Genomics of microgeographic adaptation in the Amazonian hyperdominant tree Eperua falcata Aubl. (Fabaceae).

Louise Brousseau, Paul V. A. Fine, Erwin Dreyer, Giovanni Giuseppe Vendramin, Ivan Scotti

To cite this version:

Louise Brousseau, Paul V. A. Fine, Erwin Dreyer, Giovanni Giuseppe Vendramin, Ivan Scotti. Genomics of microgeographic adaptation in the Amazonian hyperdominant tree Eperua falcata Aubl. (Fabaceae). 2018. hal-02790168

HAL Id: hal-02790168
https://hal.inrae.fr/hal-02790168

Preprint submitted on 5 Jun 2020

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 Attribution - NonCommercial 4.0 International License
Genomics of microgeographic adaptation in the Amazonian hyperdominant tree *Eperua falcata* Aubl. (Fabaceae)

Louise Brousseau\(^1,2,6\), Paul V. A. Fine\(^3\), Erwin Dreyer\(^4\), Giovanni G. Vendramin\(^5\), Ivan Scotti\(^6\)

1. INRA, UMR EcoFoG Ecologie des Forêts de Guyane, Campus Agronomique BP709, 97387 Kourou cedex, French Guiana
2. IRD, UMR DIADE Diversité - Adaptation - Développement, 911 Avenue Agropolis, BP64501, 34394 Montpellier, France.
3. Department of Integrative Biology, University of California, Berkeley, 1005 Valley Life Science Building, Berkeley, California 94720-3140, USA.
4. INRA, Université de Lorraine, UMR EEF Ecologie et Ecophysiologie Forestière, 54280 Champenoux, France.
5. Institute of Biosciences and BioResources, National Research Council (IBBR-CNR), Division of Florence, Via Madonna del Piano 10, 50019 Sesto Fiorentino (FI), Italy.
6. INRA, URFM Ecologie des Forêts Méditerranéennes, Domaine Saint-Paul, Site Agroparc CS 40509, 84914 Avignon cedex 9, France.

**Corresponding Author:**
Louise Brousseau. IRD, UMR DIADE, 911 Avenue Agropolis, BP 64501, 34394 Montpellier, France. [louise.brousseau@ird.fr](mailto:louise.brousseau@ird.fr)
Abstract

Plant populations can undergo very localized adaptation, allowing widely distributed populations to adapt to divergent habitats in spite of recurrent gene flow. Neotropical trees - whose large and undisturbed populations often span a variety of environmental conditions and local habitats - are particularly good models to study this process. Here, we carried out a genome scan for selection through whole-genome sequencing of pools of populations, sampled according to a nested sampling design, to evaluate microgeographic adaptation in the hyperdominant Amazonian tree *Eperua falcata* Aubl. (Fabaceae). A high-coverage genomic resource of ~250 Mb was assembled *de novo* and annotated, leading to 32,789 predicted genes. 97,062 bi-allelic SNPs were detected over 25,803 contigs, and a custom Bayesian model was implemented to uncover candidate genomic targets of divergent selection. A set of 290 divergence outlier SNPs was detected at the regional scale (between study sites), while 185 SNPs located in the vicinity of 106 protein-coding genes were detected as replicated outliers between microhabitats within regions. These genes potentially underlie ecologically important phenotypes and indicate that adaptation to microgeographic habitat patchiness would affect genomic regions involved in a variety of physiological processes, among which plant response to stress (for e.g., oxidative stress, hypoxia and metal toxicity) and biotic interactions. Identification of genomic targets of microgeographic adaptation in the Neotropics is consistent with the hypothesis that local adaptation is a key driver of ecological diversification, operating across multiple spatial scales, from large- (i.e. regional) to microgeographic- (i.e. landscape) scales.

Key words: Amazonia, Neotropics, *Eperua falcata*, Pool-Seq, *de novo* assembly, genome scan, SNPs, $G_{ST}$, Bayesian modelling, adaptive divergence, local adaptation.
Introduction

Plant populations can undergo extremely localized adaptation, allowing populations of widely distributed species to adapt to divergent habitats in spite of recurrent gene flow. This phenomenon has been repeatedly observed in annual plants growing in strongly contrasting yet geographically adjacent habitats, particularly in cases with large differences in soil composition (Bradshaw 1960; Jain and Bradshaw 1966; Antonovics and Bradshaw 1970; Antonovics 2006; Gould et al. 2014); in some cases, molecular evidence also supports adaptation to habitat mosaics (Turner et al. 2010; Fustier et al. 2017). In a well-known study of *Howea* palm trees, Savolainen et al. (2006) provided strong evidence of a link between adaptation to habitat mosaics and sympatric speciation. The process underlying these instances of adaptation with gene flow has been dubbed ‘microgeographic adaptation’, whereby adaptive divergence occurs at geographical scales of the same order of magnitude as gene dispersal distance (Richardson et al. 2014). Adaptation to multiple optima within a continuous population has been identified as a major mechanism for the maintenance of genetic diversity (Delph and Kelly 2014). Moreover, conditions under which adaptive divergence can occur with gene flow have been explored theoretically with models including two environmental patches connected by varying levels of gene flow (Bulmer 1972; Hendry et al. 2001; Yeaman and Guillaume 2009; Yeaman and Otto 2011; Schmid and Guillaume 2017). Surprisingly, these simulations show that a wide range of migration-selection equilibrium states can lead to adaptation to each patch with maintenance of polymorphism (i.e. maintenance of patch specialists, as opposed to the emergence of a generalist genotype).

