MR. BEYENE ZEWDIE (Orcid ID : 0000-0002-6020-916X)

Article type : Special Issue

Genetic composition and diversity of Arabica coffee in the crop's center of origin and its impact on four major fungal diseases

Beyene Zewdie[1]*‡, Yves Bawin[2,3,4,5]‡, Ayco J. M. Tack[1], Sileshi Nemomissa[6], Kassahun Tesfaye[7], Steven B. Janssens[5,8,9], Sabine Van Glabeke[3], Isabel Roldán-Ruiz[3,4], Tom Ruttink[3], Olivier Honnay[2,9], Kristoffer Hylander[1]

[1] Department of Ecology, Environment and Plant Sciences, Stockholm University, SE-106 91 Stockholm, Sweden

[2] Plant Conservation and Population Biology, KU Leuven, Leuven, Belgium

[3] Plant Sciences Unit, Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), Melle, Belgium

[4] Department of Plant Biotechnology and Bioinformatics, Ghent University, Zwijnaarde, Belgium

[5] Crop Wild Relatives and Useful Plants, Meise Botanic Garden, Meise, Belgium

[6] Department of Plant Biology and Biodiversity Management, Addis Ababa University, Addis Ababa, Ethiopia

[7] Institute of Biotechnology, Addis Ababa University, Addis Ababa, Ethiopia

[8] Department of Biology, KU Leuven, Leuven, Belgium

[9] Leuven Plant Institute, Heverlee, Belgium

*Corresponding author: beyu2009@gmail.com

‡Authors have equal contribution

Abstract

Conventional wisdom states that genetic variation reduces disease levels in plant populations. Nevertheless, crop species have been subject to a gradual loss of genetic variation through selection for specific traits during breeding, thereby increasing their vulnerability to biotic stresses such as pathogens. We explored how genetic variation in Arabica coffee sites in southwestern Ethiopia was related to the incidence of four major fungal diseases. Sixty sites were selected along a gradient of management intensity, ranging from nearly wild to intensively managed coffee stands. We used genotyping-by-sequencing of pooled leaf samples (pool-GBS) derived from 16 individual coffee shrubs in each of the sixty sites to assess the variation in genetic composition (multivariate: reference allele frequency) and genetic diversity (univariate: mean expected heterozygosity) between sites. We found that genetic composition had a clear spatial pattern and that genetic diversity was higher in less managed sites. The incidence of the four fungal diseases was related to the genetic composition of the coffee stands, but in a specific way for each disease. In contrast, genetic diversity was only related to the within-site variation of coffee berry disease, but not to the mean incidence of any of the four diseases across sites. Given that fungal diseases are major challenges of Arabica coffee in its native range, our findings that genetic composition of coffee sites impacted the major fungal diseases may serve as baseline information to study the molecular basis of disease resistance in coffee. Overall, our study illustrates the need to consider both host genetic composition and genetic diversity when investigating the genetic basis for variation in disease levels.

Key words: *Coffea arabica*; Fungal diseases; Genetic composition and genetic diversity; Genotyping-by-sequencing; Management gradient; Southwestern Ethiopia.

Introduction

Genetic variation is a key determinant of the persistence of natural populations when exposed to pathogen infections (Jump et al., 2009). Disease levels are often lower in genetically heterogeneous natural populations, which could be attributed to the presence of diversity in resistance mechanisms among individuals compared to genetically homogenous populations (Burdon & Laine, 2019; Ekroth et al., 2019). The positive relationship between genetic variation and disease resistance has also been observed in agricultural systems, where fields with cultivar mixtures or multiline cultivars often have lower disease levels than fields with single cultivars (Mundt, 2002; Reiss & Drinkwater, 2018; Zhu et al., 2000). Conventional wisdom states that parasites transmit more readily between closely related individuals and that genetic variation in the host populations reduces disease burdens. Nevertheless, the relationships between host diversity and parasite burden is context-dependent and might vary for example between experimental and wild host populations (Gibson & Nguyen, 2020). The impact of a disease on host populations may depend on many variables and the relationship between genetic variation in host populations and the intensity of a specific disease remains understudied.

Despite the potential benefits of genetic diversity in agricultural systems, modern agriculture still heavily relies on improved crop varieties cultivated in monocultures (Zhou et al., 2002). The low levels of genetic diversity in managed agricultural fields might have facilitated the spread of diseases in several crops, resulting in some cases in total crop losses. One striking example is the wipe-out of monoclonal banana plantations by *Fusarium* wilt (Dita et al., 2018). Existence of genetic variation in crops is important for resistance or tolerance to biotic stresses such as diseases (Colque-Little et al., 2021; Rodenburg et al., 2017). Therefore, a better understanding and management of genetic variation in crops is essential to optimize the conservation and use of crop genetic resources in breeding programs (Brozynska et al., 2016; Fu, 2015; Zhou et al., 2002).

Arabica coffee, *Coffea arabica* L., is widely grown across the tropics and subtropics for its beans, which are used to produce one of the most consumed beverages in the world (ICO, 2020). Nevertheless, Arabica coffee cultivation is highly vulnerable to pests and diseases, of which coffee leaf rust (causal agent *Hemileia vastatrix* Berk. & Broome) is one of the best known problems for the production of Arabica coffee due to its worldwide presence on coffee plantations (Avelino et al., 2018; McCook, 2006). The intensity of coffee leaf rust infection is predicted to increase due to global warming, posing a major threat to global coffee production (Talhinhas et al., 2017; Toniutti et al., 2017). The high susceptibility of cultivated coffee for pests and diseases can partly be ascribed to the low genetic diversity within and among coffee cultivars (Labouisse et al., 2008; Setotaw et al., 2013; Silvestrini et al., 2007; Steiger et al., 2002). To increase the resistance of cultivated Arabica coffee against biotic stressors, the introduction of genetic variation from wild *C. arabica* individuals to the cultivated genepool has recently been encouraged (Davis et al., 2019; Scalabrin et al., 2020). Interestingly, the progenitors of the most commonly cultivated Arabica coffee (i.e. Typica and Bourbon varieties) were found to be genetically differentiated from wild accessions, suggesting that wild accessions have a large potential for improvement of the globally cultivated *Coffea* genepool (Hein & Gatzweiler, 2006; Sant'Ana et al., 2018; Scalabrin et al., 2020; Silvestrini et al., 2007; Tesfaye et al., 2014).

The primary center of origin and diversity of *C. arabica* is situated in the southwestern Ethiopian highlands (Davis et al., 2012), a region characterized by a mosaic landscape with some larger areas of moist evergreen Afromontane forests, many small forest patches, and open areas for cultivation of annual crops, communal grazing areas and home gardens (Lemessa et al., 2013). Arabica coffee is endemic to the understory of the moist Afromontane forest in Ethiopia, where human disturbance is limited to the harvest of ripe coffee berries and spices (Gole et al., 2008). Arabica coffee is also the major cash crop in southwestern Ethiopia, where it is grown under native forest trees across a broad management intensity gradient. At the lowest levels of management intensity, coffee is grown by smallholder farmers under a diverse tree canopy, both in forest edges and in patches embedded in a matrix of annual crop lands (Lemessa et al., 2013). The thinning of the understory and the removal of herbaceous vegetation is a common practice in smallholder farms, while the use of plant protection agents and fertilizers is unusual due to a lack of resources. Farmers also additionally plant self-generated seedlings or seedlings from selected cultivars to increase yield (Aerts et al., 2011; Schmitt et al., 2010). At the other end of the management intensity gradient, coffee is cultivated in intensively managed plantations, which consist exclusively of selected coffee cultivars (often resistant to coffee berry disease) and a few sparsely placed shade trees. More intensive management practices like pruning, fertilization, and weeding or herbicides are applied in this system. Pesticides are not used in this landscape for control of diseases and pests. In contrast to the high number of traditional smallholder coffee farms, intensively managed coffee plantations are rare and much more recent in Ethiopia (Labouisse et al., 2008). However, in practice, coffee management has intensified during the last four decades, imposing a high pressure on the natural coffee forests. As a result, forest degradation and fragmentation in southwestern Ethiopia rapidly increased (Aerts et al., 2017; Geeraert et al., 2019; Shumi et al., 2019), threatening the wild Arabica gene pool (Berecha et al., 2014). The diversity of Arabica coffee genetic resources is also at risk due to climate change (Davis et al., 2012; Moat et al., 2017, 2019) and genetic erosion through admixture of wild individuals with cultivars (Aerts et al., 2013).

Arabica coffee stands in Ethiopia are threatened by several fungal diseases, including coffee leaf rust, coffee berry disease (causal agent *Colletotrichum kahawae* Waller & Bridge), coffee wilt disease (causal agent *Gibberella xylarioides* Heim & Saccas), and *Armillaria* root rot (causal agent *Armillaria mellea* Vahl ex and Fries) (Avelino et al., 2018; Hindorf & Omondi, 2011; Zewdie et al., 2020). Coffee leaf rust is recognized by orange powdery spores on the abaxial side of leaves. These dry spores can be dispersed over long distances by wind or insects, while local dispersal is facilitated by rain splashes or coffee workers. Coffee berry disease infects young developing berries, displaying black sunken spots that grow to cover the whole berry and eventually result in completely mummified beans that fall from the shrub (Hindorf & Omondi, 2011; Waller et al., 1993). Coffee berry disease occurs more frequently in forested areas at higher altitudes, whereas coffee leaf rust is severe in more intensively managed systems, especially at lower altitudes (Daba et al., 2019; Zewdie et al., 2020). Coffee wilt disease infects coffee shrubs through wounds and blocks the vascular system, eventually leading to the wilting of the coffee shrubs (Girma et al., 2001, 2009). *Armillaria* root rot kills infected coffee shrubs as it completely damages the roots (Gezahgne et al., 2004). Coffee wilt disease and *Armillaria* root rot spread mainly through contact with infected soil or movement of diseased plant material between sites (Waller et al., 2007). The increase in the severity of coffee berry disease in particular has forced Ethiopian coffee growers to gradually replace their landraces by resistant cultivars, a process that has drastically reduced the genetic diversity of cultivated coffee in the area (Aerts et al., 2013). While the temporal aspect of the co-evolutionary history of these fungal diseases with coffee in this landscape is not well-known, the diseases have been present in the landscape for at least several decades, if not longer. This suggest some that host-pathogen co-evolution has taken place. Coffee leaf rust is believed to have a long co-evolutionary history in East Africa (McCook, 2006), coffee berry disease likely originated in the neighbouring country Kenya (Hindorf & Omondi, 2011), and coffee wilt disease was first reported in the Central African Republic on Excelsa coffee, *Coffea liberica* var. *dewevrei*, although a different strain of the pathogen infects Arabica coffee in Ethiopia (Girma et al., 2001). Consequently, southwestern Ethiopia harbours a unique landscape to investigate the existence of host-pathogen co-evolutionary relationships in Arabica coffee. Taken together, a thorough characterisation of the incidence of these fungal diseases in Arabica coffee stands in Ethiopia in relation to their genetic variation is needed to optimally conserve and exploit Arabica coffee genetic resources for disease resistance. Nevertheless, studies that investigate the relationship between genetic composition and genetic diversity of coffee on one hand and the incidence of diseases on the other hand across the Arabica coffee landscape in Ethiopia are lacking.

In this study, we aimed to link genetic variation in Arabica coffee to the incidence of four major fungal diseases along a gradient of management intensity in southwestern Ethiopia. We collected leaf samples from 16 coffee shrubs per site in a total of 60 sites ranging from nearly wild to intensively managed coffee. We estimated genetic variation by performing Genotyping-By-Sequencing on pooled samples (pool-GBS), to capture global patterns of sequence polymorphisms at the population level. Pool-GBS is a cost-efficient library preparation method for genome-wide allele frequency fingerprinting (GWAFF) of large numbers of samples (Bélanger et al., 2016; Byrne et al., 2013; Verwimp et al., 2018). After the experimental validation of the pool-GBS method in *C. arabica*, we addressed the following questions:

1. Does the genetic composition and genetic diversity of Arabica coffee stands vary along gradients of environmental variables, coffee management intensity, and spatial location?
2. How does the incidence of coffee leaf rust, coffee berry disease, coffee wilt disease, and *Armillaria* root rot relate to the genetic composition and genetic diversity of the coffee stands?
3. Does the among-coffee shrub variation in the incidence of coffee leaf rust and coffee berry disease relate to the genetic diversity in coffee stands along the management intensity gradient?

We expected differences in genetic composition (allele frequency spectrum) among the natural and more intensively managed sites and a higher level of genetic diversity in more natural forest sites compared to more intensively managed sites. We also expected that genetic composition in the coffee stands would relate to the variation in disease levels, that higher genetic diversity in coffee stands would coincide with a lower disease incidence at site-level, and that higher genetic diversity at site-level would correlate with a higher variation in the incidence of diseases among coffee shrubs within the same site.

Materials and Methods

Site selection and environmental variables

The present study was conducted in the Gomma and Gera districts of the Jimma zone in the Oromia regional state in southwestern Ethiopia. Collection sites were located between 7°37'–7°56' N and 36°13'–36°39' E (Fig. 1a,b). The region is characterized by a unimodal rainfall pattern with the main rainy season between May and September and the main dry season between December and March. The altitude of the area ranges from 1506 to 2159 m above sea level. We selected 60 coffee sites representing a broad gradient of management intensity including nine intensively managed commercial coffee plantations. At each site, we established a plot of 50 × 50 m where we recorded biotic and abiotic environmental and management variables, and marked 16 coffee shrubs at the intersections of 10 × 10 m grid cells in the central 30 × 30 m grid (Fig. 1c). More specifically, we recorded i) altitude, ii) canopy cover, iii) number of shade trees >20 cm in diameter at breast height (DBH) in the 50 × 50 m plot, iv) coffee density as a count of coffee shrubs >1.5 m height in the central 30 × 30 m plot; and v) coffee shrub structure index. Canopy cover was based on the average of five canopy images taken above the coffee shrub layer with a Nikon Coolpix S2800 camera tied to a long stick to rise above coffee canopy and analysed separately using imageJ software (Schneider et al., 2012). Coffee structure index was created based on five attributes measured on each of the 16 coffee shrubs per site: i) number of primary and secondary orthotropic (vertical, vegetative) shoots, ii) number of plagiotropic (horizontal, fruit bearing) shoots, iii) average stem diameter at knee height, iv) average of two perpendicular diameters of the ground projection of the coffee shrub canopy, and v) proportion of the coffee height with plagiotropic branches. The index accounts for variation in coffee shrub architecture as a result of variation in management and ranges from 1 (less intensive management) to 3 (intensive management). The environmental and management variables (Table S1) were assessed in 2017 from March to May and from July to August and were also used in a previously published study (Zewdie et al., 2020).

