This is a repository copy of *Targeted re-sequencing confirms the importance of chemosensory genes in aphid host race differentiation.*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/104198/

Version: Accepted Version

Article:
Eyres, I. orcid.org/0000-0002-8413-262X, Duvaux, L., Gharbi, K. et al. (6 more authors) (2017) Targeted re-sequencing confirms the importance of chemosensory genes in aphid host race differentiation. Molecular Ecology, 26 (1). pp. 43-58. ISSN 0962-1083

https://doi.org/10.1111/mec.13818

This is an open access article under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Targeted re-sequencing confirms the importance of chemosensory genes in aphid host race differentiation

Isobel Eyres[1$], Ludovic Duvaux[1], Karim Gharbi[3], Rachel Tucker[1], David Hopkins[1], Jean-Christophe Simon[2], Julia Ferrari[4]*, Carole M. Smadja[5]*, Roger K. Butlin[1]*

* Joint last authors

1. Department of Animal and Plant Sciences, University of Sheffield, Sheffield, United Kingdom
2. INRA, Institut de Génétique, Environnement et Protection des Plantes, UMR 1349 IGEPP, Domaine de la Motte, 35653 Le Rheu Cedex, France
3. Edinburgh Genomics, Ashworth Laboratories, University of Edinburgh, Edinburgh, United Kingdom
4. Department of Biology, University of York, York, United Kingdom
5. Institut des Sciences de l'Evolution (UMR 5554 CNRS-IRD-CIRAD-Université de Montpellier), cc065, place Bataillon, Campus Triolet, Université de Montpellier, 34095 Montpellier cedex 05, France

Keywords: adaptation, speciation, *Acyrthosiphon pisum*, chemosensory genes, targeted resequencing, genome scan.

[$]Corresponding author
Isobel Eyres, Department of Animal and Plant Sciences, Alfred Denny Building, University of Sheffield, Western Bank, Sheffield, S10 2TN, UK. Email: i.eyres@sheffield.ac.uk

Abstract

Host-associated races of phytophagous insects provide a model for understanding how adaptation to a new environment can lead to reproductive isolation and speciation, ultimately enabling us to connect barriers to gene flow to adaptive causes of divergence. The pea aphid (*Acyrthosiphon pisum)* comprises host-races specialising on legume species, and provides a unique system for examining the early stages of diversification along a gradient of genetic and associated adaptive divergence. As host-choice produces assortative mating, understanding the underlying mechanisms of choice will contribute directly to understanding of speciation. As host-choice in the pea aphid is likely mediated by smell and taste, we use capture sequencing and SNP genotyping to test for the role of chemosensory genes in the divergence between eight host-plant species across the continuum of differentiation and sampled at multiple locations across western Europe. We show high differentiation of chemosensory loci relative to control loci in a broad set of pea aphid races and localities, using a model-free approach based on Principal Component analysis. Olfactory and gustatory receptors form the majority of highly differentiated genes, and include loci that were already identified as outliers in a previous study focusing on the three most closely related host races. Consistent indications that chemosensory genes may be good candidates for local adaptation and barriers to gene flow in the pea aphid open the way to further investigations aiming to understand their impact on gene flow, and to determine their precise functions in response to host plant metabolites.

Introduction

Speciation depends on the evolution of barriers to gene flow, and natural selection is now considered to be an important driver in this process (Kirkpatrick & Ravigné 2002; Nosil *et al.* 2002; Via 2009; Nosil 2012); local adaptation can lead to reproductive isolation in the face of gene flow. Contact between populations that have undergone some divergence through selection or geographical isolation is a common occurrence, and the opportunity that this provides for gene flow may cause breakdown of the initial divergence. An important challenge in current speciation research is therefore to understand how lineages can maintain differentiation and progress towards speciation despite ongoing gene exchange (Smadja & Butlin 2011).

With the exception of polyploidy, speciation tends to be a long process, requiring the progressive buildup of reproductive isolation (Abbott *et al.* 2013). Where lineages are undergoing ecological speciation in the face of gene flow, reproductive isolation can start with the action of divergent selection on locally adaptive loci. This initial divergence may then be facilitated by the association of local adaptation and assortative mating, by close linkage in the genome, by pleiotropy or where the same trait influences both components of isolation (Felsenstein 1981; Servedio 2008; Smadja & Butlin 2011). There is then the possibility that initially divergent genome regions will expand over time as gene flow diminishes between lineages (Feder *et al.* 2012).

One way to study the progress of barriers to gene flow and their role in contributing to speciation is to identify candidate loci in populations experiencing divergence based on local adaptation, early in the process of speciation, and to track the action of selection across a continuum of divergence through space or time (Jones *et al.* 2012; Martin *et al.* 2013). In order to do this effectively, we must be able to identify loci involved in the initial local adaptation with confidence. While many studies have identified highly differentiated loci that are potentially under divergent selection, successful follow-up studies to outlier scans have rarely been achieved (Rogers &

87 Bernatchez 2005, 2007; Wood *et al.* 2008; Butlin 2010; Jones *et al.* 2012; Malinsky *et*
88 *al.* 2015).
89
90 Quantitative Trait Locus (QTL) studies have traditionally been used to identify
91 genome regions connected to local adaptation (Hawthorne & Via 2001; Ungerer &
92 Rieseberg 2003; Baxter *et al.* 2008), and population genomic scans for outlier loci are
93 commonly used to identify outlier loci relating to local adaptation and the reduction
94 of gene flow between populations (Nosil *et al.* 2008; Galindo *et al.* 2010). Both of
95 these methods can now be performed with very large numbers of markers, and
96 therefore can have high resolution (Hohenlohe *et al.* 2010). However, it can still be a
97 challenge to pinpoint the specific targets of selection with confidence; there are
98 multiple reasons why outlier loci may be detected in one sample only, and it is
99 important to confirm that identified outliers are the true targets of natural selection.
100 Functional interpretation of differentiated, 'outlier', loci or QTL (Barson *et al.* 2015)
101 and experimental tests for the action of selection (Barrett *et al.* 2008; Gompert *et al.*
102 2012) are ultimately critical but can be a major investment. Given the uncertainties
103 associated with outlier detection (e.g. Hermisson 2009), an important step in many
104 systems is to confirm outliers by repeating analyses in new samples, separated in
105 time or space.
106
107 Where we have good reason to suspect the involvement of a gene category in a
108 speciation system, targeted gene sequencing can allow us to look in more specific
109 regions for signals of reduced gene flow, whilst avoiding some of the problems of the
110 population genomics and QTL methods (e.g. false positives caused by multiple
111 testing, uncertainty about the genuine target of selection) (Smadja *et al.* 2012).
112 Combining QTL mapping with outlier loci scans, to associate outliers with
113 phenotypes, can provide a powerful indication of the source of selection driving
114 speciation (Rogers & Bernatchez 2005; Via & West 2008; Via 2012; Via *et al.* 2012),
115 and reveals the enormous potential we now have to follow up on outlier scans once
116 outliers have been confidently identified.
117

Host-race formation in phytophagous insects represents an excellent model for the evolution of reproductive isolation resulting from divergent selection in the face of gene flow (Drès & Mallet 2002; Bush & Butlin 2004; Forister *et al.* 2011). Very high diversity is linked to specialization via host-switching and co-speciation in many insect taxa (Weiblen & Bush 2002). Host races (genetically distinct populations, locally adapted to different host plant species but still experiencing some level of gene flow), and their host plants (clearly defined species, but often geographically proximate), provide a very helpful set-up for examining the interplay between divergent selection and ongoing genetic exchange (Drès & Mallet 2002).

