

Intrinsic pre-zygotic reproductive isolation of distantly related pea aphid host races

Varvara Fazalova¹, Bruno Nevado², Ailsa McLean¹ and H. Charles J. Godfray¹

¹Department of Zoology, University of Oxford,

²Department of Plant Sciences, University of Oxford

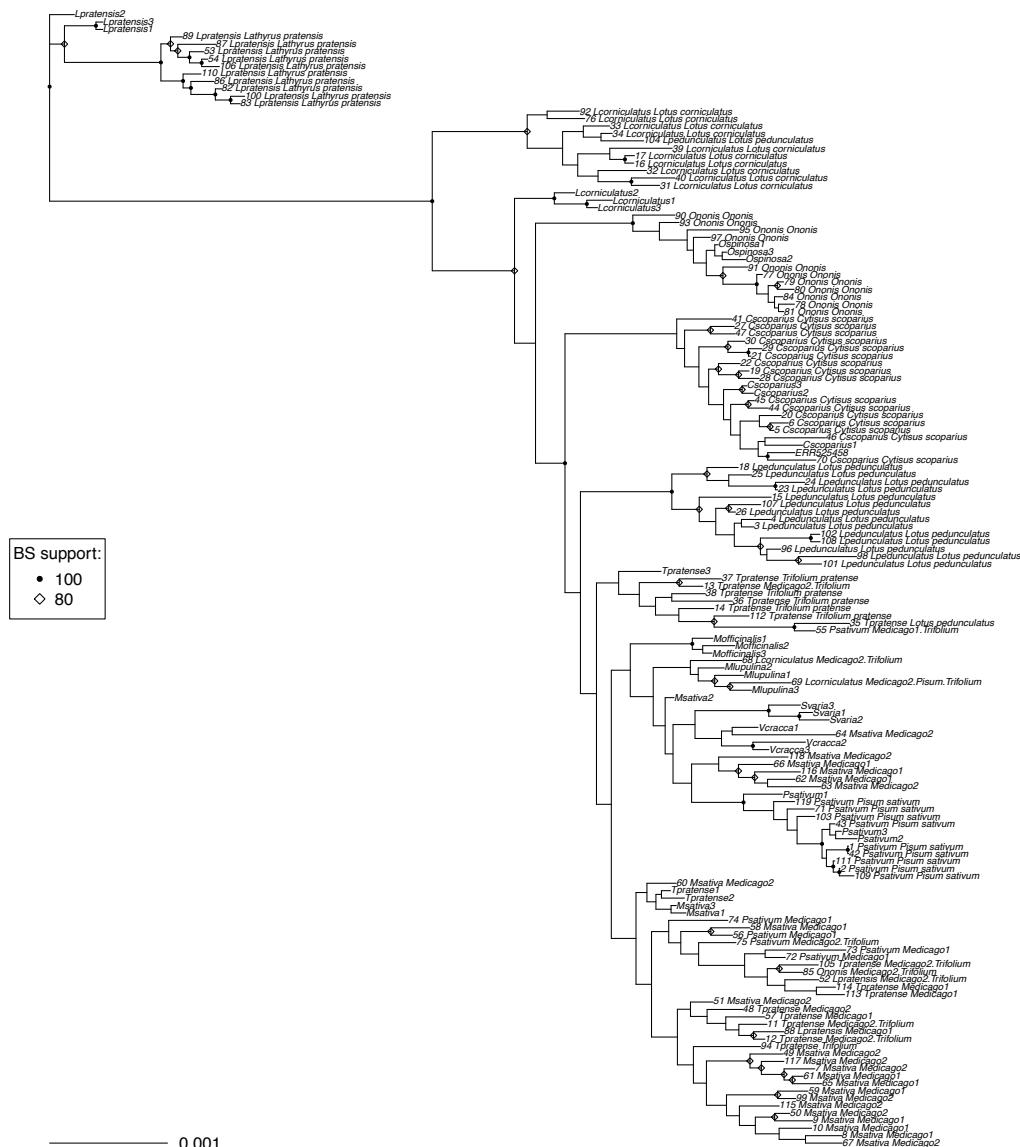
File S1. Phylogenetic reconstruction (for pea aphid host races).

