MSCA-EF-IF-ST Project PATTERNS: Periodic Technical Report Part B

Maud Fagny, Frederic Austerlitz

To cite this version:

Maud Fagny, Frederic Austerlitz. MSCA-EF-IF-ST Project PATTERNS: Periodic Technical Report Part B. CNRS. 2021. ⟨hal-04709589⟩

HAL Id: hal-04709589

https://hal.science/hal-04709589v1

Submitted on 25 Sep 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons CC BY-NC-SA 4.0 - Attribution - Non-commercial use - ShareAlike - International License

11/08/2021

Project Number: 845083

Project Acronym: PATTERNS

Project title: Detecting Polygenic Adaptation Targeting Gene Expression Regulation In Humans using eQTL Networks

Periodic Technical Report

Part B

Period covered by the report: from 01/04/2020 to 30/06/2021

Periodic report: 1st

1 Explanation of the work carried out by the beneficiaries and Overview of the progress

At the beginning of the project, the ER main aim was to pursue a career in academic research in the field of complex trait genomics. On top of developing a cutting-edge research project (see section 1.1), she participated in several training activities during her MSCA-IF-EF-ST action that helped her improving her CV and her skills (see section 4), and supervised a master student. She obtained a permanent position as "Chargée de Recherche" at INRAE (Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement) that started on July 1, 2021, and precipitated the termination of the action on June 30, 2021.

1.1 Scientific Objectives

For this part, the objectives corresponding to section 1.1 of the DoA will be highlighted in italic, and the work carried out by the ER will be described below each objective, in normal font.

The PATTERNS project aims at better understanding the evolution of complex traits in human populations. To this end, the ER will develop an innovative strategy that combines network biology approaches, specifically expression Quantitative Trait Loci (eQTL) networks, with the most advanced methods from population genomics in order to identify groups of independent regulatory mutations influencing the expression of one or a group of genes and presenting signatures of polygenic selection. Because the recent evolution of complex traits was mediated by population-specific polygenic selection targeting regulatory mutations and the regulatory landscape vary across tissues, integrating genetic and expression data from multiple tissues and populations will help us to identify these complex phenotypes under selection.

***Objective 1**: **Develop an approach to detect polygenic selection signals and assess its power.** The ER will develop an approach based on the use of eQTL networks to identify groups regulatory mutations targeting the same genes. She will then seek to assess the power of this approach using simulations.*

The ER developed an approach based on grouping mutations in genic and intergenic regions of the genome according to their level of connection within a bipartite eQTL network, and the biological function they potentially regulated. In order to evaluate the power of this approach, the ER performed a simulation study using SLiM[1]. Rapidly, she simulated polygenic selection targeting one or several phenotypes determined by several loci with various effect sizes. She then assessed the power of several existing statistics to detect signature of polygenic adaptation. The simulation study showed that, among those tested, the most powerful statistics to detect polygenic selection signals were F_{ST} and Fay and Wu's H. However, Fay and Wu's H was also sensitive to stabilizing selection, and as such, not necessarily the best statistic to use on real dataset (see 1.2.1 for more details). The power of F_{ST} to detect selection also increased with the difference between the original and the optimal phenotype values, and was only at its maximum when this difference was large. Finally, and surprisingly, H_{12}, while built to detect soft sweeps, was performing very poorly on our simulation. Using a score combining these two approaches and other with less power such as Tajima's D may help gain power to detect polygenic selection

on a larger range of phenotypic optima. All scripts necessary to reproduce the simulation have been stored on her github page: https://github.com/maudf/PATTERNS.

Related to this objective, the ER was the main advisor of a student (François Mallordy, Master 2 Biology, ENS de Lyon, France). He studied the evolution of allele frequencies and statistics based on the allele frequency spectrum under polygenic selection under a range of parameters defining the effect sizes of each mutation, initial allelic frequencies at selection onset, and individual fitness values. He found that while the simulation parameters affected strongly the values of several statistics, the values of the statistics based on the allele frequency spectrum did not differ in the polygenic selection simulation compared with the simulations under neutrality.