Data from perennial plants, and from trees in particular, suggest that this may be a general phenomenon, that can be readily observed at the phenotypic level (Yeaman and Jarvis 2006; Brousseau et al. 2013; Vizcaíno-Palomar et al. 2014; Lind et al. 2017) (but see Latreille & Pichot (2017) for a case of a complete lack of microgeographical adaptation). At the molecular level, association between alleles (or genotypes) and habitats has been detected in a variety of studies (see review by Savolainen et al. (2007)), suggesting an important role for intra-specific and intra-populational variation in the evolution of habitat specialization (Scotti et al. 2016). The advent of genome-wide approaches has increased detection power, making it possible to more precisely evaluate the genetic bases of microgeographic adaptation (Turner et al. 2010; Eckert et al. 2015; Izuno et al. 2017; Fustier et al. 2017; Lind et al. 2017). In general, those genome scan studies have found only a minority of loci (in the order of a few percent of the total) that exhibits microgeographic disruptive selection. Yet genome-wide,
sequencing-based approaches can go beyond a finer estimation of the number of loci involved in the process, because they permit formulating functional interpretations of the divergence process as sequence annotation can suggest putative function of the target loci (Turner et al. 2010; Eckert et al. 2010, 2015; Fustier et al. 2017).

Microgeographic adaptation involves the filtering of phenotypes and genotypes that favor adaptation to habitat patches, and therefore can be viewed as an initial step towards the diversification of populations, ecotypes and eventually species divergence. In particular, environmental filtering has frequently been invoked as a driver of niche partitioning and species diversification in the Neotropics (Fine et al. 2004, 2005, 2013; Baraloto et al. 2007; Kraft et al. 2008; Fine 2015); intra-populational diversification processes across environmental gradients and/or habitat patches thus may play a key role in the generation of functional diversity. Moreover, species’ short-term adaptive potential (Harrisson et al. 2014; De Kort and Honnay 2017) largely depends on standing genetic variation. Microgeographic adaptation can thus play a role in the maintenance of adaptive genetic diversity. Measuring its extent in the wild is thus of major importance, as it could provide insights into the ability of populations to evolve towards new phenotypic and genetic optima under global climate change (GCC).

Trees are particularly good models to study microgeographic adaptation: many wild tree species have large populations with widespread distribution spanning a variety of microhabitats, harbor large amounts of genetic diversity, and display large dispersal distances (Hamrick and Godt 1990; Petit and Hampe 2006; Neale and Kremer 2011; Sork et al. 2013; Fetter et al. 2017). They therefore correspond quite closely to the conditions under which selection-migration equilibrium theory has been developed. Moreover, methods to detect locus-specific divergence (and infer selection) should work quite well with microgeographic processes in populations of the most common forest trees, because their relatively shallow (spatial) population structures and large effective population sizes should minimize the introduction of biases (e.g., because of drift) and the generation of false positives.

Lowland Amazonian rainforests, whose landscapes show fine patchiness of contrasting topographic and edaphic conditions varying over microgeographic scales (Fig. 1a,c,d), represent an ideal study system to investigate microgeographic adaptation processes. In Amazonian rainforests, the recent study by ter Steege et al. (2013) reported 227 hyperdominant tree species accounting for half of all stems (ter Steege et al. 2013). Some of these widespread species are also ecological generalists, whose preferences could be explained by individual phenotypic plasticity or by microgeographic adaptation to habitat
patchiness (Fortunel et al. 2016). We focus here on dense, almost monodominant stands of *Eperua falcata* Aubl. (Fabaceae) in French Guiana that occur across environmentally heterogeneous areas, with the goal of identifying genomic loci undergoing divergence at the microgeographic scale between habitats. To do this, a genome scan was carried out through a nested sampling design combining regional and microgeographic spatial scales (i.e. replicates of microhabitats, ~300 m apart, in two study sites, ~300 km apart). We analyzed patterns of genomic divergence from large (i.e. regional) to microgeographic spatial scales within a hierarchical Bayesian framework (Brousseau et al. 2016) designed to identify locus-specific departures from genome-wide divergence in nested population samples.

**Materials and Methods**

*Species model*

*Eperua falcata* Aubl. ([Fig. S1](#)) is a reportedly non-nodulating (Villadas et al. 2007) tree species of Fabaceae (Caesalpinioideae). It is widespread in Northeast Amazonia, from North Brazil to Venezuela through the Guiana shield ([https://demo.gbif.org/species/2943522](https://demo.gbif.org/species/2943522)), and it is a ‘hyperdominant’ species, the 13th most abundant tree species of the Amazonian tree flora (ter Steege et al. 2013). It is an evergreen, late-successional canopy tree species which is often emergent. It is shade hemi-tolerant (Bonal, Barigah, et al. 2000) and can tolerate both drought (Bonal, Atger, et al. 2000; Bonal et al. 2001) and water-logging (Baraloto et al. 2007). Its distribution is very clumped and populations are commonly found at high densities. It is a generalist with respect to microgeographic edaphic conditions, although it is most abundant in water-logged forests of bottomlands (Baraloto et al. 2007). *E. falcata* flowers and fruits during the rainy season, and seeds reach maturity in April-May, synchronously among trees in all microhabitats. Its large pollen grains (>100µm) are dispersed by animals such as bats (Cowan 1975), while seed dispersal is autochorous: heavy seeds are dispersed around crowns of mother trees through explosive pod dehiscence.

*Sampling design and study sites*

We sampled populations that replicated the environmental contrast ‘bottomland versus terra-firme’ (~300 m apart) at two regional locations (i.e. study sites) in French Guiana (~300
Laussat (western French Guiana) and Regina (eastern French Guiana) study sites are located on the coastal Guiana shield, about 300 km apart, and experience different rainfall regimes (~2,500 and 3,500 mm/year, respectively) (Brousseau et al. 2015). Each regional location harbors different microhabitats (hygromorphic sandy bottomlands and ferrallitic terra-firme plateaus or hills) separated by short distances (~300 m). In Laussat, a plateau of low elevation is adjacent to a permanently water-logged bottomland, while Regina is composed of higher elevation hilltops bordered by abrupt slopes surrounding a seasonally flooded bottomland forest. *E. falcata* trees are found through the study area, ensuring gene flow between tree populations occupying different microhabitats (Brousseau et al. 2015).

