Genome scans reveal candidate regions involved in the adaptation to host plant in the pea aphid complex.

J. Jaquiéry, S. Stoeckel, P. Nouhaud, L. Mieuzet, F. Mahéo, F. Legeai, N. Bernard, A. Bonvoisin, R. Vitalis, J.-C. Simon

MOLECULAR ECOLOGY

Molecular Ecology (2012) **21**, 5251–5264 doi: 10.1111/mec.12048

Genome scans reveal candidate regions involved in the adaptation to host plant in the pea aphid complex

J. JAQUIÉRY,*[1] S. STOECKEL,*[1] P. NOUHAUD,* L. MIEUZET,* F. MAHÉO,* F. LEGEAI,*† N. BERNARD,* A. BONVOISIN,* R. VITALIS‡ and J-C. SIMON*

*INRA, UMR 1349, Institute of Genetics, Environment and Plant Protection, Domaine de la Motte, BP 35327, 35653, Le Rheu Cedex, France, †INRIA Centre Rennes – Bretagne Atlantique, GenOuest, Campus de Beaulieu, 35042, Rennes, France, ‡CNRS – INRA, UMR CBGP (INRA – IRD – CIRAD – Montpellier SupAgro), Campus International de Baillarguet, CS 30016, F-19 34988, Montferrier sur Lez Cedex, France

Abstract

A major goal in evolutionary biology is to uncover the genetic basis of adaptation. Divergent selection exerted on ecological traits may result in adaptive population differentiation and reproductive isolation and affect differentially the level of genetic divergence along the genome. Genome-wide scan of large sets of individuals from multiple populations is a powerful approach to identify loci or genomic regions under ecologically divergent selection. Here, we focused on the pea aphid, a species complex of divergent host races, to explore the organization of the genomic divergence associated with host plant adaptation and ecological speciation. We analysed 390 microsatellite markers located at variable distances from predicted genes in replicate samples of sympatric populations of the pea aphid collected on alfalfa, red clover and pea, which correspond to three common host-adapted races reported in this species complex. Using a method that accounts for the hierarchical structure of our data set, we found a set of 11 outlier loci that show higher genetic differentiation between host races than expected under the null hypothesis of neutral evolution. Two of the outliers are close to olfactory receptor genes and three other nearby genes encoding salivary proteins. The remaining outliers are located in regions with genes of unknown functions, or which functions are unlikely to be involved in interactions with the host plant. This study reveals genetic signatures of divergent selection across the genome and provides an inventory of candidate genes responsible for plant specialization in the pea aphid, thereby setting the stage for future functional studies.

Keywords: ecological speciation, gene flow, genomic divergence, host plant adaptation, insect–plant interactions, outlier loci

Received 26 July 2012; accepted 6 August 2012

Introduction

Divergent selection targeting ecologically important traits may result in adaptive population differentiation and eventually lead to reproductive isolation, a process referred to as ecological speciation (Schluter 2001; Rundle & Nosil 2005). A fundamental question in speciation studies concerns the underlying genetic mechanisms favouring linkage disequilibrium between the loci under divergent natural selection and those controlling reproductive isolation, in the face of gene flow (Via 2001; Bolnick & Fitzpatrick 2007). Deciphering the genomic architecture of ecological specialization and reproductive isolation is therefore essential to address this issue. The outstanding diversity of herbivorous insects, which probably results from their specialization to different host plants, offers a good opportunity to evaluate how natural selection promotes divergence and contributes to species diversification (Berlocher & Feder 2002; Dres & Mallet 2002; Peccoud *et al.* 2010). Partial

Correspondence: Jean-Christophe Simon, Fax: +33 2 23 48 51 50; E-mail: jean-christophe.simon@rennes.inra.fr

[1]These authors contributed equally.

reproductive isolation is crucial in studying ecological speciation, as the influence of ecological factors on genetic divergence is generally masked by the accumulation of further differences once the speciation is completed (Via 2009). However, genetically and ecologically differentiated populations that are still connected by substantial gene flow have been clearly identified in only few taxa (Dres & Mallet 2002; Mallet 2008). This occurs for the pea aphid, *Acyrthosiphon pisum*, which represents a strong case for ecological speciation by divergent selection (Via 1999; Ferrari *et al.* 2008; Peccoud *et al.* 2009a; Peccoud & Simon 2010; Smadja *et al.* 2012).

The pea aphid forms a worldwide complex of host-associated sympatric populations (or host races) that feeds on more than 20 legume genera (Peccoud & Simon 2010). Each host race is specialized on one or a few legume species and is genetically differentiated from other races (Via 1991; Ferrari *et al.* 2006, 2008, 2012). Peccoud *et al.* (2009a) recently showed that the pea aphid complex encompasses at least eight partially isolated host races and three presumable species, thereby forming a continuum of population divergence towards virtually complete speciation. The age of the most recent maternal ancestor of pea aphid biotypes was estimated, by means of sequence data from the aphid endosymbiont *Buchnera*, to lay between 8000 and 16 000 years, which suggests a recent and rapid adaptive radiation (Peccoud *et al.* 2009b). In pea aphids, the correlation between preference and performance on the same host plant species suggests a relation between divergent selection and premating isolation, which is likely to be favoured by a specific genetic architecture (Hawthorne & Via 2001). Quantitative trait locus (QTL) mapping analyses of host acceptance and performance in two North American *A. pisum* host races, one specialized on alfalfa and the other on clover, indeed suggested that pleiotropic or closely linked loci control both traits (Hawthorne & Via 2001). However, the precise localization of these QTLs on the pea aphid genome and the identity and function of genes underlying these traits remain largely unknown.

Genome-wide scan of large sets of individuals from multiple populations is a powerful approach to explore the genetic architecture underlying adaptive divergence and to identify loci or genomic regions under ecologically divergent selection (e.g. Storz & Nachman 2003; Vasemagi *et al.* 2005; Bonin *et al.* 2006). Genome scans generally rely on the assumption that loci involved in adaptation to local environmental conditions (either directly or indirectly through genetic hitchhiking) exhibit stronger differentiation among populations and lower diversity within population, as compared to selectively neutral regions of the genome (Cavalli-Sforza 1966; Lewontin & Krakauer 1973; Wu 2001; Storz 2005). Although there are some conceptual and technical limitations in their use that must be kept in mind (see Bierne 2010; Bierne *et al.* 2011; Michel *et al.* 2010; Roesti *et al.* 2012), genome scans have successfully identified genomic regions responsible for adaptation to specific environments in a wide range of organisms (Nadeau & Jiggins 2010; Stapley *et al.* 2010; Ellegren & Sheldon 2008).

Here, we scanned the genome of host-adapted populations of the pea aphid to explore the organization of the genomic divergence associated with ecological specialization and speciation and to identify candidate loci involved in the adaptation to the host plant. To achieve this aim, we developed 390 microsatellite markers located at variable distances from predicted genes in the pea aphid reference genome (IAGC 2010) and analysed their variation in replicate samples of sympatric populations of *A. pisum* from *Medicago sativa* (alfalfa), *Trifolium pratense* (red clover) and *Pisum sativum* (pea), which are three host races commonly reported in *A. pisum* (Frantz *et al.* 2006; Peccoud *et al.* 2009a,b; Ferrari *et al.* 2012). We used a hierarchical method to identify loci showing enhanced genetic differentiation between host races, as compared to the rest of the genome. Finally, we explored the genic environment of outlier loci and assessed the putative functions of neighbouring genes based on the annotated sequence of the pea aphid genome.

Methods

Aphid sampling

Pea aphid individuals were collected on three different host plants (alfalfa, red clover and pea) (Table S1, Supporting information). These samples represent three different host races, each adapted to a specific host range and genetically differentiated from the other biotypes (Frantz *et al.* 2006; Peccoud *et al.* 2009a). Although there is little genetic structure within each host race at small and large geographic scales (Frantz *et al.* 2006; Peccoud *et al.* 2008; Via & West 2008; Peccoud *et al.* 2009a; Ferrari *et al.* 2012), we used a hierarchical sampling strategy to avoid the potential confounding effect of geography: for each of the three host plants, we sampled populations from three different regions, separated by 100–200 km, in Western Europe (Table S1, Supporting information). Approximately 900 parthenogenetic females were collected from the nine populations (see Fig. S1, Supporting information, for further details). To obtain sufficient amounts of DNA, field-collected aphids were grown individually in controlled conditions ensuring continuous clonal reproduction (16 h light per day, 18 °C). A total of 566 (of ~900) individuals survived parasitism (e.g. fungi and parasitoids).

