INRAe

M2 Research Internship (2025)

Identifying subpopulation-specific associations from large-scale paired genomic-transcriptomic data

Andrea Rau and Nathalie Vialaneix

***Animal Genetics and Integrative Biology* (GABI) research unit, INRAE Jouy en Josas**
***Applied Mathematics and Informatics Toulouse* (MIAT) research unit, INRAE Toulouse**

Email: andrea.rau@inrae.fr, nathalie.vialaneix@inrae.fr
Web: https://gabi.jouy.hub.inrae.fr/, https://miat.inrae.fr/

Keywords: Association analysis; transcriptomics; genomics; expression QTL; multi-population

Project context:

Maintaining genetic diversity is crucial for allowing agriculturally important species to adapt to changing environments driven by ongoing climate change. The domestication and selection of animal and plant species have respectively exploited only a small portion of their available genetic diversity, suggesting that a substantial part of global genetic diversity remains underexploited in breeding programs. One way to address the constraints of climate change while meeting the objectives of agroecology is to effectively identify and characterize this unexploited genetic diversity. To address this gap, the PEPR Numerical Agroecology project AgroDiv is developing cutting-edge genomics and genetics approaches to functionally characterize large-scale datasets from several agriculturally important plant and animal species.

A major goal within AgroDiv is the identification of subpopulation-specific links between gene expression and genetic variation; these subpopulations correspond to breeds (animals) or ecotypes (plants) that have often undergone strong selection for agriculturally relevant traits. In this internship, our objective is to develop a **large-scale simulation study** to evaluate the performance of several **state-of-the-art strategies to identify subpopulation-specific associations between gene expression and genetic variants**:

(1) We will make use of a large-scale (n=300) set of paired whole genome sequencing (~25M genetic polymorphisms) and transcriptomic (~13k

science for people, life & earth

genes) data in 3 pig breeds[1], generated as part of the H2020 EU GENE-SWitCH project (https://www.gene-switch.eu/). Initial work in the internship will correspond to familiarization with these data (e.g. exploratory analyses, …) as well as the use of state-of-the-art tools for the analysis of genomic and transcriptomic data (e.g., PLINK, DESeq2, RRBLUP, …).

(2) We have already identified a number of candidate tools that are a promising approach to extract sub-population specific associations between gene expression and genetic variants. These include mashr[2], meta-analyses as implemented in METAL[3], and more straightforward strategies like linear mixed models with interaction terms or fitted independently in each sub-population. To allow for full genome-wide eQTL analyses, methods such as MatrixEqtl[4] will likely be considered to allow for genome scalability.

(3) To conduct a full evaluation of these different tools, a simulation strategy will be designed and will make use of the available data. It will notably include the simulation of common associations with similar or varying effect sizes across sub-populations, as well as with similar or marked differences in allelic frequency (including population-specific variants). Combined with different parameters, we aim to identify the strengths and limitations of the available modeling strategies in combination with the simulated specificity of associations.

This initial simulation study will pave the way to developing new apposite statistical approaches to efficiently identify population-specific genetic variants of interest in a large-scale omics data integration framework.

Required skills:

- Knowledge of and experience with linear models and linear mixed models; basic familiarity with Bayesian models; basics in linear algebra.
- Solid knowledge of the R programming language for coding, data analysis, and automated report generation (Rmarkdown or Quarto), as well as Linux command-line experience. Experience using Bioconductor packages (particularly for manipulating genomic and transcriptomic data) and a computing cluster (Slurm, SGE) would be highly appreciated but is not required.
- Comfortable reading/writing English, notably to read and understand scientific articles and write software documentation.

---

[1] Daniel Crespo-Piazuelo, *et al*. (2023). Identification of transcriptional regulatory variants in pig duodenum, liver, and muscle tissues. *Gigascience*. **12**, 1-14.

[2] Sarah Urbut, Gao Wang, Peter Carbonetto and Matthew Stephens (2019). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature Genetics* **51**, 187-195.

[3] Cristen J. Willer, Yun Li, Gonçalo R. Abecasis (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, **26**(17), 2190-1.

[4] Andrey Shabalin (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, **28**(10), 1353–1358.

- Candidates should generally have motivation and interest for genomic applications in agriculture, but experience in this area is not required.

The internship will take place in a research environment that brings together biostatisticians, bioinformaticians, and biologists. **Depending on the successful candidate's preference, the internship will take place at either the INRAE research center in Jouy-en-Josas (78) or in Castanet (31)** and will be jointly supervised by Dr. Andrea Rau and Dr. Nathalie Vialaneix.

For motivated candidates, a fully funded PhD position is available following this internship as part of the PEPR Numerical Agroecology AgroDiv flagship project.

Remuneration: Legal research internship salary

To apply: Please email a CV and letter of motivation to andrea.rau@inrae.fr and nathalie.vialaneix@inrae.fr.

science for people, life & earth

RÉPUBLIQUE
FRANÇAISE
Liberté
Égalité
Fraternité

INRAØ