Fungal disease assessment

We recorded coffee leaf rust on 16 coffee shrubs per site during the dry season in 2017 (March to May) and 2018 (January to February). We assessed leaves for coffee leaf rust infection on three branches per shrub and calculated coffee leaf rust incidence as the number of leaves with coffee leaf rust out of the total number of assessed leaves at the shrub level. Coffee berry disease was assessed during the wet season of 2017 and 2018 from July to August. We recorded the total number of berries and berries with coffee berry disease infection on three branches per shrub for the 16 coffee shrubs per site. Coffee berry disease incidence was calculated as the proportion of berries with coffee berry disease symptoms divided by the total number of berries counted. Coffee wilt disease and *Armillaria* root rot were assessed within the whole 50 × 50 m plot at each site during the 2017 wet season from July to August. Their respective incidence was calculated as the proportion of coffee shrubs with coffee wilt disease or *Armillaria* root rot symptoms out of the total number of coffee shrubs in the 50 × 50 m plot. For the two fungal diseases that were assessed at coffee shrub level (coffee leaf rust and coffee berry disease), we further investigated the magnitude of variation in the incidence of the diseases among coffee shrubs within a site. We calculated the standard deviation for the incidences of each disease from the 16 coffee shrubs per site.

Genotyping by sequencing (GBS) in pooled and individual samples

Leaf samples were collected from March to May 2017 from young but fully expanded leaves from the 16 selected shrubs at each of the 60 sites resulting in a total of 960 leaf samples. Per site, a single tissue pool sample was created by pooling ca. 3 mg of silica-dried leaf material from all individuals belonging to the same site. In parallel, ca. 20 mg of silica-dried leaf material of each of the 16 individual samples from two of the sites representing two different management intensities (Gera 1 and Gomma 16) were analysed as individual samples to validate the pool-GBS method. A *Pst*I single-enzyme GBS protocol slightly adapted from Elshire et al. (2011) was used to construct GBS libraries of the 60 pooled samples (two GBS ligation replicates per sample) and 32 individual DNA samples (one GBS ligation replicate per sample) (Supplemental Materials, Fig. S1).

Coffea arabica is an allotetraploid species (2n = 4x = 44) most likely originating from a single interspecific hybridization event between the diploid species *C. canephora* and *C. eugenioides* (Bawin et al., 2021; Scalabrin et al., 2020; Tesfaye et al., 2007). The *C. arabica* genome thus comprises two subgenomes, each derived from one of its progenitor species. Because sequence-based genotyping relies on mapping reads obtained by high-throughput sequencing onto a reference genome sequence and identifying read-reference polymorphisms, the choice of the reference genome sequence is critical. In allotetraploids, one may choose either one subgenome as a non-redundant reference (to avoid ambiguous read mapping) or both subgenomes (to capture the entire sequence space). In our approach, reads of *C. arabica* were mapped onto the genome sequence of *C. canephora* (Denoeud et al., 2014). Consequently, an equal number of reads derived from both *C. arabica* subgenomes may map onto their respective homoeologous region in the reference genome sequence, creating 'genome-collapsed' loci (Limborg et al., 2016). Importantly, variant calling algorithms will identify, but not discriminate between, *within*-subgenome polymorphisms (derived from the different alleles of a given locus on one of the subgenomes; 'true' Single Nucleotide Polymorphisms (SNPs); relevant for estimates of population genetic diversity and genetic composition), and *between*-subgenome polymorphisms (derived from reads of homoeologous loci; resulting from the evolutionary sequence divergence between the founder species of the allotetraploid; not relevant for population genetics). Given the assumed single interspecific hybridization event at the origin of *C. arabica* (Scalabrin et al., 2020), such *between*-subgenome polymorphisms are likely shared by all individuals of this species, whereas *within*-subgenome SNPs differentiate between individuals, and in turn, their allele frequency varies between populations. Because *between*-subgenome polymorphisms carry signals related to evolutionary genetics, they should be excluded before estimating population genetic parameters such as genetic differentiation and genetic diversity based on 'true' *within*-subgenome SNPs (See Supplementary Materials for further details). In this regard, a previous study that processed high-throughput sequencing data of individual *C. arabica* samples (one genotype per sample) removed polymorphic positions that were consistently called as 'heterozygous state' across the set of individuals (Sant'Ana et al., 2018). Nevertheless, the detection of *between*-subgenome polymorphisms based on their fixed heterozygous state is not sufficiently accurate in high-throughput sequencing data of pooled samples (multiple genotypes per sample). In particular, the allele frequency (quantitative variable between 0 and 1) of *between*-subgenome polymorphisms in pooled samples may substantially deviate from an allele frequency of 0.5 in every sample due to stochastic fluctuations in the contribution of each allele to the read depth of a given locus (Andrews et al., 2016; Limborg et al., 2016). We therefore implemented a more suitable filtering method for the removal of *between*-subgenome polymorphisms in pool-Seq data based on the relative stability of allele frequencies across populations (measured as F_{ST} values), instead of strict allele frequency thresholds as filter criterion.

Although the RAF_{pool} value of a *between*-subgenome polymorphism may differ from its expected value (0.5) in a *single* pooled sample (i.e. per locus determined by near-equal read depth derived from both subgenomes in each constituent individual), its RAF_{pool} value in *all* sixty pooled samples was expected to be stable and often centred around 0.5. Because of the high expected consistency of RAF_{pool} values per sample, the genetic differentiation on that single polymorphic position measured over all sixty pooled samples (therefore, estimated by F_{ST}) is expected to be relatively low. Consequently, *between*-subgenome polymorphisms can be identified based on their level of genetic differentiation among a large set of pooled samples and removed, by F_{ST} threshold, irrespective of their RAF_{pool} value in a single pooled sample.

The GBS read data were pre-processed and mapped onto the reference genome sequence of *C. canephora* (Denoeud et al., 2014), which was the only published reference genome sequence of the genus *Coffea* with full access to all (meta)data at the time of data processing. The Bayesian variant calling algorithm implemented in SNAPE-pooled (Raineri et al., 2012) was used to identify variant positions in pool-GBS data. For each variant position, the allele frequency per pool sample is calculated as the number of reads representing the reference allele (i.e. the allele in the *C. canephora* reference genome sequence) divided by the total number of reads mapped to that position, with a minimal read count of 30 reads (denoted as Reference Allele Frequency, RAF_{pool}). Variants in the 32 individual-GBS samples were called using the Unified Genotyper in the Genome Analysis ToolKit (GATK) v3.7 (McKenna et al., 2010). After variant filtering, the reference allele frequency per variant position across all individuals per site (for the 32 individuals from 2 sites) was calculated as the number of discrete called reference alleles divided by the total number of discrete called alleles in the set of genotypes per site (RAF_{ind}) for comparison with RAF_{pool}. A detailed overview of the individual-GBS and pool-GBS protocol and read data analyses is provided as supporting information (Supplementary Materials, part 1).

To remove *between*-subgenome polymorphisms from pool-GBS data, we calculated F_{ST} across the sixty sites for each variant separately following Nei & Chesser (1983) and variants with an F_{ST} value lower than 0.03 were removed. Calibration of the F_{ST} threshold value is described in detail in the Supplemental materials, part 2. In the individual samples, positions with heterozygous genotype calls in at least 75% of the individuals were considered as *between*-subgenome polymorphisms and subsequently removed (Sant'Ana et al., 2018). The number of *within*-subgenome SNPs in the individual-GBS data and the corresponding pool-GBS data was compared to determine the agreement between both SNP sets and effects of various parameters during variant calling and filtering. The python scripts used to discard *between*-subgenome polymorphisms in pools and individuals are available on Gitlab (see Supplemental Materials, part 2).

Validation of pool-GBS

The number of GBS loci with a minimum depth of 30 reads that was shared between a pooled sample and the corresponding individual samples was determined using BEDTools v2.27.1 (Quinlan & Hall, 2010). The reproducibility of RAFs in every pooled sample (n = 60) and the accuracy of RAF_{pool} values in the pooled sample of sites Gera 1 and Gomma 16 (n = 2) was assessed by the variance explained by predictive models based on cross-validation (VEcv) (Li, 2016, 2017). VEcv shows the percentage of variation in the reference data that is explained by the observed data. A VEcv value higher than 80% is considered as excellent (Li, 2016). The RAF_{pool} of *within*-subgenome SNPs in the first pool-GBS ligation replicate and the RAF_{pool} of *within*-subgenome SNPs in the second pool-GBS ligation replicate of each pooled sample (Fig. S1) were set as the reference and the observed data, respectively, to assess the reproducibility of RAFs between both replicates. The RAF_{ind} of *within*-subgenome SNPs were considered as reference values for the RAF_{pool} in their corresponding pool-GBS samples (Supplemental Material, Fig. S1) to assess the accuracy of RAFs in pools. The VEcv was calculated between the reference and the observed values using the '*vecv'* function in the R package SPM (Li, 2016) in R v3.6.1 (R Core Team, 2019).

Estimation of genetic composition and genetic diversity

Genetic variation was quantified with several metrics divided into two groups: multivariate genetic composition and univariate genetic diversity. Genetic composition consisted of the RAF_{pool} of all *within*-subgenome SNPs across all sites. This allele-frequency-by-site matrix was used for ordination analyses (see below). Moreover, we calculated the genetic differentiation between all pairs of sites as

F_{ST} following (Weir & Cockerham, 1984) using the '*stamppFst*' function in the R package STAMPP (Pembleton et al., 2013). The ploidy level of each site was set to 64, which equals the ploidy level of *C. arabica* (4) multiplied by the number of individuals collected per site (16). For genetic diversity, we calculated three different metrics: the mean expected heterozygosity (mean H_E), nucleotide diversity pi (π), and Watterson's estimator theta (θ). The expected heterozygosity at each SNP position was calculated as H_E = 2*RAF*(1-RAF). The mean expected heterozygosity is relatively robust to fluctuations caused by low frequency alleles, representing a conservative measure for genetic diversity in populations (Luikart & Cornuet, 1998; Nei et al., 1975). The nucleotide diversity pi and Watterson's estimator theta were first calculated for each pool-GBS ligation replicate separately. Subsequently, the mean of the two replicates was calculated to obtain one value for each site. The calculations of π and θ were also restricted to the set of *within*-subgenome SNPs that was retained after all filtering steps (Supplemental Material, Fig. S1). Both π and θ were estimated with NPstat v1 (Ferretti et al., 2013) using the filtered Samtools mpileup files that were created for SNP calling. NPstat was run with a window size of 10,000, a maximum coverage of 500, and without a minimum allele count filter (m = 0). Because the values of the three genetic diversity estimates (mean H_E, π, and θ) were highly correlated ($r \geq 0.69$, Fig. S6), we chose to conduct all further analyses with one parameter for genetic diversity (i.e. mean H_E).

Genetic composition and genetic diversity of coffee stands along gradients in environmental, management intensity, and spatial variables

The variation in genetic composition of coffee stands was assessed with a principal component analysis (PCA) on the Hellinger transformed RAF_{pool} data using the *'rda'* function in the R package VEGAN (Oksanen et al., 2019). The Hellinger transformation was used to standardize the data for the multivariate approach. To be able to visualise the variation in genetic composition across the landscape we performed a cluster analysis on the RAF_{pool} data. We did not intend to delineate distinct clusters of sites with similar genetic composition, but we aimed to visualize to what extent the variation in genetic composition also displayed spatial patterns. This was obtained by subsequent colour-marking of the cluster groups on a map. The clusters were defined with a hierarchical cluster algorithm using the '*hclust'* function in base R. To determine the appropriate clustering algorithm and optimal number of clusters, we first performed cluster validation using the '*clValid'* function in the R package CLVALID (Brock et al., 2008). This function allows the simultaneous selection of multiple clustering algorithms, validation measures, and number of clusters in a single function call. We also illustrated the spatial variation in genetic diversity of the coffee sites on a map of the study area using the *'geocode'* function in the R package GGMAP (Kahle & Wickham 2013) and colour-marked the sites according to their mean expected heterozygosity. We assessed the relationship between the genetic composition of coffee sites and environmental variables, management intensity, and altitude variation with a constrained redundancy analysis (RDA) on the Hellinger transformed $\mathrm{RAF}_{\mathrm{pool}}$ data using the '*rda'* function in the R package VEGAN. Five constraining variables (altitude, canopy cover, number of shade trees with DBH >20 cm, coffee density, and coffee structure index) were included in the RDA model. The collinearity between these variables, which was determined with the '*vif'* function in the R package CAR (Fox & Weisberg, 2019), was low (variance inflation factor < 3). Selection of variables that contributed to variation in genetic composition was performed using a forward selection method with Bonferroni correction.

To determine the presence of spatial structure in genetic composition and genetic diversity among sites, we created Moran's eigenvector maps (MEMs: Dray, 2020; Dray et al., 2006). MEMs are orthogonal vectors with a unit norm that maximize Moran's coefficient of spatial autocorrelation (Dray et al., 2006, 2012). Only MEMs with positive eigenvalues were considered for further selection to test for spatial autocorrelation (Borcard & Legendre, 2002; Dray et al., 2006). The significance of spatial autocorrelation in each MEM was tested by calculating Moran's I using the function '*moran.randtest'* in the R package ADESPATIAL (Dray, 2020). To select MEMs that might have structured the genetic composition of coffee stands, we performed forward selection (Dray, 2020) on the Hellinger transformed $\mathrm{RAF}_{\mathrm{pool}}$ values. Based on the selected MEMs and environmental and management variables, the variation in genetic composition was partitioned into environmental, management, and spatial variables using the '*varpart'* function in the R package VEGAN. We performed a partial RDA on Hellinger transformed $\mathrm{RAF}_{\mathrm{pool}}$ values, with the selected MEMs and environmental variables as explanatory variables and explored their significance using the '*anova.cca'* function in the R package VEGAN. We used the adjusted R^2 to evaluate the contribution of each fraction.

To assess the relationship between genetic diversity in coffee stands and environmental variables and management intensity, we fitted a linear model with the mean H_E as response variable and the environmental and management variables as explanatory variables using the base R function '*lm*'. The significance of the spatial structure of genetic diversity was tested by permutation, using the '*moran.randtest*' function in the R package ADESPATIAL.