Genetic clustering of individuals from the same pea aphid host races from distant locations has shown using microsatellite markers [1–3]. However, phylogenetic reconstruction based on few markers can be misleading due to incomplete lineage sorting or ongoing gene flow. Thus, in order to choose distantly and closely related biotypes, we constructed a phylogeny based on available genome-wide data (Duvaux et al. 2015 and the SpeciAphid project). The dataset of Duvaux et al. [4] contained targeted resequencing of ~500 chemosensory genes from 120 individuals representing eight host races collected in the UK, while the data from the SpeciAphid project consisted of whole-genome resequencing data (unpublished, used with permission of authors, AphidBase <http://bipaa.genouest.org/is/aphidbase/>) from 33 individuals sampled from 11 host-plant races collected in France.

We downloaded the raw Illumina data from GenBank (Bioprojects PRJEB6325 and PRJNA255937), trimmed low-quality ends and adaptors using cutadapt v 1.8.3 [5], and mapped reads to the pea aphid genome version Acyr_2.0 using bwa mem v. 0.7.12 [6] with default parameter values. Mapped data in bam format were realigned around indels using the GATK v. 3.4 [7], and used to perform per-individual SNP calling in samtools v. 1.2 [8] with a minimum quality threshold of 20 for both mapping and base-quality. Obtained SNPs were filtered to remove low quality SNP calls (< 15); SNPs with low depth (< 8 reads); SNPs near indels (< 3 bp); and heterozygous SNPs with less than 2 reads supporting each allele. We used the option to report homozygous-reference blocks for each individual with a minimum depth of 8 reads (bcftools call -g8), and the resulting vcf files were converted into fasta format using custom scripts. We concatenated the resulting fasta files for each scaffold after removing sites with more than 80% missing data, in effect selecting only the sequences for the ca. 500 chemosensory genes used in Duvaux et al. (2015). The concatenated dataset (570,278 bp) was used for Maximum-Likelihood phylogenetic inference with raxml v 8.0 [9] using the GTR model with Gamma-distributed rate variation. We performed 100 rapid bootstrap runs to gauge node support for the best-scoring maximum likelihood (ML) tree obtained. The phylogenetic tree obtained confirmed that French and British host races belong to the same genetic groups, likely reflecting the importance of long-distant dispersal rather than in-situ diversification in determining phylogeographic patterns within this species complex (**Figure S1.1**).

Figure S1.1. Compete maximum likelihood tree for pea aphid host races.

Pea aphid phylogeny based on ca. 500 chemosensory genes for clones collected in the UK and France. Clones from France have format “Plant_host Sample_number”, the UK clones: “Clone_Number Collection_Plant Host-race_assignment”. Host-race assignment (can be more than to one host race). is according to supplementary files Duvaux_CNV-PeaAphid_Sup.Tables from [4].



References

1. 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. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 7495–500. (doi:10.1073/pnas.0811117106)
2. Ferrari J, West JA, Via S, Godfray HCJ, Al ET. 2012 Population genetic structure and

- secondary symbionts in host-associated populations of the pea aphid complex. *Evolution (N. Y.)* **66**, 375–390. (doi:10.5061/dryad.8qb00)
3. Henry LM, Peccoud J, Simon J-C, Hadfield JD, Maiden MJC, Ferrari J, Godfray HCJ. 2013 Horizontally transmitted symbionts and host colonization of ecological niches. *Curr. Biol.* **23**, 1713–7. (doi:10.1016/j.cub.2013.07.029)
 4. Duvaux L, Geissmann Q, Gharbi K, Zhou J-J, Ferrari J, Smadja CM, Butlin RK. 2015 Dynamics of Copy Number Variation in Host Races of the Pea Aphid. *Mol. Biol. Evol.* **32**, 63–80. (doi:10.1093/molbev/msu266)
 5. Martin M. 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10. (doi:10.14806/ej.17.1.200)
 6. Li H, Durbin R. 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760.
 7. McKenna A *et al.* 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–303. (doi:10.1101/gr.107524.110)
 8. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009 The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079.
 9. Stamatakis A. 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313. (doi:10.1093/bioinformatics/btu033)