***Objective 2**: **Identify regulatory mutations under polygenic adaptation in different tissues and populations.** The ER will apply the developed method to different tissues from the GTEx dataset to identify groups of genetic variants regulating groups of genes and carrying signatures of polygenic adaptation.*

Objective 2 corresponded to WP2 and 3. The first part of Objective 2 was to build the eQTL networks using the GTEx v8 datasets[2] (WP2). Here some modifications have been done compared to the Research Methodology and Approaches outlined in section 1.1 of the DoA (see 6.1.1). The ER obtained the tissue-specific eQTL networks directly from collaborators from the Harvard T.H. Chan School of Public Health (Sheila Gaynor and John Quackenbush). She then identified regulatory modules in each tissue-specific network, functionally characterized them using Gene Ontology enrichment analyses. For more information about the methodology used and the results obtained, see 1.2.2.

The second part of Objective 2, no definitive results were obtained at the date of the early termination of the grant (30/06/2021, corresponding to 15 months), in line with the Gantt Chart presented in the DoA part B1 3.1, which planned the results to be ready by month 17. The research on this part of the project will be continued after the PATTERNS project has ended.

As soon as the results will be available and the article presenting them will be written, a queryable version of the tissue-specific eQTL networks and the corresponding selection scores will be made available freely to all scientists, in an interactive web app similar to this one: https://maud-fagny.shinyapps.io/TF-gene_network_Maize/.

***Objective 3**: **Identify phenotypes under polygenic adaptation**. The ER will characterize the selected mutations and their target genes to identify the biological functions under polygenic selection and quantitatively evaluate the proportion of variance of gene expression that can be explained by polygenic adaptation events. She will then seek to replicate those results on other publicly available or host-owned datasets.*

This objective included WP4 and 5, that were planned between months 11 and 19 in the Gantt chart. Due to the early termination of the grant, this objective has not yet been fulfilled (see 6.1.3).

1.2 Explanation of the work carried out per WP

1.2.1 Work Package 1

The researcher has developed an approach that consists in grouping mutations depending on the biological functions they regulate. This approach consists in (1) identifying groups of mutations regulating the expression of groups of functionally related genes using the eQTL network approach[3], (2) searching for enrichment in genetic variants carrying signatures of polygenic adaptation, (3) identifying targeted phenotypes by characterizing the regulatory role of selected variants and the biological function of their target genes. The eQTL networks was chosen to identify groups of mutations affecting the regulation of functionally related genes based on results previously obtained by the ER and published in *PNAS* in 2017[4]. The ER has argued in favor of such an approach and outlined its rational in **an Opinion piece published in *Trends In Genetics* in June 2021**[5], which also summarized the existing literature. To ensure open access to this article, the draft was first posted in the host institution repository HAL (https://hal.archives-ouvertes.fr/hal-03100982v1) and the gold access option was then chosen upon acceptance for publication in *Trends In Genetics*. The financial support of the MSCA-IF-ST-EF was highlighted in the 'Acknowledgments section' as follow: 'This work was supported by the Marie Sklodowska-Curie grant PATTERNS (to M.F.; 845083)'.

1.2.1.1 Simulations

In order to test for the power of various statistics to detect polygenic selection, the ER performed a simulation study. The aim was to simulate polygenic selection targeting a specific phenotype regulated by several independent eQTLs, each with a small effect on the phenotype. Using SLiM[1], she simulated 20 independent sequences of 100kb, each carrying an eQTL located at the center of the sequence. The eQTL was introduced at generation 1. Because she used forward-in-time simulation, she needed to avoid too much loss of eQTLs by drift during the first few generations. Consequently, the eQTLs were introduced randomly at frequency 0.05 in the population. The effect sizes of the eQTLs on the phenotype (here gene expression) were set to simulate observed eQTL effect sizes in the GTEx data set[6] using previous results obtained by the ER: they were randomly drawn from a normal distribution of mean 0 and standard deviation of 0.78[4]. She also simulated 10 independent neutral sequences of 100kb with no effect on the phenotype in each simulation.