Relationship between diseases and genetic composition and genetic diversity

To examine the relationship between the incidence of fungal diseases and genetic, environmental and management variables at the site-level, we ran separate generalized linear (mixed) effects models for each of the four fungal diseases with disease incidence as response variable. The five environmental and management variables (listed above) were included in all models as explanatory variables, as well as the first three PCA-axes from the indirect ordination to account for variation in genetic composition. In order to understand how the incidence of fungal diseases related to the genetic diversity in coffee stands, we included the mean H_E as explanatory variable along with the five environmental and management variables in a separate model. We fitted generalized linear mixed models (GLMM) with a binomial distribution and a logit link function for coffee leaf rust and coffee berry disease with the disease incidence as response variable using the '*glmer*' function in the R package LME4 (Bates et al., 2015). Generalized linear models (GLM) were fitted with a binomial distribution and a logit link function for coffee wilt disease and *Armillaria* root rot incidence using the '*glm*' function in base R. For the incidence of coffee leaf rust and coffee berry disease, which was assessed at coffee shrub level for two successive years, the parameter 'year' was included in the models as a fixed effect term to account for variation in disease incidence between different years. In addition, the parameter 'site' was included in these models as random effect.

Relationship between genetic diversity and among shrub variation in disease incidence

In order to assess if the variation in the incidence of coffee leaf rust and coffee berry disease among coffee shrubs was correlated with variation in genetic diversity, we ran a linear model with the standard deviation of the incidence of coffee leaf rust and coffee berry disease as response variables and genetic diversity (mean H_E) as explanatory variable. The season of sampling of the diseases ('year') was included in the models as a fixed effect to account for variation in disease incidence among years. Before all analyses, we evaluated the model fit using the package SJPLOT (Lüdecke, 2020). All analyses were performed in R v3.6.1 (R Core Team, 2019). For more details about our research questions, response variables and the models fitted, see Table S2.

Results

GBS summary data and validation of pool-GBS

More than 85% of the reads of every pool-GBS sample mapped onto the *C. canephora* reference genome sequence with a minimum mapping quality score of 20, indicating that a unique read mapping location could be identified. In total, 4,523 loci (mean length of 161 base pairs) with a depth of minimum 30 reads per locus were found in the 60 pools, covering 726,318 nucleotides of the 471 Mbp *C. canephora* reference genome sequence (0.15%). Of these 4,523 loci, 3,605 (79.7%) were present in at least 57 out of 60 pools. The set of 32 individual samples used for pool-GBS validation contained 4,148 loci with a read depth of minimum 30 reads, covering 653,981 nucleotides (0.14%) of the *C. canephora* reference genome sequence. The majority of those loci (3,105 out of 4,148, 74.9%) were observed in at least 31 out of the 32 individual samples, indicating a relatively low level of missing loci and fair saturation of read depth across the entire sample set. After all filtering steps, 487 SNPs in the pools and 292 in the individuals were retained (Figs. S1-5). The VEcv value of the comparisons between the RAF spectra of *within*-subgenome SNPs shared between individual-GBS and pool-GBS for the two sites was on average 90.7%, indicating a high correspondence between RAF_{ind} and RAF_{pool} and a high accuracy of the pool-GBS method. Furthermore, the VEcv values of the comparisons between pool-GBS ligation replicates at each of the sixty sites were on average 94.4 (range 88.3–97.8), showing that the pool-GBS ligation replicates displayed high reproducibility and accuracy of RAF_{pool} spectra (Table S3).

Genetic composition and genetic diversity of coffee stands along gradients in environmental, management and spatial variables

The hierarchical cluster analysis partitioned the sites into four clusters based on similarities of their genetic composition (Fig. 2), which also to a large extent is mirrored in how they occupy different areas in a two-dimensional principal component analysis (PCA) plot (Fig. 3a). However, note that the clusters do not represent genetically differentiated groups in the sense that all sites in one cluster are more similar to each other than to sites in another cluster, since the genetic composition is changing gradually (as seen by the large cloud with all sites in the PCA, Fig 3a). Rather the clusters facilitate the division of sites into groups that can be used to illustrate how the variability in genetic composition, that still is substantial, is distributed across the landscape. The sites in the less intensively managed forest in the western part of the landscape fell into one cluster (G1), while the sites dominated by smallholder's landraces in the eastern part of the landscape were grouped into another cluster (G2) (Fig. 4a). These two clusters showed some overlap in the PCA space (Fig. 3a) even though the sites were geographically separated (Fig. 4a). However, sites from the other two clusters (G3 and G4) were more spread apart in the PCA plot, and most of them were also spread out geographically across the landscape (Fig. 4a). These two clusters (G3 and G4) contain the intensively managed commercial plantation sites and some of the more intensively managed smallholder farmer sites, and tended to be more similar to the forest sites (G1) than the smallholder sites (G2) (Fig. 3a). The pairwise F_{ST} values showed similar patterns of genetic differentiation (Fig. S7). The highest and lowest F_{ST} values for the pairwise comparisons were 0.378 and 0.007, respectively (Fig. S7). A direct gradient analysis with redundancy analysis (RDA) showed that coffee structure index, canopy cover, and coffee density were related to the variation in genetic composition of the coffee stands (Fig. 3b; Table S4). The four cluster groups differed from each other in the RDA space in a similar way as in the PCA, but with less overlap between the groups (see Fig. 3a and b). The MEM global test that was used to assess spatial autocorrelation in the genetic composition was highly significant ($p < 0.001$). Environmental and management variables and spatial variables explained a nearly similar amount of variation and together, spatial and environmental variables explained 23% of the variation in genetic composition of the coffee stands (Fig. S8).

The genetic diversity (mean H_E) of coffee ranged between 0.157 and 0.253 (Table S5). The genetic diversity of the coffee stands was positively related to altitude ($F_{(1, 57)} = 12.49$, SC = 0.40, $p < 0.001$) and decreased with coffee structure index ($F_{(1, 57)} = 6.87$, SC = -0.30, $p = 0.011$) (Fig. 5, Table S6). However, we did not detect a spatial structure in the genetic diversity of coffee stands (MEM global test: observed = 0.026, $p = 0.32$; see also Fig. 4b).

The relationship between genetic composition and genetic diversity in coffee stands and incidence of fungal diseases

The incidence of each of the four fungal diseases was related to at least one of the first three PCA axes describing the genetic composition of the coffee sites (Table 1). The relationships were specific to each of the four diseases. Incidence of the fungal diseases was also related to several environmental and management variables, but the relationship varied for the different diseases (Table 1). For example, coffee leaf rust incidence decreased with altitude, whereas the incidence of coffee berry disease and *Armillaria* root rot were positively related to altitude. Coffee leaf rust incidence decreased, whereas *Armillaria* root rot incidence increased, with coffee structure index (Table 1).

Genetic diversity, as based on the pooled sample of 16 coffee shrubs per site, did not explain any of the variation in among-site incidence of any of the four fungal diseases (Table 2).

The relationship between genetic diversity and among shrub variation in disease incidence

The variation among coffee shrubs in the incidence of coffee leaf rust was not significantly related to genetic diversity in coffee sites (Fig. 6a; Table S7). However, the variation among shrubs in the incidence of coffee berry disease was positively related to genetic diversity in the coffee sites (Fig. 6b, Table S7).

Discussion

We studied to what extent genetic composition and genetic diversity of coffee stands could explain the variation in fungal diseases across a landscape where Arabica coffee is native and managed with different intensities. Genetic composition showed a clear spatial pattern, to some extent related to environmental and management variables (e.g. canopy cover, coffee structure index, coffee density), while the genetic diversity did not show a spatial pattern, but was related to for example elevation and the number of shade trees. The incidence of the four major fungal diseases on coffee was related to the genetic composition of the coffee sites, but in different ways for the different diseases. On the other hand, the incidence of the diseases was not lower in sites with a high genetic diversity even if the variability in the incidence of coffee berry disease among shrubs within sites was higher in such sites. Overall, our study illustrates the need to consider both the genetic composition and genetic diversity of the host species when investigating the genetic basis for variation in disease levels.

Pool-GBS validation

The high mapping success rate in the pool-GBS data suggests that the majority of the reads from the *C. eugenioides*-derived subgenome mapped well onto the *C. canephora* reference genome sequence, despite the relatively high genetic differentiation that was reported between the two subgenomes (Scalabrin et al., 2020). Consequently, the mapping of *C. arabica* reads onto a *C. canephora* reference genome sequence does not seem to systematically exclude reads from the *C. eugenioides*-derived subgenome, indicating that this reference genome sequence is suited for genome-wide fingerprinting studies of *C. arabica*. The number of *within*-subgenome SNPs identified in the pool-GBS data was low, yet within the expected order of magnitude based on SNP numbers reported by previous GBS studies on *C. arabica* (Sant'Ana et al., 2018; Scalabrin et al., 2020). The estimated nucleotide diversity pi in the pool-GBS samples of *C. arabica* also strongly corresponds to the estimate of Scalabrin et al. (2020), who found that this diversity was about ten times lower than the diversity in a set of samples from each progenitor species. This low amount of genetic variation can be explained by the assumed recent origin of the species after a single hybridization event (Bawin et al., 2021; Scalabrin et al., 2020).

We implemented an F_{ST}-based filter and optimized the F_{ST} stringency threshold to balance between removing most *between*-subgenome polymorphisms and retaining most *within*-subgenome SNPs. A lower F_{ST} threshold may increase the number of retained *between*-subgenome polymorphisms, thus masking patterns in the genetic variation among populations. Conversely, a higher F_{ST} threshold value may exclude too many *within*-subgenome SNPs, possibly leading to incorrect inferences of genetic relationships across populations. A careful evaluation of the filtering procedure showed that it removed the majority of the *between*-subgenome polymorphisms in the pool-GBS data, without affecting the relative differences in the genetic diversity among the sixty Arabica coffee sites.

About half of the SNPs in the individual-GBS samples of the two sites were not present in their constituent pool-GBS sample. The majority of those SNPs had alleles with a low frequency in the individual-GBS samples ($RAF_{ind} < 0.05$ or $RAF_{ind} > 0.95$), which suggests that pool-GBS is not very sensitive to the detection of low-frequency alleles. Such alleles were often missing in the pool-GBS data because they were not detected (absent in both pool-GBS ligation replicates), not reproducible (present in only one out of two pool-GBS ligation replicates), or removed during SNP filtering due to their low posterior probability value assigned by SNAPE-pooled. SNPs in a pool-Seq sample with low-frequency alleles often have a similar (low) read depth as read errors, and are therefore inevitably discarded during data cleaning (Dorant et al., 2019; Inbar et al., 2020). However, low-frequency alleles in pool-Seq data are believed to mainly provide information about recent demographic changes between closely related populations, hence they may not be required to infer broad-scale genetic patterns in the pool-GBS data of *C. arabica* (Baye et al., 2011; Génin et al., 2015). The high VEcv values of the comparison between RAF_{ind} and RAF_{pool} indicated that the RAFs in the pool-GBS data accurately reflected the RAFs in the individual-GBS data (Li, 2016). The pool-GBS approach also resulted in highly reproducible RAF_{pool} values, as shown by the high VEcv values of the comparison between the data of pool-GBS ligation replicates. Both observations are in line with previous pool-GBS studies, confirming that pool-GBS is accurate and reproducible if each individual equally contributes to the pooled DNA sample (Bélanger et al., 2016; Verwimp et al., 2018).

Variation in genetic composition and genetic diversity across the landscape

Our results show that the genetic composition of the coffee stands varied between the coffee sites across the landscape with both spatial and environmental imprints. We suggest that some of this variation in genetic composition across the landscape could be attributed to the history of coffee production in these landscapes. According to the historian McCann (1995, p. 159), active coffee cultivation in the area started in the early twentieth century in the nowadays more degraded forest patches in the eastern part of our study area (Gomma), where coffee occurs less as a wild species. Increasing demand for coffee export in the late-nineteenth century and close access to the emerging trade routes encouraged the King of Gomma to stimulate coffee cultivation in his kingdom. In contrast, the Gera king and the farmers close to the more intact forest areas in Gera relied mostly on collecting wild coffee from the large forests present in this area (McCann, 1995), which is still a common practice in some areas in the western part of our focal landscape, even if many farmers currently are introducing more intensive coffee production here. Perhaps the imprint of the early coffee cultivation by smallholder farmers in the eastern part of the landscape is the reason for the stronger genetic cluster in this area. More recently in the 1980s, intensively managed plantation coffee systems started growing different improved cultivars that are resistant to coffee berry disease in our focal landscape. Since then, some farmers have started to mix cultivars with their landraces. Some of the coffee sites in our study area were originally planted with cultivars in the 1980s, as part of a government-initiated coffee expansion plan, and then were distributed to smallholder farmers after the fall of the Derg (the socialist) regime. We suggest that the spread of cultivars across the landscape explains the wide geographical distribution of coffee sites with a genetic composition from the two cluster groups that contain more intensively managed sites (note again that the genetic composition is gradual across all sites and that the clusters are used for illustrative purposes to be able to refer to groups of sites with similar genetic composition). These cultivars were selected using germplasm from the genetic reservoirs of the forest coffee systems of southwestern Ethiopia (Labouisse et al., 2008), which could explain the tendency of the sites from the more intensively managed smallholder farmers and plantation coffee systems to be more similar to the forest type of coffee than other smallholder coffee with medium management. The intensively managed plantations were also much spread out in the PCA plot, showing that there could be more differentiation in the genetic composition among plantations than among stands from smallholders landraces and forest systems. Thus, it seems as if the variation in genetic composition across the landscape is consistent with the historic development of coffee cultivation in this region. These patterns support the notion that the genetic composition of crops varies across landscapes due to anthropogenic influences and dispersal limitation (Orsini et al., 2013; Sertse et al., 2019).

Even if we did not detect any spatial pattern of genetic diversity among the coffee stands, there was some genetic variation that was related to environmental variables. We expected a higher genetic diversity in the less intensively managed forest sites, which might be confirmed by the positive relationship between genetic diversity and altitude and the negative relationship with a management variable (i.e. coffee structure index). However, the pattern did not become evident when plotting the mean expected heterozygosity of each site on the map, with only a few intensively managed plantations towards the eastern, lower altitude areas showing relatively lower genetic diversity.

Surprisingly, both the highest (H_E = 0.247) and lowest (H_E = 0.157) amounts of genetic diversity were observed in intensively managed coffee sites, indicating that these sites could be more variable in their genetic diversity than less managed sites. While coffee management is expected to reduce the genetic diversity of coffee through allelic loss associated with genetic drift and inbreeding (Aerts et al., 2013; Frankham, 2005), this might have been compensated by farmers via the introduction of new genotypes to their plots from the forest systems or via improved cultivars (Labouisse et al., 2008).