*Molecular methods*

Genomic DNA was extracted from fresh leaves according to a modified CTAB protocol (Doyle and Doyle 1987): CTAB buffer was supplemented in proteinase K in concentrations of 80 µg/ml, and the deproteinization step was realized twice (once with Dichloromethane:Chloroform:Isoamylalcool 24:24:1 and once with Chloroform:Isoamylalcool 24:1). Genomic DNA was eluted in Tris-EDTA 0.5X with RNase A in concentration 10 µg/ml, and quantified twice using a Nanodrop 2000. Individuals belonging to the same regional location and microhabitat were pooled together in equimolarity, resulting in four libraries of pooled individuals with total DNA concentrations ranging between 180 and 240 ng/µl. The four libraries were sequenced (paired-end) on three lanes of an Illumina HiSeq 2500. Each library was distributed over the three lanes, and each lane received the different libraries in equal quantity.

*Bioinformatics pipeline*

**Reads pre-processing.** Paired-end reads (100 nt) were pre-processed (cleaned and filtered) before assembly. Individual bases of low quality were masked using the ‘Fastq_masker’ tool of the suite fastx_toolkit (http://hannonlab.cshl.edu/fastx_toolkit) under a quality threshold of 25. Reads-ends were trimmed using Fastq_trimming’ (fastx_toolkit) by trimming the first and last ten bases of each reads (resulting in reads of 80 nt). Reads were further sorted and orphan
reads were excluded using Sickle (Joshi and Fas 2011) with the following parameters: minimum per base quality threshold = 25 and minimum read length = 70. The quality of the cleaned reads was assessed using ‘fastqc’. Reads pre-processing resulted in 2,332,116,952 high-quality reads (Table S1). Cleaned reads have been deposited on the European Nucleotide Archive (ENA, EMBL), under project accession code PRJEB9879, sample accession codes ERS791957, ERS791958, ERS791959, ERS791960.

**De novo assembly and mapping.** Because *E. falcata* is a non-model species with no reference genome, cleaned reads were assembled into a crude *de novo* genome assembly which constitutes a large new pan-genomic resource. Because classical assemblers do not easily support the numerous polymorphisms contained in libraries of pooled individuals (resulting in the segregation of reads from different populations pools into different assembled contigs and scaffolds), the sequenced libraries were *de novo* assembled using RayMeta (Boisvert *et al.* 2012), an assembler originally developed for meta-communities. The different libraries were assembled together in order to maximize the coverage of the *de novo* assembly, with a kmer size of 21. The assembled contigs were then aligned together in order to re-align redundant contigs using CD-HIT-EST (Li and Godzik 2006; Fu *et al.* 2012) with a sequence identity threshold of 95%. Non-redundant contigs were retained and used as reference to map *a posteriori* each library separately using bwa (Li and Durbin 2009): ‘bwa aln’ and ‘bwa sampe’ (only paired reads were mapped). The .sam files (one by library) were converted into .bam, sorted, indexed and merged into a unique .mpileup file using the samtools suite (Li *et al.* 2009): ‘samtools view’, ‘samtools sort’, ‘samtools index’ and ‘samtools mpileup’.

**Structural and functional annotation.** The crude pangenomic reference was annotated with the suite implemented in Maker mpi v. 2.31.8 (Cantarel *et al.* 2008) and associated tools. In summary, repetitive elements were detected by repeatMasker v.4.0.3 (http://www.repeatmasker.org/) and *ab initio* gene prediction was realized by Augustus v.2.6.1 (Stanke *et al.* 2006) at the scaffold level using both the reference transcriptome of *E. falcata* (Brousseau *et al.* 2014) and a composite proteome reference composed of all published reference proteomes of Fabaceae (*Cajanus cajan, Cicer arietinum, Glycine max, Lupinus angustifolius, Medicago truncatula, Phaseolus angularis, Phaseolus vulgaris* and *Vigna radiata*) plus the reference proteomes of *Arabidopsis thaliana*. The predicted proteins were further BLASTed against the non-redundant protein database with NCBI BLASTp v.2.6.0. At last, gene ontology terms of each predicted gene were retrieved using Interproscan 5.21.60 (Jones *et al.* 2014).
SNP detection and Filtering. The absolute abundance of different nucleotides was counted for each assembled position using the PoPoolation2 mpileup2sync.jar script (Kofler et al. 2011). The .java.sync output file containing nucleotide counts for each position of the assembly was post-processed through custom perl scripts. (1) Monomorphic sites were discarded and rare SNPs rare were masked (replaced by ‘N’) under a minimum allele frequency (maf) threshold of 0.05, all libraries confounded (meaning that a minimum number of 8 alternative variants out of 160 sequenced gametes in the whole dataset was required for a polymorphism to be retained). (2) Among the remaining SNPs, only bi-allelic polymorphisms with a total maximum coverage of 200X over all libraries to avoid SNPs caused by potential misassembly of paralogs into the same contigs and scaffolds. (3) Bi-allelic SNPs with library-level coverage < 20X were discarded, resulting in a SNP dataset composed of SNPs with coverage > 20X within each of the four libraries. (4) Contiguous SNPs were further filtered out to avoid polymorphisms caused by copy number variation (CNV). (4 and 5) SNPs located within repeat elements (detected by repeatMasker) as well as SNPs located in contigs with SNP density > 0.02 were filtered out, once again to avoid SNPs located in paralogs.

Computation of basic diversity statistics

Nei’s molecular diversity $\pi$ (Nei 1987) was computed for each of the four sub-populations at the contig level with unphased SNP data, under the hypothesis of random mating, as the sum of unbiased estimates of heterozygosity at the nucleotide (SNP) level (Nei 1987):

$$\pi = \sum_i^S H_{e,i}$$

with $S =$ number of polymorphic loci in a contig, and

$$H_{e,i} = H'_{e,i} \frac{n}{n-1}$$

where $n =$ sample size and $H'_{e,i} =$ expected heterozygosity at the $i^{th}$ polymorphism in the contig. Because different contigs have different lengths, to compare across contigs we present $\pi_L = \pi/L$, where $L$ is contig length, that is, molecular diversity per base.