The ER then simulated two populations of N=10,000 evolving under various selection regimes (Figure 1A): (i) neutrality, where no selection acted on the phenotype; (ii) stabilizing selection, where the phenotype optimum value was set at 0 for the two population, and (iii) population-specific, directional selection, where after a burn-in of 1000 generations under stabilizing selection, one population had its optimum phenotype value changed to 2, 5 or 10. The ER decided to simplify the simulations by considering that the selected phenotype was directly the gene expression level, and by simulating an additive model where the phenotype value was the sum of the eQTL effect sizes for each individuals. 100 replicates were run for each scenario, and 200 individuals per population were sampled at the 10,000th generation and their genotypes were recorded in VCF files. The ER also extracted eQTL allele frequencies and phenotype values every 100 generations to check that they evolved as expected (Figure 1B-F).

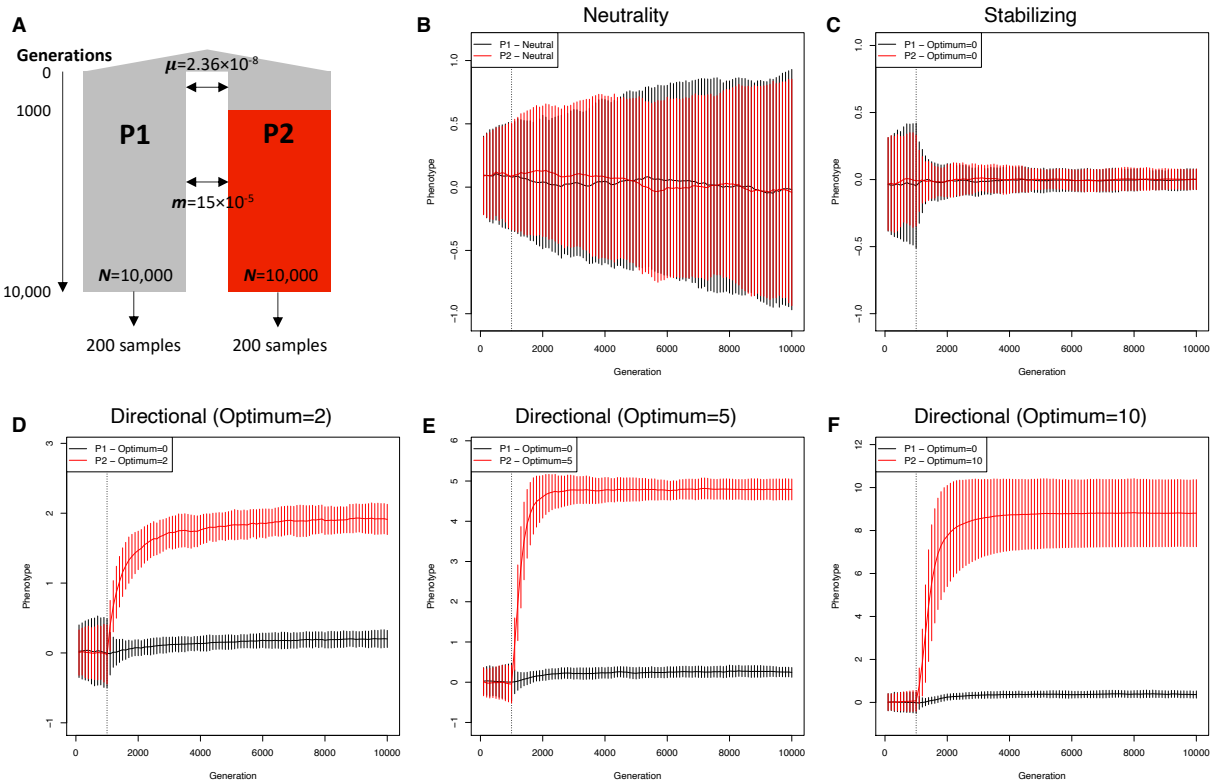

Figure 1: **Evolution of average phenotypic values.** (A) Simulation model used. When pertinent, directional selection starts at generation 1000 (red). ***m*** is the migration rate and ***μ*** the mutation rate. (B-F) Evolution of average phenotypic values with time. Population 1 is represented in grey and population 2 in red. Standard errors are represented by the small bars every 100 generation. The dotted line represents generation 1000 (onset of selection).