The pea aphid, *Acyrthosiphon pisum*, is a well-established model for the study of ecological speciation (Peccoud & Simon 2010), and was the first aphid species to have its genome sequenced (The International Aphid Genomics Consortium 2010). *A. pisum* lives and feeds on species of the bean family (Fabaceae); in Europe, at least 15 genetically distinct host-plant-associated populations (races) have been described, each associated with one or a few host plant species. *A. pisum* races show increased preference for and performance on their associated plant species in comparison to alternative host plants (Via 1991; Ferrari *et al.* 2008). Races form a continuum of divergence ranging from pairs which produce around 10% F1 hybrids up to and including strongly isolated host races with F_{ST} exceeding 0.8 in sympatry, which probably no longer experience gene flow (Peccoud *et al.* 2009, 2015). This continuum of divergence between races provides us with a rare opportunity to examine the progression of barriers to gene flow across the genome. Although pea aphid host plants have overlapping ranges (Peccoud *et al.* 2009), aphid host-races both feed and mate on their specific plants, which leads to assortative mating and the potential for the evolution of reproductive isolation.

Because assortative mating is related to host-plant, how aphids select a plant to settle and feed on has the potential to be an important component in the evolution of premating isolation. Indeed, the chemosensory system has frequently contributed to host, habitat and mate choice in a range of study systems (reviewed in Smadja & Butlin 2008). Aphid recognition of the host plant and establishment of phloem

feeding has several stages (Powell *et al.* 2006; Simon *et al.* 2015); before an aphid settles to feed it may respond to plant volatiles near the surface of the leaf (Nottingham & Hardie 1993), and undertake initial probing with the stylets (Caillaud & Via 2000). Volatile and non-volatile odor and taste molecules are recognized in insects by a set of chemoreceptors found in the chemosensory organs (antennae, mouth parts, and maxilliary palps) (Kopp *et al.* 2008; Shiao *et al.* 2013). These chemosensory genes include gustatory (GR), odorant (OR) and ionotropic (IR) receptors (Hallem *et al.* 2006; Croset *et al.* 2010), as well as odorant binding proteins (OBPs) which are involved in the transport of odorants (Leal 2005), chemosensory proteins (CSPs) and sensory neuron membrane proteins (SNMPs) (Leal 2005; Jin *et al.* 2008; Vogt *et al.* 2009). Evidence is accumulating for the key role of chemosensory genes in host specialization in insects (Visser 1986; Whiteman & Pierce 2008; Schymura *et al.* 2010). They exist in large multigene families in most insects (Sánchez-Gracia *et al.* 2009), and both their birth and death mode of evolution and the detection of positive selection on branches of these multi-gene families point to rapid evolution in specialized lineages (Matsuo 2008; Briscoe *et al.* 2013; Duvaux *et al.* 2015).

In the pea aphid, multiple lines of evidence now point to the importance of chemosensory genes as a category in underpinning feeding decisions. Behavioural studies indicate that aphids show a distinct preference for their associated host plant when presented with a choice of alternative hosts (Ferrari *et al.* 2006), as well as increased survival and fecundity. Genetic evidence from whole genome scans (Jaquiéry *et al.* 2012), targeted re-sequencing (Smadja *et al.* 2012), examination of copy number variation (Duvaux *et al.* 2015) and gene expression (Eyres *et al.* 2016) have all found indications that chemosensory genes differ between pea aphid races. Although these studies confirm the value of further investigation of chemosensory genes in pea aphids, and provide us with a set of potentially interesting target chemosensory genes, this type of broad genomic study is prone to problems of false positives, as well as questionable reliability and repeatability (François *et al.* 2016; Jensen *et al.* 2016). Before we progress to examine target genes in more detail, it is important to confirm the findings of these studies.

There is a large number of tests available for the detection of outliers relating to local adaptation (e.g. Beaumont & Nichols 1996; Beaumont & Balding 2004; Foll & Gaggiotti 2008; Whitlock & Lotterhos 2015). In general these methods evaluate the genetic differentiation between populations and identify extreme values corresponding to candidate regions of the genome. Outlier scans have proved successful in many cases at identifying loci potentially under selection (Nosil *et al.* 2009; Butlin 2010). However, a disadvantage of many outlier detection methods is their requirement for *a priori* assignment of individuals to populations (Yang *et al.* 2012; François *et al.* 2016). In populations undergoing divergence in the face of gene flow, such as pea aphid host-associated races, the potential for sampling migrants and hybrids is high, making confident assignment of individuals to populations a difficult requirement to fulfill. In this study we use PCAdapt (Duforet-Frebourg *et al.* 2014, 2015), a method for the detection of candidate loci using Principal Components analysis (PCA), which is individual-based and therefore is well suited to analysing data where population level assignment of individuals is uncertain. As our interest lies in identifying loci relating to differences in host-plant preference between aphids, rather than in analysing genetic population structure, it is useful to be able to identify outliers based on genetic divergence rather than *a priori* population classification. Because PCAdapt identifies factors underlying the major axes of genetic variation among individuals, and then searches for loci strongly influencing these factors, it also allows us to examine only the important variation among races, rather than all pairwise race comparisons, thus reducing the risk of false positives from multiple comparisons. Furthermore, unlike many model-based outlier methods, PCAdapt does not assume an island model, and so is better suited to the wide range of levels of differentiation seen among pea aphid races.

Previous work (Smadja *et al.* 2012) has identified chemosensory genes as a promising set of candidate barrier loci; in an F_{ST} outlier scan of 9889 SNPs in 172 target genes (chemosensory and control) the proportion of outlier SNPs identified in Grs and Ors was significantly higher than in non-chemosensory control genes. Furthermore, this study identified a set of 18 chemosensory genes that were

unusually divergent between host races. These chemosensory candidates were identified as outliers in comparisons between three of the more closely related, although still highly specialized, pea aphid races (feeding on *Medicago sativa*, *Trifolium pratense*, and *Lotus pedunculatus*) (Ferrari *et al.* 2008, 2012; Peccoud *et al.* 2009) in a single geographic region.

In diverging populations, alleles underlying local adaptation can differ among localities because of drift, availability of mutations or differences in selection, but repeated patterns of differentiation across the geographic range of the pea aphid races would provide evidence for loci that diverge in response to common divergent selection pressures rather than as the result of stochastic processes. In addition, we wished to test whether loci involved in differentiation between one pair of races, were also likely to contribute to differentiation between other pairs. Therefore, our intention here was to test the pattern of divergence in chemosensory genes across (a) a larger number of pea aphid races along the continuum of differentiation and (b) multiple populations covering a broader geographic distribution. Incorporating a wider selection of aphid races, including the far more divergent races associated with *Lathyrus pratensis, Cytisus scoparius* and *Ononis spinosa*, will ultimately allow us to capitalize on the continuum of divergence in pea aphids, by examining patterns relating to the extent of divergence between races and the progression of barriers to gene flow across the genome. Additional races also potentially facilitate the identification of new chemosensory outliers relating to local adaptation in previously untested races. Repeating outlier scans on independently sampled aphids allows us to exclude false positives from the initial scan, and confirm the association of outliers with host race, the target environmental variable. As argued above, this confirmation is likely to be a valuable step in many comparable studies.

Materials and Methods

SNP data from Capture Sequencing

We used the capture sequencing dataset generated in Duvaux *et al.* (2015) using SureSelect. This was generated from 120 aphids (between 12 and 17 individuals per host plant) from eight host plant species (*Lotus pedunculatus*, *Lotus corniculatus*, *Medicago sativa*, *Trifolium pratense*, *Lathyrus pratensis*, *Pisum sativum*, *Cytisus scoparius* and *Ononis spinosa*), sampled 30 m apart to ensure distinct genotypes (supplementary file 1a). Aphids were collected in south-east England over three years, all less than 100km apart. SureSelect, which uses RNA probes to capture regions of interest from genomic DNA, was used prior to sequencing. Capture targets were candidate genes potentially relating to identification and selection of host plants, including all of the chemosensory genes that had been partially or fully annotated in Assembly 1.0 of the pea aphid genome (Smadja *et al.* 2009; Zhou *et al.* 2010): 79 olfactory receptor (Or) genes, 77 gustatory receptor (Gr) genes, 11 odorant binding protein (OBP) genes and 10 chemosensory protein (CSP) genes. Other genes potentially involved in chemosensation, 11 ionotropic glutamate receptor (Ir) genes and 9 sensory neuron membrane protein (SNMP) genes, were also included as targets, along with putative cis-regulatory regions relating to all these genes (for Ors, Grs, OBPs and CSPs, 50bp predicted regions were identified upstream of target genes and for IRs and SNMPs, 500 bp upstream regions were targeted, details in Duvaux *et al* (2015)). 69 genes from the P450 gene family (Zhang *et al.* 2010) potentially relating to detoxification, five pheromone synthesis genes, and 5 salivary protein genes were also included. 211 randomly chosen genes were added as controls. After mapping reads to the pea aphid genome (assembly 2.1) using Stampy 1.0.17, 21610 SNPs were called for all 120 aphid genotypes using Platypus 0.7.9.2 (Rimmer *et al.* 2014).

