QBIO305 • GROUP RESEARCH PROJECTS

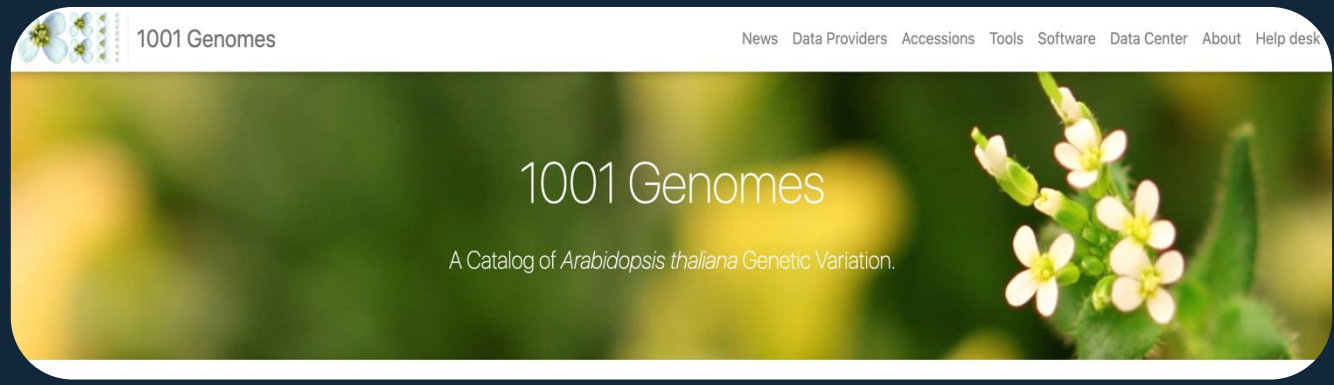
From climate maps to genomic hypotheses

Design your own *Arabidopsis thaliana* population genomics project

Scan to explore my research, teaching materials and interactive apps.

tahirali-biomics.github.io

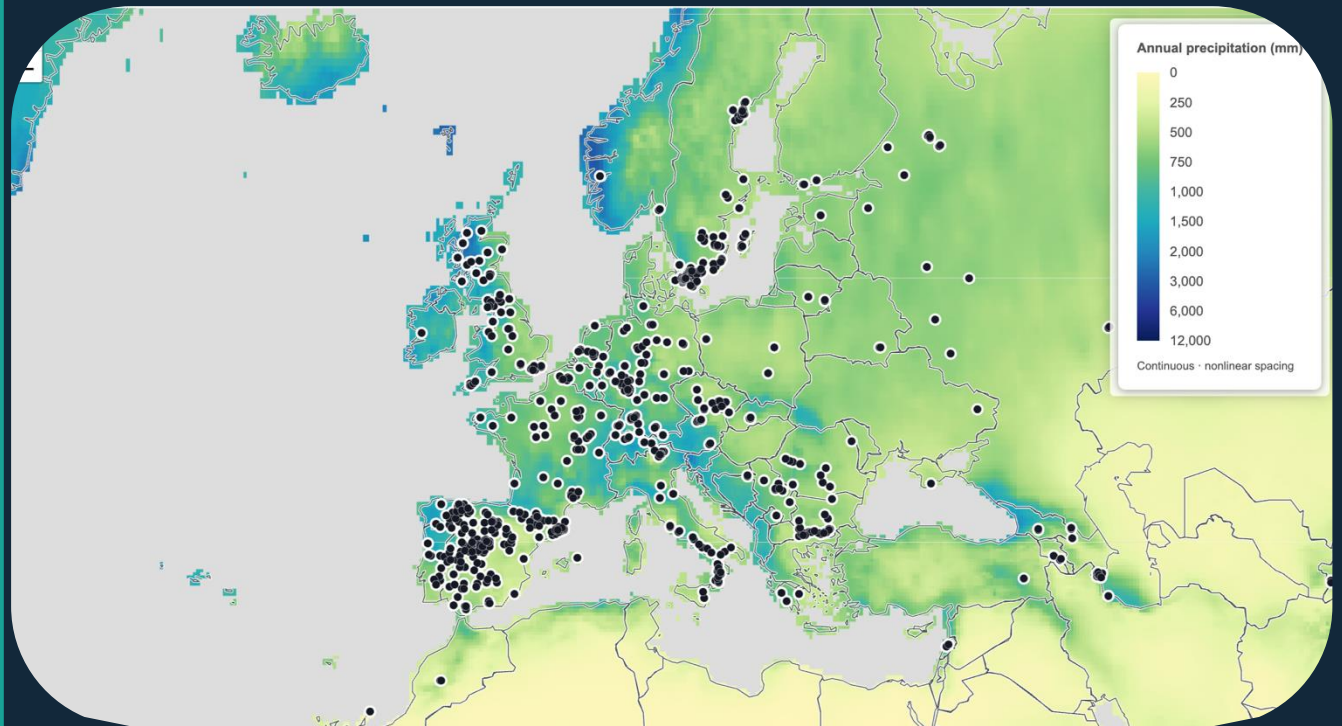
Juliette de Meaux | Tahir Ali | Swan Portalier



UNIVERSITY OF COLOGNE

AG de Meaux
Population Genetics
Dr. Tahir Ali

Arabidopsis 1001 Genomes · Climate Explorer



What will your group investigate?

By the end, you should be able to turn an ecological observation into a feasible genomic study.

01 Explore

Map accession locations and environmental contrasts before deciding which samples to compare.

02 Ask & predict

Write a biological question, hypothesis and measurable genomic prediction.

03 Request the data

- Identify accessions using the **1001 Genomes Climate Explorer,** then obtain their SRA IDs from the **1001_genome_accession_list**
- Identify candidate genes using **GFF_final_with_annotation_2.gff**
- **Submit the accession list, SRA IDs, and candidate genes** to receive your group-specific VCF.

Explore → Ask → Predict → Select accessions & genes → Request VCF

Why *Arabidopsis thaliana*?

A small plant with unusually rich genomic and geographic resources.