We then discarded from the sample all presumable copies of the same clone, the migrant individuals coming from other host races and the hybrids between host races. To do so, we genotyped the sampled individuals at seven microsatellite loci (AlA09M, AlB07M, AlB08M, AlB12M, ApF08M, ApH08M and ApH10M, see Caillaud *et al.* 2004), following the conditions described in Peccoud *et al.* (2008). Given the high polymorphism of this set of loci (Frantz *et al.* 2006), individuals bearing the same multilocus genotype were considered as copies of the same clone (see Fig. S1, Supporting information, for information on the clonal composition of each population). A single copy of each clone per population was then kept for further analyses, which resulted in a total of 426 unique multilocus genotypes. To discard presumable migrants and/or hybrids in our data set, we ran the Bayesian clustering method implemented in STRUCTURE (Pritchard *et al.* 2000) on the multilocus genotypes at the seven microsatellite markers, setting the hypothetical number of clusters to $K = 3$, which corresponds to the number of host plants. Each STRUCTURE analysis was run for 100 000 steps, after a 100 000-step burn-in period, using the admixture model that assigns individual proportions of genotypes to each cluster and discarding prior information on the collection site or the host plant. About 80% of the individuals had a large genotype membership (>90%) in a single cluster, corresponding to the host plant (and not the geographical region) from which they were collected (see Fig. S1, Supporting information). In each population, we then selected 20 individuals among those assigned to their host plant with a probability larger than or equal to 90%. This condition was not met in two populations from pea where we had to include seven individuals that were assigned to their host plant with an 80% probability on average. All further analyses were conducted on this subset of 180 individuals (three host races × three geographic populations × 20 individuals). The different steps allowing the selection of individuals for the genome-scan analysis are summarized in Fig. S1 (Supporting information).

Large-scale genotyping

The 180 selected individuals were genotyped at 390 microsatellite loci (see Table S2, Supporting information, for loci used and primer sequences) as described in Jaquiéry *et al.* (2012). These loci were picked among a larger set of 952 loci genotyped on a subsample of eight individuals from the three host races, based on the reliability of amplification, their polymorphism and the ease of scoring. The vast majority of these 390 loci was chosen to locate on different scaffolds in the V1 genome assembly of the pea aphid (IAGC 2010) to cover a wide genomic area. These loci were also selected based on their distances from genes predicted in the reference pea aphid genome. Although there is no available data on the extent of recombination and linkage disequilibrium in the pea aphid genome, we showed in a recent study that microsatellites standing at <30 kb from predicted genes displayed a reduced genetic diversity as compared to microsatellites located at more than 30 kb (Jaquiéry *et al.* 2012). This suggests that 30 kb may be considered as a relevant threshold below which neutral loci might be influenced by hitchhiking. This 30-kb threshold was, therefore, used in all analyses presented hereafter, although we insist that this particular choice shall be considered as somewhat arbitrary until confirmed by dedicated studies. Seventy-two loci were located within EST contigs (Sabater-Munoz *et al.* 2006) and 235 loci were located at less than 30 kb from a predicted gene. Among the latter markers, 25 were located at less than 14 kb (average distance: 3.4 kb, median: 1.8 kb) from genes coding for chemosensory proteins, gustative receptors, odorant-binding proteins or odorant receptors previously annotated and potentially involved in *A. pisum* plant specialization (Zhou *et al.* 2010; Smadja *et al.* 2009, 2012). The 83 remaining loci were chosen because they were located at more than 30 kb from any predicted genes in the V1 genome assembly. After genotyping, all individuals with more than 10% of missing data were eliminated, resulting in a final data set of 166 individuals genotyped at 390 loci (Table 1).

Because a new genome assembly (V2) was released after our selection of microsatellite markers, we checked for possible differences in the genomic environment of the 390 loci between the V1 and the V2 assemblies. Seven loci (of 390) could not be unambiguously located

Table 1 Genetic diversity at 390 microsatellite loci calculated within populations of the pea aphid *Acyrthosiphon pisum* collected on alfalfa, pea and red clover. N, sample size; H_e, expected heterozygosity (median); A_r, allelic richness computed on eight diploid individuals (median), F_{IS} (median) and 95% confidence interval estimated from bootstrapping over loci. Geographical origin of populations: M = Mirecourt, R = Ranspach, S = Switzerland

Host plant	Population	N	A_r	H_e	F_{IS} [95% CI]
Alfalfa	M	20	3.61	0.60	0.012 [0;0.035] ns
	R	18	3.79	0.61	0.009 [0;0.034] ns
	S	14	3.56	0.61	0.002 [−0.007;0.027] ns
Pea	M	20	3.24	0.53	−0.014 [−0.028;0] ns
	R	19	3.00	0.52	−0.029 [−0.059;−0.002]*
	S	20	3.01	0.53	−0.025 [−0.041;−0.008]*
Red clover	M	17	3.90	0.64	0.029 [0;0.05] ns
	R	20	3.84	0.63	0.024 [0.004;0.046]*
	S	18	3.81	0.63	0.026 [0.006;0.052]*

* $P<0.05$, ns, non significant.

in the V2 assembly, because of multiple blast hits in the V2 but not in the V1 assembly. For the 383 remaining loci, we calculated the distance to the closest exon of RefSeq genes (i.e. a set of 10,249 genes with empirical support, IAGC 2010, AphidBase: http://www.aphidbase.com/aphidbase) and grouped loci in two categories, those located at less than 30 kb and those at more than 30 kb from an exon of a RefSeq gene. On the basis of this procedure, only 25 loci (of 383) had different genomic environment in the V2 genome assembly relatively to the V1 assembly.

Whole-genome genetic structure

Expected heterozygosity (Nei 1987) and allelic richness (assuming a minimal sample size of eight individuals) were calculated per population for each locus with FSTAT 2.9.4 (Goudet 2005). F_{IS} was estimated for each locus in each population using FSTAT 2.9.4, and 95% confidence intervals (CI) of median F_{IS} estimates were computed by bootstrapping over loci using R (R Development Core Team 2012). F_{IS} was considered as significant if the 95% CI did not contain zero. Pairwise F_{ST} estimates (Weir & Cockerham 1984) were computed between populations using ARLEQUIN 3.5 (Excoffier & Lischer 2010), and their significance was estimated by 16 000 random permutations of individuals among populations.

We investigated the genetic structure of our data with STRUCTURE (Pritchard *et al.* 2000), using the admixture model and varying the number of putative clusters from $K = 1$ to $K = 9$. Each analysis was replicated 10 times and consisted in 100 000 steps following a 25 000-step burn-in period. Then, to test for hidden hierarchical structure, we performed independent STRUCTURE analyses on each of the three different subgroups that were identified, as recommended by Evanno *et al.* (2005).

Effect of the genomic environment on genetic structure

Natural selection is predicted to affect polymorphism not only at the targeted genes, but also at linked markers (Maynard-Smith & Haigh 1974). We thus tested for a potential effect of the genomic environment of the microsatellite loci on allelic richness, using the following linear mixed effect model (R package Lme4, Bates & Sarkar 2007) fitted with REML (restricted maximum likelihood) and implemented in R (R Development Core Team 2012):

$$\begin{aligned} Allelic\,richness \sim\ & Number\,of\,Microsatellite\,Repeats \\ & + Genomic\,environment + 1|factor(population) \\ & + 1|factor(Host\,Plant) + 1|factor(Loci) \end{aligned}$$

The *Number of Microsatellite Repeats* was measured on the V2 genome assembly of the pea aphid (IAGC 2010) as the maximal number of consecutive di-, tri- or tetranucleotide repeats found between the two primers. This provides a rough estimate of the number of repeats for each locus (as the number of repeats measured on the reference genome is used as a surrogate for the average number of repeats found in a population), but this still allows capturing essential information as the number of microsatellite repeats explains a large proportion of variance in allelic richness (see the Results section). The *Genomic environment* was either measured as (i) the number of RefSeq genes located in 60-kb window centred on the focal microsatellite loci or as (ii) the distance (in kb) to the first exon of the closest RefSeq gene. Two distinct models were therefore constructed using either measure, to test for the effect of the *Genomic environment*. The *Number of Microsatellite Repeats* and *Genomic environment* were considered as fixed effects, and *Population*, *Host Plant* and *Loci* were considered as random effects. The *Number of Microsatellite Repeats* and *Genomic environment* effects were tested by likelihood ratio tests based on the drop in Akaike's information criterion following the inclusion of the focal variable into the partial model comprising all other variables.