PCAdapt analysis of capture sequencing data

Capture data were filtered, removing SNPs based on the following criteria: quality score < 40, copy number != 1, minor allele count < 3, scored in < 60% of individuals, observed heterozygotes >10 more than expected. Removing three individuals without copy number information (Med210, Lped84, Lped82), along with duplicate clone Pisum5, left 7232 loci in 116 individuals.

These 7232 SNPs were analysed using the rapid, PCA-based method in PCAdapt (version 3.0 in R version 3.2.4) (Duforet-Frebourg *et al.* 2015). PCAdapt performs scans for natural selection using Principal Components analysis; examining correlations between SNPs and each principal component allows the detection of SNPs that strongly influence patterns of variation and are putatively involved in adaptive differentiation along these axes. PCAdapt does not require any prior definition of populations.

An initial run with K=20 principal components was used to select the correct K; a scree plot indicated K=7 (supplementary file 1b) as appropriate. After running with K=7, it was apparent that aphids in the *Lathyrus pratensis*-associated race have a large number of very influential SNPs in PC1 (supplementary file 1c and 1d); as this makes outlier identification difficult, we excluded *Lathyrus*-associated individuals from subsequent analyses. Excluding *Lathyrus*-associated individuals left 104 genotypes in 7 races. We then re-ran PCAdapt with one fewer principal component (K=6) (supplementary file 1e). Component-wise outlier scans were performed in PCAdapt, using loadings as the test statistic (corresponding to the correlation between each SNP and the principal component of interest), *p*-values were calculated based on making a Gaussian approximation for each PC and estimating the standard deviation of the null distribution, see Duforet-Frebourg *et al.* (2015), and after converting *p*-values to *q*-values, SNPs with *q*<=0.05 were considered "outliers". If outliers are randomly distributed in the genome, as might be expected for false positives, the number in any one gene will follow a Poisson distribution. The poisson.test() function in R was used to identify "outlier" genes containing significantly more SNPs with *p*<=0.05 than expected by chance, given the overall proportion of outliers and the total number of SNPs per gene, for each principal

304 component in turn (the same strategy as used by Smadja *et al.* 2012). Loci with few
305 SNPs but with a high proportion of outliers may not depart significantly from the
306 Poisson expectation. Therefore, this test may be prone to false-negatives but it is
307 expected to provide a conservative list of genes with strong differentiation.
308
309
310 *Aphid collection and DNA extraction for GoldenGate SNP genotyping*
311
312 Pea aphids were collected from the same eight host-plants as used in the capture
313 sequencing dataset: *La. pratensis*, *O. spinosa*, *C. scoparius, Lo. corniculatus*, *Lo.*
314 *pedunculatus*, *P. sativum*, *M. sativa* and *T. pratense.* In the UK, collection took place
315 over two years (2012 and 2013) in locations near Bristol, Peterborough, Sheffield and
316 the Blankney estate in Lincolnshire. Aphids from mainland Europe were collected in
317 France (Mirecourt, Volgesheim, Ranspach and Bugey) and Switzerland. Where
318 possible aphids were included from at least two UK locations and two locations in
319 mainland Europe. In total 29 location-and-race specific groups of aphids were
320 included, with a minimum of two and a maximum of seven sampling locations per
321 race, and a mean of 13.5 individuals per race per sampling location. Details of
322 sampling locations can be found in supplementary files 1f and 1g.
323
324 Aphids were grown up clonally from field-collected individuals on *Vicia faba* in the
325 laboratory to provide enough individuals for DNA extraction. Aphids were stored in
326 ethanol prior to DNA extraction. DNA was extracted from 5 aphids per genotype,
327 using NucleoSpin Tissue Kit standard protocol (Macherey-Nagel, Düren, Germany).
328
329
330 *GoldenGate SNP assay design, sample processing and allele calling*
331
332 Target SNPs were identified in control, chemosensory and detoxification genes in the
333 capture sequencing dataset. To design our custom set of 384 SNPs, flanking
334 sequences of 100bp to either side of target SNPs were processed using the Illumina
335 Assay Design Tool (ADT) in order to confirm their suitability for the assay, and the

finalized panel of 384 SNPs was ordered from Illumina (Illumina, San Diego, CA, USA). The final 384 SNPs comprised 222 in chemosensory genes, 71 in non-chemosensory genes of interest (P450s, PS and Rad51C), and 91 in control genes. 127 target SNPs were in genes identified as having a significant excess of outlier SNPs in the capture sequencing dataset. SNP IDs, chromosome positions, and flanking sequences for the panel of 384 SNPs are available in supplementary file 2.

SNP data were analysed from each plate in turn using the Genotyping module of Illumina's GenomeStudio package (Illumina, San Diego, CA, USA). SNPs were filtered for quality using standard thresholds, SNPs with no polymorphism, SNPs with no heterozygotes, and SNPs with indication of copy number variation were also removed. This left 179 high quality SNPs for further analysis (supplementary file 3). Aphids with more than 12 null SNP calls were removed from the dataset, leaving data for 373 aphids.

PCAdapt analysis of GoldenGate SNP genotyping data

The 179 SNPs in 373 aphids were analysed using the rapid, PCA-based method in PCAdapt (version 3.0 in R version 3.2.4) (Duforet-Frebourg *et al.* 2015). As with the capture sequencing data, an initial run (K=20 principal components) indicated K=7 as appropriate (supplementary file 1h). As with the capture sequencing analysis, *La. pratensis*-associated individuals had a large number of influential SNPs in PC1, so were excluded from subsequent analyses (supplementary file 1i).

Excluding individuals from the *Lathyrus*-associated race (PC1 score > 0.1) left 338 genotypes in 7 races. We then re-ran PCAdapt with one fewer principal component (K=6) (supplementary file 1j). Component-wise outlier scans were performed in PCAdapt, using loadings as the test statistic (corresponding to the correlation between each SNP and the principal component of interest), and after converting *p*-values to *q*-values significant SNPs ($q < 0.05$) were considered "outlier SNPs".

Comparing within and between race variation in GoldenGate SNP genotyping data

Based on clustering of aphid genotypes in PCAdapt, the race of 10 aphids was re-assigned from the collection host to the typical host of the genetic cluster to which they belonged (presumed migrants sampled on non-host plants), and six aphids were removed as potential early-generation hybrid individuals (details in supplementary file 1k). We then used locus-by-locus analysis of molecular variance (AMOVA), performed in Arlequin (v 3.5) (Excoffier & Lischer 2010), to examine hierarchical genetic structuring, by race and by locality within race, at each of the 179 GoldenGate SNP genotyping loci.

