

**Article (Discoveries section)**

**Low spontaneous mutation rate and Pleistocene radiation of pea aphids**

Varvara Fazalova<sup>1</sup>, Bruno Nevado<sup>2</sup>

<sup>1</sup>University of Oxford, Department of Zoology

<sup>2</sup>University of Oxford, Department of Plant Sciences

**Corresponding author**

Varvara Fazalova, varvara.fazalova@zoo.ox.ac.uk, fazalova@gmail.com

**Keywords**

Spontaneous mutation rate, pea aphid, speciation

**Abstract**

Accurate estimates of divergence times are essential to understand the evolutionary history of species. It allows linking evolutionary histories of the diverging lineages with past geological, climatic and other changes in environment and shed light on the processes involved in speciation. The pea aphid radiation includes multiple host races adapted to different legume host plants. It is thought that diversification in this system occurred very recently, over the past 8,000 to 16,000 years. This young age estimate was used to link diversification in pea aphids to the onset of human agriculture, and led to the establishment of the pea aphid radiation as a model system in the study of speciation with gene flow. Here, we re-examine the age of the pea aphid radiation, by combining a mutation accumulation experiment with a genome-wide estimate of divergence between distantly related pea aphid host races. We estimate the spontaneous mutation rate for pea aphids as  $2.7 \times 10^{-10}$  per haploid genome per parthenogenic generation. Using this estimate of mutation rate and the genome-wide genetic differentiation observed between pea aphid host races, we show that the pea aphid radiation is much more ancient than assumed previously, predating Neolithic agriculture by several hundreds of thousands of years. Our results rule out human agriculture as the driver of diversification of the pea aphid radiation, and call for re-assessment of the role of allopatric isolation during Pleistocene climatic oscillations in divergence of the pea aphid complex.

**Introduction**

The pea aphid (*Acyrtosiphon pisum*, Harris) radiation includes at least 15 host races (Peccoud et al. 2015) specialized to different legume host plants (e.g., Müller 1962; Via 1991; Ferrari et al. 2006) that represent genetically divergent lineages (e.g., Peccoud, Ollivier, et al. 2009; Ferrari et al.

2012). Previous work suggests that diversification in this host race complex occurred very rapidly, i.e. within the past 8,000 – 16,000 years ago (Peccoud, Simon, et al. 2009). This timing of divergence was used to link diversification of pea aphid host races to the increased availability of potential host plants due to climate warming and the onset of Neolithic agriculture (Peccoud, Simon, et al. 2009). Hybrids between most host races are relatively common in nature and this – together with the assumed recent divergence – has been interpreted as evidence for a scenario of divergence with gene flow in the pea aphid complex (Peccoud and Simon 2010). Pea aphids have since been established as a model system for studying speciation with gene flow (Via et al. 2012).

The pea aphid radiation includes a few host races that do not hybridise in nature (Peccoud, Ollivier, et al. 2009), and intrinsic reproductive isolation (both post-mating pre-zygotic and post-zygotic) between some races has recently been shown (Fazalova et al. 2018). This very rapid completion of speciation is at odds with a scenario of very recent divergence with gene flow, and indeed contrasts with other insect taxa for which the build-up of intrinsic reproductive isolation takes much longer, e.g. at least 1,500,000 years in *Heliconius* butterflies (Kozak et al. 2015) or over 200,000 years in *Drosophila* flies (Coyne and Orr 1997). While the number of generations per year in natural populations is difficult to estimate accurately, available estimates suggest that completion of speciation takes over 6 million generations in *Heliconius* (4 generations per year, (Keightley et al. 2015)) and over 3 million generations in *Drosophila* (15 generations per year, (Pool 2015)) but only 120-240 thousand generations in pea aphids (15 generations per year, 14 parthenogenetic (Loxdale and Balog 2018) and 1 sexual).

The mismatch between time taken to achieve strong intrinsic reproductive isolation in pea aphids compared to other insect taxa calls for re-examination of the age of this complex. To do so, we estimated genome-wide spontaneous mutation rate for several pea aphid host races with a mutation accumulation experiment; and estimated average genomic differentiation for two distantly related host races, using a simple strict-isolation model. Together, these analyses show that the pea aphid radiation is much older than previously assumed, at least 543,000 years (95% CI 419,000 – 772,000). This new estimate implies that the time needed for completion of speciation in pea aphids is similar to that estimated in other insect systems, and calls for re-interpretation of the mechanisms driving diversification in pea aphids – in particular, the role of geographic isolation in the diversification of the pea aphid radiation needs to be reassessed.

## Results and Discussion

### ***Low spontaneous mutation rate in pea aphids***

In order to obtain an estimate of the spontaneous mutation rate of the pea aphid, we performed a mutation accumulation experiment, with twelve parthenogenetic lines (representing four host races) over 28 generations (Figure S1). For each mutation accumulation line, at the 1<sup>st</sup> and 28<sup>th</sup> generation we re-sequenced genomes to a mean coverage 38×, trimmed and mapped sequence data to the

pea aphid genome, and identified de novo mutations using a strict bioinformatics pipeline (see Material and Methods). On average 91% of reads were mapped to the reference genome assembly (Table S2) and on average 44% of the genome was callable within each line (i.e. had called genotypes at both the 1<sup>st</sup> and the 28<sup>th</sup> generations; Table 1).

