A linkage disequilibrium-based statistical test for Genome-Wide Epistatic Selection Scans in structured populations

Léa Boyrie1 · Corentin Moreau2 · Florian Frugier2 · Christophe Jacquet1 · Maxime Bonhomme1

Received: 20 March 2020 / Revised: 21 July 2020 / Accepted: 21 July 2020 / Published online: 30 July 2020
© The Author(s), under exclusive licence to The Genetics Society 2020

Abstract
The quest for signatures of selection using single nucleotide polymorphism (SNP) data has proven efficient to uncover genes involved in conserved and/or adaptive molecular functions, but none of the statistical methods were designed to identify interacting alleles as targets of selective processes. Here, we propose a statistical test aimed at detecting epistatic selection, based on a linkage disequilibrium (LD) measure accounting for population structure and heterogeneous relatedness between individuals. SNP-based \(T_{rd}\) and window-based \(T_{corPC1}\) statistics fit a Student distribution, allowing to test the significance of correlation coefficients. As a proof of concept, we use SNP data from the *Medicago truncatula* symbiotic legume plant and uncover a previously unknown gene coadaptation between the *MtSUNN* (*Super Numeric Nodule*) receptor and the *MtCLE02* (*CLAVATA3-Like*) signaling peptide. We also provide experimental evidence supporting a *MtSUNN*-dependent negative role of *MtCLE02* in symbiotic root nodulation. Using human HGDP-CEPH SNP data, our new statistical test uncovers strong LD between *SLC24A5* (*skin pigmentation*) and *EDAR* (*hairs, teeth, sweat glands development*) world-wide, which persists after correction for population structure and relatedness in Central South Asian populations. This result suggests that epistatic selection or coselection could have contributed to the phenotypic make-up in some human populations. Applying this approach to genome-wide SNP data will facilitate the identification of coadapted gene networks in model or non-model organisms.

Introduction

In populations, natural selection targets genomic regions with evolutionarily conserved functions or with genetic variants contributing to adaptation to changing environments. Patterns of DNA sequence polymorphisms in these regions are expected to bear the signature of directional or balancing, positive selection on adaptive mutations, or of negative selection against deleterious mutations (Bamshad and Wooding 2003; Nielsen 2005; Vitti et al. 2013). Identifying genes showing such selection signatures has been a major goal of population genetics over the last decades. Many statistical methods have been developed, accounting for the neutral evolution expected for molecular polymorphisms in populations with varying degrees of genetic structure or with particular demographic histories (Pavlidis and Alachiotis 2017; Vitti et al. 2013; Weigand and Leese 2018). Thanks to high-throughput sequencing technologies, these methods can now be used to perform Genome-Wide Scans for Selection (GWSS) using single nucleotide polymorphism (SNP) datasets (Ahrens et al. 2018; Haasl and Payseur 2016; Oleksyk et al. 2010). Although GWSS have identified cohorts of genes associated with past or ongoing selective processes, they are not designed to identify gene coadaptation, resulting from epistatic selection on interacting genes (Otto and Whitlock 2009).

Few studies have examined the impact of epistatic selection models in population samples using DNA
polymorphisms. Simulations of two-locus epistatic models with different degrees of recombination (i.e., $0 < c < 0.5$) in a panmictic population have shown that the efficiency of epistatic selection and its statistical detection are improved if standing genetic variation already exists, rather than if selection operates on de novo mutations (Takahasi 2009; Takahasi and Tajima 2005). Simulations of two-locus coadaptation in subdivided populations have also shown that the fixation probability of the coadapted haplotype across all subpopulations increases upon moderate migration and isolation (Takahasi 2007).

Adaptive epistatic interactions between alleles at two independent loci are expected to generate linkage disequilibrium (LD). It has been shown that the correlation coefficient $r$, an LD measure related to $r^2$, can be used to detect epistatic selection between two bi-allelic loci in a population because it is a directional measure which can indicate an excess of ancestral and derived allelic associations, relative to recombinant allelic associations (Takahasi and Innan 2008). This will detect epistatic selection either in a coadaptation model where two derived alleles can form a coadapted allelic combination, or in a compensatory model where the two derived alleles are individually deleterious but compensate when combined (Piskol and Stephan 2008). More recently, a simulation study of epistatic selection in structured populations has suggested the use of the $D_\text{IS}^2$ measure of LD (Ohta 1982a; 1982b), which quantifies how the frequencies of the different two-locus haplotypes in a subpopulation depart from the average frequencies across all subpopulations (Id-Lahoucine et al. 2019). However, high $D_\text{IS}^2$ values were found in simulated models of two-locus epistatic selection but also of single-locus selection at two independent loci. This suggests that $D_\text{IS}^2$ cannot always distinguish between these two models. In addition, population structure, genetic drift, and relatedness among individuals due to nonrandom mating also act as confounders because they increase genome-wide levels of LD and generate long-distance LD (Glémin et al. 2006; Mangin et al. 2012; Nordborg 2000; Slatkin 2008), which can falsely be interpreted as signatures of epistatic selection (Zhang et al. 2004).

A significant improvement towards capturing the LD due to physical linkage in a structured population with various degrees of relatedness among individuals was the introduction of $r^2_{\text{pc}}$, an $r^2$ measure which includes the kinship matrix into the calculation in order to penalize correlated two-locus genotypic data arising from high levels of relatedness (Mangin et al. 2012). Subsequently, the $r_{\text{pc}}$ measure of LD was proposed to identify interspecific genetic incompatibilities corresponding to pairs of loci showing an excess of ancestral haplotype combinations in admixed populations (Schumer and Brandvain 2016). The $r_{\text{pc}}$ measure is a partial correlation coefficient between genotypic data at two loci given the genome-wide ancestry proportion, $a$, between two species. Although such type of measure represented an improvement, authors acknowledged that genetic relationships in populations with complex demographic histories and genetic structures were not fully taken into account (Schumer and Brandvain 2016).

In this study, we propose a statistical test to detect epistatic selection in heterogeneously structured populations (i) between two bi-allelic SNPs by using the $r_{\text{v}}$ measure, or (ii) between two genomic regions including each multiple SNPs, by using the $cor_{PC1v}$ measure. This latter measure captures the quantitative correlation between the first principal component (PC1) summarizing the multi-SNP genotypes for each genomic region. Using simulations of genome-wide SNP data in structured diploid populations with random to self-mating processes under two epistatic selection models, we show that, compared with $r$ and $cor_{PC1}$, $r_{\text{v}}$ and $cor_{PC1v}$ (i) drastically reduced the background LD generated by population structure and relatedness between individuals; (ii) showed an equivalent or a lower power to detect epistatic selection, depending on the mating process, on the dominance of selected mutations and on the selection model; and (iii) $T$ statistics ($T = \sqrt{n - 2} \frac{\bar{r}_{\text{v}} - r_{\text{v}}}{\text{SE}(\bar{r}_{\text{v}})}$) derived from $r_{\text{v}}$ or $cor_{PC1v}$ (i.e., $T_r$ and $T_{corPC1v}$) fitted a Student distribution $\tau_{0.02}$ under the null hypothesis of independence between the two tested loci. Hence, unlike $T$ statistics derived from $r$ or $cor_{PC1}$ (i.e., $T_r$ and $T_{corPC1}$), $T_{r}$ and $T_{corPC1v}$, can be used for statistical testing of the correlation coefficient between two loci, while accounting for population structure and heterogeneous relatedness between individuals.