As divergent selection acting across different environments not only affect genetic differentiation at the targeted genes but also at linked markers through hitchhiking (Cavalli-Sforza 1966; Lewontin & Krakauer 1973; Maynard-Smith & Haigh 1974), we tested for an effect of the presence of genes in the genomic environment of the microsatellite loci on the genetic structure (F_{ST}) measured at these loci. To that end, we estimated the Pearson's correlation coefficient between F_{ST} and either the number of RefSeq genes located within a 60-kb window centred on the focal microsatellite locus, or the distance between the microsatellite and the closest exon from a RefSeq gene. We also verified that the number of microsatellite repeats did not affect F_{ST}. Finally, we also tested for a difference in genetic differentiation (F_{ST}) between the 25 microsatellite loci located in the vicinity of chemosensory genes and the remaining 358 loci (Mann–Whitney two-sided test).

Detection of loci influenced by selection

To detect those loci that depart from neutral expectation, and which are therefore potentially involved in the adaptation to the host plant, we used the hierarchical method developed by Excoffier *et al.* (2009) and implemented in ARLEQUIN 3.5 (Excoffier & Lischer 2010). This method is an extension to Beaumont & Nichols' (1996) approach, which assumes a hierarchical island model of migration between structured populations. It is therefore particularly adapted to the pea aphid complex

where populations from the same host plant exchange migrants at a higher rate than populations from different host races. The distribution of the genetic differentiation among host races expected under neutrality was estimated by means of coalescent simulations. The among-race differentiation was characterized by the parameter F_{CT}, which accounts for the geographical structure within host races. 100 000 coalescent simulations were performed conditionally on the multilocus estimate of F_{CT} at the 390 microsatellite loci, assuming 50 groups and 100 demes per group. The observed data from each locus were compared with the simulated distribution, and a particular locus was classified as a significant outlier if it lay outside the 99% confidence envelope. For the purpose of this study, we focused on loci putatively involved in divergence between host races and therefore considered only the loci falling above the upper confidence limit, because such genome scan methods are not well adapted to detect balancing selection (Beaumont & Balding 2004). As we were mostly interested in identifying outlier loci involved in host plant specialization, rather than in adaptation to local environmental conditions, we checked whether the outliers identified from this global analysis (in which all three host races were included simultaneously) were not classified as outliers within host races. To that end, we ran three independent analyses for the detection of outliers within each host race. In that case, there was only one hierarchical level, that of the populations. To identify the host race(s) in which selection targeted outlier loci, we performed three additional hierarchical analyses between each pair of host races: clover *vs.* alfalfa, alfalfa *vs.* pea and pea *vs.* clover.

Note that in this study, individuals were sampled using a hierarchical design to avoid confounding geographical effects or sample-related specificities. Currently, only one method (i.e. Excoffier *et al.* 2009) can account for hierarchical structure and therefore make use of all the information contained in hierarchical data sets. In a recent simulation study, Narum & Hess (2011) showed that this hierarchical method had higher type I and II error rates than island-based methods (e.g. Bayescan, Fdist2). There is, however, no evidence that this conclusion applies to our case. Indeed, these authors compared bi-allelic markers (at which heterozygosity cannot exceed 0.5 by definition) with an envelope made of multi-allelic markers (at which heterozygosity lies in the [0,1] range). Additionally, they simulated genetic data under a stepping-stone migration model, while Excoffier *et al.* (2009) method assumes a hierarchical migration model. As an alternative to Arlequin 3.5, nonhierarchical methods such as Bayescan (Foll & Gaggiotti 2008) can be used to analyse hierarchical data set, but this causes several difficulties. (i) If a single analysis is performed in which the hierarchical structure is neglected, this increases the number of false positives (Excoffier *et al.* 2009). (ii) Alternatively, several pairwise analyses between populations with contrasted phenotypes can be performed. This, however, raises the problem of multiple testing and of loss of power (as the whole data set cannot be used in a unique analysis). For these reasons, we chose to use the method implemented in Arlequin 3.5 because it is the most adapted to analyse our hierarchical data set.

We also performed Structure and AMOVA (Analysis of MOlecular VAriance, Excoffier *et al.* 1992) analyses on the set of outlier loci and on the set of nonoutlier loci, to test whether the two groups of loci showed similar patterns of genetic structure. For the AMOVAs, populations were nested within host races, and the significance was estimated by 16 000 random permutations.

Then, all the predicted genes located at <30 kb from the outlier loci in the V2 genome assembly of the pea aphid (IAGC 2010; Legeai *et al.* 2010) were collected and blasted against nonredundant (NR) peptide database (NCBI, June 2011 version) to assess their putative functions. The NCBI NR protein database is a collection of all protein data from all species for which such information is available. Protein domains were identified with Interproscan against the Interpro database (September 2011 version) (Hunter *et al.* 2009). Gene ontology associations were performed with Blast2Go (Conesa *et al.* 2005) using Interpro and NR databases. These analyses assume similar genome organization between the reference genome (i.e. LSR1 that belongs to the alfalfa race) and the populations studied here. We cannot exclude, though, that some genomic rearrangements may have occurred in some host races.

Finally, as we specifically designed some microsatellite loci in the vicinity of the chemosensory genes described by Smadja *et al.* (2009, 2012) for their potential role in the adaptation to the host plant, we also tested whether the chemosensory function was overrepresented in the outlier category (test of proportion).

Genomic architecture of host plant adaptation

We investigated whether the among-host-race differentiation measured at the microsatellite loci located on the same scaffolds as the 11 Arlequin 3.5 outliers (based on the V2 genome assembly) was higher than the average. The pattern of linkage disequilibrium between the loci that were physically linked with outliers was also investigated. The correlations (R^2) between allele frequencies between the outlier loci and those located on the same scaffold were calculated within each population using

FSTAT 2.9.4. An overall estimate, combining R^2 values calculated within each population, was provided by FSTAT 2.9.4. *P*-values for linkage disequilibrium were calculated by 1000 randomizations of genotypes at each locus within populations.

Results

Genomic structure of pea aphid populations

Allelic richness ranged from 3.00 to 3.90 per population, the expected heterozygosity from 0.52 to 0.64, and F_{IS} varied from −0.029 to 0.029 (Table 1). Populations of the pea race stood out from the two other host races, being characterized by a reduced genetic diversity at the scale of the population and by an excess of heterozygotes at the individual level, as denoted by negative F_{IS} (significant for two of the three pea populations, see Table 1). The red clover race showed higher F_{IS} than the other races, with two populations having significantly positive values.

Analyses with STRUCTURE revealed that the most likely number of clusters was $K = 2$ (based on Evanno *et al.*'s 2005 method), with one cluster composed of individuals collected on red clover and alfalfa and the second of those collected on pea. Nevertheless, for $K = 3$ the three clusters corresponded exactly to the host plant on which individuals had been sampled (Fig. 1A). For $K = 3$, each individual had a proportion of genome that originated from the cluster corresponding to its host plant >0.71 (average was 0.96). When analyses were run separately on each of the three clusters, the most likely number of clusters was $K = 2$ for pea and red clover and $K = 3$ for alfalfa. The membership coefficients indicated some geographical structure for populations collected on alfalfa and red clover but not for those collected on pea (Fig. 1B).