1.2.1.2 Power Study

The ER then tested the power of well-known statistics based on the allele frequency spectrum such as Tajima's D[7], Fu and Li's D and F[8], Fay and Wu's H[9] and Zheng's E[10] that are frequently used in positive selection power studies but are known to have limited power of detection in the context of polygenic selection. She also used H_{12}[11], a statistic based on haplotype length and especially designed to detect soft sweeps (events of positive selection on standing variants, which are rather common in the case of polygenic selection). Finally, she used two statistics measuring population differentiation : F_{ST}[12] and *PCadapt*[13]. Scores were computed on 2.5kb around the eQTLs for the frequency-spectrum based statistics, on each SNP for the population differentiation-based statistics, and on 400 SNPs-windows centered on the selected eQTLs for H_{12}.

Because under polygenic selection we expected an increase in moderate and high frequency SNPs and a decrease in low-frequency SNPs, we were expecting Tajima's D, Fu and Li's D and F and Zheng's E to be higher under selection than under neutrality, and Fay and Wu's H to be smaller. We also expected to see higher F_{ST} values and higher χ^2 values in *PCadapt* reflecting a higher differentiation at selected sites between P1 and P2. Finally, we expected higher H12 values, because under polygenic selection with soft sweeps, the most frequent haplotypes are likely to be a little more frequent than under neutrality. The null distributions for each statistic were thus determined using the neutral simulations, and the power was then computed using a false positive rate of 5% (i.e. thresholds were set so that only 5% of the eQTLs in the neutral simulations were detected).

In order to compute the power of each statistic, the ER counted the number of eQTLs effectively detected as under selection in each simulation. The ER considered that polygenic selection was detected only if at least two eQTLs were detected as selected in a simulation. The power is the proportion of simulations for which polygenic selection was effectively detected. The obtained power for each statistic is presented in Table 1. As expected, the frequency spectrum-based statistics performed poorly to detect polygenic adaptation, with the exception of Zheng's E and Fay and Wu's H that performed surprisingly well, especially with optimal phenotypes of 5 and 10. However, Zeng's E and especially Fay and Wu's H also showed a skewed distribution in the case of stabilizing selection, potentially reflecting the increase in intermediate to high derived frequencies at some eQTLs that were necessary to maintain an average phenotype of 0. These tests must thus be used cautiously to detect polygenic adaptation, and in the case of positive signals, efforts must be done to disambiguate the results by ruling out stabilizing selection.

Table 1: **Power of different statistics to detect polygenic selection.**

$\alpha = 0.05$	***Optimum of phenotype in divergent population***			
Method	**Stabilizing**	**2**	**5**	**10**
F_{ST}	0.00	0.06	**0.60**	**0.94**
PCadapt	0.00	0.05	**0.12**	**0.12**
H_{12}	0.01	0.01	0.02	**0.06**
Tajima's *D*	**0.15**	0.15	0.26	0.22
Fu and Li's *D*	**0.27**	0.19	0.24	0.25
Fu and Li's *F*	**0.21**	0.19	0.29	0.26
Fay and Wu's *H*	**0.41**	**0.34**	**0.59**	**0.90**
Zheng's *E*	**0.20**	**0.29**	**0.52**	**0.80**

1.2.2 Work package 2

1.2.2.1 eQTL data

The ER obtained eQTL results computed on 29 tissues from the GTEx v8 data set accession phs000424.v8.p2 from Sheila Gaynor and John Quackenbush (Harvard School of Public Health). For power reasons, 29 tissues with at least 200 samples were retained for analysis. Rapidly, genotyping nad fully processed filtered and normalized RNA-Seq data obtained from the GTEx Portal (www.gtexportal.org). *cis*- and *trans*-eQTLs were then identified using the MatrixEQTL package[14]. *cis*-eQTLs were defined as SNPs within 1Mb of the gene transcription start site, and *trans*-eQTLs as all other SNPs. The ER used a linear regression model with covariates assuming an additive effect of genotypes to map eQTLs. She accounted for population stratification by using the first five principal components of the genotypes as covariates. She further adjusted for sex, PCR, genotyping platform, and the GTEx-recommended set off PEER factors based on sample size. The FDR-corrected *p*-values were obtained using the qvalue R package[15], and only eQTLs with corrected p-values lower than 0.05 were used to build the eQTL networks.