Empirical detections of epistatic selection on SNP data are scarce in the literature (Brachi et al. 2015; Caicedo et al. 2004; Csilàr et al. 2014; Grzeskowiak et al. 2014; Hu and Hu 2015; Pool 2015). We assayed our statistical test in the frame of Genome-wide Epistatic Selection Scans (GWESS), with genomic SNP data from two different model organisms: the legume plant *Medicago truncatula* and humans. As a proof of concept, we first described in *M. truncatula* the detection of epistatic selection between the Super Numeric Nodule *MtSUNN* gene, encoding a receptor which is central for the negative regulation of symbiotic root nodulation, and the CLAVATA3-like (CLE) signaling peptide *MtCLE02*. accordingly, an ectopic expression of the *MtCLE02* gene in *M. truncatula* wild-type (WT) and *sunN* mutant roots experimentally demonstrated a *MtSUNN*-dependent negative role of the *MtCLE02* gene on nodulation, hence validating functionally the genetic interaction between these two genes. In humans, we illustrated the usefulness of the approach by identifying a significant epistatic or coselection signal in Central South Asian populations between *SLC24A5* and *EDAR* genes, encoding, respectively, a cation exchanger affecting pigmentation in zebrafish and human (Lamason et al. 2005) and a receptor...
A linkage disequilibrium-based statistical test for Genome-Wide Epistatic Selection Scans in structured populations

Table 1 Two-locus epistatic selection models under coadaptation or compensation in a haploid population.

| Allele at locus A | Allele at locus B | Allelic combination (haplotype) | Fitness value | Coadaptation | Compensation |
|------------------|------------------|---------------------------------|--------------|--------------|--------------|
| A                | B                | AB                              | 1            | 1            |              |
| a                | B                | ab                              | 1            | 1 - s        |              |
| A                | b                | Ab                              | 1            | 1 - s        |              |
| a                | b                | ab                              | 1 + s        | 1            |              |

Specific fitness values are assigned to individuals depending on the allelic combination they carry at the two loci (e.g., SNP) in each epistatic selection model. Alleles A and B correspond to ancestral alleles, a and b correspond to mutated (or derived) alleles.

We follow the fact that SLC24A5 and EDAR were previously shown to be under strong positive selection in Europe and East Asian populations, respectively (Bryk et al. 2008; Sabeti et al. 2007; Speidel et al. 2019), our results highlight the role of epistatic selection or coselection in shaping gene coadaptation during the evolution of populations.

Materials and methods

Genetic models of epistatic selection

We follow fitness genotype formalization under epistatic selection models as in (Takahasi and Innan 2008; Takahasi and Tajima 2005). Two independent bi-allelic loci A and B were considered, with ancestral alleles A and B, and derived alleles a and b, in a haploid population. The coadaptation model consists in positively selecting the two-locus ab combination. The compensatory model consists in selecting against the Ab and ab two-locus combinations, but not against AB and AB. The coefficient s is used to positively or negatively select two-locus genotypes (Table 1), and in the neutral model all fitness values are set up to 1. In a diploid population, the two-locus fitness expression is more complex because it depends on the level of dominance of the derived alleles (Supplementary Table S1).

SNP-based and window-based LD measures of epistatic selection

In a diploid organism, at a given bi-allelic SNP with alleles coded 0 and 1, the three possible genotypes are (00, 01, 11), which can be coded as the allelic dose of allele 1 (0, 1, 2). The measure on unphased genotypes between two bi-allelic loci is defined by the correlation coefficient r between vectors of genotypes at the SNPs l and m, Xl and Xm (Hill and Robertson 1968; Rogers and Huff 2009; Weir 1979):

\[
r = \frac{\sum_{i=1}^{n} (X_l^i - \bar{X}_l)(X_m^i - \bar{X}_m)}{\sqrt{\sum_{i=1}^{n} (X_l^i - \bar{X}_l)^2} \sqrt{\sum_{i=1}^{n} (X_m^i - \bar{X}_m)^2}}.
\]

In the case where 0 and 1 are the ancestral and derived alleles, respectively, a positive sign of r indicates that combinations of ancestral and derived alleles (i.e., 00 and 11) preferentially segregate in individuals at the two SNPs, compared with alternative combinations (i.e., 01 and 10). At two physically unlinked loci in a panmictic population, this measure allows to detect fitness interactions between two new mutations under the coadaptation or the compensatory model (Takahasi and Innan 2008). In the context of GWESS with high-density SNP data, we propose to use the corPC1 measure of LD between two genomic regions containing multiple SNPs. The first principal component PCIl is used to summarize quantitatively the multi-SNP genotypes of the genomic region l (see McVean 2009). Then, corPC1 is the correlation coefficient between vectors of summarized multi-SNP genotypes of the two genomic regions l and m, PCIl and PCIm:

\[
cor_{PC1} = \frac{\sum_{i=1}^{n} (PCI_l^i - \bar{PCI}_l)(PCI_m^i - \bar{PCI}_m)}{\sqrt{\sum_{i=1}^{n} (PCI_l^i - \bar{PCI}_l)^2} \sqrt{\sum_{i=1}^{n} (PCI_m^i - \bar{PCI}_m)^2}}.
\]

However, as mentioned in (Mangin et al. 2012), population structure and relatedness among individuals generate nonindependence between individuals and tend to bias upwardly the LD values. This is particularly the case in highly inbred or predominantly selfing species (Glémin et al. 2006). At a given locus, Mangin and collaborators proposed to weight the observations by multiplying the vector of genotypes by \(V^{-1}\), where V is the kinship (or relatedness) matrix among individuals. V is built with the \(V_{ij}\) covariance for all pairs (i,j) of individuals. V is the average number of identical genotypes between individuals i and j over all SNPs, in a genome-wide SNP dataset. This measure of Identity In State is a good proxy of Identity By Descent as SNP markers are likely to be accurately modeled by an infinite site mutation model. Consequently, since r is the Pearson correlation coefficient, \(r_q\) can be computed as \(\text{cor}(V^{-1}X_l, V^{-1}X_m)\) (Mangin et al. 2012) and equivalently \(\text{cor}_{PC1}\) as \(\text{cor}(V^{-1}PCI_l, V^{-1}PCI_m)\).