Bayesian modelling of genomic divergence and outlier detection
A Bayesian modelling approach was used to identify outlier SNPs under selection. Brousseau et al. (2016) has described this Bayesian framework in detail and its power has been assessed through simulations, revealing low false discovery and false non-discovery rates. Briefly, allele frequencies were inferred from read counts and pairwise locus-specific $G_{ST}$ was estimated in the model, thus integrating coverage variability across the genome and libraries.

\[
n_{i,m} \sim Bin(p_{i,m}, N_{i,m})
\]

\[
q_{i,m} = 1 - p_{i,m}
\]

where $n_{i,m}$ is the read count of the reference allele and $N_{i,m}$ the total number of read counts at marker $m$ in population $i$. $p_{i,m}$ and $q_{i,m}$ are allele frequencies of the reference and the alternative alleles respectively.

\[
GST_{(i,j),m} = \frac{(H_{t_{(i,j),m}} - H_{s_{(i,j),m}})}{H_{t_{(i,j),m}}}
\]

$GST_{(i,j),m}$ is the Nei’s fixation index (i.e. genetic differentiation) between populations $i$ and $j$ at marker $m$. $H_{t_{(i,j),m}}$ and $H_{s_{(i,j),m}}$ are the expected heterozygosities in the total metapopulation and within subpopulations respectively:

\[
H_{s_{(i,j),m}} = (p_{i,m}q_{i,m}) + (p_{j,m}q_{j,m})
\]

\[
H_{t_{(i,j),m}} = 2\bar{p}_{m}\bar{q}_{m} = 2\left(\frac{p_{i,m} + p_{j,m}}{2}\right)\left(\frac{q_{i,m} + q_{j,m}}{2}\right)
\]

Pairwise $G_{ST}$ was further partitioned into genome-wide and locus-specific components according to a hierarchical framework: neutral (genome-wide) divergence both between study sites and between microhabitats within each study sites (Fig. 2a), and adaptive (locus-specific) divergence both between sites and between microhabitats common to both sites.

\[
GST_{(i,j),m} \sim N(\mu_{(i,j),m}, \sigma^2_R)
\]

\[
\mu_{(i,j),m} = kr_{(i,j)}(\mu_r + \theta r_m) + kl_{laussat_{(i,j)}}(\mu_l_{laussat} + \theta l_m)
\]

\[
+ kl_{regina_{(i,j)}}(\mu_l_{regina} + \theta l_m)
\]

\[
kr_{(i,j)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} ; kl_{laussat_{(i,j)}} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} ; kl_{regina_{(i,j)}} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}
\]

$\mu_r$ and $\mu_l$ are the global mean of differentiation (over all markers) at regional and local scales respectively: $\mu_r$ is the global mean of differentiation between sites, $\mu_l_{laussat}$ is the global mean of differentiation between microhabitats within the study site of Laussat, and $\mu_l_{regina}$ is the global mean of differentiation between microhabitats within the study site of Régina.

$kr_{ij}$ and $kl_{ij}$ are binary matrices describing whether the populations $i$ and $j$ belong to the same study site (regional scale) and to the same microhabitat within each study site (local
scale): \( k_{r_{ij}} = 1 \) if the two populations \( i \) and \( j \) belong to the different study sites, \( k_{r_{ij}} = 0 \) otherwise; \( k_{l\_laussat_{ij}} = 1 \) if the two populations \( i \) and \( j \) belong to different microhabitats within the study site of Laussat, \( k_{l\_laussat_{ij}} = 0 \) otherwise; \( k_{l\_regina_{ij}} = 1 \) if the two populations \( i \) and \( j \) belong to different microhabitats within the study site of Régina, \( k_{l\_regina_{ij}} = 0 \) otherwise.

\( \theta_{r_m} \) and \( \theta_{l_m} \) are locus-specific parameters describing the departure from the global means of differentiation at regional (i.e. between sites) and local (i.e. between microhabitats) scales. Posteriors of \( \theta_{r_m} \) and \( \theta_{l_m} \) are normally distributed allowing thus to attribute an empirical Bayesian p-value (eBP-value) to each SNP. SNPs with an eBP < 1% were considered as outliers under divergent (upper tail) or homogenizing (lower tail) selection between sites (\( \theta_{r_m} \)) or microhabitat within sites (\( \theta_{l_m} \)).

The Bayesian model was encoded in stan (Carpenter et al. 2016) and compiled into C++ using Rstan. Because Bayesian inference is computationally expensive, the model was applied to subsets of 10,000 SNPs on a Dell Power Edge CG220 Intel Xeon (CPU E5-2670 v2 @2.50Ghz) cluster node with parallelization on 5 CPU (with parameters: virtual memory vmem = 25G, virtual memory limit h_vmem=30G and parallel_fill = 5) and explicit chains parallelization in R with the package ‘parallel’. Both dataset hashing and explicit parallelization optimized computation times and the model ran in ~ 4 hours / subset. Outputs were post-processed with R, including chain convergence checking and eBP values inference from the posterior distribution of locus-specific parameters. Due to the large number of SNPs screened, a correction for multiple-testing was applied on eBP values with the package ‘fdrtool’ with a FDR threshold of 10%. The proportions of outliers under selection were compared between geographic scales with chi-squared tests (‘prop.test’ function). Enrichment tests were further conducted through \( X^2 \) tests in order to test (i) the enrichment of outliers located within exons, within a predicted gene and in the neighborhood (< 5 kb) of a predicted gene compared to the whole SNP dataset, as well as (ii) the enrichment of GO terms between predicted genes neighboring outlier SNPs compared to all predicted genes of the entire reference.