Results

PCAdapt analysis of capture sequencing data: clustering of individuals

Running PCAdapt on the capture sequencing dataset without *La. pratensis* individuals, with K=6, allowed us to define six principal axes of variation (figure 1). The first principal component separates the *O. spinosa*-associated individuals in one direction, and (to a lesser extent) *C. scoparius*-associated individuals in the other, from all other races. PC2 separates *C. scoparius* and *O. spinosa*-associated individuals from each other and from all other races. PC3 maintains the three most closely related races in our sample (*T. pratense*, *P. sativum* and *M. sativa*) in a single group, and separates *Lo. pedunculatus, O. spinosa* & *C. scoparius*, and *Lo. corniculatus*-associated clusters. PC4 separates both *Lotus*-associated aphid races (*Lo. corniculatus* and *Lo. pedunculatus*) from all others. PC5 separates half of the *P. sativum*-associated individuals from the other races, and PC6 separates half of the *T. pratense*-associated individuals from all other races. *M. sativa*-associated aphids are slightly separated from others in both PC5 and PC6. On the basis of these six axes we can therefore distinguish aphids from the more divergent races, as well as some individuals (but not all) from the two highly similar races of *T. pratense* and *P. sativum*.

F_{ST} distributions (supplementary file 1m), calculated according to groupings defined by the *La. pratensis*-associated principal component and by the other six principal components (PC1-6), showed a large number of SNPs with $F_{ST} = 1$ in both the *La. pratensis* and the *C. scoparius*-associated axes of variation, in agreement with PCAdapt findings. As expected, mean F_{ST} was lower in later principal components. The later principal components, which separate the *Medicago, Trifolium* and *Pisum*-associated races studied by Smadja et al. (2012) show F_{ST} distributions compatible with the values previously reported (multilocus $F_{ST} = 0.019$-0.084).

PCAdapt analysis of capture data: outlier analysis

411

412 Component-wise outlier scans were performed in PCAdapt, using loadings as the test
413 statistic (corresponding to the correlation between each SNP and the principal
414 component of interest), *p*-values were converted to *q*-values to control for false
415 discovery rate, and SNPs with *q*<=0.05 were considered "outlier SNPs". A total of
416 503, 557, 299, 274, 111 and 5 SNPs with *q*<=0.05 were identified in PC1 to PC6,
417 respectively. These correspond to SNPs with the highest loadings in each
418 component, i.e. they are the most influential SNPs in each axis of variation. Loadings,
419 *p*-values and *q*-values of SNPs can be found in supplementary file 4, tab 1. A
420 significantly higher proportion of outlier SNPs were in chemosensory genes than in
421 non-chemosensory genes in principal components 1-4 (z-tests for equality of
422 proportions: PC1 control=0.138, chemosensory=0.168, *p*=0.018; PC2 control=0.147,
423 chemosensory=0.192, *p*=0.001; PC3 control=0.068, chemosensory=0.127, *p*=2e-07;
424 PC4 control=0.072, chemosensory=0.101, *p*=0.004). 614 significant global outliers
425 were identified using the Mahalanobis distance (*q*-value <=0.05), and again a
426 significantly higher proportion of outlier SNPs were in chemosensory genes than in
427 non-chemosensory genes (control=0.160, chemosensory=0.248,p=1e-09).
428

429 For each principal component in turn, and in the global analysis, genes were then
430 considered "outlier genes" when they contained significantly more SNPs with
431 *q*<=0.05 than expected by chance (Poisson test), giving: 25 outlier genes in PC1, 24 in
432 PC2, 35 in PC3, 29 in PC4, 15 in PC5 and 5 in PC6, of which 11, 14, 26, 17, 8 and 3,
433 respectively, were chemosensory (figure 2; supplementary file 4, tab 2). 35 outlier
434 genes were identified in the global analysis, of which 18 were chemosensory. Outlier
435 counts and significance test values for all genes can be found in supplementary table
436 4, tab 2. The majority of chemosensory outlier genes identified in each principal
437 component were receptor genes.
438

439 Chemosensory outlier genes tended to be identified in blocks of close similarity and
440 physical distance (figure 2), for example gustatory receptors Gr1-Gr4 (all present on
441 scaffold GL350420) are all outliers in PC3, and where putative promoters were also
442 identified they were often present as outliers along with their downstream gene, for

example Or18 and an Or18 putative promoter region are both outliers in PC3, and Gr8 and Gr45 are both outliers in PC4 along with their putative promoter regions. Positioning of scaffolds on a linkage map would provide a more robust understanding of the proximity of these outlier genes in the genome. Of the 18 outlier genes identified by Smadja *et al.* (2012) ($p < 0.05$, 3 or more outlier SNPs per gene), 14 were present in the filtered capture sequencing dataset, and nine were confirmed as outliers in this new eight-race comparison (Or17, Or18, Or20, Or21, Or36, Gr8, Gr20, Gr45 and Gr47), along with Gr15 ($p < 0.05$ but with < 3 outlier SNPs).

SNP data from GoldenGate SNP genotyping

After removing low quality SNPs and individuals (see methods), we were left with 391 unique aphid genotypes sampled from eight host plants, from between two and seven sampling locations per race, distributed across the UK, France and Switzerland. The retained set of 179 SNPs included 10 in SNMP genes (4 genes), 46 control SNPs (44 genes), 43 in Grs (22 genes), 9 in IRs (3 genes), 31 in Ors (23 genes), 34 in P450 genes (24 genes), 1 in a CSP gene, 1 in an OBP gene, 2 in a PS gene and 2 in Rad51C, a control gene identified as an outlier by Smadja *et al.* (2012) (details in supplementary file 3).

PCAdapt analysis of GoldenGate SNP genotyping data: clustering of individuals

Running PCAdapt on the GoldenGate SNP genotyping dataset after excluding *La. pratensis*-associated individuals, with K=6, allowed us to define six principal axes of variation (figure 3). The first principal component separates half of the *Lo. corniculatus* individuals in one direction, and the *C. scoparius*-associated individuals in the other direction, from all other races. PC2 separates *O. spinosa*-associated individuals in one direction, and half of the *Lo. corniculatus* individuals in the other direction, from all other races. PC3 separates *O. spinosa*-, *Lo. corniculatus*- and *C.*

475 *scoparius*-associated individuals in one direction, and *P. sativum*-associated
476 individuals in the other, from all other races. PC4 separates the *Lo. pedunculatus*
477 race from all others. *T. pratense* and *M. sativa*-associated individuals consistently
478 have the most negative values in axis 5, and are separated from the other races in
479 opposing directions in PC6.
480
481 Apart from individuals sampled from *Lo. corniculatus*, which broadly split into two
482 clusters based on whether they were sampled in the UK or in mainland Europe on all
483 axes (supplementary figure 1l, figure 3), individuals tend to fall into groups on the
484 basis of host-plant association and not on the basis of geography. A number of
485 individuals in the GoldenGate SNP genotyping dataset appeared to be migrants, i.e.
486 they were collected on one plant species, but are genetically most similar to aphids
487 collected from a different host (e.g. two individuals sampled on *La. pratensis* cluster
488 with other races, one with *Lo. corniculatus*-associated individuals, and one with *P.*
489 *sativum*-associated individuals). A number of individuals may also be hybrids
490 between two races, as they fall into different host-associated clusters on different
491 axes of variation (i.e. they have some SNP alleles typical of one race and other SNP
492 alleles typical of a different race, e.g. one individual collected from *C. scoparius*
493 clusters firmly with *O. spinosa*-associated individuals in PC4). Although aphids from
494 *T. pratense* and *M. sativa*, two of the most closely related races, are not so discretely
495 separated, all other races form distinguishable clusters on at least one principal
496 component.
497
498
499 *PCAdapt analysis of GoldenGate SNP genotyping data: outlier analysis*
500
501 Component-wise outlier scans were performed in PCAdapt, using loadings as the test
502 statistic (corresponding to the correlation between each SNP and the principal
503 component of interest), *p*-values were converted to *q*-values, and SNPs with $q<=0.05$
504 were considered "outlier SNPs". A total of 14, 17, 16, 5, 2 and 1 SNPs with $q<=0.05$
505 were identified in PC1 to PC6 respectively. These correspond to SNPs with the
506 highest loadings in each component, and are the influential SNPs in each axis of