Statistical test of epistatic selection based on LD

Under the hypothesis that observations within \(X_l\) and \(X_m\) (respectively within \(PCI_l\) and \(PCI_m\)) are independent,
then $r$ (respectively $\text{cor}_{PC1}$) can be used to obtain the $T$ statistics:

$$T = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}},$$

which follows a Student distribution $\tau_{(r-2)}$. However, in the case where the genotypes at a given locus are correlated within the population due to nonrandom mating and/or between populations due to structure, we then expect that only the $T$ statistics obtained using $r$, or $\text{cor}_{PC1}$, follow the $\tau_{(r-2)}$ distribution. In the case where the ancestral/derived allele status is known at the SNPs, a positive sign of $r$ (or $r_v$) strictly reflects the coadaptation and compensatory epistatic models (see the previous section), and a unilateral test can be performed with alternative hypothesis being “$r$ (or $r_v$) > 0”. If the ancestral/derived allele status is not known, the sign of $r$ is not interpretable, but the $p$ value of the test can be computed on either side of the null distribution. Likewise, whether the ancestral/derived allele status of the SNPs is known or not, the sign of $\text{cor}_{PC1}$ (or $\text{cor}_{PC1v}$) is not interpretable since PC1 or PC1 with opposite signs imply an identical ranking of individuals genotypes (or relatedness) in a given genomic region (see for instance (Li and Ralph 2019)).

GWESS with SNP data in M. truncatula and human

GWESS was performed in M. truncatula using a raw dataset of 22,079,533 SNP markers identified on the eight chromosomes of the species by the Medicago HapMap Project on a collection of 262 accessions (see http://www.medicagohapmap.org/downloads/mt40). The collection has already been screened for GWAS for different traits (Bonhomme et al. 2014; 2019; Burgarella et al. 2016; Kang et al. 2015; Le Signor et al. 2017; Rey et al. 2017; Stanton-Geddes et al. 2013; Yoder et al. 2014) but also for GWSS (Bonhomme et al. 2015; Branca et al. 2011; Paape et al. 2013). This highly self-mating species (95% selfing rate), originating from the Mediterranean basin, is structured in two major subpopulations, the Far-West population (FW) concentrated on the West part under Atlantic influence, and the Circum population (C) that spreads over the rest of the Mediterranean basin (Bonhomme et al. 2014; Burgarella et al. 2016; De Mita et al. 2011; Ronfort et al. 2006). Samples from the FW subpopulation and from the C subpopulation consist of 80 and 182 accessions, respectively. We used a bait approach for GWESS, in which $T_{\text{corPC1}}$ and $T_{\text{corPC1v}}$ were calculated for a given candidate gene, here MtSUNN, with the Medtr4g070970 gene identifier in the genome version 4.0—http://www.medicagogenome.org/—(Tang et al. 2014) or MrunA17Chr4g0035451 in the v5 version—https://medicago.toulouse.inra.fr/MrunA17r5.0-ANR/—(Pecrix et al. 2018), against each of the remaining 48,339 genes of the genome. The $T_{\text{corPC1}}$ and $T_{\text{corPC1v}}$ statistics were calculated based on PC1 values from SNP data located in 10 kbp windows spanning each M. truncatula gene, and PC1 values from MtSUNN. $P$ values were then obtained from the $\tau_{(r-2)}$ null distribution. For the PC1 calculation on each gene, imputed SNP data were required. Gene-based imputation was performed using the TASSEL software (Bradbury et al. 2007), where each missing base was imputed with the accession that shares the longest haplotype surrounding the base, on a window of 30 SNPs maximum (Bonhomme et al. 2014). For the calculation of $T_{\text{corPC1v}}$, the kinship matrix $V$ of the 262 individuals was estimated based on 7,252,792 SNPs with a 5% Minor Allele Frequency (MAF).

In human, GWESS was performed by using the dataset of 644,257 SNPs (431,951 SNPs with a 5% MAF) with no missing data from the HGDP-CEPH Human Genome Diversity Panel on a world-wide sample (America, Asia, Europe, Middle East, North Africa, Sub-Saharan Africa and Oceania) of 940 individuals belonging to 57 populations from 23 countries (Cann et al. 2002; Li et al. 2008). The genome version (i.e., the gene positions) B36 was used for this analysis in order to fit with SNP positions in the HGDP-CEPH dataset, similarly to (Daub et al. 2013). We used a bait approach in which $T_r$ and $T_v$ were calculated for SNPs.
located in or near SLC24A5 and EDAR genes (chromosome 15 and 2, respectively), against each of the remaining SNPs of the genome.

Functional genetic validation of the relationship between MtCLE02 and MtSUNN in M. truncatula

The M. truncatula sunn mutant, the MtCLE02 cloning strategy for overexpression, the root transformation protocol, and the conditions for phenotyping the nodulation are described in the Supplementary Method File 2.

Results

Quality control of simulations

To ensure that simulations produced consistent between-population structure and within-population inbreeding levels, the $F_{ST}$ and $F_{IS}$ parameters were calculated (Supplementary Fig. S1). At the outcome of the simulations (i.e., generation 300), the average $F_{IS}$ in self-mating and panmictic populations was equal to 0.92 and 0.07, respectively, while the average $F_{ST}$ was equal to 0.19 and 0.10, respectively. In order to quantify the fixation of coselected alleles, we tracked down the evolution of the frequency of the derived alleles $a$ and $b$ at the two SNPs $A$ and $B$ intended to be targeted by selection, located on two different chromosomes. Selection efficiency was measured by the co-fixation rate of $a$ and $b$ at each generation in each selection model (Supplementary Fig. S2). The first observation was that the COAD epistatic model generally induced a higher speed of co-fixation than the ADD positive selection model, while, as expected, the COMP model tended to maintain a higher polymorphism due to selection of both $AB$ and $ab$ combinations at the two selected loci. The second observation was that co-fixation rates of the derived alleles in self-mating populations reached more rapidly an equilibrium value than in panmictic populations, but more importantly that in self-mating populations the dominance level of the selected mutations had few effect on the co-fixation dynamics because of the very low heterozygosity level ($F_{IS} = 0.92$ at the onset of selection). A similar result was observed in a simulation study of selective sweeps in self-mating populations (Hartfield and Bataillon 2020). On the other hand, the dominance level in panmictic populations strongly impacted co-fixation dynamics due to more complex fitness patterns in the presence of heterozygotes (Supplementary Table S1). However, despite starting from standing variation at SNPs under selection, values of the co-fixation rates were moderate, and this result must be interpreted in light of the small size of the simulated population ($N = 250$ in each population), as selection efficiency increases with population size according to a factor $N_s$ (Glémin 2007).

Two-locus LD under epistatic selection models

We focused hereafter on the evolution of the two-locus average LD across simulations, in self-mating and panmictic populations under selection models with codominance. Codominance of the selected mutations indeed produced the highest co-fixation rates in panmictic populations, while the dominance level had no effect in self-mating populations (Supplementary Fig. S2). Dominance or recessivity of the selected mutations mainly impacted negatively the two-locus average LD under the COMP epistatic model in panmictic populations (Supplementary Fig. S3). In the codominant mutations model, we first observed that under the neutral model, the population structure with or without nonrandom mating generated LD between two independent loci, as measured using $r$ or $cor_{PC1}$, that could reach 0.25–0.5 at the final generation (Fig. 1). This background LD was lowered to zero or close to zero on average, when correcting these statistics by the $V$ matrix, as measured using $r_v$ or $cor_{PC1v}$.