1.2.2.2 Network module identification

The ER analyzed the structure of each 29 tissue-specific networks using the CONDOR R package[3], in order to identify the modules. She found that the network had an overall high modularity ([0.65-0.98], average 0.83, see Table 2), which means that the Networks are well-structured, with groups of SNPs densely linked to groups of genes within modules, and only loosely connected to the rest of the network.

Table 2: **Modularity structure of eQTL networks**

Abbreviation	Tissue description	Modularity	Number of modules
ADS	Adipose subcutaneous	0.96	190
ADV	Adipose Visceral Omentum	0.92	91
ADG	Adrenal Glands	0.67	6
ATA	Artery Aorta	0.977	64
ATC	Artery Coronary	0.70	6
ATT	Artery Tibial	0.96	173
BCE	Brain Cerebellum	0.68	7
BCO	Brain Cortex	0.71	9
BNA	Brain Nucleus Accumbens Basal Ganglia	0.71	6
FIB	Cell cultured Fibroblasts	0.98	239
CLS	Colon sigmoid	0.67	8
CLT	Colon Tranverse	0.66	9
EGS	Esophagus Gastroesophageal junction	0.80	28
EMC	Esophagus Mucosa	0.96	150
EMS	Esophagus Muscularis	0.94	117
HRA	Heart Atrial Appendage	0.71	11
HRV	Heart Left Ventricle	0.67	10
LIV	Liver	0.65	6
LNG	Lung	0.94	115
MSK	Muscle Skeletal	0.96	148
NRV	Nerve tibial	0.98	261
PAN	Pancreas	0.77	19
PIT	Pituitary Gland	0.73	9
SNE	Skin Not Sun-Exposed (Suprapubic)	0.96	160
SSE	Skin Sun-Exposed (lower leg)	0.97	223
SPL	Spleen	0.68	7
STM	Stomach	0.68	8
THY	Thyroid	0.98	275
WBL	Whole Blood	0.95	162

1.3 Impact

The Part B1 section 2.1 outlined the expected impact of this project in terms of career prospects and training. The main aim of the ER was: '*to build on her unique multi-disciplinary background in population genetics and system biology to pursue a career in academic research in the field of evolution of complex traits.*' She was also planning to '*apply to independent researcher positions both in France (both at the CNRS as Chargée de Recherche and at the University as Maitre de Conférences) and in Europe (Assistant Professor positions in Universities and Research Institutes).*' PATTERNS had a major and positive impact on the ER career, as she applied to and got a position as independent researcher at INRAE in Gif-sur-Yvette, France. Her capacity to get the MSCA-IF-ST-EF, and the opinion piece she co-wrote with her advisor in *Trends In Genetics* exposing the rational of her research was definitively a major factor in her being hired by INRAE.

In terms of scientific skills, the expected impact of PATTERNS was to 'broaden the field of expertise of the ER, in particular by increasing her knowledge in theoretical population

genetics and evolutionary anthropology'. This was done during WP1, for which she had to strengthen her knowledge in the field of polygenic selection detection, and reflect in the Opinion piece published in *Trends In Genetics*, that contained a literature review on this field. In terms of training, the expected impacts of PATTERNS were twofold. First, she aimed to '*improve her research data and project management skills*'. The training in data management (see section 4) she followed during the PATTERNS project and the discussion on project management and practice with both this project and the project of the Master student she supervised allowed her to reach this goal. Second, she aimed to acquire new 'skills in transferring cutting-edge research results to the general public', which was done by participating to several outreach actions (see section 2.2).
Finally, while the results of this project have not been published yet, due to the early termination of the project, the impact of the ER research on her field is already measurable: she obtained important results by providing a new approach to detect polygenic adaptation targeting multiple loci with weak impact on the phenotype. Thanks to her interest and implication in reproducible and open science, she also provided several pipelines with tutorial that can be used by researchers in her field to replicate her results and use her approach on other data, species or research questions (see 2.1).