507 variation. Loadings, *p*-values and *q-values* of SNPs can be found in supplementary file
508 3. Of these 55 outlier SNPs, 42 (76%) are in chemosensory genes, while only three
509 (5%) are in control genes.
510
511
512 *Arlequin analysis of GoldenGate SNP genotyping data*
513
514 The AMOVA analysis revealed that a large percentage of total genetic variation was
515 between the 8 host-associated races (47.79%, $p<0.005$), while a much smaller
516 percentage of total variation was attributable to between-locality differences within
517 each race (5.63%, $p<0.005$). Examining the mean percentage of total genetic
518 variation explained by among-group and between-geographical-location variation in
519 the locus-by-locus analysis allowed us to compare chemosensory and control SNPs.
520 Among-group variation was lower (20.88%) and between-location variation was
521 higher (8.18%) in control loci in comparison to chemosensory loci (45.80% and
522 6.90%, respectively), demonstrating the importance of between race differences in
523 chemosensory genes in comparison to neutral loci.
524
525
526 *Comparison of capture sequencing and SNP genotyping results*
527
528 The capture sequencing dataset contained far more SNPs (7232), examined in fewer
529 individuals (116) in eight races sampled in close proximity, whilst the GoldenGate
530 SNP genotyping dataset contained fewer SNPs (179), examined in a larger number of
531 individuals (391) and covering multiple populations from a far larger European
532 sampling distribution. Nevertheless, the axes identified in our capture sequencing
533 and SNP genotyping datasets are broadly equivalent, although the order differs
534 between analyses (as might be expected from the different composition of the
535 samples): PC1 in the capture sequencing analysis and PC2 in the GoldenGate SNP
536 genotyping analysis both distinguish *O. spinosa*-associated individuals, PC2 in the
537 capture sequencing and PC1 in the GoldenGate SNP genotyping analysis both
538 distinguish *C. scoparius*-associated individuals, while PC3 in the capture sequencing

and PC1 in the GoldenGate SNP genotyping analysis both distinguish *Lo. corniculatus* individuals. PC4 separates the *Lo. pedunculatus*-associated population in both analyses, and *M. sativa* and *T. pratense*-associated individuals are distinguished by PC5 in both analyses. *P. sativum*-associated individuals can be distinguished from other races in capture sequencing PC5 and GoldenGate SNP genotyping PC3.

SNPs with a high loading in their significant GoldenGate SNP genotyping component often have a high loading in the equivalent capture sequencing component; for example, the same SNP (Or36.1_17759) has the top loading in PC4 in both capture sequencing and SNP genotyping analyses, and the top SNP in the capture sequencing PC5 (Gr21.1_461003) has the 11th highest loading in GoldenGate SNP genotyping PC5. Of the 27 significant GoldenGate SNP genotyping outlier SNPs present in the capture sequencing dataset, 23 also make a significant contribution to a capture sequencing factor. The four SNPs not contributing include one control SNP (Control_g84.3_29958), two SNPs in P450 genes (P450_g33.1_53017 and P450_g48.9_38206), and one Gustatory Receptor SNP (Gr1.2_96172). To simulate a null expectation for overlap between the two datasets, the total number of significant GoldenGate SNPs (34) was randomly re-assigned to the set of 179 GoldenGate SNPs with 100,000 permutations, and for each permutation we calculated the overlap between capture SNP significance at each SNP and the randomly assigned significant GoldenGate SNP. The real overlap of 23 SNPs significant in both datasets was significantly greater than this expectation ($p <$ 0.0001).

There were significant, strong positive correlations between squared loadings of SNPs in the two datasets (figure 4), both when comparing broadly equivalent components (see figure 4), and when looking at maximum loadings per SNP across axes in each dataset (Pearson's correlation = 0.52, $p<0.0001$).

There is substantial overlap in the genes identified between axes of variation: 389 outlier SNPs are related to more than one principal component in the capture sequencing dataset. Furthermore, 14 chemosensory outlier genes (Poisson test

571 $p<0.05$) and 5 putative chemosensory promoter outliers (Poisson test $p<0.05$) are
572 identified in more than one principal component. In the GoldenGate SNP genotyping
573 dataset, 12 outlier SNPs are present in more than one axis of variation.

Discussion

Pea aphids provide a promising system for examining the process of speciation with gene-flow, and the progression from initial natural selection acting on adaptive loci to complete genomic differentiation and reproductive isolation between races. Chemosensory genes appear to be important targets of natural selection in this system; examining the differentiation of these genes between races, and how this changes as divergence between races increases, will enable follow up work looking at the genetic architecture of speciation with gene-flow (making good use of the continuum of divergence between races seen in the pea aphid). Previous studies have indicated the value of further investigation of chemosensory genes in pea aphids. However, the types of large-scale genomic study used to identify targets of natural selection are prone to false positives, and have questionable reliability and repeatability (François *et al.* 2016; Jensen *et al.* 2016). Before we progress to examine target genes in more detail, it is important to confirm the findings of these studies. Here, we have undertaken a comprehensive follow-up to previous work, over a broader geographical range than previously examined to confirm the presence of chemosensory outlier genes and their relationship to host-plant adaptation. This is a step that could usefully be applied in many other comparable systems. We have also incorporated additional, more divergent host races giving insight into the role of the same genes at different stages in differentiation.

We have analysed genetic information from two datasets not previously used to detect outlier SNPs. The capture sequencing dataset (Duvaux *et al.* 2015) included pea aphids from eight races: the three originally looked at by Smadja *et al.* (2012), and five more. As well as confirming the repeatability of outliers among the original three races (table 1), extending the outlier scans to additional races allowed us to test whether the same chemosensory loci were implicated in multiple host shifts. The GoldenGate SNP genotyping dataset included pea aphids from the same eight races, this time sampled from locations across the UK, France and Switzerland. Sampling aphids from more localities across a broader geographical range enabled us to check that outlier genes relate directly to host plant species: there will have

been other environmental variables correlated with race where only single geographic regions were examined, whereas replication across different localities and years tends to confirm the relationship between chemosensory gene differences and adaptation to host plants. Given ongoing gene flow among races (Peccoud & Simon 2010), consistent patterns of differentiation are unlikely to be explained by genomic regions of low recombination, whose effect on differentiation is greatest where gene flow is low or absent, but additional evidence for the action of divergent selection is still desirable (Jensen *et al.* 2016). Our results identify good targets for this future work.

Chemosensory genes confirmed as targets of selection

We were able to confirm the findings of (Smadja *et al.* 2012), that a significantly higher proportion of outlier SNPs lie in chemosensory genes than in control genes, and that this is true in different samples of aphids, from more races and localities. In both datasets analysed here we again show that Gr and Or genes form the majority of chemosensory outlier loci. We specifically re-identify ten of the outlier genes found in Smadja *et al.* (2012) in the capture sequencing dataset (table 1; supplementary file 6), and three genes (Gr15, Or21 and Or36) were identified in the Smadja *et al* (2012) analysis and in both the capture sequencing and GoldenGate SNP genotyping datasets. The correlation between the two analyses undertaken in this study was also strong; all chemosensory outlier SNPs identified in the GoldenGate SNP genotyping analysis (incorporating multiple populations per race), that were present in the capture sequencing dataset, were also identified as outliers there.