Pairwise F_{ST} comparisons between host races confirmed a strong effect of the host plant on genetic structure: the pea race was highly differentiated from red clover and alfalfa races, with an average F_{ST} between the two host races of 17%. Red clover and alfalfa races were less divergent ($F_{ST} = 6.9\%$) (Fig. S2, Supporting information). Contrastingly, there was little differentiation among populations within host races: populations collected on pea were not genetically differentiated ($F_{ST} < 0.006$, NS), whereas most red clover and alfalfa populations were slightly differentiated within each host race (Fig. S2, Supporting information).

Effect of the genomic environment on the genetic structure

The number of microsatellite repeats was positively correlated with allelic richness (linear mixed effect model [LMEM], $P < 10^{-16}$, $R^2 = 41\%$, Fig. S3A, Supporting information). The genomic environment of the loci also influenced allelic richness: allelic richness was negatively correlated with the number of genes in a 60-kb window (LMEM, $P < 0.0001$, $R^2 = 11\%$, Fig. S3B, Supporting information) and positively correlated with the distance to the closest gene (LMEM, $P < 0.001$, $R^2 = 7\%$, Fig. S3C, Supporting information). Contrastingly, genetic structure (F_{ST}) was affected neither by the number of microsatellite repeats (Pearson correlation, $P = 0.86$, $r = -0.009$, Fig. 2A) nor by the distance to the closest gene ($P = 0.12$, $r = -0.08$, Fig. 2C), but slightly increased with the number of genes within a 60-kb window centred on each locus ($P = 0.0024$, $r = 0.15$, Fig. 2B). We found no significant difference between F_{ST} measured at microsatellite loci located near a chemosensory gene ($F_{ST} = 0.102$, $n = 25$) and F_{ST} measured at the remaining loci ($F_{ST} = 0.100$; $n = 358$) (Mann–Whitney test: $w = 4601$, $P = 0.81$).

Detection of outlier loci

We found 11 outlier loci at the $\alpha = 0.01$ threshold from the global analysis with ARLEQUIN 3.5 (Table 2, Fig. 3). The hierarchical analyses run between pairs of host races also supported these outliers: all these loci but one were identified as significant outliers (at $\alpha = 0.01$) in at least one pairwise comparison (and often in two

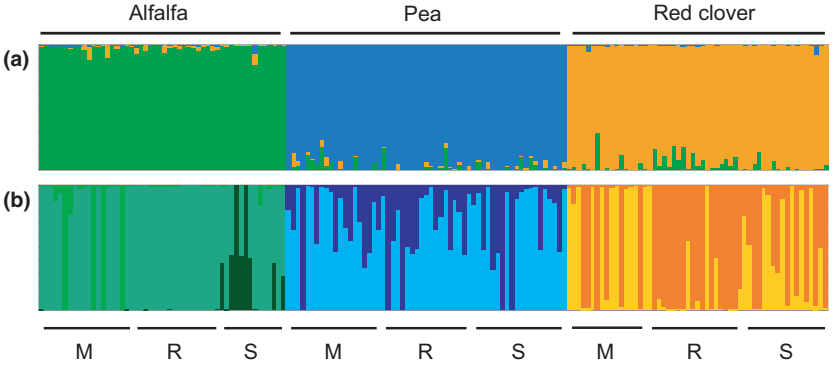

Fig. 1 Clustering of pea aphids by STRUCTURE analyses for $K = 3$ clusters (panel A). Each individual is represented by a vertical line, divided into up to $K = 3$ coloured segments representing the individual's ancestry to each cluster. Panel B: genetic substructure of each of the three clusters identified in A) (with $K = 3$ for each). Host plants and localities are respectively labelled on the top and bottom of the figure (M = Mirecourt, R = Ranspach, S = Switzerland).

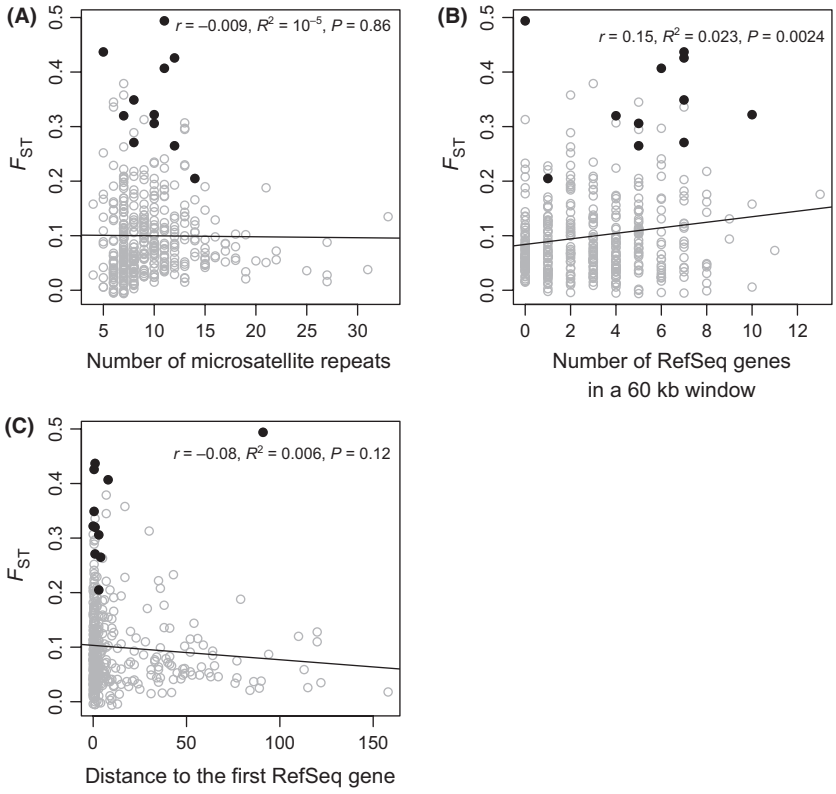

Fig. 2 Effects of the number of microsatellite repeats (panel A) and of the genomic environment of the microsatellite loci (panels B and C) on the genetic differentiation (F_{ST}) tested with Pearson's correlations. Loci identified as significant outliers at $\alpha < 0.01$ with ARLEQUIN 3.5 are shown in black.

Table 2 List of the 11 outlier loci detected with ARLEQUIN 3.5 at $\alpha = 0.01$. F_{CT} values (and significance level as outlier) are provided for the analysis over all three races and for analyses are conducted between pairs of host races. In these analyses, geographical populations were nested within host races. The ranking of these F_{CT} values is given in bracket (1 = most differentiated locus, 390 = less differentiated locus, with pairs ranked among the 15% highest F_{CT} showed in bold)

	Overall hierarchical analysis	Pairwise hierarchical analyses between host races		
Locus ID	Red clover *vs.* alfalfa *vs.* pea F_{CT}, significance	Red clover *vs.* alfalfa F_{CT}, (rank), significance	Alfalfa *vs.* pea F_{CT}, (rank), significance	Pea *vs.* red clover F_{CT}, (rank), significance
EQ118272_44	0.54***	0.084 (96) ns	0.75 (**1**)***	0.53 (**7**)*
T_125804_1	0.52***	0.078 (104) ns	0.21 (**11**)**	0.63 (**3**)***
Q_116274_17	0.50**	−0.016 (383) ·	0.63 (**4**)**	0.66 (**2**)**
T_112827_1	0.50**	0.004 (328)*	0.57 (**8**)*	0.67 (**1**)**
T_121287_2	0.43**	0.12 (**57**) ns	0.68 (**2**)**	0.37 (**26**) ns
D_114374_2	0.41***	0.38 (**1**)***	0.47 (**15**)**	0.37 (**25**)*
D_128625_1	0.40**	0.103 (72) ns	0.44 (**19**) ·	0.60 (**5**)**
D_115338_1	0.40**	0.16 (**33**) ns	0.31 (**41**) ns	0.62 (**4**)*
T_110814_1	0.36**	0.22 (**14**)**	0.23 (75) ns	0.55 (**6**)**
D_121300_2	0.35**	0.053 (147) ns	0.55 (**9**)**	0.44 (**15**) ·
AlB08M	0.25**	0.17 (**26**)**	0.28 (**53**) ns	0.30 (**46**) ·