Variation in fungal diseases in relation to genetic composition and genetic diversity

We know from our previous studies that the different diseases have different niches in relation to various environmental and management variables (Zewdie et al., 2020, 2021). Here, we show that additional variation in the incidence of the fungal diseases across the landscape was attributed to the difference in genetic composition of the coffee stands. The strong relationship between genetic composition and incidence of coffee berry disease across the sites could be attributed to the fact that varieties resistant to this disease have been widely used in the plantations as well as spread to smallholder farmers (Labouisse et al., 2008). These cultivars were mainly selections from the forest systems that were established as an immediate response to the catastrophic effects of coffee berry disease outbreak in the main coffee growing areas in the early 1970s (Van der Graaff, 1978; Van der Graaff & Pieters, 1983). Yet, there is no clear information on what proportion of the coffee sites contain cultivars that have been introduced. The three other fungal diseases (coffee leaf rust, coffee wilt disease and *Armillaria* root rot) also showed a genetic signal in how their incidences varied across the landscape, even though the relationships were weak compared to the relationship between genetic composition and coffee berry disease incidence. There are also indications of a genetic component for resistance to other diseases than coffee berry disease, e.g. coffee wilt disease (Pieters & van der Graaff, 1980; Van der Graaff & Pieters, 1978) and coffee leaf rust (Barka et al., 2020; Eskes, 1983; Hindorf & Omondi, 2011; Ribas et al., 2011; Silva et al., 2006). Studies that link disease dynamics to genetic variation in coffee are limited in this landscape. On the main crops, host genetic variation has been highlighted to have a (strong) link to the resistance response to diseases. As an example, genetic variation in tomato has been shown to alter virulence of the generalist pathogen *Botrytis cinerea* (Soltis et al., 2019); maize inbred lines with different genetic background are shown to have varying levels of resistance to *Striga hermonthica,* the severe disease of maize limiting production in the sub-Saharan Africa (Stanley et al., 2020); and genetic variation in several barely landraces is highlighted to have contributed favourable alleles for improvement of the crop for resistance of diseases such as barely yellow dwarf virus and barely powdery mildew (Muñoz-Amatriaín et al., 2014). The existence of a genetic component in the incidence of the diseases found in this study adds interesting aspects to the opportunity for breeding programs to screen for genotypes that are resistant to these diseases.

Unlike the genetic composition of coffee stands, the genetic diversity of the coffee stands did not relate to the mean incidence of any of the fungal diseases. In contrast, a negative relationship between host genetic diversity and the level of pathogen infection has been found both in other natural (Ekroth et al., 2019; Gibson & Nguyen, 2020) and crop systems such as wheat (Huang et al., 2011; Ben M'Barek et al., 2020), rice (Zhu et al., 2000) and potato (Garrett & Mundt 2000). Despite the lack of a relationship in our study, the use of mixtures of cultivars or multiline cultivars have been reported to have greater significance in the management of crop diseases (Mundt, 2002; Wolfe, 1985). For example, susceptible and resistant variety mixtures had 94% lesser rice blast caused by *Magnaporthe grisea* severity as compared to single variety (monoculture) rice (Zhu et al., 2000). Using two-cultivar mixtures with different resistance profiles, Cox et al. (2004) showed that the severity of wheat leaf rust caused by *Puccinia triticina* and tan spot caused by *Pyrenophora tritici-repentis* was lower on the susceptible cultivars when resistant and susceptible cultivars are grown in mixture compared with monoculture. Introduction of 25% disease-resistant cultivars into a pure stand of durum wheat cultivar susceptible to *Septoria tritici* blotch has resulted in 50% reduction in the severity of the disease compared to the susceptible pure stand (Ben M'Barek et al., 2020). Disease suppressions by the use of cultivar mixtures have also been reported on wheat strip rust caused by *Puccinia striipformis* (Chaulagain et al., 2017; Huang et al., 2011; Sapoukhina et al., 2013) and potato late blight caused by *Phytophthora infestans* (Andrivon et al., 2003; Garrett & Mundt, 1999; Yang et al., 2019). However, such negative relationships between host genetic diversity and disease levels could also be inconsistent among study systems specifically for observational studies (Gibson & Nguyen, 2020; Mundt, 2002). As an important caveat, in host-pathogen systems, the effect of host diversity could be cancelled out by the diversity in the pathogen population (Ganz & Ebert, 2010;

Jeger et al., 1981), an effect which we did not consider in this study. Another important information that we lack in this study, but which could have been very important, is information on the historical outbreak of the diseases. This could be possible in future studies through gathering information through extensive interviews with local smallholder farmers.

Variability within sites in fungal diseases

We expected that variation in incidence of the fungal diseases among coffee shrubs would be higher in sites with higher genetic diversity, due to the variation in resistance among the genetically variable individual coffee shrubs. Such pattern was detected for coffee berry disease, but not for coffee leaf rust. This could be explained by the introduction of cultivars resistant to coffee berry disease in some smallholder sites, which might have reduced the infection on some shrubs, but also increased the genetic variation between shrubs at site level. Yet, the pattern is not strong and might be driven by a few intensively managed coffee plantations that grow a selected resistant cultivar (i.e. low within-site genetic diversity but almost no incidence of coffee berry disease). An increase in genetic diversity in some sites, for e.g. by mixing cultivars and original plants, did not increase the within-site variation of coffee leaf rust incidence, which could be due to lack of resistance in cultivars to coffee leaf rust unlike to the coffee berry disease (Daba et al., 2019). Several mechanism might lead to reduced disease levels in mixtures of host populations, *namely*: i) occurrence of higher proportion of resistant cultivars in mixtures and reduced disease inoculums, ii) resistant cultivars could serve as barriers for the dispersal of the disease inoculums, and iii) shallow pathogen dispersal gradients (Andrivon et al., 2003; Garrett & Mundt, 1999; Mundt, 2002). Our finding on the positive relationship between within-site variation in coffee berry disease and genetic diversity is in line with the first two mechanisms. On the other hand, wind-dispersed pathogens with shallow dispersal gradients (e.g. *Hemileia vastatrix)* are suggested to result in higher effect of host-diversity on disease suppression as compared to splash-dispersed pathogens (e.g. *Colletotrichum kahawae*) (Garrett & Mundt, 1999, 2000), which is not the case for coffee leaf rust in our study. The incidence of coffee leaf rust could thus be driven more by environmental variables than by the genetic diversity of coffee *per se*. For example, the low coffee leaf rust incidence in the little managed forest sites was mainly related to elevation (Zewdie et al., 2020).

Implications for the conservation and breeding of Arabica coffee

Coffee production in the world is challenged by the outbreak of several pests and diseases (Avelino et al., 2018; Bedimo et al., 2008; Hindorf & Omondi, 2011; Jaramillo et al., 2011; McCook, 2006). To increase resistance against diseases in cultivated coffee, wild genetic variation is believed to be of utmost importance for coffee breeding (Davis et al., 2019; Scalabrin et al., 2020). The forests of southwestern Ethiopia represent the native habitats for Arabica coffee and are unique reservoirs of the genetic diversity of coffee (Labouisse et al., 2008; Tesfaye et al., 2007, 2014). The finding that the genetic composition of coffee in different sites was related to the incidence of major fungal diseases indicate that these landscapes harbor valuable genetic resources for breeding disease resistance in Arabica coffee. Nevertheless, the spread of cultivars across the landscape could be problematic in terms of introgression of genetic material from cultivars into wild *C. arabica* plants (Aerts et al., 2013). With the increasing intensification of coffee management in the region, coffee genetic resources might be even more under threat (Aerts et al., 2013; Labouisse et al., 2008). This suggests that there is an urgent need to safeguard the genetic diversity of coffee in its native state for a sustainable future of coffee production (Aerts et al., 2017). Among the four fungal diseases, coffee berry disease is the major challenge in Ethiopia, as it causes severe yield losses, forcing farmers to gradually replace their original landraces that are susceptible to the disease by coffee berry disease resistant cultivars. Smallholder farmers must obtain revenue to sustain themselves, and just advising them to refrain from using cultivars in their plots to avoid loss of genetic diversity is unlikely to work. An alternative option is to use strict protection measures on the less intensively managed natural forests that restrict the introduction of cultivars, in combination with *ex situ* conservation of the sources of coffee genetic diversity in the farmers landraces for potential use in future breeding work (Aerts et al., 2015). We hope that our study will provide a path forward for more detailed analyses, surpassing the individual and population level genetics, and moving from reduced-representation sequencing to screenings of the entire genome complement. This would enable to delve deeper into the underlying genetic mechanisms and understand which regions of the genome are linked to the variation in disease levels. Moreover, the focus of future analyses may particularly lie on the distribution of cultivars among the different sites, the degree of crop-wild introgression and its consequence for genetic variation in coffee. Such studies would be useful to inform decision makers to improve management and facilitate the use of crop genetic resources to combat future climate related disease outbreaks.

Acknowledgements

This work was supported by the Swedish Research Council (Grant no. VR2015-03600 to KH) and by Research Foundation – Flanders (FWO) (Grant no. G056517N to OH, IRR, and TR). We thank the coffee stand owners for allowing us to work in their coffee plots, the local and regional administrations for providing necessary permits and Tijani Ibrahim and Raya A/Oli for their assistance during data collection. We are also grateful to the Ethiopian institute of Biodiversity for the permit to export coffee leaf material for the genetic analysis.

References

Aerts, R., Berecha, G., Gijbels, P., Hundera, K., Van Glabeke, S., Vandepitte, K., Muys, B., Roldán-Ruiz, I., & Honnay, O. (2013). Genetic variation and risks of introgression in the wild *Coffea arabica* gene pool in south-western Ethiopian montane rainforests. *Evolutionary Applications*, *6*(2), 243–252.

Aerts, R., Berecha, G., & Honnay, O. (2015). Protecting coffee from intensification. *Science*, *347*(6218), 139.

Aerts, R., Geeraert, L., Berecha, G., Hundera, K., Muys, B., De Kort, H., & Honnay, O. (2017). Conserving wild Arabica coffee: Emerging threats and opportunities. *Agriculture, Ecosystems & Environment*, *237*, 75–79.

Aerts, R., Hundera, K., Berecha, G., Gijbels, P., Baeten, M., Van Mechelen, M., Hermy, M., Muys, B., & Honnay, O. (2011). Semi-forest coffee cultivation and the conservation of Ethiopian Afromontane rainforest fragments. *Forest Ecology and Management*, *261*(6), 1034–1041. https://doi.org/10.1016/j.foreco.2010.12.025

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*(2), 81–92.

Andrivon, D., Lucas, J., & Ellisseche, D. (2003). Development of natural late blight epidemics in pure and mixed plots of potato cultivars with different levels of partial resistance. *Plant Pathology*, *52*(5), 586–594.

Avelino, J., Allinne, C., Cerda, R., Willocquet, L., & Savary, S. (2018). Multiple-disease system in coffee: From crop loss assessment to sustainable management. *Annual Review of Phytopathology*, *56*, 611–635.

Barka, G. D., Caixeta, E. T., Ferreira, S. S., & Zambolim, L. (2020). In silico guided structural and functional analysis of genes with potential involvement in resistance to coffee leaf rust: A functional marker based approach. *PLoS ONE*, *15*(7), e0222747.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bawin, Y., Ruttink, T., Staelens, A., Haegeman, A., Stoffelen, P., Mwanga Mwanga, J. I., Roldán-Ruiz, I., Honnay, O., & Janssens, S. B. (2021). Phylogenomic analysis clarifies the evolutionary origin of *Coffea arabica*. *Journal of Systematics and Evolution*, *59*(5), 953–963.

Baye, T. M., He, H., Ding, L., Kurowski, B. G., Zhang, X., & Martin, L. J. (2011). Population structure analysis using rare and common functional variants. *BMC Proc*. 2011; 5(Suppl 9): S8. https://doi.org/10.1186/1753-6561-5-S9-S8

Bedimo, J. A. M., Njiayouom, I., Bieysse, D., Nkeng, M. N., Cilas, C., & Notteghem, J.-L. (2008). Effect of shade on Arabica coffee berry disease development: Toward an agroforestry system to reduce disease impact. *Phytopathology*, *98*(12), 1320–1325.

Bélanger, S., Esteves, P., Clermont, I., Jean, M., & Belzile, F. (2016). Genotyping-by-sequencing on pooled samples and its use in measuring segregation bias during the course of androgenesis in barley. *The Plant Genome*, *9*(1), plantgenome2014-10.

Ben M'Barek, S., Karisto, P., Abdedayem, W., Laribi, M., Fakhfakh, M., Kouki, H., Mikaberidze, A., & Yahyaoui, A. (2020). Improved control of *Septoria tritici* blotch in durum wheat using cultivar mixtures. *Plant Pathology*, *69*(9), 1655–1665.

Berecha, G., Aerts, R., Vandepitte, K., Van Glabeke, S., Muys, B., Roldán-Ruiz, I., & Honnay, O. (2014). Effects of forest management on mating patterns, pollen flow and intergenerational transfer of genetic diversity in wild Arabica coffee (*Coffea arabica* L.) from Afromontane rainforests. *Biological Journal of the Linnean Society*, *112*(1), 76–88.

Borcard, D., & Legendre, P. (2002). All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling*, *153*(1), 51–68. https://doi.org/10.1016/S0304-3800(01)00501-4

Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). clValid: An R package for cluster validation. *Journal of Statistical Software*, *25*(1), 1–22. https://doi.org/10.18637/jss.v025.i04

Brozynska, M., Furtado, A., & Henry, R. J. (2016). Genomics of crop wild relatives: Expanding the gene pool for crop improvement. *Plant Biotechnology Journal*, *14*(4), 1070–1085. https://doi.org/10.1111/pbi.12454

Burdon, J. J., & Laine, A.-L. (2019). *Evolutionary Dynamics of Plant Pathogen Interactions*. Cambridge University Press.

Byrne, S., Czaban, A., Studer, B., Panitz, F., Bendixen, C., & Asp, T. (2013). Genome wide allele frequency fingerprints (GWAFFs) of populations via genotyping by sequencing. *PLoS ONE*, *8*(3), e57438.

Chaulagain, B., Chhetri, G. B. K., Shrestha, S. M., Sharma, S., Sharma-Poudyal, D., & Lamichhane, J. R. (2017). Effect of two-component cultivar mixtures on development of wheat yellow rust disease in the field and greenhouse in the Nepal Himalayas. *Journal of General Plant Pathology*, *83*(3), 131–139.