By repeating outlier analyses on eight races we confirm that differences in chemosensory genes are important to the divergence of the broader spectrum of pea aphid races, and incorporating more localities in our GoldenGate SNP genotyping dataset allowed us to confirm a direct link between plant choice and chemosensory differences, distinct from other environmental variables that might be correlated with differences between single populations. AMOVA showed large contributions of race and small contributions of locality to genetic variation, a

638 pattern that was more pronounced in chemosensory than in control genes,
639 supporting the relationship between chemosensory gene divergence and race in the
640 face of gene flow. The congruence we observed between multiple independent
641 samples of aphids provides support for the individual outlier genes identified, the
642 general importance of chemosensory genes in between race differences, and more
643 specifically the potential role of Grs and Ors. Although the repeated identification of
644 specific outlier loci could relate to underlying genomic architecture at these sites
645 (Jensen *et al.* 2016) (which one would expect to be the same between aphids in
646 different datasets), comparisons of gene categories are particularly informative
647 indications of the validity of our results as there is little reason to expect that all
648 chemosensory receptor genes will share an unusual feature such as a distinct
649 mutation rate or low diversity, given that they are widely distributed in the genome.
650
651 ORs and GRs in insects tend to be activated in combinations to signal the presence of
652 specific compounds (Hallem *et al.* 2006). As previously suggested (Smadja *et al.*
653 2012), mutations in these genes could potentially lead to changes in sensitivity or
654 specificity of nerve activation, and combinations of mutations in different receptor
655 genes might be required for a complex modification of response to multiple
656 compounds differing between host plants. As ORs and GRs belong to large, fast
657 evolving gene families, and are the main peripheral discriminators, they are the best
658 *a priori* targets for involvement in host shifts. This assumption is supported by a
659 number of studies highlighting the involvement of chemoreceptors in differences
660 between host-associated races (McBride 2007; Smadja *et al.* 2012; McBride *et al.*
661 2014; Duvaux *et al.* 2015). In contrast OBPs and CSPs are smaller more conserved
662 families, more involved in presenting ligands to receptors. Although they have been
663 implicated in some host-shift cases (Matsuo *et al.* 2007; Dworkin & Jones 2009), we
664 find little evidence of their importance in pea aphid host-race formation. Duvaux *et*
665 *al.* (2015) and McBride (2007) both relate gain and loss of chemoreceptors to
666 between-race differences. Our finding of physical clusters of outlier genes (likely to
667 result from recent tandem duplications; Smadja *et al.* 2009) may suggest that
668 divergence after duplication is critical for the evolution of new response patterns
669 (figure 2).

The genetic architecture of divergence between races

Incorporating more races into our analysis has allowed us to identify a large number of new chemosensory genes as potential targets of selection. It is clear that the same set of genes is not necessarily involved in each adaptive host shift in the pea aphid; different chemosensory genes were outliers on different axes of variation although some axes separate multiple races. Identifying outlier genes in the more distinct aphid races, such as those very divergent in *O. spinosa* and *C. scoparius*-associated races, will be useful for follow-up work, as we cannot necessarily expect to examine divergence at the same chemosensory outliers in all race comparisons. The identification of multiple targets of selection relating to the same adaptive shifts suggests an polygenic basis of local adaptation in the pea aphid, fitting with the findings of Hawthorne and Via (2001), Caillaud and Via (2012) and Jaquiéry *et al.* (2012) who all identified multiple QTLs relating to aphid plant choice. The overlap in some cases between chemosensory genes identified on different axes of variation (see supplementary file 6 and figure 2) also shows how the same chemosensory genes can be involved in different adaptive host shifts, consistent with a combinatorial model of chemoreceptor activation (Hallem & Carlson 2004; Carey & Carlson 2011), and with the possibility that combinations or varying concentrations of plant compounds may act to trigger host acceptance or rejection.

Smadja *et al.* (2012) found that outlier genes could be divided into those with mainly non-synonymous substitutions and those with mainly synonymous site substitutions, and took this to imply a role for regulatory changes in the loci with mainly synonymous outliers, the high divergence at synonymous sites reflecting divergent selection in closely linked regulatory regions. Consistent with this, we detected a large number of putative promoter regions as outliers. Often an outlier putative promoter was present upstream of a gene that was also identified as an outlier (e.g. Gr4, Gr8 and Or18 in PC3 and Gr45 and Gr8 in PC4 of capture sequencing data). This could be the result of hitchhiking in regions surrounding targets of selection, or it

could relate to evolution of gene expression in some receptors. Although chemosensory genes as a class do not show more differential expression between races than other genes, some chemosensory genes are significantly differentially expressed between pea aphid races (Eyres *et al.* 2016); however, there is almost no overlap between these differentially expressed genes and the putative promoters identified here (ApisSNMP8 being the only exception).

The number of loci with extremely high loadings, equivalent to fixed differences, between *La. pratensis*-associated aphids and the others was notably high, suggesting the possibility of the accumulation of extensive neutral divergence between the *La. pratensis*-associated race and the other races. In accordance with this, outlier SNPs relating to *La. pratensis* contained a considerable number of control SNPs. Peccoud *et al* (2009) suggested that the highly genetically differentiated *La. pratensis*-associated race was nearing complete speciation, as no hybrid was detected with sympatric races, and our results are consistent with the minimal gene flow estimated between this more divergent race (Peccoud *et al.* 2009), in comparison to the higher gene flow between more genetically-similar races, where loci experiencing barriers to gene flow will stand out more clearly against a background of low differentiation (Nosil *et al.* 2009; Butlin 2010). This pattern of increased neutral divergence in the race with the lowest ongoing gene-flow supports the pea aphid host-race system as a promising one for examining the genomic architecture of speciation with gene-flow as races progress towards complete reproductive isolation.

PCAdapt is a useful tool for analysing data with uncertain population assignment

PCAdapt needs no prior information about assumed population membership of samples. On the whole, aphids clearly clustered on the basis of the plant that they were collected from. However, we detected multiple possible hybrids, as well as migrant aphids that clustered with individuals sampled from a different host plant, in our large SNP genotyping dataset. Because we were interested in identifying loci relating to differences in host-plant preferences between races, using methods that require *a priori* knowledge of population structure would have required us to exclude or reclassify these individuals. Not doing this removed any artificial

734 population structuring, which could arise from removing intermediate or non-
735 conforming genotypes. In some cases (*P. sativum* and *T. pratense* in capture
736 sequencing data and *Lo. corniculatus* in GoldenGate SNP data), PCAdapt was able to
737 identify unexpected substructure within races, which are normally found to have
738 little substructure, regardless of spatial scale (Frantz *et al.* 2006; Peccoud *et al.* 2008;
739 Ferrari *et al.* 2012). The splitting of *P. sativum* and *T. pratense* individuals in the
740 capture sequencing data is particularly interesting, as it was not identified in the
741 analysis of the same data set (Duvaux *et al.* 2015) on the basis of 1777 SNPs and
742 random forest clustering. Instead, Duvaux *et al*. identified two clusters of the *M.*
743 *sativa*-associated individuals. For the most similar races, presumably with the most
744 recent origin and/or the highest gene flow, these findings may indicate that there is
745 some overlap between spatial and host-associated structure. This emphasizes the
746 value of avoiding prior classification, especially where this is based on a small
747 number of markers chosen for their ability to separate host races in a single region
748 (as with microsatellites often used in pea aphid studies (Jaquiéry *et al.* 2012)).
749
750 By looking at outliers relating to the principal components of genetic variation,
751 rather than looking for global or pairwise F_{ST} outliers, we get a more biologically
752 realistic insight into divisions between races, presumably looking at variation
753 reflecting historical population sub-division by host switching, and reflecting true
754 adaptive differences between relevant groups of races. This method also enabled us
755 to carry out far fewer comparisons – with six principal components of variation
756 rather than 28 pairwise comparisons between races – thus reducing the problems
757 associated with multiple testing.
758
759 *Conclusions*
760 The positive identification of outliers based on differences within and between
761 populations can be caused by many factors other than divergent selection.
762 Population size change (Teshima *et al.* 2006), population structure (Excoffier *et al.*
763 2009), and background selection (Stephan 2010) can all affect the detection of F_{ST}
764 outliers. Furthermore, once candidate loci have been identified, there are few cases
765 where their status has been confirmed in relation to phenotype or fitness impacts

766 (Jensen *et al.* 2016). The exceptions are in cases where, like in our analyses,
767 candidate loci were defined *a priori* (e.g. Colosimo *et al.* 2005; Hoekstra *et al.* 2006).
768 We have sound biological reasons for looking at these candidates, we have followed
769 them up in varied datasets and can confirm the outlier status of chemosensory
770 genes as a category as well as some specific Or and Gr genes. It is now important to
771 link these loci to behavioural differences between races, and to their assortative
772 mating, and to examine the genomic context of these potential targets of selection.

