization Tutorial
Session 2: Full Toolbox with Docker Compose

Dr. İsmail K. Sağlam

November 15, 2025

Contents

1 Overview

In this session you will:

- Extend the base container into a full population-genomics toolbox.
- Use `docker compose` to define reusable services.
- Run:
  - ANGSD genotype calling.
  - PCAngsd structure, inbreeding, and selection scan.
  - R-based plotting scripts for PCA and admixture.
- Learn how to run commands interactively and override defaults.

We assume:

- `isophya-course:0.1` is built (Session 1).
- Your directory layout is:

```
bioinf-containers-course/
  containers/     # Dockerfile, environment.yml
  data/           # BAMs, metadata (.clst, .info), filters (.sites, .chr)
  results/        # outputs
  scripts/        # 01_call_genotypes.sh, 02_pcangsd_pipeline.sh, R scripts
  workflow/       # compose.yaml, main.nf, nextflow.config
  hpc/            # Apptainer.def, run_apptainer.sbatch
```

2 Build or verify the toolbox image

If you have not built the image yet (or rebuilt it):

```
cd ~/bioinf-containers-course

docker build -t isophya-course:0.1 -f containers/Dockerfile .
```

This image contains:

- `ANGSD`, `samtools`, `bcftools`
- `R` and required R packages
- `pcangsd` and dependencies

3 Compose file layout

Create `workflow/compose.yaml` with a common base and services:

```
version: "3.9"

x-common: &common
  image: isophya-course:0.1
  user: "${UID}:${GID}"
```

```
  working_dir: /workspace
  volumes:
    - ../workflow:/workspace:rw
    - ../data:/data:ro
    - ../scripts:/workspace/scripts:ro
    - ../results:/results:rw
  tty: true

services:
  toolbox:
    <<: *common
    container_name: isophya-toolbox
    command: ["bash"]

  angsd-call:
    <<: *common
    container_name: isophya-angsd-call
    environment:
      POP: "isophya71"
      THREADS: "8"
      REF: "/data/isophya_contigs_CAYMY.fasta"
    command:
      - bash
      - -lc
      - |
        ./scripts/01_call_genotypes.sh

  pcangsd-pipeline:
    <<: *common
    container_name: isophya-pcangsd-pipeline
    environment:
      POP: "isophya71"
      THREADS: "8"
      BEAGLE: "/results/genotypes/isophya71.beagle.gz"
    command:
      - bash
      - -lc
      - |
        ./scripts/02_pcangsd_pipeline.sh

  r:
    <<: *common
    container_name: isophya-r
    command: ["bash","-lc","R"]
```

4 ANGSD genotype calling service

4.1 Script layout

Your `scripts/01_call_genotypes.sh` is responsible for running ANGSD and writing outputs under `/results`.

Typical structure (simplified):

```
#!/usr/bin/env bash
set -euo pipefail
```

```
POP="${POP:-isophya71}"
THREADS="${THREADS:-4}"
REF="${REF:-/data/isophya_contigs_CAYMY.fasta}"

BAMLIST="/data/${POP}.bamlist"
OUT_DIR="/results/genotypes"
mkdir -p "${OUT_DIR}"

angsd \
  -bam "${BAMLIST}" \
  -ref "${REF}" \
  -out "${OUT_DIR}/${POP}" \
  -GL 1 \
  -doMajorMinor 1 \
  -doMaf 1 \
  -doGlf 2 \
  -doGeno 5 \
  -doBcf 1 \
  -only_proper_pairs 1 \
  -doPost 1 \
  -postCutoff 0.80 \
  -minMapQ 10 \
  -minQ 20 \
  -SNP_pval 1e-12 \
  -minMaf 0.05 \
  -nThreads "${THREADS}"

# Example: convert BCF to VCF
bcftools convert \
  -O b \
  -o "${OUT_DIR}/${POP}.bcf" \
  "${OUT_DIR}/${POP}.bcf"
```

4.2 Run the ANGSD step

From the project root:

```
cd ~/bioinf-containers-course

docker compose -f workflow/compose.yaml run --rm angsd-call
```

Outputs appear under:

```
ls results/genotypes
```

5 PCAngsd + selection + plotting service

5.1 Script layout

`scripts/02_pcangsd_pipeline.sh` coordinates:

- PCAngsd structure and inbreeding.
- PCAngsd pcadapt selection scan.