Colque-Little, C., Abondano, M. C., Lund, O. S., Amby, D. B., Piepho, H.-P., Andreasen, C., Schmöckel, S., & Schmid, K. (2021). Genetic variation for tolerance to the downy mildew pathogen *Peronospora variabilis* in genetic resources of quinoa (*Chenopodium quinoa*). *BMC Plant Biology*, *21*(1), 1–19.

Cox, C., Garrett, K., Bowden, R., Fritz, A., Dendy, S., & Heer, W. (2004). Cultivar mixtures for the simultaneous management of multiple diseases: Tan spot and leaf rust of wheat. *Phytopathology*, *94*(9), 961–969.

Daba, G., Helsen, K., Berecha, G., Lievens, B., Debela, A., & Honnay, O. (2019). Seasonal and altitudinal differences in coffee leaf rust epidemics on coffee berry disease-resistant varieties in Southwest Ethiopia. *Tropical Plant Pathology, 44*, 1–7. doi: 10.1007/s40858-018-0271-8

Davis, A. P., Chadburn, H., Moat, J., O'Sullivan, R., Hargreaves, S., & Lughadha, E. N. (2019). High extinction risk for wild coffee species and implications for coffee sector sustainability. *Science Advances*, *5*(1). https://doi.org/10.1126/sciadv.aav3473

Davis, A. P., Gole, T. W., Baena, S., & Moat, J. (2012). The impact of climate change on indigenous arabica coffee (*Coffea arabica*): Predicting future trends and identifying priorities. *PLOS ONE*, *7*(11), e47981.

Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., Zheng, C., Alberti, A., Anthony, F., Aprea, G., Aury, J.-M., Bento, P., Bernard, M., Bocs, S., Campa, C., Cenci, A., Combes, M.-C., Crouzillat, D., Silva, C. D., … Lashermes, P. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science*, *345*(6201), 1181–1184. https://doi.org/10.1126/science.1255274

Dita, M., Barquero, M., Heck, D., Mizubuti, E. S. G., & Staver, C. P. (2018). Fusarium wilt of banana: Current knowledge on epidemiology and research needs toward sustainable disease management. *Frontiers in Plant Science*, *9*, 1468. https://doi.org/10.3389/fpls.2018.01468

Dorant, Y., Benestan, L., Rougemont, Q., Normandeau, E., Boyle, B., Rochette, R., & Bernatchez, L. (2019). Comparing Pool-seq, Rapture, and GBS genotyping for inferring weak population structure: The American lobster (*Homarus americanus*) as a case study. *Ecology and Evolution*, *9*(11), 6606–6623. https://doi.org/10.1002/ece3.5240

Dray, S. (2020). *Moran's Eigenvector Maps and related methods for the spatial multiscale analysis of ecological data*. https://cran.r-project.org/web/packages/adespatial/vignettes/tutorial.html

Dray, S., Legendre, P., & Peres-Neto, P. R. (2006). Spatial modelling: A comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological Modelling*, *196*(3), 483–493. https://doi.org/10.1016/j.ecolmodel.2006.02.015

Dray, S., Pélissier, R., Couteron, P., Fortin, M.-J., Legendre, P., Peres-Neto, P. R., Bellier, E., Bivand, R., Blanchet, F. G., Cáceres, M. D., Dufour, A.-B., Heegaard, E., Jombart, T., Munoz, F., Oksanen, J., Thioulouse, J., & Wagner, H. H. (2012). Community ecology in the age of multivariate multiscale spatial analysis. *Ecological Monographs*, *82*(3), 257–275. https://doi.org/10.1890/11-1183.1

Ekroth, A. K. E., Rafaluk-Mohr, C., & King, K. C. (2019). Host genetic diversity limits parasite success beyond agricultural systems: A meta-analysis. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1911), 20191811. https://doi.org/10.1098/rspb.2019.1811

Elshire, R. J., Glaubitz, J. C., Sun, Q., Pol, J. A., Kawamoto, K., & Buckler, E. S. (2011). A Robust, Simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, *6*(5), e19379. doi: 10.1371/journal.pone.0019379

Eskes, A. B. (1983). Incomplete resistance to coffee leaf rust. In F. Lamberti, J. M. Waller, & N. A. Van der Graaff (Eds.), *Durable Resistance in Crops* (pp. 291–315). Springer New York. https://doi.org/10.1007/978-1-4615-9305-8_26

Ferretti, L., Ramos-Onsins, S. E., & Pérez-Enciso, M. (2013). Population genomics from pool sequencing. *Molecular Ecology*, *22*(22), 5561–5576. https://doi.org/10.1111/mec.12522

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression*. https://cran.r-project.org/web/packages/car/citation.html

Frankham, R. (2005). Genetics and extinction. *Biological Conservation*, *126*(2), 131–140. https://doi.org/10.1016/j.biocon.2005.05.002

Fu, Y.-B. (2015). Understanding crop genetic diversity under modern plant breeding. *Theoretical and Applied Genetics*, *128*(11), 2131–2142. https://doi.org/10.1007/s00122-015-2585-y

Ganz, H. H., & Ebert, D. (2010). Benefits of host genetic diversity for resistance to infection depend on parasite diversity. *Ecology*, *91*(5), 1263–1268.

Garrett, K., & Mundt, C. (1999). Epidemiology in mixed host populations. *Phytopathology*, *89*(11), 984–990.

Garrett, K., & Mundt, C. (2000). Host diversity can reduce potato late blight severity for focal and general patterns of primary inoculum. *Phytopathology*, *90*(12), 1307–1312.

Geeraert, L., Hulsmans, E., Helsen, K., Berecha, G., Aerts, R., & Honnay, O. (2019). Rapid diversity and structure degradation over time through continued coffee cultivation in remnant Ethiopian Afromontane forests. *Biological Conservation*, *236*, 8–16. https://doi.org/10.1016/j.biocon.2019.05.014

Génin, E., Letort, S., & Babron, M.-C. (2015). Population Stratification of Rare Variants. In E. Zeggini & A. Morris (Eds.), *Assessing Rare Variation in Complex Traits* (pp. 227–237). Springer New York. https://doi.org/10.1007/978-1-4939-2824-8_16

Gezahgne, A., Coetzee, M. P. A., Wingfield, B. D., Wingfield, M. J., & Roux, J. (2004). Identification of the *Armillaria* root rot pathogen in Ethiopian plantations. *Forest Pathology*, *34*(3), 133–145.

Gibson, A. K., & Nguyen, A. E. (2020). Does genetic diversity protect host populations from parasites? A meta-analysis across natural and agricultural systems. *Evolution Letters*, *5*(1), 16–32.

Girma, A., Hulluka, M., & Hindorf, H. (2001). Incidence of tracheomycosis, *Gibberella xylarioides* (*Fusarium xylarioides*), on Arabica coffee in Ethiopia. *Journal of Plant Diseases and Protection*, 136–142.

Girma, A., Million, A., Hindorf, H., Arega, Z., Teferi, D., & Jefuka, C. (2009). Coffee wilt disease in Ethiopia. *Coffee Wilt Disease*, 50–68.

Gole, T. W., Borsch, T., Denich, M., & Teketay, D. (2008). Floristic composition and environmental factors characterizing coffee forests in southwest Ethiopia. *Forest Ecology and Management*, *255*(7), 2138–2150.

Hein, L., & Gatzweiler, F. (2006). The economic value of coffee (*Coffea arabica)* genetic resources. *Ecological Economics*, *60*(1), 176–185. https://doi.org/10.1016/j.ecolecon.2005.11.022

Hindorf, H., & Omondi, C. O. (2011). A review of three major fungal diseases of *Coffea arabica* L. in the rainforests of Ethiopia and progress in breeding for resistance in Kenya. *Journal of Advanced Research*, *2*(2), 109–120.

Huang, C., Sun, Z., Wang, H., Luo, Y., & Ma, Z. (2011). Spatiotemporal effects of cultivar mixtures on wheat stripe rust epidemics. *European Journal of Plant Pathology*, *131*(3), 483–496.

ICO. (2020). *Country Data on the Global Coffee Trade.* http://www.ico.org/profiles_e.asp

Inbar, S., Cohen, P., Yahav, T., & Privman, E. (2020). Comparative study of population genomic approaches for mapping colony-level traits. *PLoS Computational Biology*, *16*(3), e1007653.

Jaramillo, J., Muchugu, E., Vega, F. E., Davis, A., Borgemeister, C., & Chabi-Olaye, A. (2011). Some like it hot: The influence and implications of climate change on coffee berry borer

(*Hypothenemus hampei*) and coffee production in East Africa. *PLoS ONE*, *6*(9), e24528. https://doi.org/10.1371/journal.pone.0024528

Jeger, M. J., Griffiths, E., & Jones, D. G. (1981). Disease progress of non-specialised fungal pathogens in intraspecific mixed stands of cereal cultivars. I. Models. *Annals of Applied Biology*, *98*(2), 187–198.

Jump, A. S., Marchant, R., & Peñuelas, J. (2009). Environmental change and the option value of genetic diversity. *Trends in Plant Science*, *14*(1), 51–58. https://doi.org/10.1016/j.tplants.2008.10.002

Kahle, D., & Wickham, H. (2013). ggmap: Spatial visualization with ggplot2. *The R Journal,* 5(1), 144–161.

Labouisse, J.-P., Bellachew, B., Kotecha, S., & Bertrand, B. (2008). Current status of coffee (*Coffea arabica* L.) genetic resources in Ethiopia: Implications for conservation. *Genetic Resources and Crop Evolution*, *55*(7), 1079. https://doi.org/10.1007/s10722-008-9361-7

Lemessa, D., Hylander, K., & Hambäck, P. (2013). Composition of crops and land-use types in relation to crop raiding pattern at different distances from forests. *Agriculture, Ecosystems & Environment*, *167*, 71–78. https://doi.org/10.1016/j.agee.2012.12.014

Li, J. (2016). Assessing spatial predictive models in the environmental sciences: Accuracy measures, data variation and variance explained. *Environmental Modelling & Software*, *80*, 1–8. https://doi.org/10.1016/j.envsoft.2016.02.004

Li, J. (2017). Assessing the accuracy of predictive models for numerical data: Not r nor r2, why not? Then what? *PLOS ONE*, *12*(8), e0183250. https://doi.org/10.1371/journal.pone.0183250

Limborg, M. T., Seeb, L. W., & Seeb, J. E. (2016). Sorting duplicated loci disentangles complexities of polyploid genomes masked by genotyping by sequencing. *Molecular Ecology*, *25*(10), 2117–2129.

Lüdecke, D. (2020). *sjPlot: Data Visualization for Statistics in Social Science. R package version 2.8.6*. https://CRAN.R-project.org/package=sjPlot

Luikart, G., & Cornuet, J.-M. (1998). Empirical evaluation of a test for identifying recently bottlenecked populations from allele frequency data. *Conservation Biology*, *12*(1), 228–237. https://doi.org/10.1111/j.1523-1739.1998.96388.x

McCann, J. C. (1995). *People of the plow: An agricultural history of Ethiopia, 1800–1990*. University of Wisconsin Press.

McCook, S. (2006). Global rust belt: *Hemileia vastatrix* and the ecological integration of world coffee production since 1850. *Journal of Global History*, *1*(2), 177–195.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Moat, J., Gole, T. W., & Davis, A. P. (2019). Least concern to endangered: Applying climate change projections profoundly influences the extinction risk assessment for wild Arabica coffee. *Global Change Biology*, *25*(2), 390–403. https://doi.org/10.1111/gcb.14341

Moat, J., Williams, J., Baena, S., Wilkinson, T., Gole, T. W., Challa, Z. K., Demissew, S., & Davis, A. P. (2017). Resilience potential of the Ethiopian coffee sector under climate change. *Nature Plants*, *3*(7), 1–14. https://doi.org/10.1038/nplants.2017.81

Mundt, C. C. (2002). Use of multiline cultivars and cultivar mixtures for disease management. *Annual Review of Phytopathology*, *40*, 381–410. https://doi.org/10.1146/annurev.phyto.40.011402.113723

Muñoz-Amatriaín, M., Cuesta-Marcos, A., Hayes, P. M., & Muehlbauer, G. J. (2014). Barley genetic variation: Implications for crop improvement. *Briefings in Functional Genomics*, *13*(4), 341–350.

Nei, M., & Chesser, R. K. (1983). Estimation of fixation indices and gene diversities. *Annals of Human Genetics*, *47*(3), 253–259. https://doi.org/10.1111/j.1469-1809.1983.tb00993.x

Nei, M., Maruyama, T., & Chakraborty, R. (1975). The Bottleneck effect and genetic variability in populations. *Evolution*, *29*(1), 1–10. https://doi.org/10.1111/j.1558-5646.1975.tb00807.x

Oksanen, J., Blanchet, F. G., Michael, F., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, L. G., Solymos, P., Stevens, M. H., Szoecs, E., & Wagner, H. (2019). *vegan: Community Ecology Package version 2.5-6 from CRAN*. https://rdrr.io/cran/vegan/

Orsini, L., Vanoverbeke, J., Swillen, I., Mergeay, J., & De Meester, L. (2013). Drivers of population genetic differentiation in the wild: Isolation by dispersal limitation, isolation by adaptation and isolation by colonization. *Molecular Ecology*, *22*(24), 5983–5999.

Pembleton, L. W., Cogan, N. O., & Forster, J. W. (2013). St AMPP: An R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Molecular Ecology Resources*, *13*(5), 946–952.

Pieters, R., & van der Graaff, N. A. (1980). Resistance to *Gibberella xylarioides* in *Coffea arabica*: Evaluation of screening methods and evidence for the horizontal nature of the resistance. *Netherlands Journal of Plant Pathology*, *86*(1), 37–43. https://doi.org/10.1007/BF02650392

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

R Core Team. (2019). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. http://www.r-project.org/

Raineri, E., Ferretti, L., Esteve-Codina, A., Nevado, B., Heath, S., & Pérez-Enciso, M. (2012). SNP calling by sequencing pooled samples. *BMC Bioinformatics*, *13*(1), 1–8.

Reiss, E. R., & Drinkwater, L. E. (2018). Cultivar mixtures: A meta-analysis of the effect of intraspecific diversity on crop yield. *Ecological Applications*, *28*(1), 62–77. https://doi.org/10.1002/eap.1629

Ribas, A. F., Cenci, A., Combes, M.-C., Etienne, H., & Lashermes, P. (2011). Organization and molecular evolution of a disease-resistance gene cluster in coffee trees. *BMC Genomics*, *12*(1), 240.

Rodenburg, J., Cissoko, M., Kayongo, N., Dieng, I., Bisikwa, J., Irakiza, R., Masoka, I., Midega, C. A., & Scholes, J. D. (2017). Genetic variation and host–parasite specificity of *Striga* resistance and tolerance in rice: The need for predictive breeding. *New Phytologist*, *214*(3), 1267–1280.