References

Abbott R, Albach D, Ansell S *et al.* (2013) Hybridization and speciation. *Journal of Evolutionary Biology*, **26**, 229–246.

Barrett RDH, Rogers SM, Schluter D (2008) Natural selection on a major armor gene in threespine stickleback. *Science*, **322**, 255–257.

Barson NJ, Aykanat T, Hindar K *et al.* (2015) Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. *Nature*, **528**, 405–408.

Baxter SW, Johnston SE, Jiggins CD (2008) Butterfly speciation and the distribution of gene effect sizes fixed during adaptation. *Heredity*, **102**, 57–65.

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: Biological Sciences*, **263**, 1619–1626.

Briscoe AD, Macias-Muñoz A, Kozak KM *et al.* (2013) Female behaviour drives expression and evolution of gustatory receptors in butterflies. *PLoS Genet*, **9**, e1003620.

Bush GL, Butlin RK (2004) Sympatric speciation in insects. In: *Adaptive speciation*, eds Dieckmann U,. Doebeli M, Metz JAJ & Tautz D, pp 229–248.

Butlin RK (2010) Population genomics and speciation. *Genetica,* **138(4)**, 409-418.

Caillaud MC, Via S (2000) Specialized feeding behavior influences both ecological specialization and assortative mating in sympatric host races of pea aphids. *The American Naturalist*, **156**, 606–621.

Caillaud MC, Via S (2012) Quantitative genetics of feeding behavior in two ecological races of the pea aphid, *Acyrthosiphon pisum*. *Heredity*, **108**, 211–218.

Carey AF, Carlson JR (2011) Insect olfaction from model systems to disease control. *Proceedings of the National Academy of Sciences*, **108**, 12987–12995.

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

Croset V, Rytz R, Cummins SF *et al.* (2010) Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction (DL Stern, Ed,). *PLoS Genetics*, **6**, e1001064.

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

Duforet-Frebourg N, Bazin E, Blum MGB (2014) Genome scans for detecting footprints of local adaptation using a Bayesian factor model. *Molecular Biology and Evolution*, msu182.

Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MGB (2015) Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 Genomes data. *arXiv:1504.04543 [q-bio]*.

Duvaux L, Geissmann Q, Gharbi K *et al.* (2015) Dynamics of copy number variation in host races of the pea aphid. *Molecular Biology and Evolution*, msu266.

Dworkin I, Jones CD (2009) Genetic changes accompanying the evolution of host specialization in *Drosophila sechellia*. *Genetics*, **181**, 721–736.

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

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.

Eyres I, Jaquiéry J, Sugio A *et al.* (2016) Differential gene expression according to race and host plant in the pea aphid. *Molecular Ecology*, n/a–n/a.

Feder JL, Egan SP, Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends in Genetics*, **28**, 342–350.

Felsenstein J (1981) Skepticism towards Santa Rosalia, or why are there so few kinds of animals? *Evolution*, **35**, 124.

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

Forister ML, Dyer LA, Singer MS, Stireman III JO, Lill JT (2011) Revisiting the evolution of ecological specialization, with emphasis on insect–plant interactions. *Ecology*, **93**, 981–991.

François O, Martins H, Caye K, Schoville SD (2016) Controlling false discoveries in genome scans for selection. *Molecular Ecology*, **25**, 454–469.

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

Galindo J, Grahame JW, Butlin RK (2010) An EST-based genome scan using 454 sequencing in the marine snail *Littorina saxatilis*. *Journal of Evolutionary Biology*, **23**, 2004–2016.

Gompert Z, Lucas LK, Nice CC *et al.* (2012) Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. *Evolution*, **66**, 2167–2181.

Hallem EA, Carlson JR (2004) The odor coding system of *Drosophila*. *Trends in Genetics*, **20**, 453–459.

Hallem EA, Dahanukar A, Carlson JR (2006) Insect odor and taste receptors. *Annual Review of Entomology*, **51**, 113–135.

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

Hoekstra HE, Hirschmann RJ, Bundey RA, Insel PA, Crossland JP (2006) A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science*, **313**, 101–104.

Hohenlohe PA, Bassham S, Etter PD *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet*, **6**, e1000862.

Jaquiéry J, Stoeckel S, Nouhaud P *et al.* (2012) Genome scans reveal candidate regions involved in the adaptation to host plant in the pea aphid complex. *Molecular Ecology*, **21**, 5251–5264.

Jensen JD, Foll M, Bernatchez L (2016) The past, present and future of genomic scans for selection. *Molecular Ecology*, **25**, 1–4.

Jin X, Ha TS, Smith DP (2008) SNMP is a signaling component required for pheromone sensitivity in *Drosophila*. *Proceedings of the National Academy of Sciences*, **105**, 10996–11001.

Jones FC, Grabherr MG, Chan YF *et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55–61.

Kirkpatrick M, Ravigné V (2002) Speciation by natural and sexual selection: models and experiments. *The American Naturalist*, **159**, S22–S35.

Leal WS (2005) Pheromone Reception. In: *The Chemistry of Pheromones and Other Semiochemicals II* Topics in Current Chemistry. (ed Schulz S), pp. 1–36. Springer Berlin Heidelberg.

Malinsky M, Challis RJ, Tyers AM *et al.* (2015) Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science*, **350**, 1493–1498.

Martin SH, Dasmahapatra KK, Nadeau NJ *et al.* (2013) Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Research*, **23**, 1817–1828.

Matsuo T (2008) Genes for host-plant selection in *Drosophila*. *Journal of Neurogenetics*, **22**, 195–210.

Matsuo T, Sugaya S, Yasukawa J, Aigaki T, Fuyama Y (2007) Odorant-Binding Proteins OBP57d and OBP57e affect taste perception and host-plant preference in *Drosophila sechellia*. *PLoS Biol*, **5**, e118.

McBride CS (2007) Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*. *Proceedings of the National Academy of Sciences*, **104**, 4996–5001.

McBride CS, Baier F, Omondi AB *et al.* (2014) Evolution of mosquito preference for humans linked to an odorant receptor. *Nature*, **515**, 222–227.

Nosil P (2012) *Ecological Speciation*. Oxford University Press, Oxford.

Nosil P, Crespi BJ, Sandoval CP (2002) Host-plant adaptation drives the parallel evolution of reproductive isolation. *Nature*, **417**, 440–443.

Nosil P, Egan SP, Funk DJ (2008) Heterogeneous genomic differentiation between walking-stick ecotypes: “isolation by adaptation” and multiple roles for divergent selection. *Evolution*, **62**, 316–336.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375–402.

Nottingham SF, Hardie J (1993) Flight behaviour of the black bean aphid, *Aphis fabae*, and the cabbage aphid, *Brevicoryne brassicae*, in host and non-host plant odour. *Physiological Entomology*, **18**, 389–394.

Nouhaud P, Peccoud, J, Mahéo, F *et al* (2014) Genomic regions repeatedly involved in divergence among plant-specialized pea aphid biotypes. *Journal of evolutionary biology* **27**(9) 2013-2020.

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, de la Huerta M, Laurence L, Simon J (2015) Genetic characterization of new host-specialized biotypes and novel associations with bacterial symbionts in the pea aphid complex. *Insect Conservation and Diversity,* **8**(5), 484-492.

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

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

Powell G, Tosh CR, Hardie J (2006) Host plant selection by aphids: behavioral, evolutionary, and applied perspectives. *Annual Review of Entomology*, **51**, 309–330.

Rimmer A, Phan H, Mathieson I *et al.* (2014) Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, **46**, 912–918.

Rogers SM, Bernatchez L (2005) FAST-TRACK: Integrating QTL mapping and genome scans towards the characterization of candidate loci under parallel selection in the lake whitefish (*Coregonus clupeaformis*). *Molecular Ecology*, **14**, 351–361.