5 chromosomes

Compact, annotated reference genome

1,135 accessions

Public database

1001 Genomes metadata

Many environments

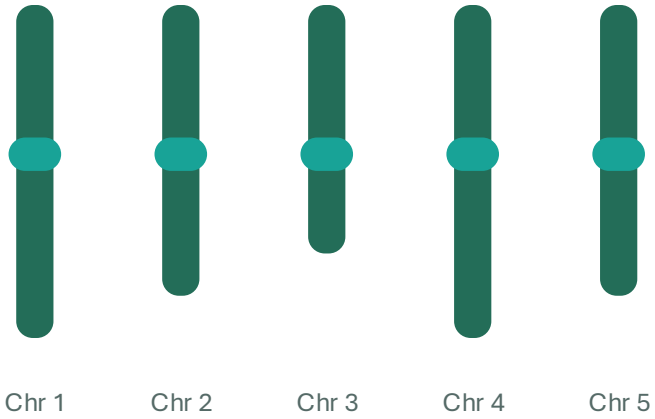
Natural accessions span strong climatic gradients

Gene knowledge

Published literature and phenotypes

TAIR

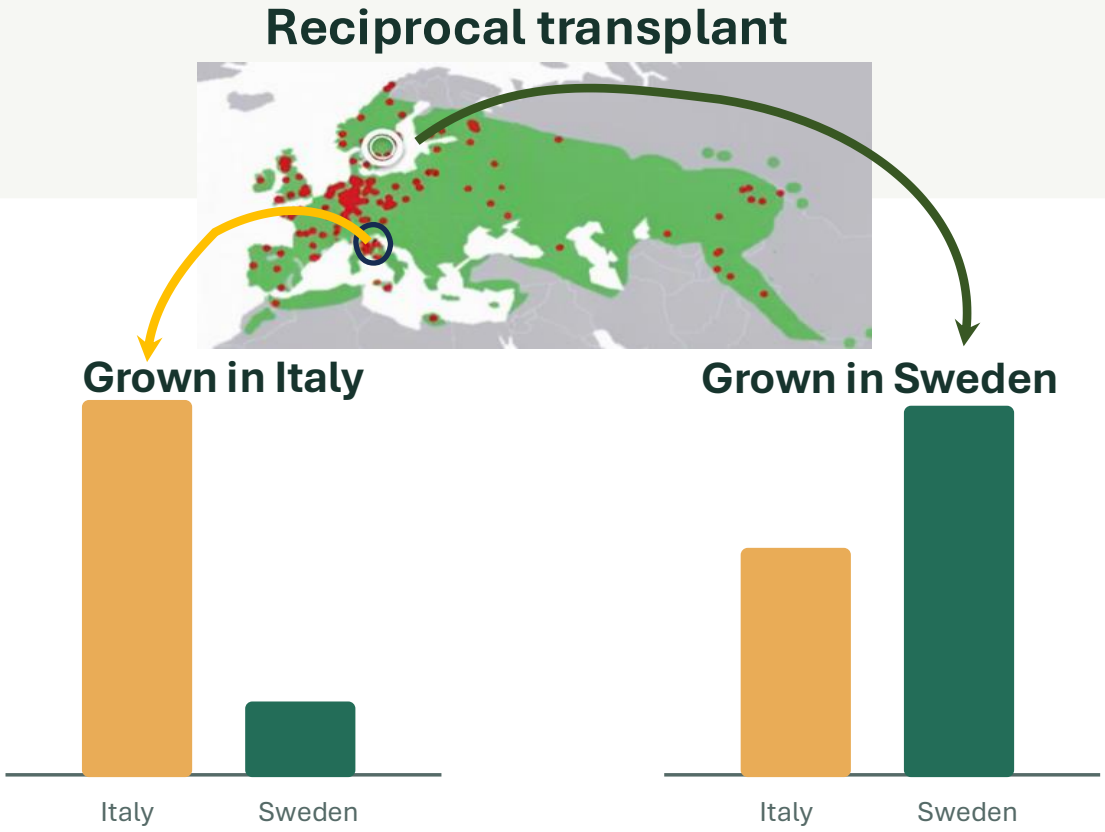
The nuclear genome



Source: 1001 Genomes Project (1,135 genomes); existing QBio305 introduction, slides 13–16.

Local adaptation is a biological question

Genomic differences alone are not direct evidence that a population is better adapted.



The real test

Plants from Italy and Sweden were grown at both sites. Local plants generally produced more fruit in their home environment.

Genomic differentiation can suggest adaptation, but it does not demonstrate higher fitness

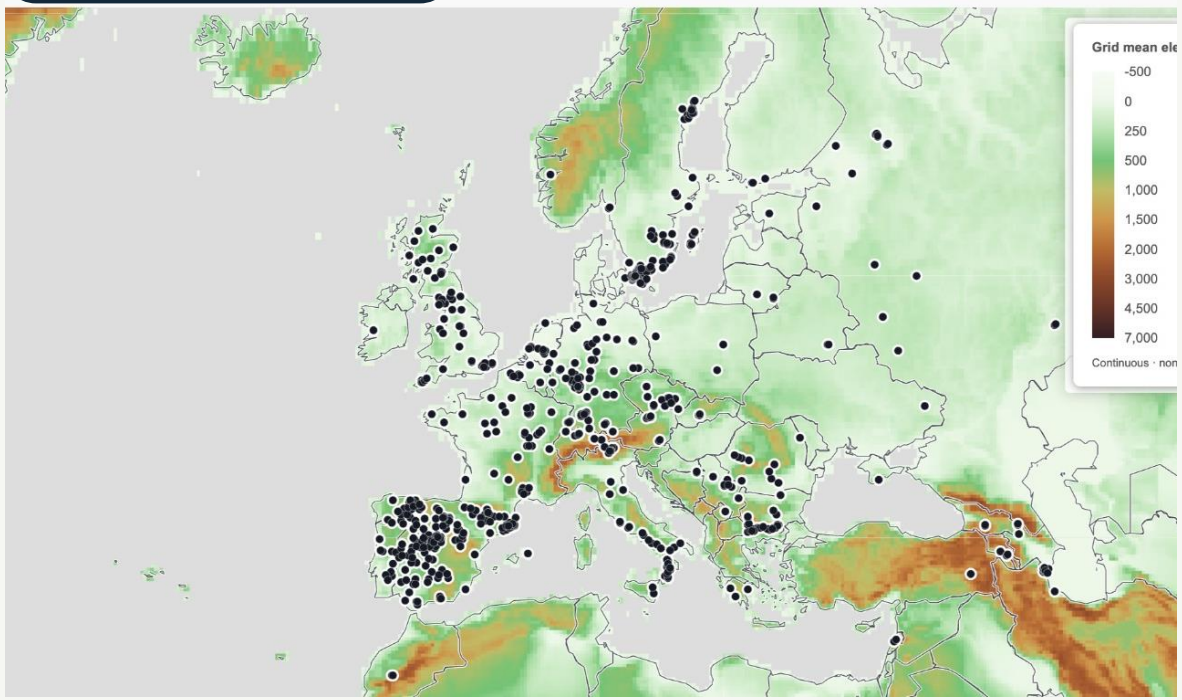
Local adaptation—each population performs best in its home environment; moving them reduces fitness

Ågren & Schemske 2012, New Phytologist, doi:10.1111/j.1469-8137.2012.04112.x; bars are conceptual, not published estimates.