**Results**

The crude reference genome produced here, although incomplete and fragmented, constitutes a new, valuable pan-genomic resource containing key information about population-level genetic diversity (i.e. tens of thousands of loci scattered throughout the
The four population pools were assembled de novo into a high-coverage pan-genomic reference: de novo assembly led to 325,249 contigs in 272,865 scaffolds for a total assembly length of about 250 Mb (252,364,908 bases assembled de novo of which 249,308,901 were successfully mapped a posteriori by the different libraries, Table S1) and a mapping coverage of 57.81X on average, Fig. S3. The structural annotation predicted 32,789 genes and 117,278 exons, and detected 124,287 repeated elements. The functional annotation of the predicted genes resulted in 32,075 successfully BLASTed genes, and only 732 predicted genes did not return any BLAST hit. Additionally, gene-ontology terms (GO terms) were retrieved for 18,577 predicted genes, while the remaining 14,212 genes were not associated to any GO term.

A total of 97,062 bi-allelic SNPs scattered across 25,803 contigs and 23,276 scaffolds were identified, according to a stringent procedure, with a final SNP density of 0.003/base (sd=0.004) with a Ti/Tv=1.80, Table S2 and Fig. S4. Normalized molecular diversity $\pi_L$ varied between $9 \times 10^{-6}$ and $8 \times 10^{-3}$ (Table S3), with little variation among populations. 41,050 SNPs (42.3%) were located in scaffolds with at least one predicted gene while 56,012 SNPs (57.7%) were located in scaffolds without any predicted gene. 26,435 SNPs (27.2%) were located within a predicted gene, of which 10,863 (11.2%) within exons, while 14,615 SNPs were in the neighborhood of a predicted gene with a mean distance to a predicted gene of 737 bp (with min = 1bp and max = 27.2 kb).

The Bayesian model revealed low genome-wide differentiation between populations and significant departures from neutral expectations for a small subset of SNPs. Genome-wide genetic differentiation (average $G_{ST}$ across loci) ranged between 0.026 and 0.032 at both geographical scales, Fig. S5. Accordingly, genome-wide parameters of divergence estimated through $G_{ST}$ partitioning were not higher between regions than between microhabitats within regions, as confirmed by overlapping credible intervals, Table S4. Locus-specific departures from genome-wide divergence (i.e. outlier SNPs) were identified from the distribution of locus-specific parameters ($\theta_r m$ and $\theta l_m$). Between 375 and 2,045 outlier SNPs were detected to be under selection with an eBP threshold of 1%: 375 SNPs under homogenizing selection at regional scale, 2,045 SNPs under divergent selection at regional scale, 510 SNPs under homogenizing selection at local scale and 1,770 SNPs under divergent selection at local scale. After multiple-testing correction, the number of significant SNPs fell to 475 outlier SNPs (0.49% of all SNPs screened): 290 SNPs (in 268 contigs) under divergent selection at the regional scale and 185 SNPs (in 168 contigs) under divergent selection at the local scale, with no outlier shared between the regional and the local scales. Outlier SNPs under divergent
selection were characterized by clines in allele frequencies between regional locations and/or local microhabitats. Clines in allele frequency across sites and microhabitats are illustrated in Fig. 4 for different kinds of neutral and outlier SNPs. A total of 140 outlier SNPs (29.5% out of 475 outlier SNPs) were located within predicted genes, among which 39 (8.2%) were within exons, and 86 are in the neighborhood (< 5kb) of one or several predicted genes (Supplementary files S1 and S2). However, outlier SNPs were not more likely to be located within exons ($X^2 = 3.9368$, df = 1, p-value = 0.05), within predicted genes ($X^2 = 1.0846$, df = 1, p-value = 0.30), or in the neighborhood of predicted genes ($X^2 = 3.1964$, df = 1, p-value = 0.07), compared to any SNP in the entire dataset. Nevertheless, 87 GO-terms were significantly enriched in the subset of genes surrounding outlier SNPs (84 genes) compared to all predicted genes of the entire reference at the regional scale, and 100 GO-terms were significantly enriched in the subset of genes surrounding outlier SNPs (106 genes) at the local scale, Table S5.

**Discussion**

This study analyzed patterns of genomic divergence in wild and undisturbed forest stands of the Amazonian hyperdominant tree *Eperua falcata* in French Guiana. For this species, former studies have already reported high heritability in quantitative growth and leaf traits in a common garden experiment (Brousseau et al., 2013), suggesting high levels of standing genetic variation for functional phenotypic traits. In addition, seedlings native to different microhabitats displayed genetically-driven differences in phenotypic traits probably due to a combination of neutral processes (i.e. short-distance seed flow and local inbreeding) and local adaptation (Audigeos et al. 2013; Brousseau et al. 2015). However, past studies were not able to conclusively investigate the evolutionary drivers behind microgeographic genetic divergence because they were based on datasets restricted to few (e.g. candidate gene approach, Audigeos et al. (2013)) or anonymous markers (AFLPs, Brousseau et al. (2015)), and the study of microgeographic adaptation lacked a comprehensive survey at the genomic scale. Here, we provide a new, genomic-scale, evidence of microgeographic adaptation associated with (topographic and edaphic) habitat patchiness in this species.

*Patterns of genome-wide divergence.*
Genome-wide (supposedly neutral) differentiation was low at both the local and regional geographic scales, ranging from 2 to 3% (Fig. S5), similarly to what previously found by Brousseau et al. (2015), who reported $F_{ST}$ values between 2 and 3% between microhabitats within sites, and about 3% between these study sites (based on a screening of about 1,100 AFLPs). In two additional study sites in French Guiana (Paracou and Nouragues), another study (Audigeos et al. 2013) provided comparable estimates: $F_{ST}$ $\cong$ 1% between study sites and $F_{ST}$ from 1 to 2% between microhabitats within sites. The SNP density that we estimated here (a mean of 3 SNP every 1,000 bp, Table S2) was quite low: Neale and Kremer (2011) for example reported that SNP frequency in trees is generally in the range 1 SNP per 100 bp. However, in our study the stringent SNP filtering probably filtered out many actual polymorphisms, leading to an under-estimation of genome-wide SNP density.