2 Update of the plan for exploitation and dissemination of result

No update of the plan for exploitation and dissemination of results is necessary. Here is a summary of the dissemination of results that have already been undertaken. Most of the dissemination will occur after the termination of the grant, once the results have been finalized, given that the combination of the covid-19 pandemic and the early termination of the grant did not allow the ER to perform many dissemination actions during the 15 months of the PATTERNS project.

2.1 Dissemination actions towards the scientific community

The rational of the project, and the interest to combine systems biology with quantitative and population genetics was described in an opinion piece published in *Trends in Genetics*[5].

All scripts and pipelines have been made available on the github of the ER (https://github.com/maudf/PATTERNS).

In addition, and despite the limitations brought by the covid-19 pandemic, the ER managed to participate to several online conferences to disseminate her results to the community:

- The Virtual Maize Meeting 2020 (June 25 – 26, 2020, https://documents.maizegdb.org/maizemeeting/abstracts/2021Program.pdf): it allowed her to reinforce her professional network by meeting researchers working in the field of maize genetics and adaptation. She also presented (oral presentation) previously obtained results[16] that proves the rational of WP4, and used this opportunity to introduce gene regulatory approaches and their utility in terms of intergenic region annotation and understanding of adaptation at the molecular level to the scientific community.
- The Virtual CSHL Network Biology Meeting (March 16 - 19, 2021, https://meetings.cshl.edu/abstracts.aspx?meet=NETWORK&year=21): it allowed her to reinforce her professional network by meeting researchers working in the field of gene regulatory networks and adaptation. She also presented (poster presentation) previously obtained results[16] that proves the rational of WP4.

- The virtual SMBE Meeting 2021 (July 3 – 8, 2021, https://www.smbe.org/smbe/MEETINGS/SMBE2021.aspx): This meeting allowed her to extend her professional network in the field of population genetics and evolution.

2.2 Dissemination actions towards the general public

- A webpage was built on the host laboratory web site to promote the project and related actions: https://www.ecoanthropologie.fr/en/patterns-project-9139.
- The ER participated to a cycle of online conferences on evolutionary biology organized by the association "Fête le savoir" that organize conferences and scientific activities around a different theme each year (https://fetelesavoir.com/page39.html).
- The ER participated as an alumni to an in-person "Career development seminar" towards middle school students (Collège Irène et Frédéric Joliot-Curie, Pantin, France, https://joliotcurie-pantin.webcollege.seinesaintdenis.fr/), with the aim to promote jobs in science and academic research.

3 Update of the data management plan

No update of the data management plan is necessary.

4 Training objectives

The main aim of this project in terms of training was to enhance the ER prospects to develop a career in academic research. It was successful because she obtained a permanent position at INRAE in early April 2021.

A Career Development Plan was outlined with the supervisor at the beginning of the project that consisted in a 5-steps program:

- Training for job application, including supervision for scientific project writing and training for the job interviews.
- Publishing an opinion piece in a high impact factor review journal to expose her views on existing research and the possible future developments in her field of research.
- Applying to several positions in 2021 including (1) Chargée de Recherche (research associate) at INRAE in the Biology and Plant Adaptation Department, and (2) Chargée de Recherche position at three CNRS departments: Vegetal Biology; Ecology and Evolution; and Math, Physics and Computer Sciences.
- In case no position was obtained in 2021, several actions were planned: (1) re-application for similar positions in 2022 after careful consideration of the jury comments and application to other research institutes and universities depending on possible openings; (2) application to other grants in order to obtain funding to continue the postdoc after the end of the MSCA-IF-ST-EF project; (3) re-discussing the career development plan and looking for positions in the private sector.
- Applying for starting grants in 2022 such as the ERC Starting Grant and the ATIP grant at CNRS in order for the ER to develop her own project and start her team.

A series of training actions were also undertaken. While most formal training actions through institution-provided courses planned for 2020 were cancelled due to the covid-19 pandemic, the ER managed to find mentors, including her supervisor, other members of the host institution, and collaborators from her own network, to improve her skills in data management, student advising and mentoring, job application and grant application writing.

Student advising training: The ER had the opportunity to co-advise a Master 2 student with her supervisor Frédéric Austerlitz. It allowed her to discuss several aspects of supervising such as how to plan an internship and manage the student's time, how to best leverage the student's strong and weak spots and how to best train them to become autonomous. These discussions occurred before the beginning of the internship and during the 6 months of the internship (January-June 2021).