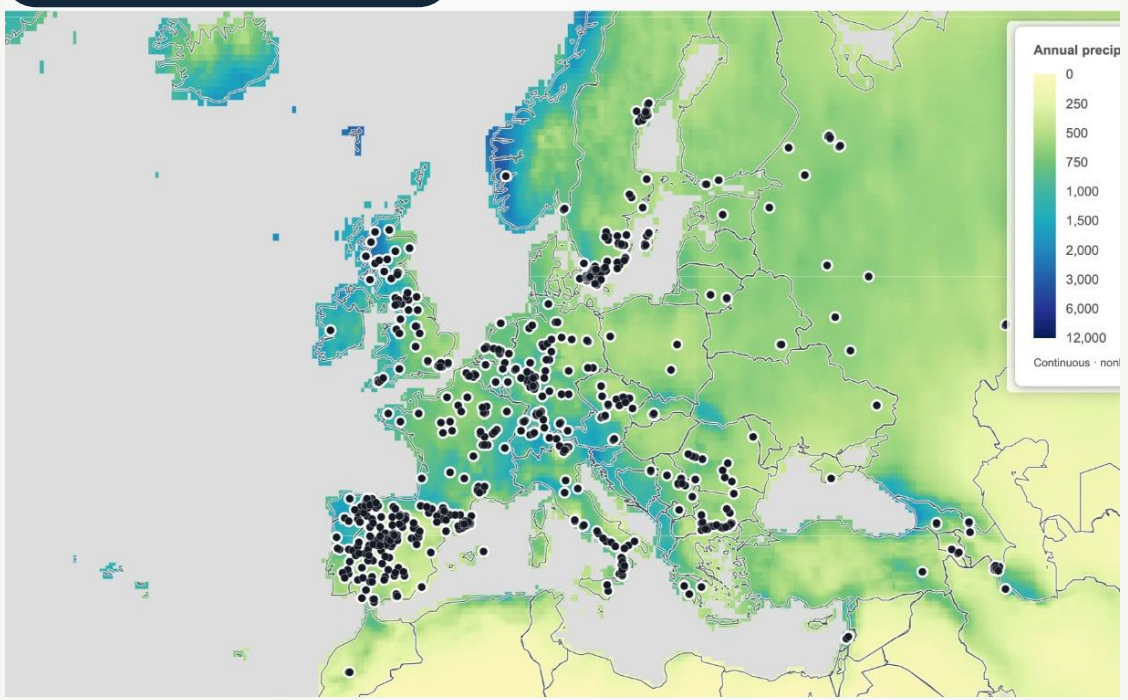
One map. Several possible questions.

The same accessions can be explored against precipitation, temperature, elevation and other layers.

ELEVATION



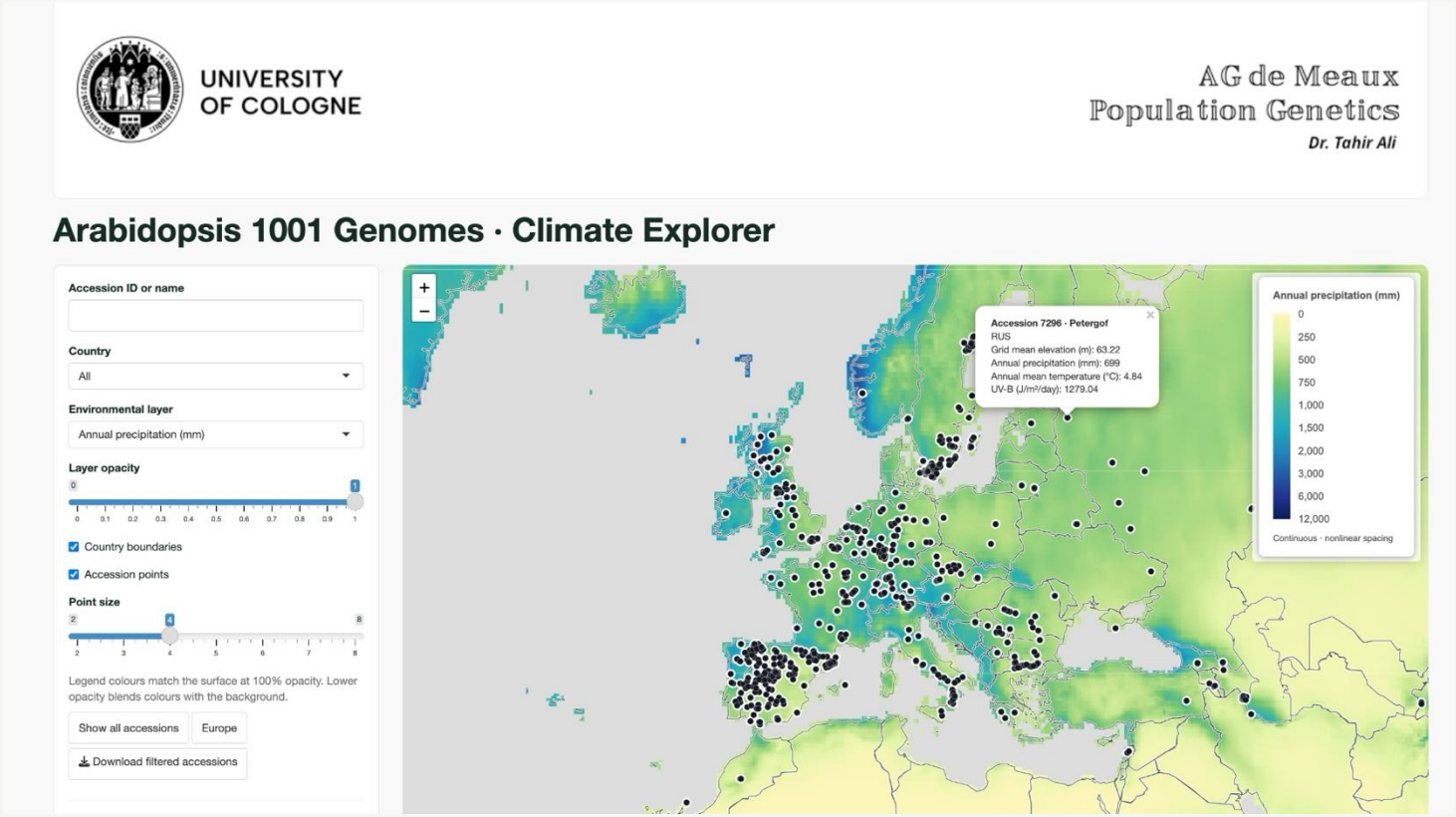
PRECIPITATION



Map-based observation → research question, not yet a causal explanation.

Explore *Arabidopsis* 1001 Genomes . Climate Explorer

Use the mapped accessions and climate layers to propose a meaningful comparison.