Low genome-wide differentiation suggests that high levels of gene flow in *E. falcata* regularly homogenize genome-wide genetic diversity both between regions and between microhabitats within regions. The consequences of recurring inter-habitat gene flow on population adaptability could be paradoxical. On one hand, extensive gene flow can allow the spread of advantageous alleles between populations, thus granting high adaptive potential for populations established in heterogeneous environments. On the other hand, gene flow may also counteract the effect of natural selection, with migration load halting adaptive divergence of populations into locally adapted subpopulations (Lenormand 2002; Le Corre and Kremer 2012; De Kort et al. 2015; Pujol et al. 2018). The deleterious effect of gene flow on adaptive divergence could explain the low proportion of outliers detected (see below). Nevertheless, high gene flow (and resulting low genome-wide differentiation) did not completely swamp the signal of adaptive divergence between populations.

Microgeographic adaptation.

Locus-specific (putatively adaptive) departures from genome-wide divergence were detected for 0.49% of the 97,062 SNPs analyzed. We detected significantly more outliers at the regional than at the microgeographic scale: 290 versus 185 outlier SNPs ($X^2 = 22.826$, df = 1, $p = 1.77 \times 10^{-6}$). This result can be explained as the logical consequence of an order of magnitude difference in the geographical distances between populations: 300 meters between microhabitats within sites versus 300 km between study sites. Indeed, ecological conditions are probably more variable - and divergent selection stronger - at the regional scale than divergent selective pressures between microhabitats separated by hundreds of meters, and the
effect of divergence would be reinforced by more limited gene flow. The proportion of outliers detected over microgeographic scales, although lower, remains nevertheless non trivial. Because the Bayesian model focuses on replicated footprints of adaptation only, it is extremely unlikely that the detected outliers could have been structured between microhabitats in two different regions purely by chance (Savolainen et al. 2013). Moreover, no outlier was detected jointly at both spatial scales, as shown by the heart-shaped joint distribution of regional- and local-scale locus-specific parameters (Fig. 3). Instead, outliers were always specific to only one spatial scale, suggesting that different portions of the genome respond to different types of environmental constraints; in future studies, local adaptation in wild populations should be investigated by integrating different spatial scales, including microgeographic habitat patchiness, into the sampling design. It is however important to stress that our Bayesian approach may miss footprints of polygenic adaptation, if adaptation acts in different ways in the two study sites (through different selection strengths and/or different loci affected) which is known to be common in wild trees (Berg and Coop 2014; Lind et al. 2017; Harrisson et al. 2017). Indeed, $F_{ST}$-based methods are more prone to detect adaptation caused by loci of large effects than by loci with small effects, especially since adaptation occurs with ongoing gene flow (Hedrick 2006; Yeaman and Whitlock 2011; Le Corre and Kremer 2012; Savolainen et al. 2013). In other words, the set of outliers detected is probably enriched with loci of large effects. Even if the Bayesian model employed here proved its robustness to detect footprints of polygenic adaptation caused by weak selection strengths ($s=0.05$) in simulations (Brousseau et al. 2016), it may still miss footprints of polygenic adaptation if the strength of selection is variable among loci (Sork et al. 2013).

By providing evidence of adaptation to microgeographic habitat patchiness, our results agree with previous studies showing soil variation as a driver of adaptation in trees. Indeed, previous studies have reported examples of adaptation to soil nutrient composition such as potassium deficiency in *Cornus florida* (Cornaceae) (Pais et al. 2017), to aluminum and phosphate ion concentrations in *Pinus contorta* (Pinaceae) (Eckert et al. 2012), to substrate age in *Metrosideros polymorpha* (Myrtaceae) (Izuno et al. 2017), to soil types in *Abies religiosa* (Pinaceae) (Méndez-González et al. 2017), or to soil water availability in *Pinus albicaulis* (Pinaceae) (Lind et al. 2017). Preliminary evidence of microgeographic adaptation has been recently reported in Neotropical tree species, including *E. falcata* (Audigeos et al. 2013; Brousseau et al. 2013, 2015; Torroba-Balmori et al. 2017). At the community level, the study by Fortunel et al. (2016) suggests a mechanism through which microgeographic adaptation may take place: seedling mortality of *terra firme* species was high when planted in
seasonally flooded habitats. Topography-linked microgeographic patchiness is likely driven by a complex interaction of abiotic and biotic factors (including soil types (Misiewicz and Fine 2014), water-logging and flooding events, water availability (Clark et al. 1999), seasonal soil drought (Markewitz et al. 2010), soil fertility (John et al. 2007), light availability (Ferry et al. 2010), competition and predation (Fine et al. 2004)), hence many different classes of loci may respond to selection for microgeographic adaptation. It is, however, difficult to identify the exact genomic targets of natural selection, particularly in non-model species, because of potential hitchhiking between causal mutations and the analyzed SNPs, and because of the lack of chromosome-level genome assembly as well as no reliable genome-wide estimates of linkage disequilibrium in this species. Despite these limitations, the identity of the genes neighboring outlier SNPs may provide useful information, as many ‘divergence’ outlier SNPs detected here were located within or near (< 5 kb) protein-coding genes involved in a variety of biological functions (Supplementary files S1 and S2), with 106 genes neighboring a genomic footprint of microgeographic divergence (i.e. one or several outlier SNPs) and 100 Gene Ontology terms (corresponding to 49 genes) significantly enriched in the subset of genes neighboring outlier SNPs, Table S5. Some of them are particularly meaningful regarding their biological significance in terms of survival, growth, defense and reproduction (Supplementary files S1 and S2).