Second, we observed that selection models tended to generate more LD than the neutral model, as measured using $r_v$ or $cor_{PC1v}$. For instance, in self-mating species, the COAD, COMP, and ADD selection models all tended to generate more LD than the neutral model, with COAD and COMP generating more LD than the ADD model (Fig. 1). In panmictic populations, however, only the COMP model generated a consistent LD, compared with the COAD and ADD models which generated low LD. Nevertheless, despite correcting for population structure, it remained difficult to distinguish epistatic selection from additive selection in self-mating populations at the haplotype level, as different haplotypes were under selection in each sub-population. This artifact was less prominent when focusing on the SNPs targeted by selection.

Finally, SNP-based LD measures ($r_{fr}$) were more efficient than haplotype-based LD measures ($cor_{PC1v}$) to detect epistatic selection (Fig. 1). However, these measures could not capture any signal once allele fixation at one SNP or co-fixation at the two SNPs occurred. On the other hand, $cor_{PC1v}$ relied on SNP polymorphisms in the genomic region surrounding SNPs under selection, so that they could benefit from the hitch-hiking effect even after allele fixation at the selected SNPs.

False positive control and power of two-locus correlation statistics

On the assumption that values of the correlation coefficient $r$ follow a Student distribution $\tau_{n-2}$ under the null
hypothesis of independence between the two variables tested, we examined the fit of the statistics $T_r$, $T_{\text{corPC1}}$, $T_{\text{rv}}$, and $T_{\text{corPC1v}}$, to such a distribution. False positive (FP) proportions of these statistics are given for different rejection quantiles of the Student distribution $\tau_{n-2}$ in Table 2. Two time-points were considered in neutral simulations, at generation 140 in the midst of the time course, and at the last generation 300. At generation 140, in both the self-mating and random-mating models, $T_r$ and $T_{\text{corPC1}}$ showed excessively large FP proportions. For instance, FP proportions ranging from 55 to 81% were observed for a 1% type I error, while $T_r$ and $T_{\text{corPC1}}$ showed adequate, conservative FP proportions ranging from 0.1 to 3% for the same 1% type I error (Table 2). At generation 300, a similar behavior was observed, with FP proportions ranging from 74 to 91% for a 1% type I error, while $T_r$ and $T_{\text{corPC1}}$ showed adequate—though less conservative in the case of the self-mating model—FP proportions ranging from 1.1 to 22% for the same 1% type I error. This indicates that corrections for population structure and heterogeneous relatedness are necessary in order to perform statistical tests of the neutral hypothesis for a null correlation between two independent loci (two SNPs or two genomic regions), accounting for “noisy” neutral processes.

A power analysis of the $T_r$, $T_{\text{corPC1}}$, $T_{\text{rv}}$, and $T_{\text{corPC1v}}$ statistics was then performed by using simulated data (i)
under the null hypothesis of neutrality and independence between loci and (ii) under each of the selection models and independence between loci. At both time-points (genera-

False positive proportions are calculated as the proportion of simulations in which the statistics has a value greater than the defined rejection quantile of the \( \tau_{(\alpha/2)} \) distribution, for different type I errors: 10, 5, and 1%. In our simulations, the sample size \( n \) was equal to 500. Since the sign of the correlation coefficient is not interpretable, especially for \( T_{corPC1} \) and \( T_{corPC1} \), the absolute values of \( T_{r} \), \( T_{corPC1} \), \( T_{corPC1} \), and of the Student distribution \( \tau_{(\alpha/2)} \) were used for false positive proportion calculation.

Bold values correspond to values of \( T_{r} \), and \( T_{corPC1} \), the \( T \) statistics corrected for population structure and relatedness between individuals.

under the null hypothesis of neutrality and independence between loci and (ii) under each of the selection models and independence between loci. At both time-points (generations 140 and 300), a general trend was that the detection power with \( r_{TV} \) and \( corPC1/corPC1 \) was higher for the COMP model than for the COAD or the ADD models (i.e., 25–50%, 10–65%, and 10–30%, respectively, for \( \alpha = 5\% \) with \( r_{TV} \) or \( corPC1 \) statistics), especially when considering random mating (Fig. 2). In addition, the correction of LD-based measures by the kinship matrix (\( r_{TV/corPC1} \)) did not increase the detection power of epistatic selection; rather, it tended to reduce power, especially in the COMP model, but not in the COAD model. This is due to the fact that the fixation of the \( AB \) allelic combination was more frequent in subpopulations than in the whole population in the COMP model (see Supplementary Fig. S2) — a consequence of unequal initial frequencies of the ancestral/derived alleles in the simulations, leading to high LD values when population structure was not taken into account. Finally, \( corPC1 \) tended to show less power than \( r_{TV} \) because of haplotype heterogeneity, namely when the same selected allele was associated to different haplotypes within a subpopulation (Fig. 2a–d).

Detection of two-locus coadaptation in the \( M. \) truncatula plant

The \( MtsSUNN \) gene encodes a leucine-rich repeats-receptor-like kinase, whereas the \( MtCLE02 \) genes encode a CLAVATA-like secreted signaling peptide. The SUNN receptor function, which is crucial in the systemic negative regulation of nodulation, was previously associated to other CLE secreted signaling peptide encoding genes, \( MtCLE12 \) and \( MtCLE13 \) (Mortier et al. 2012; 2010), but not with \( MtCLE02 \). Whereas their expression was induced by the inoculation with symbiotic rhizobia bacteria initiating nodulation, this was not the case for \( MtCLE02 \) (Supplementary Fig. S4A, B). These previously documented CLE/SUNN relationships pointed us to test for a putative

| Table 2 | False positive (FP) proportions for \( T_r \), \( T_{corPC1} \), \( T_{corPC1} \), and \( T_{corPC1} \) statistics in comparisons with the Student distribution (\( \tau_{(\alpha/2)} \)) used for testing the significance of the correlation coefficient. |
|---------|---------------------------------------------------------------|
| Generation | Mating scheme | Statistics | FP proportions |
|----------|----------------|-------------|-----------------|
|          |                | \( T_r \)    | \( 10\% \) \( 5\% \) \( 1\% \) | \( 10\% \) \( 5\% \) \( 1\% \) |
| 140      | Self-mating    | \( T_r \) 85\% 82\% 74\% | \( T_{corPC1} \) 89\% 86\% 81\% | \( T_{corPC1} \) 13\% 8\% 3\% |
|          | Random-mating  | \( T_r \) 72\% 66\% 55\% | \( T_{corPC1} \) 83\% 78\% 70\% | \( T_{corPC1} \) 4\% 1\% 0\% |
| 300      | Self-mating    | \( T_r \) 92\% 91\% 87\% | \( T_{corPC1} \) 95\% 93\% 91\% | \( T_{corPC1} \) 31\% 26\% 20\% |
|          | Random-mating  | \( T_r \) 85\% 81\% 74\% | \( T_{corPC1} \) 93\% 91\% 87\% | \( T_{corPC1} \) 4\% 1\% 0\% |

Experimental evidence for the genetic relationship between \( MtsSUNN \) and \( MtCLE02 \) in \( M. \) truncatula