***$P < 0.001$. **$P < 0.01$, *$P < 0.05$, · $P < 0.1$, ns $P \geq 0.1$.

comparisons) from the hierarchical analyses run between pairs of host races (Table 2). The only exception concerned the marker D_115338_1 ($P = 0.013$ in pea *vs.* red clover comparison). The results for all loci are provided in Table S3 (Supporting information). On the basis of their F_{CT} values, ranking and significance as outlier in pairwise analyses, it appeared that the pea race largely contributed to the outlier status of several of these 11 loci. When ARLEQUIN 3.5 was run within each of the three host races, we detected four loci (of 390) departing from the neutral expectation (at $\alpha = 0.01$) in the alfalfa race, five in the clover race and nine in the pea race (see Table S3, Supporting information). There was only one locus in common between this set of 18

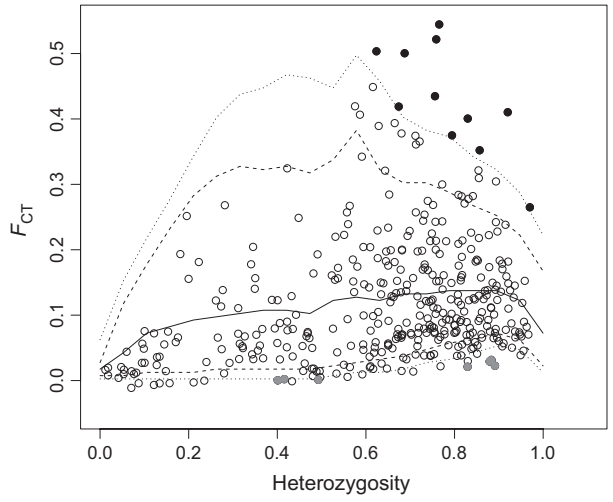

Fig. 3 Genetic differentiation (F_{CT}) among host races of the pea aphid (when controlling for within host plant structure) as a function of heterozygozity for each of the 390 microsatellite loci estimated with ARLEQUIN 3.5. Plain line, median; dashed lines, 5th and 95th quantiles of the neutral envelope; dotted lines, 1st and 99th quantiles; circles, nonoutlier loci; black dots, outlier loci putatively involved in divergence between host races; grey dots, outlier loci putatively under balancing selection.

within-host races outliers and the set of 11 between-host races outliers. This locus (EQ118272_44) departed from neutrality within the clover race, with $F_{ST} = 0.12$ ($P = 0.007$). Nevertheless, this locus remained a significant outlier between host races when the analyses were performed between alfalfa and pea races (i.e. when clover populations were excluded, Table 2).

AMOVA analyses performed separately on the set of 11 outliers and that of nonoutlier loci showed that in both cases, populations were significantly ($P < 0.01$) more structured by the host plant ($F_{CT\ outliers} = 0.40$, $F_{CT\ nonoutliers} = 0.12$) than by the geography ($F_{SC\ outliers} = 0.02$, $F_{SC\ nonoutliers} = 0.01$). Similar results emerged from STRUCTURE analyses (see Fig. S4, Supporting information).

Genic environment of outlier loci

More than 55% of the 73 genes near outlier loci corresponded to genes with unknown functions, owing to the absence of homology with any other genes in public databases (Table S4, Supporting information). Two of the outliers (D_115338_1 and T_110814_1) were close to the genes corresponding to the olfactory receptors Or12 and Or22, respectively (Smadja *et al.* 2009, 2012). As we specifically developed some microsatellite loci close to chemosensory genes, it might be that this gene function is overrepresented in the outlier category just by chance. Indeed, this proportion (2 of 11) was not significantly different from that measured for nonoutlier loci (23 of 372; test of proportion: $P = 0.33$). Interestingly, two other outlier loci (D_121300_2, Q_116274_17) were located in the vicinity of three genes encoding salivary proteins (Bos *et al.* 2010; Carolan *et al.* 2011). D_121300_2 is, indeed, located at 27 kb of gene ACYPI089376 (gene name on gene prediction 2.1), which corresponds to part of the gene ACYPI002172 (previous gene name on gene prediction 1.2) identified in the salivary secretome of the pea aphid (Carolan *et al.* 2011). Another gene, ACYPI54712, distant from 4.5 kb from D_121300_2 is homolog to a protein found in the salivary glands of the green peach aphid *Myzus persicae* (Bos *et al.* 2010). The outlier locus Q_116274_17 was also found to be at 12 kb from the salivary protein ACYPI000472 (that conserved the same gene ID on the 1.2 and 2.1 gene predictions). The putative function of these three genes (ACYPI089376, ACYPI54712 and ACYPI000472) is unknown. Another outlier (AlB08M) was close to a cadherin gene (ACYPI073870), a gene involved in calcium binding, a function expressed in salivary glands too (Will *et al.* 2007; Carolan *et al.* 2011; Hogenhout & Bos 2011). We also found three genes interacting with ubiquitin (ACYPI086543 and ACYPI005897 near outlier T_110814_1, and ACYPI47822 near outlier D_114374_2), a function often found in salivary proteins (e.g. Bos *et al.* 2010; Carolan *et al.* 2011). Hypothetical functions of other genes located at less than 30 kb from outliers include protein binding, regulator of transcription, oxydo-reduction process amongst other functions, with no obvious role in host plant adaptation (Table S4, Supporting information).

Genomic architecture of host plant adaptation

The 11 outlier loci were located on 11 distinct scaffolds, five of which also containing some nonoutlier microsatellite loci (Table 3). The physical distance between outliers and these neighbouring markers ranged from 49 to 1067 kb. Only one locus, located at 49.8 kb from the outlier locus T_112827_1, showed high genetic divergence between host races ($F_{CT} = 0.30$) and was identified as an outlier at the $\alpha = 0.05$ significance threshold in our analyses with ARLEQUIN 3.5. By contrast, two loci located at approximately 170 kb from outliers showed extremely low F_{CT} values (Table 3). In spite of the physical proximity between these loci based on the V2 genome assembly, we found no significant linkage disequilibrium between outlier loci and the neighbouring markers, except for outlier D_121300_2 and one of its neighbour (D_126141_4) distant from 1067 kb (Table 3).

Discussion

Genomic structure of pea aphid populations

Our population genetic analyses based on an extensive set of microsatellite loci revealed a strong genetic

Table 3 Microsatellite loci physically linked to outlier loci from the V2 genome assembly of the pea aphid. Genetic differentiation between host races (F_{CT}, estimated with ARLEQUIN 3.5 in a hierarchical model where geographical populations are nested within the three host races), the ranking of the loci based on this value (1 = most differentiated locus, 390 = less differentiated locus) and the significance level as outlier are shown. Correlations between allelic frequencies between focal outliers and linked loci (and significance) are also shown

						Structuration			Linkage disequilibrium	
Focal outlier	Genomic location (V2): Scaffold and position	F_{CT}	Loci physically linked based on the V2 genome assembly	Genomic location (V2): Scaffold and position	Distance to the focal outlier (kb)	F_{CT}	Rank	P	R^2	P
D_128625_1	GL349727: 418833	0.40	D_112905_1	GL349727: 831826	413.0	0.19	77	0.11	0.039	0.44
T_112827_1	GL349916: 186700	0.50	D_122425_3	GL349916:136891	49.8	0.30	26	0.034	0.083	0.10
T_110814_1	GL349896: 271058	0.36	D_117281_4	GL349896: 209685	61.4	0.08	216	0.18	0.055	0.24
D_121300_2	GL349691: 1174046	0.35	D_126141_4	GL349691: 106832	1067.2	0.21	62	0.16	0.144	0.0001
			T_127540_1	GL349691: 165594	1008.5	0.19	79	0.20	0.076	0.81
			126536_5	GL349691: 1339346	165.3	0.05	308	0.04	0.091	0.13
			D_114323_1	GL349691: 1584993	410.9	0.18	92	0.26	0.073	0.55
118272_44	GL349996: 392509	0.54	10433	GL349996: 218859	173.6	0.022	349	0.05	0.084	0.15