We detected several genes involved in plant response to stress near outlier SNPs. For example, we detected a glutaredoxin-C11 (predgene_010110) involved in managing the level of reactive oxygen species (ROS) in plant cells (Mittler et al. 2004), a H$_2$O$_2$-responsive gene encoding a metallopeptidase M24 family protein which is up-regulated under stress in Arabidopsis thaliana (Ordoñez et al. 2014), and a gene encoding a plant cysteine oxidase 4-like (predgene 003921) involved in response to flooding and soil hypoxia (Weits et al. 2014; White et al. 2017). Oxidative stress is very common in bottomlands, where water-logging and seasonal flooding events cause soil hypoxia and oxygen deprivation in root cells, leading to the production of ROS (Blokhina et al. 2003). This observation is concordant with a previous candidate-gene approach in which a catalase (i.e. a gene involved in ROS detoxification) has been reported under selection in this species (Audigeos et al. 2013). We also detected five genes with possible implications in lipid biosynthesis (predgenes 006075, 019500, 002577, 001066, 000689), as for example a palmitoyl-acyl carrier protein thioesterase which is known to play an essential role for plant growth and seed development (Dörmann et al. 2000; Bonaventure et al. 2003). Lipid metabolism is of high importance for plant development, from early germination to reproduction. Indeed, lipids are the primary material for cutin, and
they are largely involved in stress signaling and resistance (Okazaki and Saito 2014). We also detected two genes involved in proline amino acid biosynthesis (predgenes 007023 and 007024), which is accumulated in response to drought stress in plants (Fu et al. 2018). A calmodulin-like protein (probable CML48, predgene 022002) was also detected in the neighborhood of a microgeographic outlier. Calmodulin and calmodulin-like plant calcium sensors have evolved unique functions of high importance in the regulation of plant development and plant response to stress (Perochon et al. 2011). We also detected a probable cellulose synthase A (predgene 000745) whose mutations can alter morphology or induce resistance to cellulose biosynthesis inhibitors as reported in A. thaliana (Richmond 2000), a gene for plant metal tolerance (predgene 014347), and a mRNA-decapping enzyme-like (predgene 011170) which is of importance for early seedling development (Goeres et al. 2007).

We also detected two genes mediating biotic interactions - through plant defense against pathogens and recognition of symbiotic bacteria - as potential targets of microgeographic adaptation. In particular, plant chemical defense against herbivores and pathogens is likely to play a key role in microgeographic adaptation; previous studies have experimentally demonstrated that a trade-off between growth and antiherbivore defenses strengthens habitat specialization between congeneric species within large Neotropical plant families (e.g. Annonaceae, Burseraceae, Euphorbiaceae, Fabaceae, Malvaceae) (Fine et al. 2004, 2006, 2013). Here, we detected a chitinase (predgene 000254), a family of enzymes known for their role in resistance to pathogens (Punja and Zhang 1993; Grover 2012). We also discovered a gene involved in recognition of symbiotic bacteria (LysM receptor-like kinase, predgene 002974) which enables the model legume Lotus japonicus to recognize its bacterial symbiont (Madsen et al. 2003; Radutoiu et al. 2003; Ryals et al. 2008). Symbiotic nitrogen fixation is a specific feature of legumes, and nutrient availability largely varies across microhabitats (Ferry et al. 2010). The presence of genes involved in the perception of symbiotic bacteria near genomic footprints of microgeographic adaptation in this supposedly non-nodulating legume (Villadas et al. 2007) is particularly interesting. However, accurate functional characterizations of the proteins surrounding outlier SNPs are required to draw robust conclusions about the physiological processes involved in microgeographic adaptation.
Although largely exploratory, our results open a new window into the evolutionary processes underlying the generation of (Neo)tropical biological diversity, and support the hypothesis of niche evolution as a driver of habitat specialization and diversification (Leibold 2008). Microgeographic adaptation driven by a complex interaction of abiotic and biotic factors may be an initial step towards habitat specialization and ecological (sympatric) speciation by trees (Savolainen et al. 2006; Feder et al. 2012). In this study we provide support for this process at the intra-populational level in an Amazonian tree species. However, the efficacy of microgeographic adaptive divergence at a small but non negligible fraction of loci in driving ecological speciation (as opposed to the effect of genome-wide neutral divergence, caused by strong restrictions in gene flow, in assisting allopatric speciation) warrants further investigation. It is also possible that microgeographic adaptation contributes to the generalist and hyperdominant distribution of *E. falcata* (ter Steege et al. 2013). In this view, local adaptation to habitat patches would maintain multiple adaptation optima (i.e. multiple genetic combinations of alleles at different loci), thus pre-adapting the populations to a multitude of environmental situations. Indeed, two evolutionary processes can theoretically allow populations to cope with environmental changes, such as GCC, and avoid extinction (Aitken et al. 2008; Harrisson et al. 2014): populations can migrate to new areas to track ecological niches, causing shifts in the species’ distribution, and/or adapt locally to the new conditions through the action of natural selection (Hoffmann and Sgro 2011). These processes are likely to occur simultaneously and populations’ response to GCC can be viewed as a ‘race where populations are tracking the moving optima both in space and time’ (Kremer et al. 2012). From the standpoint of the entire metapopulation of *E. falcata* in the region, local adaptation to multiple optima would be tantamount to population-level balancing selection, which is considered to be a major component of adaptive potential (Barrett and Schluter 2008; Delph and Kelly 2014). Such a reservoir of adaptively useful genetic diversity may help the species deal with environmental challenges.

We provided here a new, genomic-scale, evidence of divergent selection at both regional and microgeographic scales in the Amazonian hyperdominant *Eperua falcata*, suggesting that microgeographic environmental heterogeneity caused by variations in topography and edaphic factors creates the conditions for local adaptation in lowland Neotropical rainforests. We notably detected signals of adaptive divergence at many SNPs, whose surrounding genes may ultimately constitute the genomic basis for local adaptation. Accurate functional annotation further gave some interesting clues as to the biological processes targeted by divergent selection, emphasizing the potential roles of abiotic stresses...
related to water-logging and drought, and of biotic interactions, in driving local adaptation.
We firmly believe that new advances in next-generation sequencing technologies and in
Bayesian methods will contribute to fill the gaps in our understanding of trees evolution in the
Neotropics, and we strongly recommend the integration of different spatial scales (including
microgeographic habitat patchiness) in future studies addressing the processes of adaptation
and diversification in lowland rainforests of Amazonia.