1 Choose a climate layer

e.g. annual precipitation

2 Zoom and filter

Inspect geographic groups

3 Click accessions

Read local climate values

4 Record a contrast

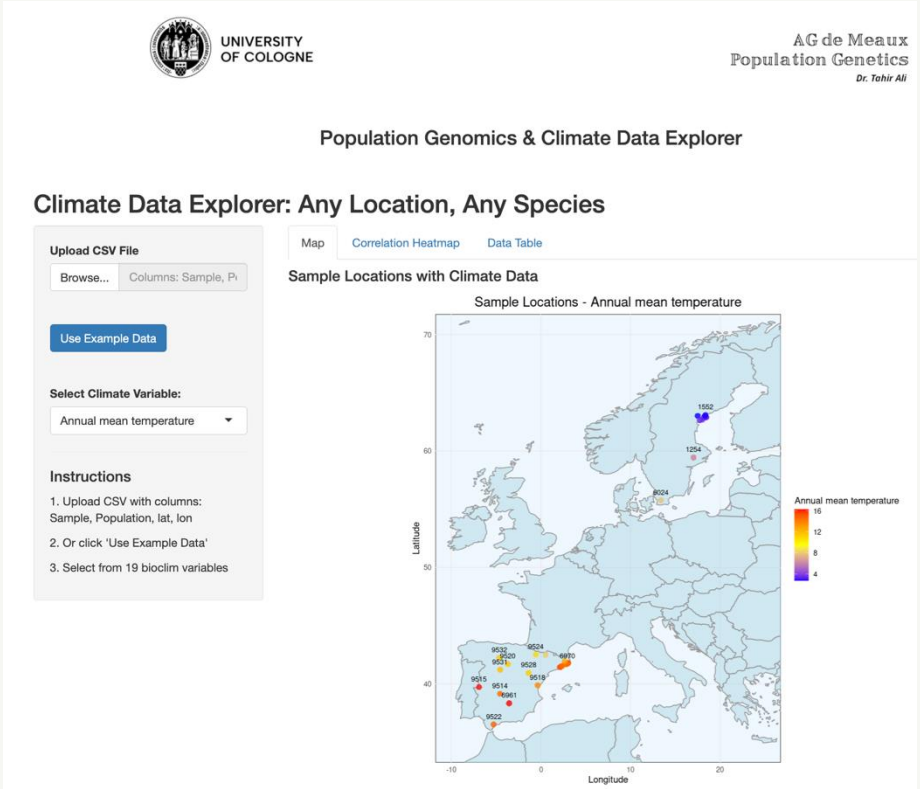
Save the IDs of accessions from two contrasting regions.

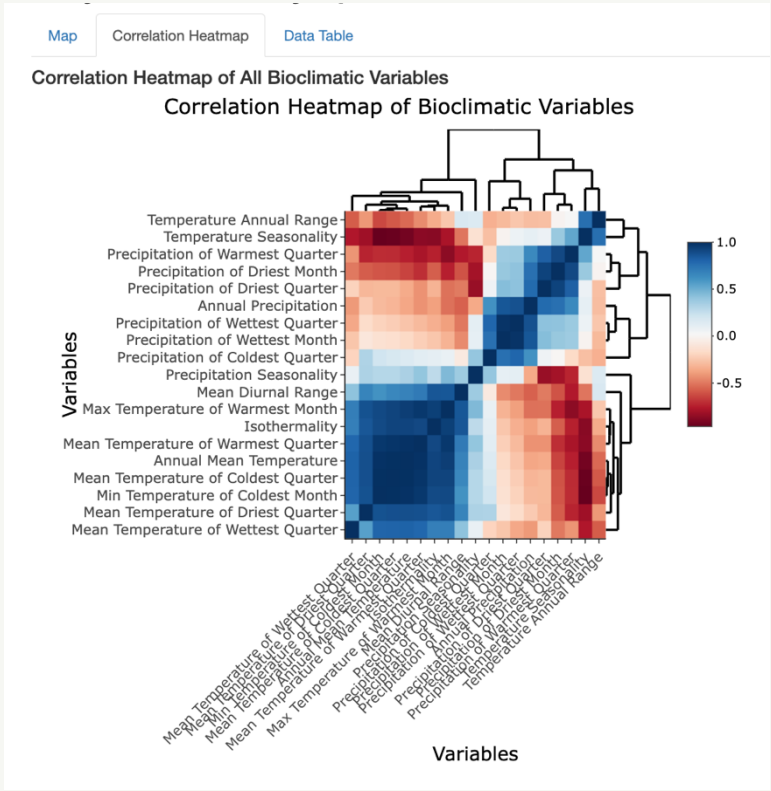
OPEN 1001 GENOMES CLIMATE EXPLORER

Scan to explore 1,135 accessions. Find two collection sites with contrasting climates.

Climate Data Explorer: Any Location, Any Species

Explore bioclimatic variables across sampling locations to identify meaningful environmental contrasts





1 Load Sample Locations

Upload a CSV with latitude and longitude, or select Use Example Data.

2 Map a climate variable

Choose from 19 bioclimatic variables, such as annual precipitation.

3 Compare two populations

Find two sites with contrasting colours. Use the data table to record their values

4 Explore climate correlations

Use the heatmap to find two variables that rise together or vary in opposite directions

[OPEN CLIMATE EXPLORER]

Scan to open the Climate Data Explorer and extract climate variables for any geographic location

Observation → Question → Hypothesis → Prediction

Example: a Spain–Sweden climate comparison.

Observation	Warm/dry and cold/wet regions differ in their climatic conditions.
Question	Do drought-response genes show greater genetic differentiation between these regions than the genomic background?
Hypothesis	Drought-related genes may show elevated differentiation between the two accession groups
Prediction	Prespecified drought genes will have higher F_{ST} than comparable genomic regions.

Null: Drought genes are not more differentiated than comparable genomic regions.

Choose two contrasting groups of accessions

Use the course accession-to-SRA mapping; instructors will confirm genotype availability.

GROUP A · e.g., warm/dry sites

Choose 20 accessions from a warm/dry region that appear in the supplied accession-to-SRA mapping.

Record accession IDs, SRA IDs, coordinates and relevant climate values.

GROUP B · e.g., colder/wetter sites

Choose 20 accessions from a contrasting colder/wetter region, also listed in the course mapping.

Keep group sizes balanced and confirm that the environmental contrast fits your question.

Important: temperature, precipitation, geography and population ancestry can be confounded.

What must you collect before requesting your VCF?

Collect sample identifiers and a biological reason for each group; you do not need to download sequence files.