Job application training: The ER received mentoring from several researchers both remotely and in-person between July 2020 and March 2021 on how to apply for permanent researcher positions in France. She applied to three jobs: Chargée de recherche CRCN INRAE (Written application dead-line November 2020, interviews March 2021), Chargée de Recherche CRCN CNRS in the Ecology/Evolution department and in the Math/Computer Sciences/Physics for Biology interdisciplinary department (Written application dead-line January 2021, interviews March-April 2021). She had many discussions with her supervisor Frédéric Austerlitz, and other researchers from the lab (Paul Verdu, Guillaume Achaz), and from other labs (Maud Tenaillon, Clémentine Vitte, Chloé Girard, Benoit Landrein) on how to write the research project part of the application (July-December 2020), and to prepare the job interviews (February-March 2021).

Theoretical population genetics training: The ER had the opportunity to improve her scientific skills through discussions on population genetics with the supervisor, Frédéric Austerlitz, and other members of the host laboratory (Paul Verdu, Bruno Toupance).

Data Management Training: The data management training was performed by Paul Verdu, who organized a half-day data management training session for the members of the AGENE team (host team) on November 25, 2020. The training was focused on how to protect and deal with personal data, in particular genetic data.

5 Supervisor activity

The supervisor first made sure that all administrative procedures were fulfilled upon the start of the ER's contract, allowing to start the contract in time, while France was in full lockdown at that time because of the sanitary crisis. He helped her also in various tasks, such as obtaining her email account and her access to the computer cluster of the MNHN.

Beside numerous informal discussions, the supervisor and the ER had a meeting every week during which they discussed in particular the scientific advancement of the project, the opportunities to disseminate the ER's results (through scientific articles and international conferences), and her career development. Because of the sanitary crisis, these meetings were either in person or by videoconference. It was quite important to keep them during the lockdown periods, in order to avoid any feeling of isolation. Moreover, as stated above, the ER was the

main supervisor of a master student (F. Mallordy). In this regard, during the internship of this student, a weekly meeting was organised between the student, the ER and the supervisor, in order to discuss the advancement of the student and provide the supervisor's experience in mentoring students. Again, these meetings were either in person or by videoconference, depending on sanitary restrictions.

6 Deviations from Annex 1

Some deviations from Annex 1 of the Grant agreement have occurred. They fall within two main categories: modifications of actions such as training activities due to the covid-19 pandemic, and non-completion of some WP due to the early termination of the project. However, note that most of these WP will be terminated in the near future, because the position that the ER obtained as a permanent researcher at INRAE allows her to wrap up previous research projects.

6.1 Tasks

6.1.1 Work package 1 and deliverables 8.1 and 8.2

While the WP 1 was wrapped up during the reporting period, the presentation of the results at international conferences and in an article has been delayed, mainly by the fact that the ER was applying and interviewing for permanent researcher positions during late winter and early spring 2021. The article is currently in preparation, and the results will be also presented international conferences in late 2021-beginning of 2022.

6.1.2 Work Package 2

For administrative reasons linked to the affiliation of the head of the host laboratory, Evelyne Heyer, Professor at the MNHN, the GTEx data request had to be processed by the MNHN and not the host institution. Due to the disruption in the administration of the MNHN caused by the Covid-19 pandemic, the ER was unable to access the raw data from the GTEx project v8 (https://gtexportal.org/home/) on time for the project. In order to circumvent this difficulty, she obtained already processed data from collaborators in the United States (John Quackenbush and Sheila Gaynor), namely the tissue-specific eQTL results, that contain only summary statistics, and as such do not necessitate any ethic clearance.

6.1.3 Work Packages 4-5 and deliverables 8.3 and 8.4

These WPs were pertaining to objective 3 and were not finished at the time of the grant termination. However, the data necessary to complete WP4, including the epigenomic annotations, have been downloaded, and the WP4 should be completed in the future, once the results from WP3 are finalized (they should soon be available). Results from WP5 may be more difficult to obtain now that the ER has left her host laboratory and is no longer supposed to be the PI of new projects on human data. In this case, a solution could be to hire a student or a postdoc that would be co-supervised by the supervisor (Frédéric Austerlitz) and the ER (Maud Fagny) in order to wrap up WP5. Related deliverables 8.3 and 8.4 (presentation of the results

of objective 2 and 3 at international conferences and redaction of an article) will be delivered as soon as the study is wrapped up.