The \( MtsSUNN \) gene encodes a leucine-rich repeats-receptor-like kinase, whereas the \( MtCLE02 \) genes encode a CLAVATA-like secreted signaling peptide. The SUNN receptor function, which is crucial in the systemic negative regulation of nodulation, was previously associated to other CLE secreted signaling peptide encoding genes, \( MtCLE12 \) and \( MtCLE13 \) (Mortier et al. 2012; 2010), but not with \( MtCLE02 \). Whereas their expression was induced by the inoculation with symbiotic rhizobia bacteria initiating nodulation, this was not the case for \( MtCLE02 \) (Supplementary Fig. S4A, B). These previously documented CLE/SUNN relationships pointed us to test for a putative
functional interaction between CLE02 signaling peptides and the SUNN receptor. As previously performed for MtCLE12 or MtCLE13 (Mortier et al. 2012; 2010), we used a genetic approach consisting in overexpressing comparatively the MtCLE02 gene in M. truncatula WT and sunn mutant roots (Fig. 4). First, a quantification of the nodule number in WT versus sunn mutant roots highlighted the well-known supernodulation phenotype of the sunn mutant (Mann and Whitney–Wilcoxon test, $p$ value $= 2 \times 10^{-8}$). Second, the nodule number was significantly decreased when MtCLE02 was overexpressed in WT roots, as validated by real time RT-PCR (Supplementary Fig. S4C), indicating a negative role of MtCLE02 on nodulation (Fig. 4, Mann and Whitney–Wilcoxon test, $p$ value $= 2 \times 10^{-6}$). Third, MtCLE02 overexpression in sunn mutant roots did not impact the nodule number (Mann and Whitney–Wilcoxon test, $p$ value $= 0.66$), in contrast to what was observed in the WT, indicating that the negative role of CLE02 on nodulation relies on the SUNN receptor.

Detection of two-locus coadaptation in human populations

In human, two GWESS were performed on the world-wide sample of 940 individuals with SNP data from SLC24A5 and EDAR genes, two major drivers of the external appearance which have been subjected to strong positive selection in human populations according to different studies (Basu Mallick et al. 2013; Beleza et al. 2013; Bryk et al. 2008; Sabeti et al. 2007; Speidel et al. 2019). Two bait
SNPs, 15_46172199 (rs2250072) and 2_108973688 (rs6749207) located in SLC24A5 and EDAR genes, respectively, were chosen for the GWESS. For each bait SNP, SNP-based statistics $T_r$ and $T_{rv}$ were calculated for 431,950 genome-wide SNPs. Scans implemented in the world-wide population with SNPs 15_46172199 and 2_108973688, each as bait, were flattened towards small $p$ values when using the $T_r$ statistic (Fig. 5a, b), compared with scans implemented with the $T_{rv}$ statistic (Fig. 5c, d). Using the SNP 15_46172199 as bait for SLC24A5 gene, a peak corresponding to the EDAR gene was detected, with the SNP 2_108946170 as the top significant SNP (Fig. 5a, c; $T_{rv}$-based $p$ value $= 2.29 \times 10^{-9}$). Conversely, when the scan was performed with the SNP 2_108973688 as bait from the EDAR gene, a peak corresponding to the SLC24A5 gene was detected, with the SNP 15_46179457 as the top significant SNP (Fig. 5b, d; $T_r$-based $p$ value $= 1.2 \times 10^{-12}$). Genome-wide LD distributions between each bait SNP and all other SNPs (Fig. 5c, d; top left of each panel) showed very high LD values between SNPs from EDAR and SLC24A5, thus indicating extremely significant signals (EDAR SNPs ranked among the top 0.01–0.18% SNPs in LD with SLC24A5; and SLC24A5 SNPs ranked among the top 0.006–0.03% SNPs in LD with EDAR). The world-wide geographic distribution of genotypes at SNPs 15_46172199—SLC24A5—and 2_108973688—EDAR—(Fig. 6c) correlated substantially with the global human population structure, as depicted by a phylogenetic tree based on the kinship matrix among individuals (Fig. 6a, b). Indeed, the derived allele at SNP 15_46172199, associated with the light skin allele at the SLC24A5 gene, was present in Europe, North Africa, Middle East and Central South Asia; and
the derived allele at SNP 2_108973688, associated with the thick hair allele at the EDAR gene, was present in East-Asia, America, and Oceania (Fig. 6c). The strong LD signature observed in the world-wide samples between SLC24A5 and EDAR, as measured with Tr, therefore reflected the selection of derived alleles in different geographic regions, and thus a correlation with the global population structure. However, the LD measured with Trv was still highly significant between SLC24A5 and EDAR, indicating that epistatic selection may have occurred between both genes at the level of geographic sub-regions.

In order to localize the geographic origin of such selection signature, GWESSs were performed within six geographic regions of the world-wide sample: Central South Asia, East Asia, Sub-Saharan Africa, Middle East, Europe, and America (Supplementary Figs. S5–S10). Only the GWESS performed in Central South Asia indicated a significant LD between SNPs at the SLC24A5 and EDAR genes (Supplementary Fig. S5C, Trv-based p value = 6.7 × 10−6 at SNP 2_108973688; Supplementary Fig. S5D, Trv-based p value = 2.8 × 10−6 at SNP 15_46174380).

Human population samples from the HGDP-CEPH dataset in Central South Asia are composed of eight different ethnic groups from Pakistan. To search for local signals, LD tests were performed with Trv between two candidate SNPs within SLC24A5 (15_46179457 (rs1834640) and 15_46172199), and three candidates SNPs within EDAR (2_108962124 (rs260607), 2_108982808 (rs17034770), and 2_108973688), for the 50 ethnic groups or populations distributed within eight geographic regions and showing polymorphism at all five SNPs. Average and standard deviation of −log10(p value) across six pairwise SNP comparisons (Fig. 6d) strongly supported a high LD between SLC24A5 and EDAR in the Burusho ethnic group from Pakistan (3.2 and 0.36, respectively), as also highlighted by genotypes of Burusho individuals (Fig. 6c, rv = 0.63 for genotypes between the two dotted lines). This pattern of high LD between SLC24A5 and EDAR in Burusho did not seem to be generated by any population sub-structure in this ethnic group, since LD tests performed with Tr resulted in a similar average (3.18) and standard deviation (0.36) of −log10(p value) (Supplementary Fig. S11).