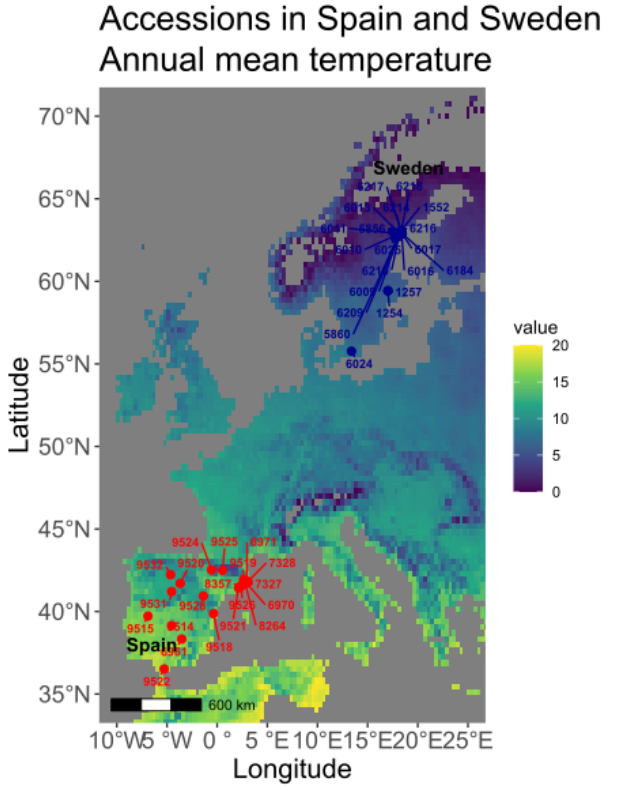
Sample identities	Accession ID · latitude · longitude	**1001 Genomes Climate Explorer** **1001 Genomes metadata** **Excel file (Available in ILIAS)**
Climate contrast	Chosen variable + units per accession	**Climate Data Explorer**
SRA identifier	Exact SRA ID linked to each accession	**Course accession-to-SRA list**
Candidate genes	10-20 gene IDs and their biological roles	**TAIR and GFF file (Available in ILIAS)**

Your instructors will prepare the selected-accession VCF once you submit the accession/SRA list.

Send us the exact accessions and SRA IDs

Use the 1001 Genomes map for accession IDs, then match each to the course accession-to-SRA file.

Group	Accession ID / name	SRA accession	Latitude, longitude	Climate value
A · warm/dry	your accession ID	from SRA list	coordinates	Annual precipitation (mm) / Annual mean temperature (°C)
B · cold/wet	your accession ID	from SRA list	coordinates	



e.g., Investigate whether drought-related genes show unusually high genetic differentiation between accessions from contrasting environments.

Required: 20 accessions per group (40 total) + exact SRA IDs; use your OWN accessions, not example IDs.

Submit 10–20 biologically justified genes

Choose genes from published functions or course annotations BEFORE inspecting genome-wide FST outliers.

Gene ID	Gene / function	Why this gene?
AT5G05410	DREB2A — Regulates drought-responsive genes	Controls plant responses to water stress
AT5G52310	RD29A — Dehydration- and cold-responsive protein	Activated during drought and dehydration
AT3G14440	NCED3 — Enzyme involved in ABA biosynthesis	Helps regulate drought responses through ABA

For every gene

- TAIR-style gene ID
- Name and function
- One literature/source citation
- Predicted pattern vs background

Source: GFF_final_with_annotation_2.gff, TAIR and relevant literature. The three named genes are examples only.

You select the samples; we provide the VCF

Use the course accession-to-SRA mapping; instructors will confirm genotype availability.

1 · YOUR GROUP

Send the 20 + 20 accession/SRA IDs, coordinates, environmental contrast and candidate genes. Do not download raw reads.

→

2 · INSTRUCTORS

Confirm genotype availability for the chosen accessions, then prepare the group-specific VCF with sample labels.

→

3 · YOUR ANALYSIS

Use the provided VCF for PCA, population FST and candidate-versus-background analyses.

Keep the accession-to-VCF sample mapping. Check group labels before analysing.

From genomic data to drought-related candidate genes

Run the workflow with your selected accessions after receiving the group-specific VCF.

DATA

Selected-accession VCF

Two groups of 20 accessions

Verify sample IDs and group labels.

→

POPULATION STRUCTURE

PCA

Visualise genetic similarity and group separation.

Are the groups genetically distinct?

→

GENETIC DIVERSITY

π / Tajima's D*

Compare within-group variation and allele-frequency patterns.

***Conditional on callable sites**

→

GENETIC DIFFERENTIATION

Genome-wide FST

Overall differentiation between groups.

Drought-gene FST

Prespecified genes vs matched genomic background.

→

GO ENRICHMENT

High-FST genes

Which biological functions are overrepresented?

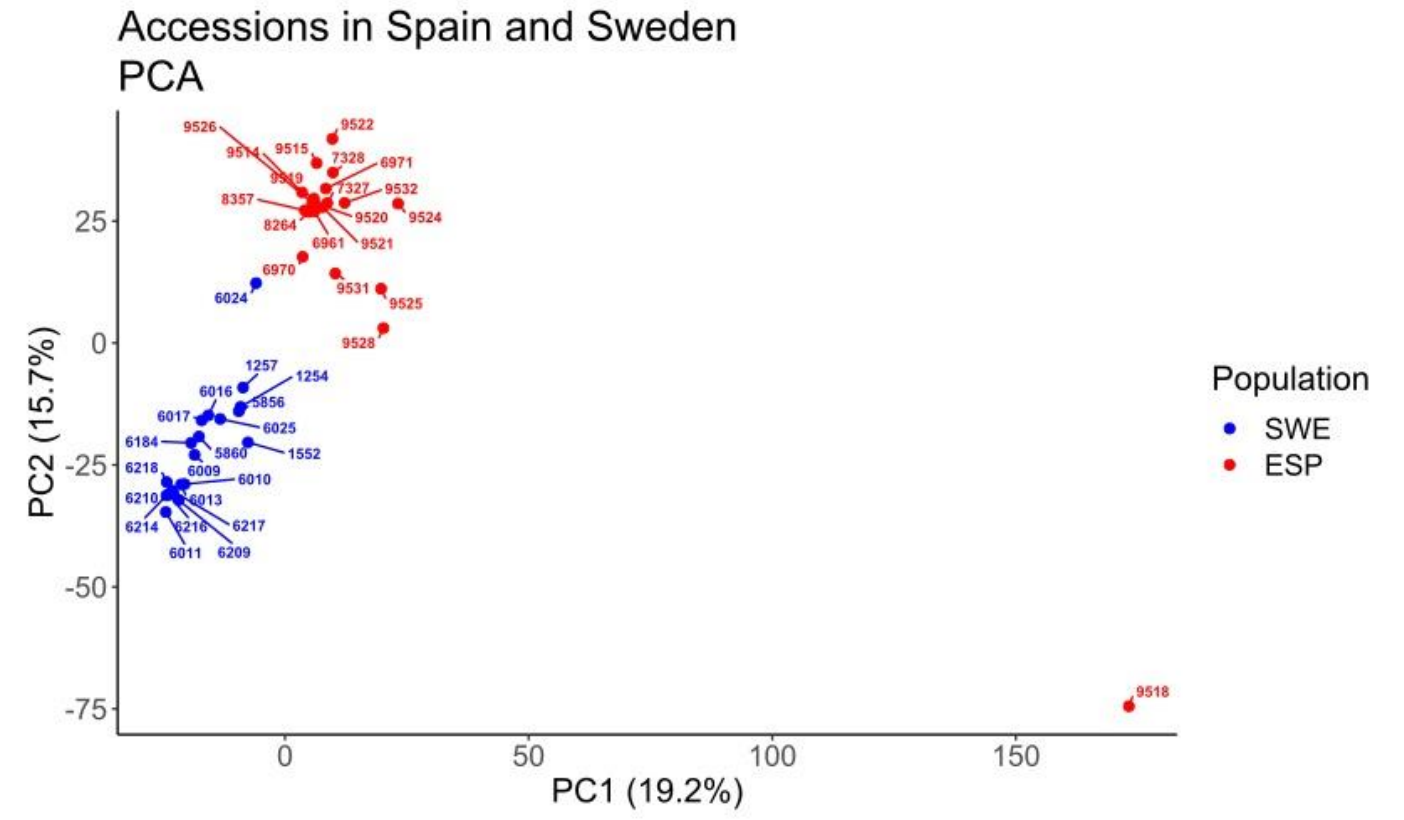
Exploratory, not a test of prespecified drought genes.