Sant'Ana, G. C., Pereira, L. F. P., Pot, D., Ivamoto, S. T., Domingues, D. S., Ferreira, R. V., Pagiatto, N. F., da Silva, B. S. R., Nogueira, L. M., Kitzberger, C. S. G., Scholz, M. B. S., de Oliveira, F. F., Sera, G. H., Padilha, L., Labouisse, J.-P., Guyot, R., Charmetant, P., & Leroy, T. (2018). Genome-wide association study reveals candidate genes influencing lipids and diterpenes contents in *Coffea arabica* L. *Scientific Reports*, *8*(1), 465. https://doi.org/10.1038/s41598-017-18800-1

Sapoukhina, N., Paillard, S., Dedryver, F., & de Vallavieille-Pope, C. (2013). Quantitative plant resistance in cultivar mixtures: Wheat yellow rust as a modeling case study. *New Phytologist*, *200*(3), 888–897.

Scalabrin, S., Toniutti, L., Di Gaspero, G., Scaglione, D., Magris, G., Vidotto, M., Pinosio, S., Cattonaro, F., Magni, F., Jurman, I., Cerutti, M., Suggi Liverani, F., Navarini, L., Del Terra, L., Pellegrino, G., Ruosi, M. R., Vitulo, N., Valle, G., Pallavicini, A., … Bertrand, B. (2020). A single polyploidization event at the origin of the tetraploid genome of *Coffea arabica* is responsible for the extremely low genetic variation in wild and cultivated germplasm. *Scientific Reports*, *10*(1), 4642. https://doi.org/10.1038/s41598-020-61216-7

Schmitt, C. B., Senbeta, F., Denich, M., Preisinger, H., & Boehmer, H. J. (2010). Wild coffee management and plant diversity in the montane rainforest of southwestern Ethiopia. *African Journal of Ecology*, *48*(1), 78–86.

Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, *9*, 671. http://dx.doi.org/10.1038/nmeth.2089 http://10.0.4.14/nmeth.2089

Sertse, D., You, F. M., Ravichandran, S., & Cloutier, S. (2019). The genetic structure of flax illustrates environmental and anthropogenic selections that gave rise to its eco-geographical adaptation. *Molecular Phylogenetics and Evolution*, *137*, 22–32.

Setotaw, T. A., Caixeta, E. T., Pereira, A. A., Oliveira, A. C. B. de, Cruz, C. D., Zambolim, E. M., Zambolim, L., & Sakiyama, N. S. (2013). Coefficient of parentage in *Coffea arabica* L. cultivars grown in Brazil. *Crop Science*, *53*(4), 1237–1247. https://doi.org/10.2135/cropsci2012.09.0541

Shumi, G., Rodrigues, P., Schultner, J., Dorresteijn, I., Hanspach, J., Hylander, K., Senbeta, F., & Fischer, J. (2019). Conservation value of moist evergreen Afromontane forest sites with different management and history in southwestern Ethiopia. *Biological Conservation*, *232*, 117–126. https://doi.org/10.1016/j.biocon.2019.02.008

Silva, M. do C., Várzea, V., Guerra-Guimarães, L., Azinheira, H. G., Fernandez, D., Petitot, A.-S., Bertrand, B., Lashermes, P., & Nicole, M. (2006). Coffee resistance to the main diseases: Leaf rust and coffee berry disease. *Brazilian Journal of Plant Physiology*, *18*(1), 119–147.

Silvestrini, M., Junqueira, M. G., Favarin, A. C., Guerreiro-Filho, O., Maluf, M. P., Silvarolla, M. B., & Colombo, C. A. (2007). Genetic diversity and structure of Ethiopian, Yemen and Brazilian *Coffea arabica* L. accessions using microsatellites markers. *Genetic Resources and Crop Evolution*, *54*(6), 1367–1379. https://doi.org/10.1007/s10722-006-9122-4

Soltis, N. E., Atwell, S., Shi, G., Fordyce, R., Gwinner, R., Gao, D., Shafi, A., & Kliebenstein, D. J. (2019). Interactions of tomato and *Botrytis cinerea* genetic diversity: Parsing the contributions of host differentiation, domestication, and pathogen variation. *The Plant Cell*, *31*(2), 502–519.

Stanley, A., Menkir, A., Paterne, A., Ifie, B., Tongoona, P., Unachukwu, N., Meseka, S., Mengesha, W., & Gedil, M. (2020). Genetic diversity and population structure of maize inbred lines with varying levels of resistance to *Striga hermonthica* using agronomic trait-based and SNP markers. *Plants*, *9*(9), 1223.

Steiger, D., Nagai, C., Moore, P., Morden, C., Osgood, R., & Ming, R. (2002). AFLP analysis of genetic diversity within and among *Coffea arabica* cultivars. *Theoretical and Applied Genetics*, *105*(2), 209–215. https://doi.org/10.1007/s00122-002-0939-8

Talhinhas, P., Batista, D., Diniz, I., Vieira, A., Silva, D. N., Loureiro, A., Tavares, S., Pereira, A. P., Azinheira, H. G., Guerra-Guimarães, L., Várzea, V., & Silva, M. do C. (2017). The coffee leaf rust pathogen *Hemileia vastatrix*: One and a half centuries around the tropics. *Molecular Plant Pathology*, *18*(8), 1039–1051. https://doi.org/10.1111/mpp.12512

Tesfaye, K., Borsch, T., Govers, K., & Bekele, E. (2007). Characterization of *Coffea* chloroplast microsatellites and evidence for the recent divergence of *C. arabica* and *C. eugenioides* chloroplast genomes. *Genome*, *50*(12), 1112–1129.

Tesfaye, K., Govers, K., Bekele, E., & Borsch, T. (2014). ISSR fingerprinting of *Coffea arabica* throughout Ethiopia reveals high variability in wild populations and distinguishes them from landraces. *Plant Systematics and Evolution*, *300*(5), 881–897.

Toniutti, L., Breitler, J.-C., Etienne, H., Campa, C., Doulbeau, S., Urban, L., Lambot, C., Pinilla, J.-C. H., & Bertrand, B. (2017). Influence of environmental conditions and genetic background of

Arabica coffee (*C. arabica* L) on leaf rust (*Hemileia vastatrix)* pathogenesis. *Frontiers in Plant Science*, *8*. https://doi.org/10.3389/fpls.2017.02025

Van der Graaff, N. A. (1978). Selection for resistance to coffee berry disease in Arabica coffee in Ethiopia. Evaluation of selection methods. *Netherlands Journal of Plant Pathology*, *84*(6), 205–215.

Van der Graaff, N. A., & Pieters, R. (1978). Resistance levels in *Coffea arabica* to *Gibberella xylarioides* and distribution pattern of the disease. *Netherlands Journal of Plant Pathology*, *84*(4), 117–120.

Van der Graaff, N. A., & Pieters, R. (1983). Resistance to coffee berry disease in Ethiopia. In *Durable Resistance in Crops* (pp. 317–334). Springer.

Verwimp, C., Ruttink, T., Muylle, H., Glabeke, S. V., Cnops, G., Quataert, P., Honnay, O., & Roldán-Ruiz, I. (2018). Temporal changes in genetic diversity and forage yield of perennial ryegrass in monoculture and in combination with red clover in swards. *PLoS ONE*, *13*(11), e0206571. https://doi.org/10.1371/journal.pone.0206571

Waller, J. M., Bigger, M., & Hillocks, R. J. (2007). *Coffee pests, diseases and their management*. CABI.

Waller, J. M., Bridge, P. D., Black, R., & Hakiza, G. (1993). Characterization of the coffee berry disease pathogen, *Colletotrichum kahawae* sp. Nov. *Mycological Research*, *97*(8), 989–994.

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 1358–1370.

Wolfe, Ms. (1985). The current status and prospects of multiline cultivars and variety mixtures for disease resistance. *Annual Review of Phytopathology*, *23*(1), 251–273.

Yang, L.-N., Pan, Z.-C., Zhu, W., Wu, E.-J., He, D.-C., Yuan, X., Qin, Y.-Y., Wang, Y., Chen, R.-S., & Thrall, P. H. (2019). Enhanced agricultural sustainability through within-species diversification. *Nature Sustainability*, *2*(1), 46–52.

Zewdie, B., Tack, A. J. M., Adugna, G., Nemomissa, S., & Hylander, K. (2020). Patterns and drivers of fungal disease community on Arabica coffee along a management gradient. *Basic and Applied Ecology, 47*, 95-106. https://doi.org/10.1016/j.baae.2020.05.002

Zewdie, B., Tack, A. J. M., Ayalew, B., Adugna, G., Nemomissa, S., & Hylander, K. (2021). Temporal dynamics and biocontrol potential of a hyperparasite on coffee leaf rust across a landscape in Arabica coffee's native range. *Agriculture, Ecosystems & Environment*, *311*, 107297. https://doi.org/10.1016/j.agee.2021.107297

Zhou, X., Carter, T. E., Cui, Z., Miyazaki, S., & Burton, J. W. (2002). Genetic diversity patterns in Japanese soybean cultivars based on coefficient of parentage. *Crop Science*, *42*(4), 1331–1342. https://doi.org/10.2135/cropsci2002.1331

Zhu, Y., Chen, H., Fan, J., Wang, Y., Li, Y., Chen, J., Fan, J., Yang, S., Hu, L., Leung, H., Mew, T. W., Teng, P. S., Wang, Z., & Mundt, C. C. (2000). Genetic diversity and disease control in rice. *Nature*, *406*(6797), 718–722. https://doi.org/10.1038/35021046

Data accessibility

Data available via the Dryad Digital Repository https://doi.org/10.5061/dryad.70rxwdc0n. FASTQ read files of all GBS libraries were deposited at the NCBI sequence read archive (SRA) in BioProject PRJNA783522.

Author's contributions

BZ, AJMT, KH, YB, TR, SJ, SN, KT, OH, IRR conceived the idea. BZ assessed the fungal diseases, environmental variables and collected coffee leaf samples for genetic analysis. YB, BZ, SVG conducted the DNA extraction and GBS library preparation, prepared samples for sequencing. YB, SVG, IRR and TR performed bioinformatics for genotype calling. BZ, AJMT, KH and YB analysed the data. BZ and YB lead the writing of the manuscript, and all authors contributed to the development of the manuscript.

Table 1. Incidence of each of the four fungal diseases on coffee as a function of genetic composition (PCA scores of the reference allele frequencies for the 487 *within*-subgenome SNPs), environmental and management variables, and sampling year*. Shown are standardized coefficients (SC) and chi-squared values (χ^2) for the minimum adequate model ($p < 0.05$) as estimated from the GL(M)Ms.

Response variable	Estimate	Year*	PC1	PC2	PC3	Altitude	Canopy cover	Shade trees >20 cm DBH	Coffee density	Coffee structure index
Coffee leaf rust incidence (GLMM)	SC	1.43	-	-0.26	-	-0.38	-	-	-	-0.27
	χ^2	10077	-	7.63	-	17.16	-	-	-	8.27
	p	<0.001	-	0.006	-	<0.001	-	-	-	0.004
Coffee berry disease incidence (GLMM)	SC	0.85	-1.20	0.82	-	2.11	-	-	-	-
	χ^2	1571	11.02	7.18	-	42.52	-	-	-	-
	p	<0.001	<0.001	0.007	-	<0.001	-	-	-	-
Coffee wilt disease incidence (GLM)	SC		0.18	-0.30	0.13	-	0.24	0.26	-0.16	0.11
	χ^2		9.83	49.11	10.96	-	25.94	28.96	18.74	5.90
	p		0.002	<0.001	<0.001	-	<0.001	<0.001	<0.001	0.015
Armillaria root rot incidence (GLM)	SC		-	-0.16	-0.36	0.51	-0.20	0.70	-	0.95
	χ^2		-	3.90	20.83	35.09	4.62	74.73	-	183.84
	p		-	0.048	<0.001	<0.001	0.032	<0.001	-	<0.001

GLMM – generalized linear mixed model; GLM – generalized linear model.

* Sampling year was added to coffee leaf rust and coffee berry disease models for which we used data from two seasons.

Table 2. Incidence of each of the four fungal diseases on coffee as a function of genetic diversity (mean expected heterozygosity), environmental and management variables, and sampling year*. Shown are standardized coefficients (SC) and chi-squared values (χ^2) for the minimum adequate model ($p < 0.05$) as estimated from the GL(M)Ms.

Response variable	Estimate	Year*	Genetic diversity (mean H_E)	Altitude	Canopy cover	Shade trees >20 cm DBH	Coffee density	Coffee structure index
Coffee leaf rust incidence (GLMM)	SC	1.43	-	-0.36	-	-	-	-0.21
	χ^2	10076	-	14.16	-	-	-	4.82
	p	<0.001	-	<0.001	-	-	-	0.028
Coffee berry disease incidence (GLMM)	SC	0.85	-	2.03	-	-	-	-0.82
	χ^2	1570	-	36.33	-	-	-	6.74
	p	<0.001	-	<0.001	-	-	-	0.009
Coffee wilt disease incidence (GLM)	SC		-	-	0.12	0.34	-0.25	-
	χ^2		-	-	8.73	67.21	61.57	-
	p		-	-	0.003	<0.001	<0.001	-
Armillaria root rot incidence (GLM)	SC		-	0.43	-0.35	0.80	-0.33	1.00
	χ^2		-	27.57	15.22	94.73	17.53	211.44
	p		-	<0.001	<0.001	<0.001	<0.001	<0.001

GLMM – generalized linear mixed model; GLM – generalized linear model

* Sampling year was added to coffee leaf rust and coffee berry disease models for which we used data from two seasons.

List of Figures

Fig. 1. Study area and plot design. Panel (a) shows a map of Ethiopia with the study area marked with a red polygon. Panel (b) shows an aerial view of the study landscape with the sixty study sites (white dots) in Gomma and Gera districts (black stars). The sites are overlaid on Google Maps (Map data ©2021 Google) using the '*geocode*' function in GGMAP library in R. Panel (c) shows layout of an individual plot, each 50 × 50 m. The numbers 1–16 indicate the sixteen coffee shrubs selected at the intersections of 10 m gridlines in the central 30 × 30 m of the plot. Shade canopy pictures were taken from five locations at the center of the 10 × 10 m quadrant, as indicated by the blue dots.