Rogers SM, Bernatchez L (2007) The genetic architecture of ecological speciation and the association with signatures of selection in natural lake Whitefish

(Coregonus sp. Salmonidae) species pairs. *Molecular Biology and Evolution*, **24**, 1423–1438.

Sánchez-Gracia A, Vieira FG, Rozas J (2009) Molecular evolution of the major chemosensory gene families in insects. *Heredity*, **103**, 208–216.

Schymura D, Forstner M, Schultze A *et al.* (2010) Antennal expression pattern of two olfactory receptors and an odorant binding protein implicated in host odor detection by the malaria vector *Anopheles gambiae*. *International Journal of Biological Sciences*, **6**, 614–626.

Servedio MR (2008) The role of linkage disequilibrium in the evolution of premating isolation. *Heredity*, **102**, 51–56.

Simon J-C, d' Alençon E, Guy E *et al.* (2015) Genomics of adaptation to host-plants in herbivorous insects. *Briefings in Functional Genomics*, elv015.

Smadja C, Butlin RK (2008) On the scent of speciation: the chemosensory system and its role in premating isolation. *Heredity*, **102**, 77–97.

Smadja CM, Butlin RK (2011) A framework for comparing processes of speciation in the presence of gene flow. *Molecular Ecology*, **20**, 5123–5140.

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

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.

Stephan W (2010) Genetic hitchhiking versus background selection: the controversy and its implications. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **365**, 1245–1253.

Teshima KM, Coop G, Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? *Genome Research*, **16**, 702–712.

The International Aphid Genomics Consortium (2010) Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biol*, **8**, e1000313.

Ungerer MC, Rieseberg LH (2003) Genetic architecture of a selection response in *Arabidopsis thaliana*. *Evolution*, **57**, 2531–2539.

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

Via S (2012) Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 451–460.

Via S, Conte G, Mason-Foley C, Mills K (2012) Localizing FST outliers on a QTL map reveals evidence for large genomic regions of reduced gene exchange during speciation-with-gene-flow. *Molecular Ecology*, **21**, 5546–5560.

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

Visser JH (1986) Host odor perception in phytophagous insects. *Annual Review of Entomology*, **31**, 121–144.

Vogt RG, Miller NE, Litvack R *et al.* (2009) The insect SNMP gene family. *Insect Biochemistry and Molecular Biology*, **39**, 448–456.

Weiblen GD, Bush GL (2002) Speciation in fig pollinators and parasites. *Molecular Ecology*, **11**, 1573–1578.

Whiteman NK, Pierce NE (2008) Delicious poison: genetics of *Drosophila* host plant preference. *Trends in Ecology & Evolution*, **23**, 473–478.

Whitlock MC, Lotterhos KE (2015) Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of FST. *The American Naturalist*, **186**, S24–S36.

Wood HM, Grahame JW, Humphray S, Rogers J, Butlin RK (2008) Sequence differentiation in regions identified by a genome scan for local adaptation. *Molecular Ecology*, **17**, 3123–3135.

Yang W-Y, Novembre J, Eskin E, Halperin E (2012) A model-based approach for analysis of spatial structure in genetic data. *Nature Genetics*, **44**, 725–731.

Zhang Y, Wang Q, Liu J, Zhang P, Chen J (2010) A genome-wide analysis of P450 gene families in the pea aphid, Acyrthosiphon pisum (Hemiptera: Aphidoidea). *Acta Entomologica Sinica*, **53**, 849–856.

Figure captions

Figure 1. PCAdapt scores for all pairwise combinations of principal components 1 to 6, after excluding *L. pratensis* associated aphids (K=6). Analysis based on 7232 SNPs from 104 aphid genotypes in 7 host-associated aphid races.

Figure 2. Loadings for each aphid genotype plotted for each principal component in turn. Outlier genes (Poisson test, $p<0.05$) in boxes are associated with each principal component in the capture sequencing dataset (7232 SNPs, 104 aphid genotypes, 7 host-associated aphid races). Genes on the same scaffold (pea aphid genome V2.1) are bracketed together, genes with >2 outlier SNPs are in bold, and genes identified as outliers in Smadja *et al* (2012) are in red.

Figure 3. PCAdapt scores for all pairwise combinations of principal components 1 to 6, after excluding *L. pratensis* associated aphids (K=6). Analysis based on 179 GoldenGate SNPs from 373 aphid genotypes in 7 host-associated aphid races.

Figure 4. squared loadings for SNPs in each principal component of the GoldenGate SNP genotyping dataset plotted against squared loadings for the most strongly correlated principal component in the capture sequencing dataset (left to right, top to bottom: capture PC1 vs. SNP genotyping PC2, capture genotyping PC2 vs. SNP genotyping PC1, capture PC3 vs. SNP genotyping PC1, capture PC4 vs. SNP genotyping PC4, capture genotyping PC5 vs. SNP genotyping PC5, and maximum

squared loading capture genotyping vs. maximum squared loading SNP genotyping). Black = control, pink = P450, green = chemosensory.

Acknowledgements: IE, LD, CS, RB and JF were supported by NERC grants NE/H004521/1 and NE/J021660/1, LD and RB by Leverhulme Trust project RPG-2013-198, JCS and CS by ANR-11-BSV7-005-01 'Speciaphid'. We would also like to thank NBAF Edinburgh for performing the capture sequencing, NBAF Sheffield for performing the GoldenGate SNP genotyping, and Anne-Lise Liabot for help with aphid collection. We also thank Sara Via and two anonymous reviewers for valuable comments on the initial version of the manuscript.

Data accessibility: Capture sequencing reads are deposited in the EBI Sequence Read Archive (SRA) with project accession no. PRJEB6325. GoldenGate SNP genotypes are available in Supplementary File 5 tab 1, and the filtered set of capture sequencing SNPs used for PCAdapt analysis are available in Supplementary File 5 tab 2.

Author contributions: RKB, CS, JF and LD designed the study. JF, LD, DH and JCS collected and reared aphids for DNA extraction. IE, LD, KG and RT generated the capture sequencing and GoldenGate SNP genotype data. IE and RKB designed and performed the analyses. IE and RKB wrote the article. All authors commented on draft versions of the manuscript. Authors declare no conflict of interests.

Tables

Table 1: Outliers in each data set (Smadja *et al*, Capture Sequencing and GoldenGate SNP genotyping), for genes present in all data sets, two data sets and just one data set each. Smadja *et al* (2012) and Capture sequencing outliers with $p < 0.05$ Poisson probability of the observed or a greater number of SNP outliers given the number of SNPs in the gene and the overall proportion of outliers. Outliers from GoldenGate SNP genotyping are genes containing a SNP with a significant loading ($q<0.05$) in PCAdapt.

		Analysed in:		
		All 3 data sets	**2 data sets**	**1 data set**
Outliers in:	**All 3 data sets**	Gr15, Or21, Or36		
	Smadja *et al* + capture	Gr45, Or17	Gr20, Gr47, Gr8, Or18, Or20	
	Smadja *et al* + SNP	Rad51C	-	
	Capture + SNP	Gr1, Gr2, Gr3, Gr33, Gr4, Gr6, Gr9	ApisSNMP4_ref	
	Smadja *et al* only	Or29	Gr39, Or11, Or13, Or14, Or15, Or51, Or56	Gr59, Or6, Or61, Or62, Or73
	Capture only	Gr17, Gr31, Gr65, Gr68, Or22, Or32, Or7	Gr10, Gr12, Gr19, Gr37, Gr42, Gr63, Gr66, OBP11, Or25, Or41, Or71, IR40a	Gr7, Gr74, OBP1, OBP4, ApisSNMP8_ref, IR8a
	SNP only	Gr21, Gr25, Gr26, Or16, Or26, Or3, Or47	Gr60, ApisSNMP3_ref	-
	Never an outlier	15	70	37