structure by the host plant all along the nuclear genome of the pea aphid. Contrastingly, the impact of geography within a host race was negligible although the sampled populations were separated by 100–200 km. Among the three host races studied here, the pea host race was the most differentiated, as already shown with a much smaller set of loci (Frantz *et al.* 2006; Peccoud *et al.* 2009a). The red clover and alfalfa host races were less divergent from each other. The pea adapted race might have diverged before the two other races started to, although this is not supported by sequence data of the maternally inherited aphid endosymbiont *Buchnera aphidicola* (Peccoud *et al.* 2009b). This pattern of differentiation could also be due to a stronger reproductive isolation, whether pre- and/or postzygotic, in the pea race. Migrants from other races are, indeed, very rare on pea, and this race shows very few hybrids in field populations (~3%), against 4 to 9% for other races within the *A. pisum* complex (Peccoud *et al.* 2009a). The pea race also differed from the other two by a lower genetic diversity, an absence of within-race genetic structure and an excess of heterozygotes. The pea-adapted race uses annual pea and annual vetches as alternative hosts, which are both unstable resources that might cause strong fluctuations in aphid population sizes. This may explain the reduced genetic diversity and the negative F_{IS} (possibly because of low founding numbers of males and females) within the pea race populations. Instable resources might favour greater dispersal abilities compared to clover and alfalfa races as these two feed on perennial plants (Frantz *et al.* 2009, 2010). Lower constraints on dispersal in clover and alfalfa host plants would account for the greater geographical differentiation between populations adapted to these plants and might favour local inbreeding, resulting in the observed homozygote excess observed in some populations of the clover race.

Interestingly, the genomic environment of our microsatellite loci affected both genetic diversity and structure of pea aphid populations. After correcting for the number of repeats, which is known to influence diversity (e.g. Ellegren 2000), we found a lower diversity and an increased genetic differentiation for microsatellite loci located in genomic regions with high gene density than for those located in regions with low gene density. These patterns are expected under genetic hitchhiking, with a reduction of diversity in the vicinity of the genomic regions targeted by selection, along with an increased divergence between environments selecting for different optima (Maynard-Smith & Haigh 1974; Barton 2000).

Genomic regions under divergent selection

Our genome scan approach identified 11 outlier loci (at the $\alpha = 0.01$ threshold) when the hierarchical analysis

was performed on the three races simultaneously. Hierarchical analyses performed between pairs of host races also supported these outliers. Whether the outlier locus EQ118272_44 is influenced by the host plant or local (geographical) effects remains unclear, *a fortiori* as this locus is far from any predicted gene.

All 11 outliers were characterized by a high heterozygosity, and there are two reasons for this pattern. First, heterozygozity is estimated overall populations in ARLEQUIN analyses, so the larger the divergence among populations, the higher overall heterozygosity can reach. Second, the power to detect outliers increases with increasing variance in allelic frequencies. Note that similar results have been reported in earlier works (e.g. Fig. 1 in Beaumont & Nichols 1996; and Smadja *et al.* 2012).

We showed that both outlier loci and nonoutlier loci were more structured by the host plant (outliers: $F_{CT} = 0.40$, nonoutliers: $F_{CT} = 0.12$) than by the geography (outliers: $F_{SC} = 0.02$, nonoutliers: $F_{SC} = 0.01$, Fig. S4, Supporting information). If the populations sampled from the same host plant derived from the same specialization event, we would then expect these populations to cluster together in analyses based on nonoutlier loci. Conversely, if the populations sampled from the same host plant derived from independent and/or recurrent specialization events (i.e. parallel adaptation), we would then expect some populations at least to cluster into geographic groups (see, for example, Colosimo *et al.* 2005). Our results support the hypothesis that adaptation to each host plant occurred once in the populations studied here, as *A. pisum* populations cluster together into host plant groups and not into geographic groups. Similar results were previously documented at larger geographical scales (Peccoud *et al.* 2008, 2009a; Ferrari *et al.* 2012). The pea aphid complex, thus, differs from the well-documented case of parallel adaptation in sticklebacks (Colosimo *et al.* 2005; Hohenlohe *et al.* 2010), presumably because aphids have strong dispersal abilities (Loxdale *et al.* 1993; Margaritopoulos *et al.* 2009).

The large genetic differentiation found here between the three host races of *A. pisum* at both outlier and nonoutlier loci contrasts with the results of Via & West's (2008), showing that American pea aphid populations adapted to red clover and alfalfa were not structured (neither geographically nor by their hosts) at nonoutlier loci. The most probable explanation for these opposing results relies on how outlier and nonoutlier loci were defined here and in Via & West (2008). The latter authors combined a QTL map for plant specialization (Hawthorne & Via 2001) with data on genetic structure between two host races on mapped loci. They showed that the QTLs affected genetic markers located up to 20–30 cM far (as differentiation was on average higher at these loci). They identified only five AFLP markers (of 45) that did not hitchhike with the QTLs and estimated genetic structure between host races with these five markers. Contrastingly, here, we identified loci as outlier only if they were highly significant with ARLEQUIN 3.5 ($\alpha = 0.01$), and all remaining loci were grouped in the nonoutlier class. If, as suggested by Via & West (2008), most markers are affected by QTLs controlling host plant specialization, this could explain why we observe an important genetic structure even on the loci that we classified as nonoutliers (see STRUCTURE and AMOVA's results). Thus, we are currently unable to resolve whether the large effect of the host plant on *A. pisum* genetic structure on both the outlier and the nonoutlier groups of loci results from drift at neutral loci as a consequence of adaptive divergence, or from hitchhiking of large genomic portions with selected genes because of low recombination in some genomic regions. Having a high-quality genetic map in the pea aphid would certainly help discriminating between these hypotheses, as they predict different distributions of F_{ST} values for nonoutlier loci across the chromosomes. Under drift at neutral loci resulting from adaptive divergence, we expect a random distribution of F_{ST} values from the nonoutlier group of loci across chromosomes, while the latter hypothesis predicts an aggregated distribution of markers according to their F_{ST} values, with only a few genomic areas (those far from any QTL) showing low differentiation between host races. Whether the outliers we identified (that each belongs to a different scaffold) map to the genomic regions identified through the QTL analysis of host acceptance and performance by Hawthorne & Via (2001) remains an open question, as basic information about the molecular markers used in their study was not provided.

Here, we chose many markers close to expressed genes and focused on outliers detected using a stringent threshold to increase the probability to target genomic regions of adaptive significance. Hence, the frequency of outliers (11 of 390) might not reflect the frequency of randomly drawn markers. Furthermore, if there are large differences in the rate of recombination in different regions of the aphid genome, in particular if strong hitchhiking effects occur in some genomic areas, this could lead to a biased estimation of the frequency of outliers. We also acknowledge that the number of markers used here is still low compared with the genome size of the pea aphid (530 Mb, IAGC 2010) and that we have probably only picked out a small part of the genomic regions (or genes) involved in host plant specialization.

Candidate genes involved in plant adaptation

The majority of the genes located at less than 30 kb of the outlier loci have unknown functions and showed no homology with genes from other model organisms. However, a few genes with functions likely to be involved in plant specialization were at 30 kb or less from several of the 11 outlier microsatellites, suggesting that this arbitrarily chosen threshold was reasonable. In particular, two outliers were located in the vicinity of three genes recently identified as encoding secreted proteins from the salivary glands of the pea aphid and the green peach aphid (Carolan *et al.* 2011). A third outlier was close to a cadherin gene, which encodes calcium-binding proteins and which has a copy encoding secreted salivary proteins in the pea aphid (Carolan *et al.* 2011). Salivary proteins are important effectors of aphid–plant interactions, which can suppress actively plant defences (Will *et al.* 2007; Bos *et al.* 2010; Hogenhout & Bos 2011). In particular, calcium-binding proteins secreted by salivary glands, and which are injected into the sieve tube following insertion of aphid stylets, can counteract plant defences by clogging of sieve elements and callose formation (Will *et al.* 2007). The fact that three of the genes previously identified by Carolan *et al.* (2011) and Bos *et al.* (2010) locate close to two of our outliers and that we also identified a cadherin gene close to a third outlier strongly suggests that these salivary proteins play a role in plant use by the different host races.