6.1.4 Communication/Outreach activities

All communication/outreach activities planned, “Fête de la Science”, “Science academy” and “Ma science infuse/Chercheur au balcon” are usually in-person activities, sometimes with mini-experiments to be performed by the public. They were all cancelled in 2020 and the first semester of 2021 due to the covid-19 pandemic. The ER replaced them by other, mostly online communication/outreach activities presented in section 2.2 of this report.

6.1.5 Training activities

Three training activities were scheduled in the Gantt chart in section 3.1 of the Annex 1 of the Grant agreement. Due to the covid-19, most of the in-person training activities were cancelled. In particular, those planned in 2020, TR 7.1: ‘Data Management Summer School’ usually organized by the Institut Polytechnique de Paris, and TR7.2: ‘Technology transfer of research results in bioscience’ usually organized by the host institution CNRS, were cancelled for 2020. The TR7.1 was replaced by a discussion with the researcher responsible for issues pertaining to Data Management in the host team, Paul Verdu, who explained the bases of data management, in particular of personal data allowing the identification of the subjects such as genotyping data, and explained the interest of a Data Management Plan and how to write one. TR7.3, a ‘Career development seminar’ at the beginning of the second year of the project was planned to help the ER train for job applications and interviews in academia and industry, but the ER had already secured a Chargée de Recherche position, so this training was not necessary anymore.

7 Bibliography

1. Haller, B. C. & Messer, P. W. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Mol. Biol. Evol.* **36**, 632–637 (2019).
2. Keen, J. & Moore, H. The Genotype-Tissue Expression (GTEx) Project: Linking Clinical Data with Molecular Analysis to Advance Personalized Medicine. *J. Pers. Med.* **5**, 22–29 (2015).
3. Platig, J., Castaldi, P. J., DeMeo, D. & Quackenbush, J. Bipartite Community Structure of eQTLs. *PLOS Comput. Biol.* **12**, e1005033 (2016).
4. Fagny, M. *et al.* Exploring regulation in tissues with eQTL networks. *PNAS* **114**, E7841–E7850 (2017).
5. Fagny, M. & Austerlitz, F. Polygenic Adaptation: Integrating Population Genetics and Gene Regulatory Networks. *Trends Genet.* **37**, 631–638 (2021).
6. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
7. Tajima, F. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* **123**, 585–595 (1989).
8. Fu, Y. X. & Li, W. H. Statistical Tests of Neutrality of Mutations. *Genetics* **133**, 693–709 (1993).
9. Fay, J. C. & Wu, C. I. Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405–1413 (2000).
10. Zheng, L. *et al.* Prolonged expression of the BX1 signature enzyme is associated with a recombination hotspot in the benzoxazinoid gene cluster in Zea mays. *J. Exp. Bot.* **66**, 3917–3930 (2015).
11. Garud, N. R., Messer, P. W., Buzbas, E. O. & Petrov, D. A. Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps. *PLOS Genet.* **11**, e1005004 (2015).
12. Wright, S. The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. *Evolution* **19**, 395 (1965).
13. Duforet-Frebourg, N., Luu, K., Laval, G., Bazin, E. & Blum, M. G. B. Detecting Genomic Signatures of Natural Selection with Principal Component Analysis: Application to the 1000 Genomes Data. *Mol. Biol. Evol.* (2015) doi:10.1093/molbev/msv334.

14. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinforma. Oxf. Engl.* **28**, 1353–1358 (2012).
15. Storey, J. D., Bass, A. J., Dabney, A. & Robinson, D. *qvalue: Q-value estimation for false discovery rate control.* (2021).
16. Fagny, M. *et al.* Identification of Key Tissue-Specific, Biological Processes by Integrating Enhancer Information in Maize Gene Regulatory Networks. *Front. Genet.* **11**, 1703 (2021).