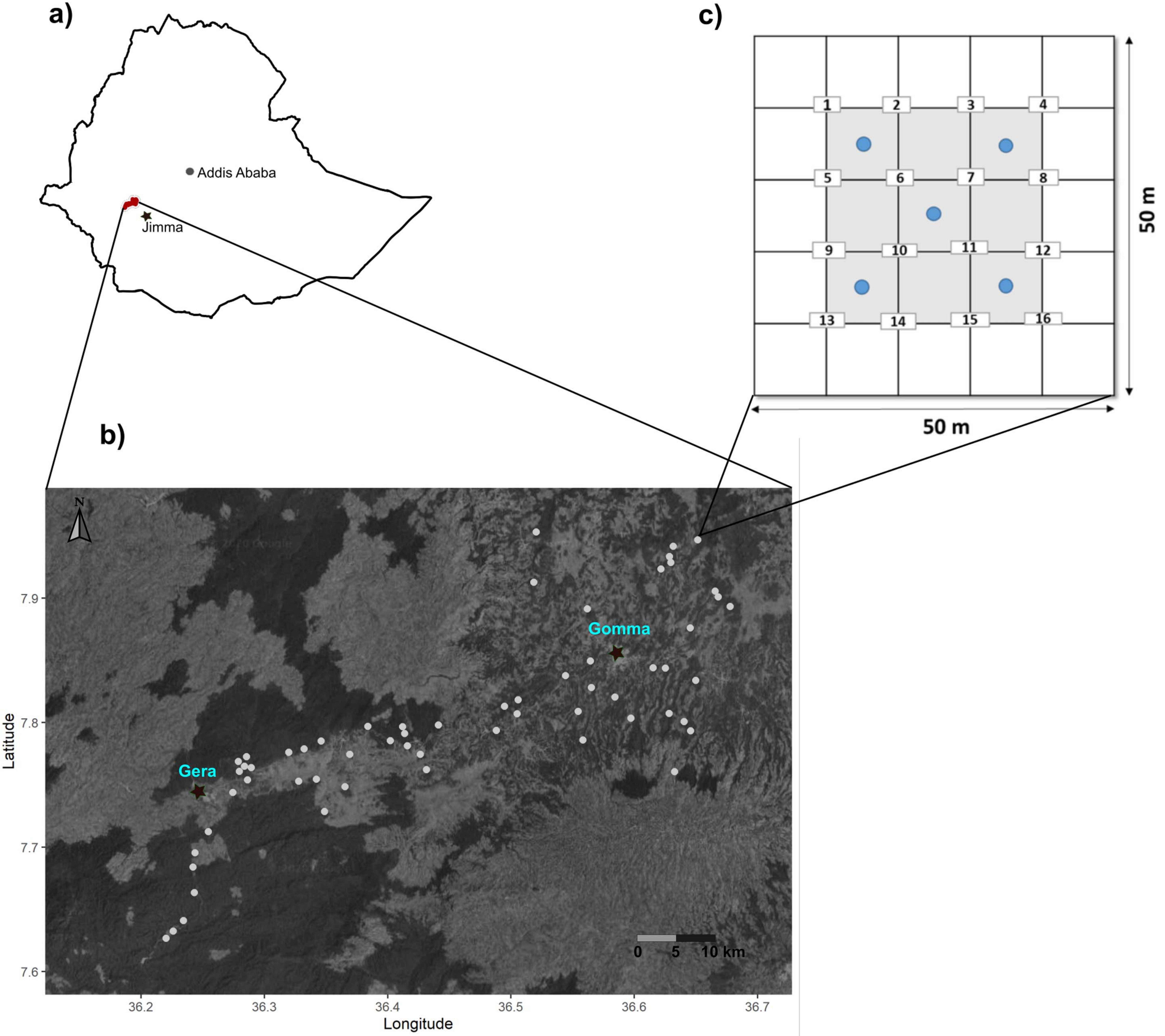
Fig. 2. Dendrogram based on hierarchical clustering of the pool-GBS reference allele frequency ($\mathrm{RAF}_{\mathrm{pool}}$) data across sixty Arabica coffee sites. Cluster results validation using the *'clValid'* function in the R package CLVALID*,* showed that the stability measure with the hierarchical clustering algorithm was the most appropriate method. The optimum number of clusters suggested with this clustering algorithm was four. The different groupings are based on the similarities in the genetic composition of coffee stands. The clusters (identified by the colours) are represented in ordination space in Fig. 3a and b and in geographic space in Fig. 4a.

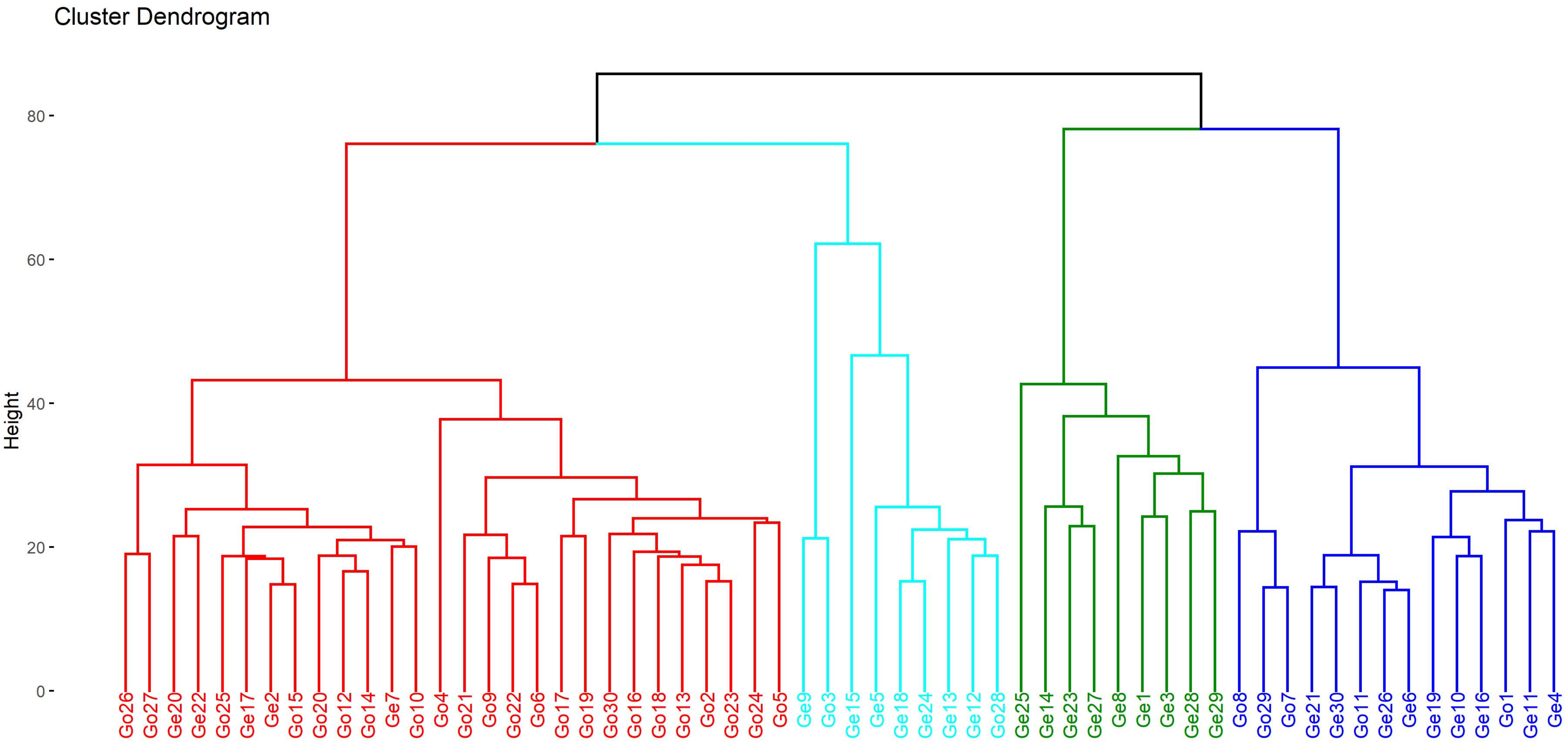
Fig. 3. Principal component analysis (PCA) (a) and redundancy analysis (RDA) (b) plot showing the ordination of 60 coffee stands based on the reference allele frequency ($\mathrm{RAF}_{\mathrm{pool}}$) on 487 variant positions per coffee stand. Coffee stands are labelled by group (G1-G4) based on hierarchical clustering. G1 (green, n = 9) mainly include less intensively managed forest sites; G2 (red, n = 28) mainly include smallholder farmers sites using landraces in the eastern part of the landscape with long history of coffee management; G3 (blue, n = 14) and G4 (light blue, n = 9) mainly include commercial plantations and some intensively managed smallholder farmers sites (See also Fig. 4a). The symbols differentiate commercial plantations from other sites. GePlant (star) refers to commercial plantations in the Gera district that were recently established (2010/2011) in montane forests/agroforest sites; GoPlant (crossed square) refers to commercial plantations in the Gomma district that were established in the early 1980s, and Others refers to smallholder farmers sites. The arrows in the RDA plot indicate environmental and management variables that significantly explain part of the variation in genetic composition of coffee. Arrowheads point to the direction of increasing gradient (for the corresponding variable) in the ordination space.

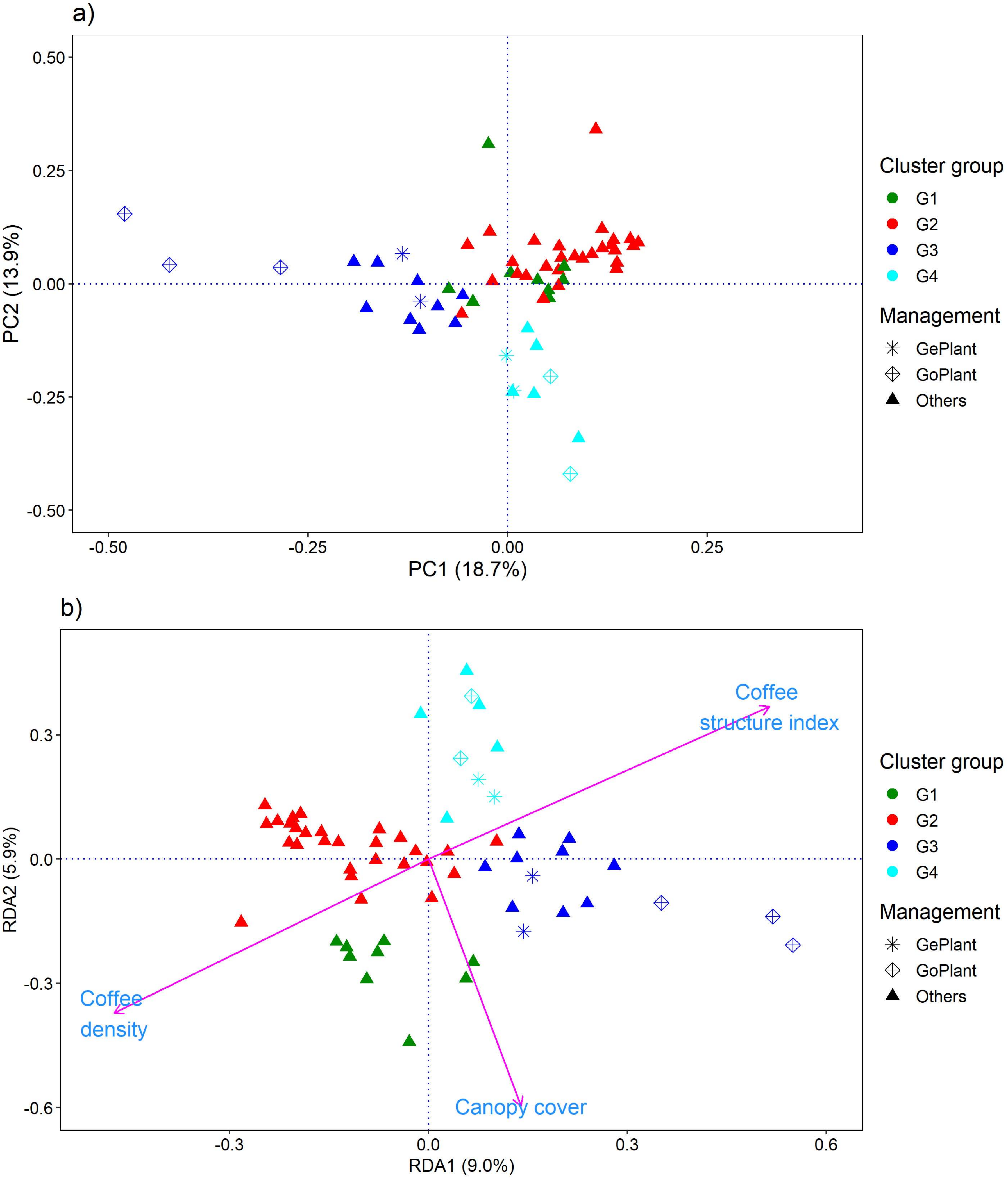
Fig. 4. Variation in a) genetic composition (reference allele frequencies, RAF_{pool}) and b) genetic diversity (mean expected heterozygosity, mean H_E) in coffee stands across the landscape. Sites in (a) are colored according to their cluster groups in the hierarchical cluster analysis on RAF_{pool} values (Fig. 2). G1 (green, n = 9) mainly include less intensively managed forest sites; G2 (red, n = 28) mainly include smallholder farmers sites using landraces in the eastern part of the landscape with long history of coffee management; G3 (blue, n = 14) and G4 (light blue, n = 9) mainly include intensively managed plantations and some intensively managed smallholder farmers sites. Sites in (b) are colored according to mean expected heterozygosity. Sites marked with triangles represent less intensive to medium level of management while those marked with crossed squares and stars represent plantation coffee systems in Gomma and Gera districts, respectively. The sites are overlaid on Google Maps (Map data ©2021 Google) using the '*geocode*' function in the GGMAP library in R.