Acknowledgements
This study has been funded by the ANR ‘BIOADAPT FLAG’ (ref. ANR-12-ADAPT-0007-01)
and benefited from an ‘Investissement d’Avenir’ grant (CEBA, ref. ANR-10-LABX-25-01),
both managed by the French Agence Nationale de la Recherche. Louise Brousseau has been
funded by a young investigator grant ‘Contrat Jeune Scientifique’ (CJS) of the French
National Institute for Agronomical Research (INRA). We are grateful to the Genotoul
bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul) for providing computing
and storage resources. We thank Saint-Omer Cazal and Julien Engel for technical assistance
and for botanical identifications. We also thank Dr Anna Stavrinides for helpful discussions
about the functional interpretation of the results.

Author contribution
LB, ED and IS designed the experiment. LB sampled leaf material, extracted and conditioned
DNA, analyzed the data including the development of the Bayesian model under the
supervision of IS and GVV. All authors wrote the article.

References
migration or extirpation: climate change outcomes for tree populations. Evolutionary
Applications 1: 95–111.
Antonovics J. 2006. Evolution in closely adjacent plant populations X: long-term persistence
Antonovics J, Bradshaw AD. 1970. EVOLUTION IN CLOSELY ADJACENT PLANT
Audigeos D, Brousseau L, Traissac S, Scotti-Saintagne C, Scotti I. 2013. Molecular


ONE 11: e0158216.


Cowan RS. 1975. *A monograph of the genus Eperua (Leguminosae-Caesalpinioideae).* Smithsonian Institution Press.


Ferry B, Morneau F, Bontemps J-DD, Blanc L, Freycon V. 2010. Higher treefall rates on slopes and waterlogged soils result in lower stand biomass and productivity in a tropical rain


**Fine PVA, Daly DC, Cameron KM. 2005.** The contribution of edaphic heterogeneity to the evolution and diversity of Burceraceae trees in the Western Amazon. *Evolution* 59: 1464–1478.


De Kort H, Vandepitte K, Mergeay J, Mijnsbrugge K V., Honnay O. 2015. The


**Latreille AC, Pichot C. 2017.** Local-scale diversity and adaptation along elevational gradients assessed by reciprocal transplant experiments: lack of local adaptation in silver fir populations. *Annals of Forest Science* 74: 77.


Yeaman S, Otto SP. 2011. Establishment and maintenance of adaptive genetic divergence


**Author information, data availability**

Sequencing data have been deposited on the European Nucleotide Archive (ENA, EMBL), under the project accession code PRJEB9879. It contains the raw reads (accessions ERS791957, ERS791958, ERS791959, ERS791960) and the draft *de novo* genome (accession ERS791961). SNPs data are available on Figshare under the DOI 10.6084/m9.figshare.4212126. Data are under embargo until formal publication of this manuscript. In the meanwhile, correspondence and requests should be addressed to Louise Brousseau (louise.brousseau@ird.fr).

**Online supporting information**

**Supplementary figures S1 to S5.**

**Supplementary tables S1 to S5.**

**Supplementary file S1.** Interactive gene network to easily navigate through the functional annotation of genes neighboring ‘divergence’ outlier SNPs (< 5 kb) at the microgeographic scale. Green bubbles indicate predicted genes; blue and pink bubbles indicate Gene ontology terms, with size depending on the number of predicted genes neighboring an outlier SNP flagged with each GO term; pink bubbles indicate GO terms significantly enriched between the subset of predicted genes neighboring outlier SNPs compared to the all predicted genes.

**Supplementary file S2.** BLASTp and functional annotation of genes neighboring (< 5 kb) ‘divergence’ outlier SNPs at the microgeographic scale.
**Figures legend**

**Fig. 1.** (a) Relief map of forest landscapes in the region of Régina, French Guiana. The map was created through the French ‘Géoportail’ (https://www.geoportail.gouv.fr/): official geospatial data were provided by the French Institute of Geographic and Forest Information ©IGN 2017. (b) *Eperua falcata* sapling. (c) and (d) Forest communities associated with topography and related soil types: *terra-firme* forests on ferralitic soils at the top of hills (c), water-logged forests on hydromorphic soils in bottomlands (d).

**Fig. 2.** (a) Sampling design scheme and correspondence with genome-wide parameters of divergence estimated through the hierarchical Bayesian model at regional ($\mu_r$) and local scales ($\mu_{Laussat}$ and $\mu_{Régina}$ respectively). (b) Distribution of geographic distances between trees in the study sites of Laussat (Western French Guiana) and Régina (Eastern French Guiana). Vertical dotted bars indicate the range of habitat turnover in each study site.

**Fig. 3.** (a) Locus-specific parameters ($\theta_{rm}$ and $\theta_{lm}$) and normal distribution inferred by the Bayesian model. Vertical bars indicate the location of outliers under homogenizing (lower tail) and divergent (upper tail) selection. Left: regional scale ($\theta_{rm}$), right: local scale ($\theta_{lm}$). (b) Proportion of outlier SNPs under selection at regional (i.e. between study sites) and at local (i.e. between microhabitats within sites) scales. (c) Joint distribution of regional and local-scale locus-specific parameters inferred through the Bayesian model. Blue dots indicate outliers under divergent selection at the regional scale, green dots indicate outliers under divergent selection at the local scale.

**Fig. 4.** Clines in allele frequency between regional locations (i.e. study sites) and microhabitats for the different kinds of (randomly sampled) neutral and outlier SNPs at different geographical scales.