Candidate-locus signals need genomic context; drift and structure can mimic selection.

* π and Tajima's D require reliable callable-site information (including invariant sites)

Do dry and wet regions differ at stress-response genes?

Another group compared Spain and Sweden using climate maps, candidate genes and population genomics.



What makes this useful?

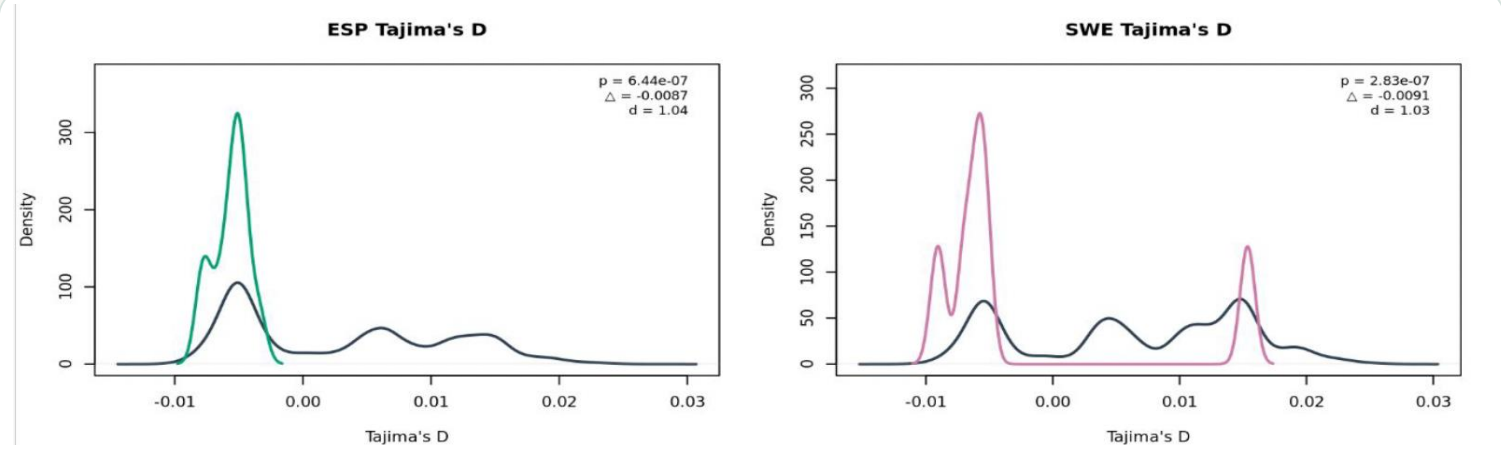
- PCA separated many Spanish and Swedish accessions, with one very distant sample.
- reveals geographic genetic structure, not necessarily drought adaptation

Follow-up question

Is differentiation genome-wide, or elevated at prespecified drought genes?

Do candidate genes differ from the genomic background?

Earlier Spain (ESP) and Sweden (SWE) population-genomics analyses



Colour = candidate loci · dark = genome-wide reference (within each population)

QUESTION

Do focal genes show unusual diversity patterns in both regions?

TAJIMA'S D

Different from background in both populations.

ESP $p = 6.44 \times 10^{-7}$ · SWE $p = 2.83 \times 10^{-7}$

NUCLEOTIDE DIVERSITY (π)

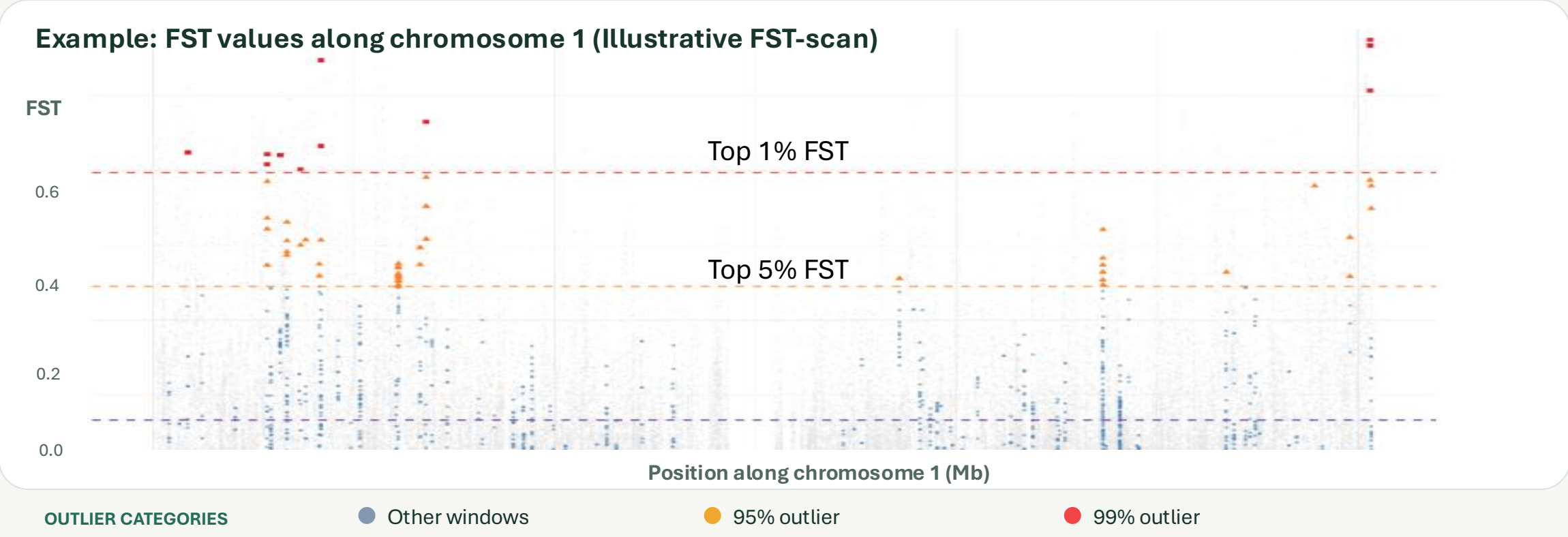
No clear difference from background.