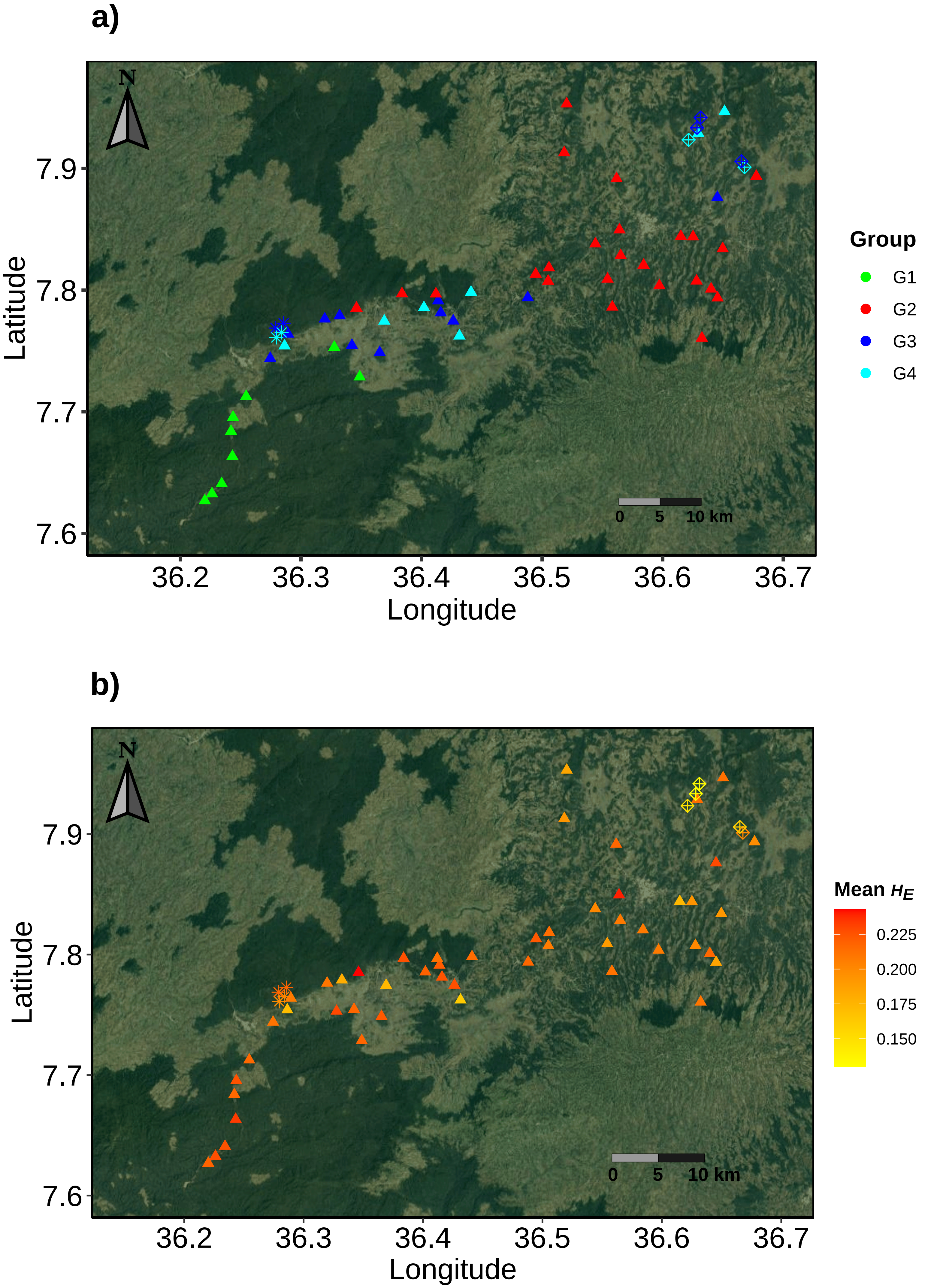
Fig. 5. The relationship between genetic diversity (mean expected heterozygosity, mean H_E) of coffee among sites as a function of elevation and coffee structure index. a) Genetic diversity of coffee among sites as a function of altitude as proxy for environment. b) Genetic diversity of coffee among sites as a function of coffee structure index as proxy for coffee management. The red, broken lines indicate regression lines for significant relationships and the grey shaded areas indicate the 95 % confidence limits for the fitted regression lines (panel a) SC = 0.40, p < 0.001; panel b) SC = -0.30, p = 0.011, See Table S6).

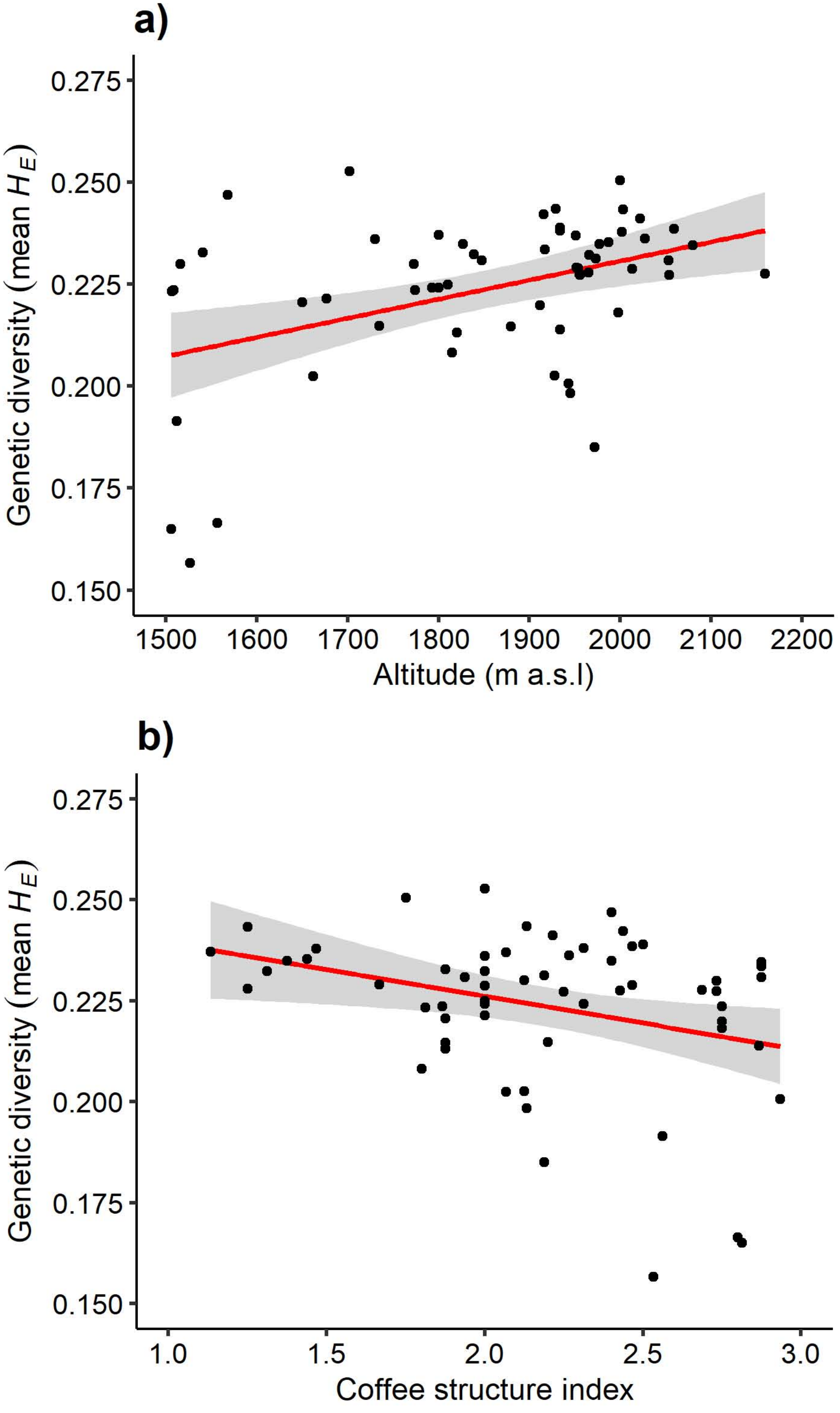
Fig. 6. Within-site variation in incidence of fungal diseases as a function of genetic diversity (mean expected heterozygosity, mean H_E). a) Standard deviation of coffee leaf rust incidence as a function of mean expected heterozygosity. b) Standard deviation of coffee berry disease incidence as a function of mean expected heterozygosity. The colored dots represent the standard deviation of a) coffee leaf rust and b) coffee berry disease from 60 sites during the two-year assessments. Regression lines are fitted for a significant relationship (panel a) Year: SC = 1.41, $p < 0.001$; Mean H_E: SC = 0.06, $p = 0.337$; panel b) Year: SC = 0.37, $p = 0.041$; Mean H_E: SC = 0.21, $p = 0.017$).

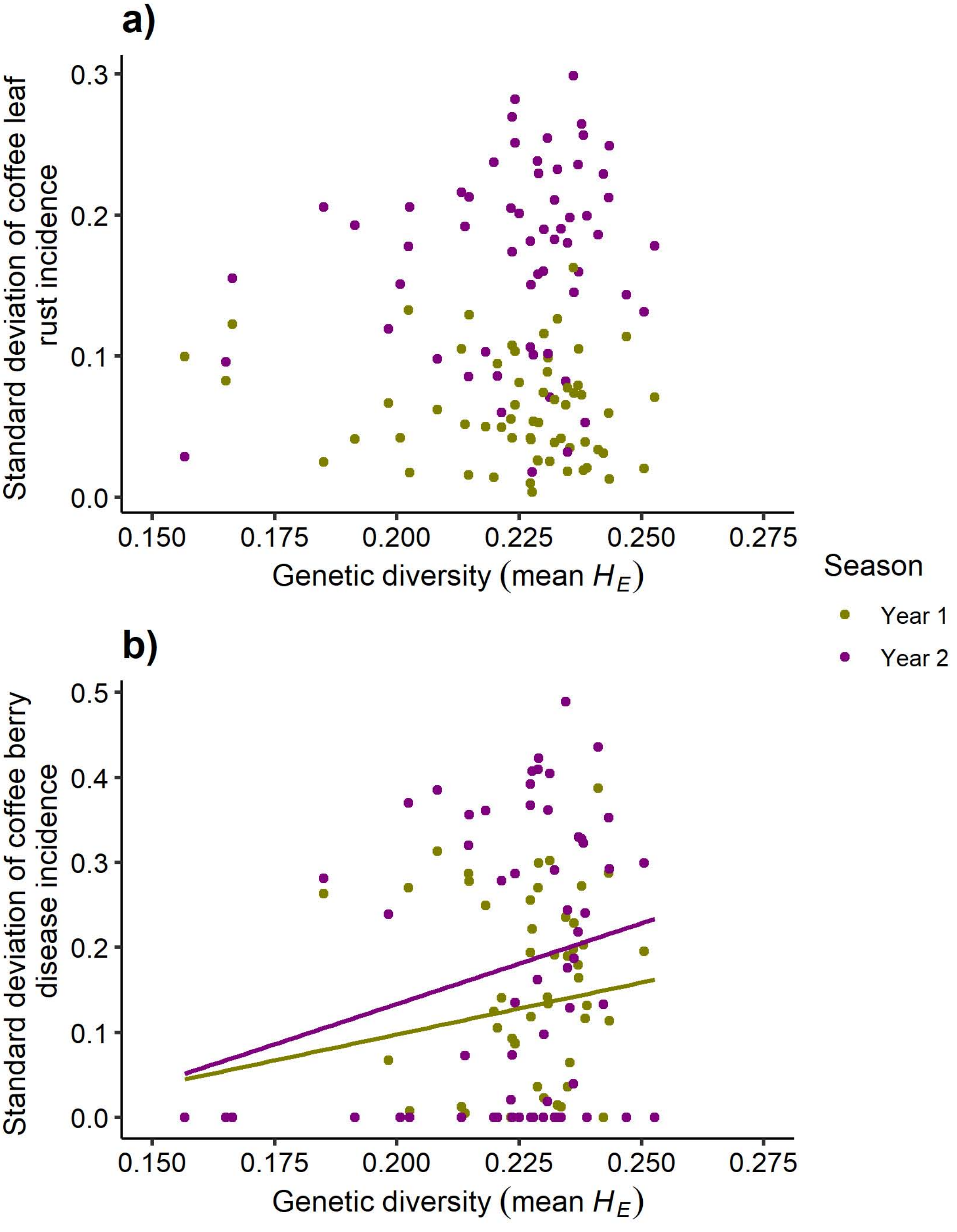

a)
Addis Ababa
Jimma
b)
N
Gomma
Gera
0 5 10 km
Latitude
7.6
7.7
7.8
7.9
Longitude
36.2
36.3
36.4
36.5
36.6
36.7
c)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
50 m
50 m


Cluster Dendrogram
Height
0
20
40
60
80
Go26
Go27
Ge20
Ge22
Go25
Ge17
Ge2
Go15
Go20
Go12
Go14
Ge7
Go10
Go4
Go21
Go9
Go22
Go6
Go17
Go19
Go30
Go16
Go18
Go13
Go2
Go23
Go24
Go5
Ge9
Go3
Ge15
Ge5
Ge18
Ge24
Ge13
Ge12
Go28
Ge25
Ge14
Ge23
Ge27
Ge8
Ge1
Ge3
Ge28
Ge29
Go8
Go29
Go7
Ge21
Ge30
Go11
Ge26
Ge6
Ge19
Ge10
Ge16
Go1
Ge11
Ge4


a)
PC1 (18.7%)
PC2 (13.9%)
-0.50
-0.25
0.00
0.25
0.50
-0.50
-0.25
0.00
0.25
Cluster group
G1
G2
G3
G4
Management
GePlant
GoPlant
Others
b)
RDA1 (9.0%)
RDA2 (5.9%)
-0.3
0.0
0.3
0.6
-0.6
-0.3
0.0
0.3
Coffee
structure index
Coffee
density
Canopy cover
Cluster group
G1
G2
G3
G4
Management
GePlant
GoPlant
Others


a)
N
7.9
7.8
7.7
7.6
Latitude
36.2
36.3
36.4
36.5
36.6
36.7
Longitude
0
5
10 km
Group
G1
G2
G3
G4
b)
N
7.9
7.8
7.7
7.6
Latitude
36.2
36.3
36.4
36.5
36.6
36.7
Longitude
0
5
10 km
Mean H_E
0.225
0.200
0.175
0.150


a)
Genetic diversity (mean H_E)
0.275
0.250
0.225
0.200
0.175
0.150
1500
1600
1700
1800
1900
2000
2100
2200
Altitude (m a.s.l)
b)
Genetic diversity (mean H_E)
0.275
0.250
0.225
0.200
0.175
0.150
1.0
1.5
2.0
2.5
3.0
Coffee structure index


a)
Standard deviation of coffee leaf rust incidence
0.3
0.2
0.1
0.0
0.150
0.175
0.200
0.225
0.250
0.275
Genetic diversity (mean H_E)
Season
● Year 1
● Year 2
b)
Standard deviation of coffee berry disease incidence
0.5
0.4
0.3
0.2
0.1
0.0
0.150
0.175
0.200
0.225
0.250
0.275
Genetic diversity (mean H_E)