We identified two other outliers close to chemosensory genes Or12 and Or22 (at 6.7 kb and 0.7 kb, respectively) and showed that this gene function was not significantly enriched in the outlier category. Our results differ from a recent study by Smadja *et al.* (2012), based on sequence data from populations of eight individuals each, which suggested that Or12 and Or22 did not evolve under divergent selection in *A. pisum* host races collected in UK on red clover, alfalfa and lotus (*Lotus pedunculatus*). Conversely, seven of our microsatellite loci were located in the vicinity (<14 kb) of four of the 19 chemosensory genes identified as outliers by Smadja *et al.* (2012) (d_127119_2 and T_127119_1: 0.5kb and 11.5 kb from Gr8, D_119783_3: 10.5 kb from Or15, t_125317_2 and D_112775_1: 9.5 kb and 13.6kb from Or29, D_128051_6 and 10114: 2.5 kb and 3.1kb from Or36), but none of these microsatellite loci were identified as outlier in the present study. The differences between our results and those in Smadja *et al.* (2012) may have several causes. First, the two studies compared different sets of host races, and it is possible that different chemosensory genes are involved in host plant specialization in the different host races. Second, genomic differentiation along chromosomes is noisy (e.g. Weir *et al.* 2005; Cao *et al.* 2011; Nadeau *et al.* 2012), so that differentiation observed at a particular locus might not reflect that of neighbouring loci. Third, our microsatellite loci are possibly too far from chemosensory genes identified as under divergent selection by Smadja *et al.* (2012) to be influenced by hitchhiking effect. This hypothesis nevertheless contradicts results of Via & West (2008), suggesting that neutral markers are influenced by host-specialization QTL over large distance. Furthermore, two of our microsatellite loci were close from Gr8 and Or36 (0.5 and 2.5 kb), so they would have hitchhiked if these genes had undergone strong directional selection. Fourth, Smadja *et al.* (2012) analysed eight individuals from a single population. Estimating genetic structure with less than 15 diploid individuals is not recommended, as allelic frequencies might be poorly estimated (Kalinowski 2005).

Conclusion

By scanning the genome of multiple pea aphids adapted to distinct host plants, we confirmed the primary influence of plant specialization on overall genetic differentiation. Considering that the adaptive radiation in the pea aphid complex is recent (Peccoud *et al.* 2009b), our results suggest a rapid genome-wide divergence promoted by ecological factors and reproductive barriers between host-adapted races. Our approach also allowed identifying with a strong statistical support a handful of loci under divergent selection. Whether differentiation among host races at nonoutlier loci results from drift as a consequence of adaptive divergence or from extensive genetic hitchhiking remains to be investigated. Nearly half of the outliers we identified were close to the genes of biological relevance in the context of molecular plant–aphid interactions. This is an encouraging result for further studies on population and functional genomics of plant adaptation and ecological speciation in the pea aphid complex, but more generally for the discovery of candidate genes controlling adaptive traits. There is now a need to improve our knowledge about the physical organization of the pea aphid genome to localize the outlier loci as well as to characterize the extent of genomic hitchhiking, in particular in relation to the recombination rate of different genomic regions. Investigating the occurrence of chromosomal rearrangements among host races is also highly relevant, as this could account for the high diversification rate observed in the pea aphid complex. Finally, functional studies on the candidate genes identified here and in the related work of Smadja *et al.* (2012) are required to determine their role in plant use by pea aphid races.

Acknowledgements

We thank L. Buechi, A. Juilland, G. Evanno, Y. Outreman, the Jaquiéry and Evanno families for help with sampling and J. Bonhomme for rearing aphids. J. Peccoud, L. Orsini, and three anonymous referees contributed constructive comments on a previous draft of this manuscript. This work was supported by INRA-AIP BioRessources, the Département INRA Santé des Plantes et Environnement, the Fondation pour la Recherche sur la Biodiversité, the ANR GW-Aphid, the ANR Speciaphid and the Swiss National Science Foundation (grants number PBLAA-122658 and PA00P3_139720 to J.J).

References

Barton NH (2000) Genetic hitchhiking. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, **355**, 1553–1562.

Bates D, Sarkar D (2007) lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-28. Available from: http://lme4.r-forge.r-project.org.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London Series B-Biological Sciences*, **263**, 1619–1626.

Berlocher SH, Feder JL (2002) Sympatric speciation in phytophagous insects: moving beyond controversy? *Annual Review of Entomology*, **47**, 773–815.

Bierne N (2010) The distinctive footprints of local hitchhiking in a varied environment and global hitchhiking in a subdivided population. *Evolution*, **64**, 3254–3272.

Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, **20**, 2044–2072.

Bolnick DI, Fitzpatrick BM (2007) Sympatric speciation: models and empirical evidence. *Annual Review of Ecology, Evolution and Systematics*, **38**, 459–487.

Bonin A, Taberlet P, Miaud C, Pompanon F (2006) Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). *Molecular Biology and Evolution*, **23**, 773–783.

Bos JIB, Prince D, Pitino M, Maffei ME, Win J, Hogenhout SA (2010) A functional genomics approach identifies candidate effectors from the aphid species *Myzus persicae* (green peach aphid). *Plos Genetics*, **6**, e1001216.

Caillaud MC, Mondor-Genson G, Levine-Wilkinson S *et al.* (2004) Microsatellite DNA markers for the pea aphid *Acyrthosiphon pisum*. *Molecular Ecology Notes*, **4**, 446–448.

Cao J, Schneeberger K, Ossowski S *et al.* (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genetics*, **43**, 956–U960.

Carolan JC, Caragea D, Reardon KT *et al.* (2011) Predicted effector molecules in the salivary secretome of the pea aphid (*Acyrthosiphon pisum*): a dual transcriptomic/proteomic approach. *Journal of Proteome Research*, **10**, 1505–1518.

Cavalli-Sforza LL (1966) Population structure and human evolution. *Proceedings of the Royal Society of London B - Biological Sciences*, **164**, 362–379.

Colosimo PF, Hosemann KE, Balabhadra S *et al.* (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science*, **307**, 1928–1933.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Dres M, Mallet J (2002) Host races in plant-feeding insects and their importance in sympatric speciation. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **357**, 471–492.

Ellegren H (2000) Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in Genetics*, **16**, 551–558.

Ellegren H, Sheldon BC (2008) Genetic basis of fitness differences in natural populations. *Nature*, **452**, 169–175.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes - application to human mitochondrial-DNA restriction data. *Genetics*, **131**, 479–491.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*, **103**, 285–298.

Ferrari J, Godfray HCJ, Faulconbridge AS, Prior K, Via S (2006) Population differentiation and genetic variation in host choice among pea aphids from eight host plant genera. *Evolution*, **60**, 1574–1584.

Ferrari J, Via S, Godfray HCJ (2008) Population differentiation and genetic variation in performance on eight hosts in the pea aphid complex. *Evolution*, **62**, 2508–2524.

Ferrari J, West JA, Via S, Godfray HCJ (2012) Population genetic structure and secondary symbionts in host-associated populations of the pea aphid complex. *Evolution*, **66**, 375–390.

Foll M, Gaggiotti OE (2008) A genome scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, **180**, 977–993.

Frantz A, Plantegenest M, Mieuzet L, Simon JC (2006) Ecological specialization correlates with genotypic differentiation in sympatric host-populations of the pea aphid. *Journal of Evolutionary Biology*, **19**, 392–401.

Frantz A, Calcagno V, Mieuzet L, Plantegenest M, Simon J-C (2009) Complex trait differentiation between host-populations of the pea aphid *Acyrthosiphon pisum* (Harris): implications for the evolution of ecological specialisation. *Biological Journal of the Linnean Society*, **97**, 718–727.

Frantz A, Plantegenest M, Simon JC (2010) Host races of the pea aphid *Acyrthosiphon pisum* differ in male wing phenotypes. *Bulletin of Entomological Research*, **100**, 59–66.

Goudet J (2005) FSTAT, A Program to Estimate and Test Gene Diversities and Fixation Indices. Available at: http://www2.unil.ch/popgen/softwares/fstat.htm, Lausanne.

Hawthorne DJ, Via S (2001) Genetic linkage of ecological specialization and reproductive isolation in pea aphids. *Nature*, **412**, 904–907.