- R postprocessing (`pcadapt.R`).
- PCA/admixture plotting (`plotPCA.R`, `plotAdmix.R`).

Sketch (abridged):

```
#!/usr/bin/env bash
set -euo pipefail

POP="${POP:-isophya71}"
THREADS="${THREADS:-8}"
BEAGLE="${BEAGLE:-/results/genotypes/${POP}.beagle.gz}"

STRUCT_OUT="/results/structure/${POP}"
SEL_PREFIX="/results/selection/${POP}"
PLOT_DIR="/results/plots"

mkdir -p "$(dirname "${STRUCT_OUT}")" \
         "$(dirname "${SEL_PREFIX}")" \
         "${PLOT_DIR}"

RSCRIPT_PCADAPT="/workspace/scripts/pcadapt.R"
RSCRIPT_PCA="/workspace/scripts/plotPCA.R"
RSCRIPT_ADMIX="/workspace/scripts/plotAdmix.R"

echo "[1/4] pcangsd: structure + inbreeding"
pcangsd \
  --beagle "${BEAGLE}" \
  --admix \
  --inbreed_samples \
  --inbreed_sites \
  --threads "${THREADS}" \
  --out "${STRUCT_OUT}"

echo "[2/4] pcangsd: pcadapt selection"
pcangsd \
  -b "${BEAGLE}" \
  --hwe "${STRUCT_OUT}.lrt.sites" \
  --pcadapt \
  --sites_save \
  -o "${SEL_PREFIX}"

echo "[3/4] R: pcadapt z-scores → statistics"
Rscript "${RSCRIPT_PCADAPT}" \
  "${SEL_PREFIX}.pcadapt.zscores" \
  "${SEL_PREFIX}"

echo "[4/4] R: PCA and admixture plots"
Rscript "${RSCRIPT_PCA}" \
  -i "${STRUCT_OUT}.cov" \
  -c 1-2 \
  -a "/data/${POP}.clst" \
  -o "${PLOT_DIR}/${POP}.pca.pdf"

Rscript "${RSCRIPT_ADMIX}" \
  "${STRUCT_OUT}.admix.2.Q" \
  "/data/${POP}.info"
```

5.2 Run the PCAngsd pipeline

```
docker compose -f workflow/compose.yaml run --rm pcangsd-pipeline
```

Check results:

```
ls results/structure
ls results/selection
ls results/plots
```

6 Interactive use and overriding defaults

6.1 Start an interactive toolbox shell

```
docker compose -f workflow/compose.yaml run --rm toolbox
```

Inside:

```
angsd -h
pcangsd -h
R --version
ls /data
ls /results
```

6.2 Run a one-off ANGSD command via service

```
docker compose -f workflow/compose.yaml run --rm angsd-call \
  -- angsd -h
```

6.3 Run PCAngsd manually

```
docker compose -f workflow/compose.yaml run --rm toolbox
# inside:
pcangsd \
  --beagle /results/genotypes/isophya71.beagle.gz \
  --admix \
  --threads 8 \
  --out /results/pcangsd/isophya71_demo
```

6.4 Plotting only

PCA plot

```
docker compose -f workflow/compose.yaml run --rm toolbox
# inside:
Rscript /workspace/scripts/plotPCA.R \
  -i /results/structure/isophya71.cov \
  -c 1-2 \
```

```
  -a /data/isophya71.clst \
  -o /results/plots/isophya71.pca.pdf
```

Admixture barplot

```
docker compose -f workflow/compose.yaml run --rm toolbox
# inside:
Rscript /workspace/scripts/plotAdmix.R \
  /results/structure/isophya71.admix.2.Q \
  /data/isophya71.info
```

---

7 Summary

After Session 2 you can:

- Use `docker compose` to launch an interactive toolbox shell.
- Run ANGSD and PCAngsd pipelines via services.
- Override defaults and run arbitrary commands inside the container.
- Generate PCA and admixture plots using R scripts.