Discussion

We introduced a statistical method which can detect the signature of epistatic selection using LD between two loci. SNP-based (Trv) and window-based (TcorrC1) statistics, which take into account the underlying population structure
and relatedness among individuals, are shown to fit a Student distribution $\tau_{n-2}$, allowing to easily and quickly test for significance of correlation coefficients in the frame of GWESS using either a candidate SNP, a gene, or a short genomic region as bait. Simulations have shown that $T_{\nu}$ and $T_{\text{corPC1}}$ showed equivalent or less power than $T_r$ or $T_{\text{corPC1}}$ to detect epistatic selection occurring simultaneously in all subpopulations, ranging from 10 to 65% (assuming a 5% type I error) depending on the epistatic selection model and mating process. Thus, selection signals in local populations could be more difficult to detect with $T_r$ and $T_{\text{corPC1}}$, because in these cases, selection can be correlated with population structure. In addition, the impact of the kinship matrix on LD correction changes depending on the scale of the sampling, with a stronger impact for large scales of geographic population structure than for smaller, less structured, and less heterogeneous geographic scales. These features suggest that GWESS should be performed not only on a global sample comprising individuals from different populations, but also on samples from different sub-populations in order to search for more population-specific patterns of epistatic selection. Despite the power was not increased when using $T_r$ and $T_{\text{corPC1}}$, simulations as well as analyses performed both in M. truncatula and human genomes strongly supported their use, in order to efficiently control for FP. Interestingly, although SNP-based statistics ($T_r, T_{\nu}$) may tend to show an increased power than window-based statistics ($T_{\text{corPC1}}, T_{\text{corPC1}}$) because they are not sensitive to haplotype heterogeneity and because mutations under epistatic selection at both loci can be directly tested by SNP-based statistics, window-based statistics also show several advantages. First, window-based statistics are faster to implement at the genome scale, notably in two-dimensional GWESS. Second, the window size can be fixed at a value that fits best the average LD decay in the species studied, even though a standard 10 kbp size can be used by default. Finally, window-based statistics allow detecting coevolving genes even after the putative

Fig. 5 LD distribution between the bait SNPs of SLC24A5 and EDAR genes and all other HGDP-CEPH SNPs in the whole human population samples ($n = 952$). LD between SNP 15_46172199 (SLC24A5) or SNP 2_108973688 (EDAR), respectively, and all other SNPs of the genome is tested using $T_r$ (a, b) or $T_{\nu}$ (c, d). The x-axis corresponds to SNP positions spanning the 22 human autosomes, each point corresponds to an SNP and the black points depict SNPs at candidate genes in epistatic selection with one SNP at the bait gene (vertical dotted line) in each figure. The y-axis is the $-\log_{10}$ ($p$ value) of the test of the correlation coefficient. Plots at the top left of each figure show the distribution of LD between each bait SNP and all other SNPs of the genome. LD values between the bait SNP of SLC24A5 and the target top SNPs of EDAR (respectively the bait SNP of EDAR and the target top SNPs of SLC24A5) is represented by an arrow.
The applications of our method to SNP data from human populations or from populations of the model plant *M. truncatula* allowed in both cases to identify a couple of genes most probably under epistatic selection, or at least under coselection. In *M. truncatula*, additional experiments revealed a genetic interaction likely shaped by epistatic selection between the *MtSUNN* and *MtCLE02* genes. The polymorphism at the *MtSUNN* gene could be driven by balancing selection at the local level because the *H* statistic (Fay and Wu, 2000) is 1.45 in the FW population and ranks among the highest 8.41% in the entire genome. On the other

---

**Fig. 6** Schematic human population structure inferred from the kinship matrix, the geographic distribution of alleles, and the LD between SLC24A5 and EDAR. **a** Neighbor-joining tree inferred from the molecular kinship matrix based on 431,951 SNPs from the HGDP-CEPH database showing the global human population structure. **b** Same tree as in (a) showing the clustering of the different sub-populations or ethnic groups sampled. **c** Bar plots depicting geographic distributions of genotypes at SNP 2_108973688 (*EDAR*) and SNP 15_46172199 (*SLC24A5*), highlighting LD patterns mainly due to the global population structure and to the selection of the derived alleles (coded 1) at the two genes. **d** Average and standard error of LD significances based on *Trv* statistics between SNPs of SLC24A5 (SNP 15_46172199, SNP 15_46179457) and EDAR (SNP 2_108962124, SNP 2_108973688, SNP 2_108982808) within each human sub-population. The bar plot pinpoints Central South Asia as the main source of within population LD probably due to coselection of alleles in Pakistan ethnic groups (mainly from Burusho, delimited by two vertical dotted lines in (b) and (c)).

fixation of coselected SNPs, because surrounding SNPs within genes or windows also carry a selection signal by hitchhiking.

The applications of our method to SNP data from human populations or from populations of the model plant *M. truncatula* allowed in both cases to identify a couple of genes most probably under epistatic selection, or at least under coselection. In *M. truncatula*, additional experiments revealed a genetic interaction likely shaped by epistatic selection between the *MtSUNN* and *MtCLE02* genes. The polymorphism at the *MtSUNN* gene could be driven by balancing selection at the local level because the *H* statistic (Fay and Wu, 2000) is 1.45 in the FW population and ranks among the highest 8.41% in the entire genome. On the other
hand, $H = 0.52$ based on the whole species, ranking among the highest 39.5%, according to a set of 47,875 genes. The polymorphism at the $MtCLE02$ gene seems more affected by an ongoing soft sweep that can be detected at the level of the whole species ($H = −2.25$, ranking among the lowest 8.98% of the genome). Still, both genes maintain polymorphisms, and epistatic selection could drive this pattern. As a proof of concept, a genetic approach was used to demonstrate the functional interaction between the CLE02 signaling peptide and the SUNN receptor in the context of symbiotic root nodulation. Indeed, the CLE02 signaling peptide negatively affected the number of nodule organs on the plant root system depending on the SUNN receptor, as previously shown for other CLE peptide encoding genes, $MtCLE12$ and $MtCLE13$ (Gautrat et al. 2019; Mortier et al. 2012; 2010). Interestingly, it should be noted that the $MtCLE02$ gene is, in contrast to $MtCLE12$ and $MtCLE13$, not regulated by symbiotic nodulation conditions and not phylogenetically closely related to previously characterized CLE peptide encoding genes shown to have a related negative impact on nodule number (Hastwell et al. 2017). The identification of a novel CLE peptide acting in this genetic pathway highlights the discovery power of our method to functionally associate gene pairs independently of their expression pattern or of a coregulation pattern criterion.

In human, the $SLC24A5$ gene, a major driver of variation in skin pigmentation, has been shown to be under positive selection in the European population (Deng and Xu 2018; Izagirre et al. 2006; Sabeti et al. 2007). The causal mutation for the light skin phenotype was not present in HGDP data (SNP rs1426654, position: 15_46213776), but SNPs used (i.e., 15_46179457 and 15_46172199) were located on the same $SLC24A5$ haplotype background that was previously characterized (Basu Mallick et al. 2013; Beleza et al. 2013; Crawford et al. 2017). In addition, the V370A mutation in the $EDAR$ gene coding for a receptor related to TNFα receptors and involved in driving hair structure, as well as teeth and sweat glands development, was shown to be under positive selection in East Asia and in native Americans, and to increase hair thickness (Bryk et al. 2008; Sadier et al. 2014; Speidel et al. 2019). However, just as for $SLC24A5$, the causal mutation in the $EDAR$ gene was not present in HGDP data, but SNPs used (i.e., 2_108962124, 2_108973688, and 2_108982808) were located within the genomic sequence of $EDAR$. The geographic distribution of genotypes at SNPs 15_46172199 and 2_108973688 strongly correlated with the world-wide human population structure, which explained the high LD observed at this level in the scans implemented with $T_r$. However, scans implemented with $T_c$ indicated a persistence of top SNPs in LD between $SLC24A5$ and $EDAR$ in the world-wide sample, which might be due to epistatic selection or coselection. We identified such selection signature in Central South Asia, with the Burusho ethnic group from Pakistan as being one possible geographic origin. The GWESS performed in Central South Asia, and subsequent LD tests performed between $SLC24A5$ and $EDAR$ in subpopulations with $T_r$ or $T_c$ statistics showed similar results, indicating a weak effect of population structure, as previously observed within this geographic region and in India (Rosenberg et al. 2006). The Burusho showed a predominant association between the derived alleles at $SLC24A5$ and the ancestral alleles at $EDAR$, which is indicative of the persistence of a typical European light skin and thin hair structure in this ethnic group. To date, no functional link is yet demonstrated between the critical skin pigmentation-related $SLC24A5$ gene and the ectodysplasin pathway to which $EDAR$ belongs. While the long-distance LD between $EDAR$ and $SLC24A5$ largely reflects strong positive selection acting independently on these two genes in different geographic regions, our results also suggest that coselection of these genes may have locally contributed to the phenotypic makeup of human populations.