ESP $p = 0.138$ · SWE $p = 0.064$

Different patterns ≠ proof of selection

Do drought-related loci show unusually high differentiation?

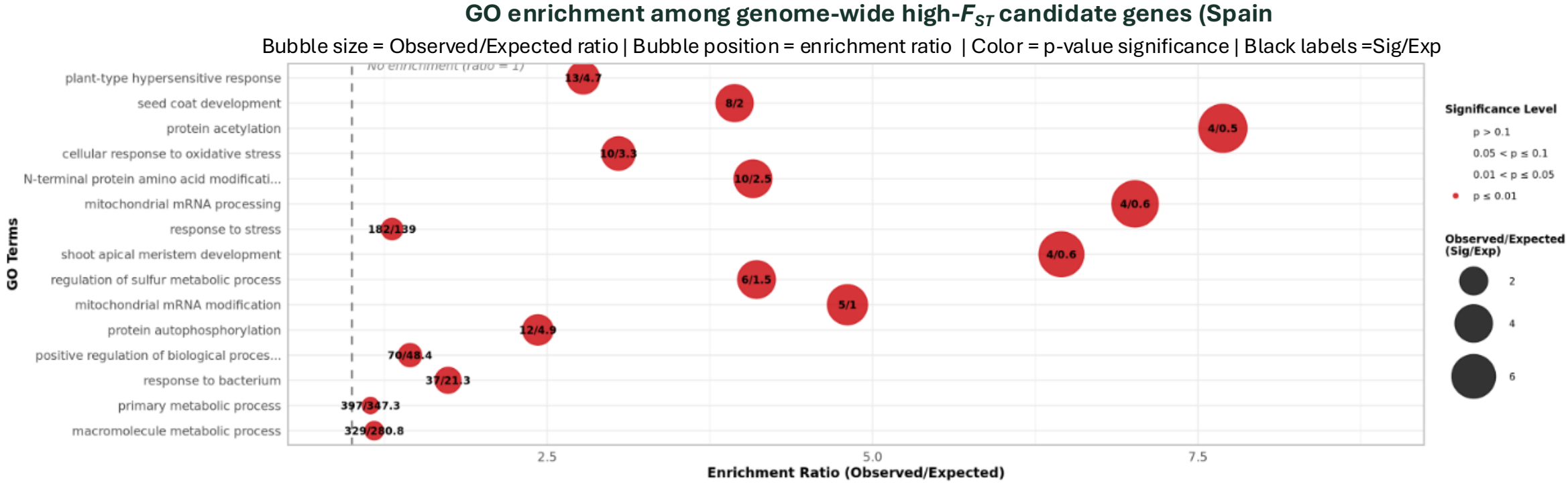
Compare population differentiation at candidate loci with the genomic background.



Answer: A few regions have high FST. Test whether drought genes are unusually enriched for outliers.

Are stress-related functions enriched among high-FST genes?

Explore GO enrichment among genome-wide high-F_{ST} genes, using eligible genes as the background



Read the plot: Y = biological functions · X = observed / expected genes · Farther right = greater enrichment

Genome-wide GO enrichment
Identify high-FST genes first, then test which biological functions are overrepresented.

Prespecified drought genes
Select drought-related genes first, then test whether their F_{ST} is elevated relative to the genomic background.

High-F_{ST} genes show enrichment of certain biological functions; whether prespecified drought genes show unusually high differentiation must be tested separately.

Do drought-related genes differ between contrasting environments?

Example QBio305 project: compare 20 warm/dry and 20 cold/wet Arabidopsis accessions.

A **GROUP A · WARM / DRY**

20 selected accessions

B **GROUP B · COLD / WET**

20 selected accessions

Research approach

01 **Compare population structure**

Use PCA to visualise genetic similarities and differences.

02 **Measure genome-wide differentiation**

Estimate FST between the two selected accession groups.

03 **Test predefined drought genes**

Compare drought-gene FST with matched genomic background loci.

04 **Investigate high-FST gene functions**

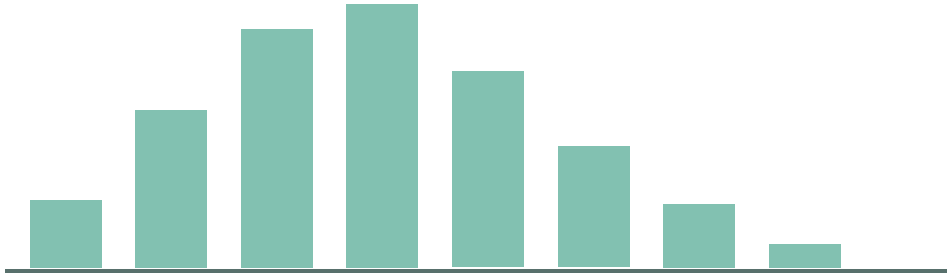
Use GO enrichment to explore functions among genome-wide outlier genes.

Interpret differences in light of population history, geographic structure and sampling.

Candidate regions vs the whole-genome background

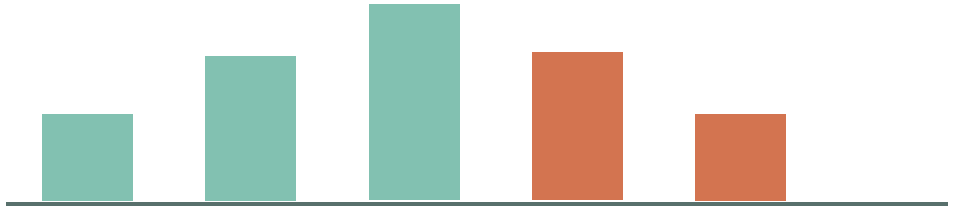
Ask whether loci linked to your ecological question stand out beyond the differentiation already present genome-wide.