Hogenhout SA, Bos JIB (2011) Effector proteins that modulate plant-insect interactions. *Current Opinion in Plant Biology*, **14**, 422–428.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *Plos Genetics*, **6**, e1000862.

Hunter S, Apweiler R, Attwood TK *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Research*, **37**, D211–D215.

IAGC (2010) Genome sequence of the pea aphid *Acyrthosiphon pisum*. *Plos Biology*, **8**, e1000313.

Jaquiéry J, Stoeckel S, Rispe C, Mieuzet L, Legeai F, Simon JC (2012) Accelerated evolution of sex chromosomes in aphids, an X0 system. *Molecular Biology and Evolution*, **29**, 837–847.

Kalinowski ST (2005) Do polymorphic loci require large sample sizes to estimate genetic distances? *Heredity*, **94**, 33–36.

Legeai F, Shigenobu S, Gauthier JP *et al.* (2010) AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome. *Insect Molecular Biology*, **19**, 5–12.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.

Loxdale HD, Hardie J, Halbert S, Foottit R, Kidd NAC, Carter CI (1993) The relative importance of short-range and long-range movement of flying aphids. *Biological Reviews of the Cambridge Philosophical Society*, **68**, 291–311.

Mallet J (2008) Mayr's view of Darwin: was Darwin wrong about speciation? *Biological Journal of the Linnean Society*, **95**, 3–16.

Margaritopoulos JT, Kasprowicz L, Malloch GL, Fenton B (2009) Tracking the global dispersal of a cosmopolitan insect pest, the peach potato aphid. *BMC Ecology*, **9**, 13.

Maynard-Smith J, Haigh J (1974) Hitch-hiking effect of a favorable gene. *Genetical Research*, **23**, 23–35.

Michel AP, Sim S, Powell THQ, Taylor MS, Nosil P, Feder JL (2010) Widespread genomic divergence during sympatric speciation. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 9724–9729.

Nadeau NJ, Jiggins CD (2010) A golden age for evolutionary genetics? Genomic studies of adaptation in natural populations. *Trends in Genetics*, **26**, 484–492.

Nadeau NJ, Whibley A, Jones RT *et al.* (2012) Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **367**, 343–353.

Narum SR, Hess JE (2011) Comparison of F_{ST} outlier tests for SNP loci under selection. *Molecular Ecology Resources*, **11** (Suppl. 1), 184–194.

Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Peccoud J, Simon JC (2010) The pea aphid complex as a model of ecological speciation. *Ecological Entomology*, **35**, 119–130.

Peccoud J, Figueroa CC, Silva AX *et al.* (2008) Host range expansion of an introduced insect pest through multiple colonizations of specialized clones. *Molecular Ecology*, **17**, 4608–4618.

Peccoud J, Ollivier A, Plantegenest M, Simon JC (2009a) A continuum of genetic divergence from sympatric host races to species in the pea aphid complex. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 7495–7500.

Peccoud J, Simon JC, McLaughlin JH, Moran N (2009b) Post-pleisctocene radiation of the pea aphid complex reveals by rapidly evolving endosymbionts. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 16315–16320.

Peccoud J, Simon J-C, von Dohlen C *et al.* (2010) Evolutionary history of aphid-plant associations and their role in aphid diversification. *Comptes Rendus Biologies*, **333**, 474–487.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

R Development Core Team (2012) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Roesti M, Hendry AP, Salzburger W, Berner D (2012) Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickelback population pairs. *Molecular Ecology*, **21**, 2852–2862.

Rundle HD, Nosil P (2005) Ecological speciation. *Ecology Letters*, **8**, 336–352.

Sabater-Munoz B, Legeai F, Rispe C *et al.* (2006) Large-scale gene discovery in the pea aphid *Acyrthosiphon pisum* (Hemiptera). *Genome Biology*, **7**, R21.

Schluter D (2001) Ecology and the origin of species. *Trends in Ecology & Evolution*, **16**, 372–380.

Smadja C, Shi P, Butlin RK, Robertson HM (2009) Large gene family expansions and adaptive evolution for odorant and gustatory receptors in the pea aphid, *Acyrthosiphon pisum*. *Molecular Biology and Evolution*, **26**, 2073–2086.

Smadja C, Canbäck B, Vitalis R *et al.* (2012) Large-scale candidate gene scan reveals the role of chemoreceptor genes in host plant specialisation and speciation in the pea aphid. *Evolution*, **66**, 2723–2738.

Stapley J, Reger J, Feulner PGD *et al.* (2010) Adaptation genomics: the next generation. *Trends in Ecology & Evolution*, **25**, 705–712.

Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology*, **14**, 671–688.

Storz JF, Nachman MW (2003) Natural selection on protein polymorphism in the rodent genus *Peromyscus*: evidence from interlocus contrasts. *Evolution*, **57**, 2628–2635.

Vasemagi A, Nilsson J, Primmer CR (2005) Expressed sequence tag-linked microsatellites as a source of gene-associated polymorphisms for detecting signatures of divergent selection in Atlantic salmon (*Salmo salar* L.). *Molecular Biology and Evolution*, **22**, 1067–1076.

Via S (1991) The genetic structure of host plant adaptation in a spatial patchwork - demographic variability among reciprocally transplanted pea aphid clones. *Evolution*, **45**, 827–852.

Via S (1999) Reproductive isolation between sympatric races of pea aphids. I. Gene flow restriction and habitat choice. *Evolution*, **53**, 1446–1457.

Via S (2001) Sympatric speciation in animals: the ugly duckling grows up. *Trends in Ecology & Evolution*, **16**, 381–390.

Via S (2009) Natural selection in action during speciation. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 9939–9946.

Via S, West J (2008) The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Molecular Ecology*, **17**, 4334–4345.

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.
Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG (2005) Measures of human population structure show heterogeneity among genomic regions. *Genome Research*, **15**, 1468–1476.
Will T, Tjallingii WF, Thonnessen A, van Bel AJE (2007) Molecular sabotage of plant defense by aphid saliva. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 10536–10541.
Wu CI (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851–865.
Zhou JJ, Vieira FG, He XL *et al.* (2010) Genome annotation and comparative analyses of the odorant-binding proteins and chemosensory proteins in the pea aphid *Acyrthosiphon pisum*. *Insect Molecular Biology*, **19**, 113–122.

This works represents a contribution to the understanding of trophic adaptation in herbivores from an evolutionary and genetic perspective. J.J., S.S. and R.V. are population geneticists interested in evolutionary biology. P.N. participated to different steps of this study and is currently doing a PhD thesis on the genetics of plant adaptation in the pea aphid under the supervision of J.C.S. and S.S. N.B. and A. B. contributed to data acquisition and analysis. L. M. and F.M. are experts in high-throughput genotyping technologies. F. L. is a bioinformatician focusing on insect genomics, genome assembly and genomic analyses. J.C.S. is a scientist with long-standing interest in the evolution of complex life-history traits, using aphids as model systems.

Data accessibility

Microsatellite genotypes and sampling information are deposited in DRYAD along with input files for running Arlequin and Structure softwares (doi:10.5061/dryad.pf5cg).

Supporting information

Additional Supporting Information may be found in the online version of this article.

Fig. S1 Description of the different steps for selecting the individuals used for extensive genotyping in each of the nine populations sampled.

Fig. S2 Pairwise F_{ST} estimates for populations and host races of the pea aphid and computed across 390 microsatellite markers.

Fig. S3 Effects of the number of microsatellite repeats (panel A) and of the genomic environment of the microsatellite loci (panels B and C) on residuals of allelic richness from linear-mixed effects models.

Fig. S4 Clustering of pea aphids by STRUCTURE analyses for $K = 3$ clusters.

Table S1 Origin of populations used (*M. sativa*: alfalfa, *T. pratense*: red clover, *P. sativum*: pea). S = Switzerland, M = Mirecourt, R = Ranspach.

Table S2 Microsatellite loci used.

Table S3 Identification of outlier loci with the hierarchical method implemented in ARLEQUIN 3.5 (Excoffier & Lischer 2010).

Table S4 Putative functions of the predicted genes located in a 60-kb windows centered on the outlier microsatellite loci (detected with ARLEQUIN 3.5) on the V2 genome assembly (IAGC 2010; Legeai *et al.* 2010).

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.