Our method was fast to implement on a computer equipped with an Intel Xeon E5-2660 v4 processor (10 Cores and 2.4–3.4 GHz performance, supplied by 256 Go memory), as only few minutes were needed to estimate the kinship matrix, depending on the sample size (e.g., 30s and ~3 min for Medicago and human, respectively), and to implement one GWESS (e.g., 2 and 12 min for 48,339 Medicago genes with $T_{corPC1}$ and 431,950 human SNPs with $T_{r}$, respectively). Although functional analyses tools for genetic interaction are still limited to a few model species, GWESS can easily be performed on non-model organisms, as long as sufficiently dense SNP data are available, to identify candidate genes under epistatic selection that may be relevant to study in related model species for instance. If GWESS is not directly applicable, possible relationships between candidate genes could be directly tested using dedicated SNP markers, but one should be cautious about the lack of benchmark with other genes. A natural extension of the one-dimension use of $T_{r}$- or $T_{corPC1}$-based tests is the implementation of two-dimensional GWESS in which the correlation of each polymorphic locus (SNP, gene or genomic region) in the genome would be tested against all remaining polymorphic loci, thanks to parallel computations on bioinformatics platforms. We anticipate that such an approach will open the way towards exploring evolutionary coadapted gene networks.

**Data availability**

The *M. truncatula* SNP dataset (hapmap format) used in this study can be retrieved at [http://www.medicagohapmap.org/downloads/mt40](http://www.medicagohapmap.org/downloads/mt40). The Human SNP dataset from the HGDP-CEPH Human Genome Diversity Panel can be retrieved at
ftp://ftp.ceph.fr/hgdp_suppl (http://www.ceph.fr/hgdp/). R scripts to implement the statistical test based on $T_r$, $T_{rc}$, $T_{corPC1}$, or $T_{corPC1}^*$, along with an example dataset, are available at https://github.com/leaboyrie/LD_corpc1.

Acknowledgements This work was supported by the “DeCoD” project funded by the French Agence Nationale de la Recherche (grant number ANR-16-CE20-0017-01). The PhD position of LB was funded by the “DeCoD” project. We thank the bioinformatics platform Toulouse Midi-Pyrénées (Genotoul). This work was performed in the LR$^2$V (Toulouse, France), part of the “Laboratoire d’Excellence” (LABEX) entitled TULIP (grant number ANR-10-LABX-41). We thank Carole Laffont (IPS2, CNRS, Grif-sur-Yvette, France) for providing results about MtcLexpressions. Work in the Florian Frugier laboratory has benefited from a French State grant (Saclay Plant Sciences, grant number ANR-17-EUR-0007, EUR SPS-GSR) and an ANR grant (“PSYCHE”, grant number ANR-16-CE20-0009-01). We thank Thomas Bataillon, two other anonymous reviewers, and Pierre-Marc Delaux for useful criticisms and comments to improve the manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Ahrens CW, Rymer PD, Stow A, Bragg J, Dillon S, Umbers KDL et al. (2018) The search for loci under selection: trends, biases and progress. Mol Ecol 27(6):1342–1356

Bamshad M, Wooding SP (2003) Signatures of natural selection in the human genome. Nat Rev Genet 4(2):99–111

Basu Mallick C, Iliescu FM, Möls M, Hill S, Tamang R, Chauhey G et al. (2013) The light skin allele of SLC24A5 in South Asians and Europeans shares identity by descent. PLoS Genet 9(11):e1003912

Beleza S, Santos AM, McEvoy B, Alves I, Martinho C, Cameron E et al. (2013) The timing of pigmentation lightening in Europeans. Mol Biol Evol 30(1):24–35

Bonhomme M, André O, Badis Y, Ronfort J, Burgarella C, Chantret N et al. (2014) High-density genome-wide association mapping implicates an F-box encoding gene in Medicago truncatula flowering time genes generates a latitudinal cline in a life history trait. Proc Natl Acad Sci USA 108(42):E864–E870

Bryk J, Hardouin E, Pugach I, Hughes D, Strotmann R, Stoneking M et al. (2008) Positive selection in East Asians for an EDAR allele that enhances NF-kappaB activation. PLoS ONE 3(5):e2209

Burgarella C, Chantret N, Gay L, Prosperi JM, Bonhomme M, Tiffin P et al. (2016) Adaptation to climate through flowering phenology: a case study in Medicago truncatula. Mol Ecol 25(14):3397–3415

Caicedo AL, Stinchcombe JR, Olsen KM, Schmitt J, Parunggan MD (2004) Epistatic interaction between Arabidopsis FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. Proc Natl Acad Sci USA 101(44):15670–15675

Cann HM, de Toma C, Cavalli L, Legrand MF, Morel V, Piouffle L et al. (2002) A human genome diversity cell line panel. Science 296(5566):261–262

Crawford NG, Kelly DE, Hansen MEB, Beltrame MH, Fan S, Bowman SL et al. (2017) Loci associated with skin pigmentation identified in African populations. Science 358(6365):1–14

Csilléry K, Lalagüe H, Vendramin GG, González-Martínez SC, Fady B, Oudou-Muratorio S (2014) Detecting short spatial scale local adaptation and epistatic selection in climate-related candidate genes in European beech (Fagus sylvatica) populations. Mol Ecol 23(19):4696–4708

Daub JT, Hofer T, Cutivet E, Dupanloup I, Quintana-Murci L, Robinson-Rechavi M et al. (2019) Evidence for polygenic adaptation to pathogens in the human genome. Mol Biol Evol 36(7):1544–1558

De Mita S, Chantret N, Loridon K, Ronfort J, Bataillon T (2011) Molecular adaptation in flowering and symbiotic recognition pathways: insights from patterns of polymorphism in the legume Medicago truncatula. BMC Biol Evol 11:229

Deng L, Xu S (2018) Adaptation of human skin color in various populations. Hereditas 155:1