Genome-wide empirical background



Many SNPs / matched windows

Candidate regions / genes



Preselected, biologically relevant loci

Schematic distributions — not observed data

Genome-wide is a reference, NOT a perfectly neutral null. Match sites and account for LD, gene length, missingness and ancestry.

Empirical genome-wide background vs truly neutral evolutionary model.

What does an unusually high candidate FST mean?

A population comparison can reflect selection, but also drift, colonisation, structure or sampling effects.

Candidate > background

A focal gene set stands out relative to the overall genomic differentiation.

Interpret as candidates for further tests—not proof of adaptation.

Candidate ≈ background

No excess differentiation is detected at the selected genes

No special candidate signal is necessary.

Genome-wide elevation

Strong differentiation across the genome can reflect isolation, bottlenecks or history.

Discuss these alternatives before selection.

Use comparable windows/SNP sets, report uncertainty, and do not call empirical high-FST tails selection tests.

Before saying “selection” — ask what else fits

Strong projects explicitly discuss alternatives and data limitations.

Geography

Could shared ancestry explain the difference?

Sampling

Are group size, missingness and outliers comparable?

Environment

Are temperature, rainfall, altitude and latitude correlated?

Statistics

Were gene sets chosen before testing? Was FDR controlled?

Can we separate drought from geography and temperature?

A useful negative or inconclusive result is still a valid scientific result.

What your group must submit to the instructors

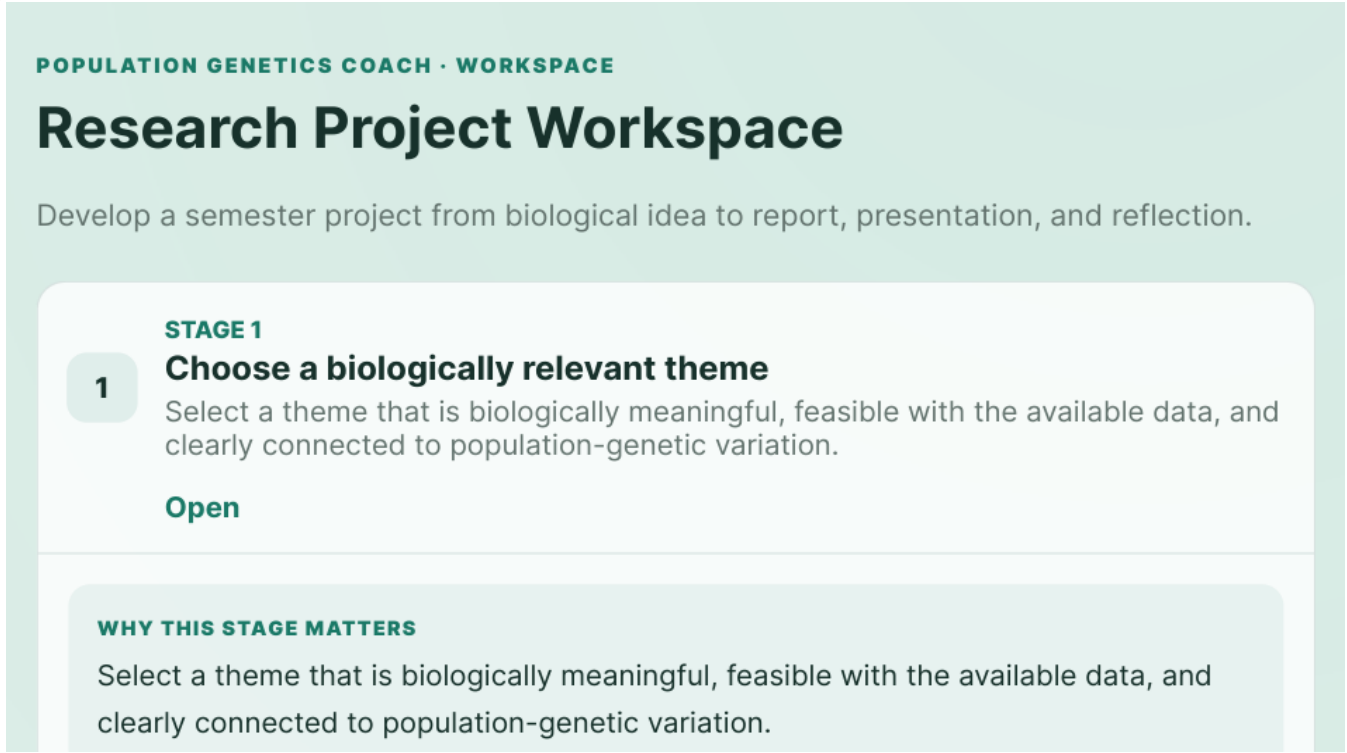
Submit the biological plan first, then send a complete VCF request with identifiers.

- **Title**
  Short, specific title
- **Research question**
  One answerable ecological/genetic contrast
- **Hypotheses**
  H0, H1 and rationale
- **Prediction**
  Candidate vs genome-wide prediction
- **Accession + SRA list**
  20 per group: accession ID, SRA, site
- **Candidate gene list**
  10–20 genes: IDs, function, hypothesis

11 Nov: title + question • 21 Nov: accession/SRA lists + candidate genes (confirm in ILIAS)

Use Population Genetics Coach as your project notebook

The workspace guides the project from biological idea to hypothesis, sample design, analysis and report.



Examples from the 12 guided stages

1. State the question
2. Write H0 / H1
3. Audit data + design
4. Match tests to predictions
5. Record and interpret results

Scan the code and select Workspaces to open your Research Project Workspace.

Tools and reference data for your project

Explore locations in one app; develop and review the research design in the other.

1001 GENOMES CLIMATE EXPLORER	Explore climate layers and accession locations
1001 Genomes metadata	Match accession IDs to SRA identifiers using the course list
Population Genetics Coach	12-stage workspace for questions, hypotheses and analysis plans
TAIR and GFF file (Available ILIAS)	Look up gene function, coordinates and GO annotation

Use the course-provided SRA mapping and annotation files when preparing the lists.

From research question to analysis plan

QBIO305 project checkpoints · Winter semester 2026/27

11 Nov
2026

Project title and research question

Describe the warm/dry versus cold/wet comparison and your main biological question.

21 Nov
2026

Accession list and dataset request

Submit 20 accessions per group, exact SRA IDs, coordinates, climate values, and a preliminary drought-gene list.

08 Dec
2026

Initial Materials and Methods

Explain the sampling comparison, genomic predictions, and analyses you plan to perform.

Please check ILIAS for any updates to project deadlines.