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155(3):1405–1413

Gautier P, Mortier V, Laffont C, De Keyser A, Fromentin J, Frugier F et al. (2019) Unraveling new molecular players involved in the autoregulation of nodulation in Medicago truncatula. J Exp Bot 70(4):1407–1417

Glémin S (2007) Mating systems and the efficacy of selection at the molecular level. Genetics 177(2):905–916

Glémin S, Bazin E, Charlesworth D (2006) Impact of mating systems on patterns of sequence polymorphism in flowering plants. Proc Biol Sci 273(1604):3011–3019

Grzeskowiak L, Stephan W, Rose LE (2014) Epistatic selection and coadaptation in the Pfr resistance complex of wild tomato. Infect Genet Evol 27:456–470

Haasl RJ, Paysier BA (2016) Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. Mol Ecol 25(1):5–23

Hartfield M, Bataillon T (2020) Selective sweeps under dominance and inbreeding. G3 10(3):1063–1075

Hastwell AH, de Bang TC, Gresshoff PM, Ferguson BJ (2017) CLE peptide-encoding gene families in Medicago truncatula and Lotus japonicus, compared with those of soybean, common bean and Arabidopsis. Sci Rep 7(1):9384

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38(6):226–231

Hu XS, Hu Y (2015) Genomic scans of zygotic disequilibrium and epistatic SNPs in HapMap phase III populations. PLoS ONE 10(6):e0131039
Id-Lahoucine S, Molina A, Cánovas A, Casellas J (2019) Screening for epistatic selection signatures: a simulation study. Sci Rep 9 (1):1026
Izaguirre N, Garcia I, Junquera C, de la Rúa C, Alonso S (2006) A scan for signatures of positive selection in candidate loci for skin pigmentation in humans. Mol Biol Evol 23(9):1697–1706
Kang Y, Sakiroglu M, Krom N, Stanton-Geddes J, Wang M, Lee YC et al. (2015) Genome-wide association studies with proteomics data reveal genes important for synthesis, transport and packaging of globulins in legume seeds. N. Phytol 214(4):1597–1613
Li H, Ralph P (2019) Local PCA shows how the effect of population structure differs along the genome. Genetics 211(1):289–304
Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319(5866):1100–1104
Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, Cierco-Ayrolles C (2012) Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity 108(3):285–291
McVean G (2009) A genealogical interpretation of principal components analysis. PLoS Genet 5(10):e1000686
Mortier V, De Wever E, Vuytsteka M, Holsters M, Goormachtig S (2012) Nodele numbers are governed by interaction between CLE peptides and cytokinin signaling. Plant J 70(3):367–376
Mortier V, Den Herder G, Whitford R, Van de Velde W, Rombauts S, Holsters M et al. (2010) CLE peptides control Medicago truncatula nodulation locally and systemically. Plant Physiol 153(1):222–237
Nielsen R (2005) Molecular signatures of natural selection. Annu Rev Genet 39:197–218
Nordborg M (2000) Linkage disequilibrium, gene trees and selling: an ancestral recombination graph with partial self-fertilization. Genetics 154(2):923–929
Ohta T (1982a) Linkage disequilibrium due to random genetic drift in finite subdivided populations. Proc Natl Acad Sci USA 79 (6):1940–1944
Ohta T (1982b) Linkage disequilibrium with the island model. Genetics 101(1):139–155
Oleksyk TK, Smith MW, O’Brien SJ (2010) Genome-wide scans for footprints of natural selection. Philos Trans R Soc Lond B Biol Sci 365(1537):185–205
Otto SP, Whitlock MC (2009) The impact of epistatic selection on the mosaic ancestry of the Drosophila genetic reference panel and the D. melanogaster reference genome reveals a network of epistatic fitness interactions. Mol Biol Evol 32(12):3236–3251
Pool JE (2015) The mosaic ancestry of the Drosophila genetic reference panel and the D. melanogaster reference genome reveals a network of epistatic fitness interactions. Mol Biol Evol 32 (10):2326–2351
Rey T, Bonhomme M, Chatterjee A, Gavrin A, Toullette J, Yang W et al. (2017) The Medicago truncatula GRAS protein RAD1 supports arbuscular mycorrhiza symbiosis and Phytophthora palmivora susceptibility. J Exp Bot 68(21-22):5871–5881
Rogers AR, Huff C (2009) Linkage disequilibrium between loci with unknown phase. Genetics 182(3):839–844
Ronfort J, Bataillon T, Santoni S, Delalande M, David JL, Prosperi JM (2006) Microsatellite diversity and broad scale geographic structure in a model legume: building a set of nested core collection for studying naturally occurring variation in Medicago truncatula. BMC Plant Biol 6:28
Rosenberg NA, Mahajan S, Gonzalez-Quevedo C, Blum MG, Nino-Rosales L, Ninis V et al. (2006) Low levels of genetic divergence across geographically and linguistically diverse populations from India. PLoS Genet 2(12):e215
Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449(7164):913–918
Sadier A, Viriot L, Pantalacci S, Lauvet V (2014) The ectodysplasin pathway: from diseases to adaptations. Trends Genet 30(1):24–31
Schumer M, Brandvain Y (2016) Determining epistatic selection in admixed populations. Mol Ecol 25(11):2577–2591
Slatkin M (2008) Linkage disequilibrium-understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9 (6):477–485
Speidel L, Forest M, Shi S, Myers SR (2019) A method for genome-wide genealogy estimation for thousands of samples. Nat Genet 51(9):1321–1329
Stanton-Geddes J, Paape T, Epstein B, Briskine R, Yoder J, Mudge J et al. (2013) Candidate genes and genetic architecture of symbiotic and agronomic traits revealed by whole-genome, sequence-based association genetics in Medicago truncatula. PLoS ONE 8(5):e56588
Takahasi K (2009) Coalescent under the evolution of coadaptation. Mol Ecol 18(24):5018–5029
Takahasi KR (2007) Evolution of coadaptation in a subdivided population. Genetics 176(1):501–511
Takahasi KR, Innan H (2008) The direction of linkage disequilibrium: a new measure based on the ancestral-derived status of segregating alleles. Genetics 179(3):1705–1712
Takahasi KR, Tajima F (2005) Evolution of coadaptation in a two-locus epistatic system. Evolution 59(11):2324–2332
Tang H, Krishnakumar V, Bidwell S, Rosen B, Chan A, Zhou S et al. (2014) An improved genome release (version Mt4.0) for the model legume Medicago truncatula. BMC Genomics 15:312
Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting natural selection in genomic data. Annu Rev Genet 47:97–120
Weigand H, Leese F (2018) Detecting signatures of positive selection in non-model species using genomic data. Zool J Linn Soc 184(2):528–533
Weir BS (1979) Inferences about linkage disequilibrium. Biometrics 35(1):235–254
Yoder JB, Stanton-Geddes J, Zhou P, Briskine R, Young ND, Tiffin P (2014) Genomic signature of adaptation to climate in Medicago truncatula. Genetics 196(4):1263–1275
Zhang W, Collins A, Gibson J, Tapper WJ, Hunt S, Deloukas P et al. (2004) Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. Proc Natl Acad Sci USA 101(52):18075–18080