We identified 43 high confidence and 24 low confidence putative mutations (see Materials and Methods). We attempted validation with Sanger sequencing for all candidate mutations, but we managed to design primers with high specificity targeting 18 high confidence and 8 low confidence mutations (other primers were not suitable as they mapped to multiple scaffolds of the genome). We obtained amplicons of expected lengths for all 18 high confidence mutations, and six low confidence mutations. Sanger sequencing results revealed that all six low confidence mutations were false positives (examples are shown in Figure S3), confirming the suitability of our filtering criteria. All low confidence mutations were assumed to be false positives and excluded from calculation of mutations rate. For high confidence mutations, Sanger sequencing failed for six amplicons despite repeated attempts. From the remaining twelve candidate mutations, we confirmed ten and identified two false-positives (Figure S3 and Table S4). This results in false-positive rate of 16.7%, which is similar to previously reported rates in other insects (Keightley et al. 2014; Keightley et al. 2015; Oppold and Pfenninger 2017). With correction for the false positive rate, we estimate the mutation rate per site per haploid genome per generation across all 12 mutation accumulation lines (Table 1) as:

$$\mu = \frac{N_{mutations}}{N_{callable\ sites} \times ploidy \times N_{generations}} \times True\ positive\ rate =$$

$$= \frac{43}{2468379192 \times 2 \times 27} \times 0.833 = 2.7 \times 10^{-10} \text{ (95\% CI } 1.9 \times 10^{-10} - 3.5 \times 10^{-10}\text{)}.$$

This might be a slight underestimate of the true spontaneous mutation rate in pea aphids, both because high throughput sequencing always carries a chance of missing true variants and because we used stringent filtering criteria. However, we expect the magnitude of this bias to be small, especially as the depth of coverage we used is relatively high (minimum of 20 reads). The transition/transversion ratio is 1.87 (Table S5).

Our estimate of the mutation rate in pea aphids is the lowest reported for any insect so far (Keightley et al. 2014; Keightley et al. 2015; Yang et al. 2015; Liu et al. 2017; Oppold and Pfenninger 2017). We are still developing an understanding on what drives the differences in mutation rate between species (Bromham 2009), thus we can only speculate as to the mechanisms behind the low mutation rate in pea aphids. One possible explanation is that the low mutation rate is related to the peculiar life cycle of aphids. Aphid females reproduce by apomictic parthenogenesis (i.e., without meiosis) during most of the year (around 10-15 generations), and with the onset of cold conditions a single sexual reproduction event takes place, after which overwintering eggs are laid. Thus, recombination in pea aphids is rare (roughly, 1 meiosis event

every 10-15<sup>th</sup> generation). This may cause increased selection for high fidelity of DNA polymerase in order to alleviate the mutation load resulting from accumulation of deleterious mutations during the parthenogenetic phase. A similar argument has recently been made to explain the low mutation rate observed in giant duckweed *Spirodela polyrhiza* (Xu et al. 2019), a species that reproduces mostly by asexual budding and exhibits the lowest mutation rate of any plant ( $2.4 \times 10^{-10}$ ). On the other hand, water flea *Daphnia pulex* – whose life cycle includes up to 5 apomictic parthenogenetic generations between each sexual reproduction event – exhibits a mutation rate 6.7 times higher than our estimate in pea aphids (Flynn et al. 2017). However, because estimates of mutation rate from other crustaceans are currently missing, it is impossible to judge if this is a low mutation rate compared to other crustaceans. Additional estimates from other species will be needed to test the role of asexual reproduction in mutation rate variation between species.

The size of the pea aphid X-chromosome (around 1/3 of the entire genome (Manicardi et al. 2015; Mandrioli et al. 2017)) provides a rare opportunity to test whether mutations rates in sex chromosomes and autosomes differ. This could potentially explain peculiarities of sex chromosome evolution such as faster differentiation or a preponderant role on speciation (Presgraves 2018). Using available annotation of X-linked region in the pea aphid genome (Jaquierey et al. 2019), we classified de novo mutations as X-linked or autosomal. We found that 11 de novo mutations (out of 42 that could unambiguously be mapped to either X or autosomes, i.e. 26%) occurred on X chromosome (Table S5), a number not significantly smaller than the expected if mutations occur at the same rate in all chromosomes (permutation test,  $P = 0.2$ ). This suggests similar mutation rates of sex-chromosomes and autosomes, and is consistent with previous results in nematodes (Denver et al. 2012).

#### **Genomic differentiation suggests more ancient onset of the pea aphid radiation**

In order to obtain an estimate of the beginning of the pea aphid radiation, we chose distantly related host races according to the phylogeny in (Fazalova et al. 2018), re-sequenced the genomes of 13 individuals of *L. pratensis* and 12 individuals of *V. cracca* to mean coverage of 20 $\times$ , and trimmed and mapped sequence reads to the pea aphid genome (on average 88.3% of reads were mapped to the reference genome, Table S2). To alleviate the effects of selection along the genome, and sex-chromosome specific biases, we extracted only 4-fold degenerate sites from autosomal genes for analysis (2,116,329 sites). Similar levels of genome-wide synonymous polymorphism were found in both host races:  $\theta = 0.0042$  and  $0.0043$ , for *L. pratensis* and *V. cracca*, respectively. Average genomic divergence between the two host races was estimated as  $d_{XY} = 0.0086$ , and average differentiation along the genome as  $d_a = d_{XY} - (\theta_X + \theta_Y)/2 = 0.0086 - (0.0043 + 0.0042)/2 \approx 0.0044$ . We can use this simple formula which accounts for ancestral polymorphism because gene flow at present is very unlikely between *L. pratensis* and *V. cracca*, with evidence of strong intrinsic reproductive isolation (Fazalova et al.

2018) and no hybrids found in the wild (Peccoud, Ollivier, et al. 2009). Furthermore, if gene flow occurred during the early stages of divergence, our estimates of divergence time would be conservative as unaccounted-for gene flow results in under-estimation of the time of divergence.

Using our estimates of the spontaneous mutation rate in pea aphid and genetic divergence between *L. pratensis* and *V. cracca*, we calculated the age of divergence of the pea aphid complex as  $T = d_a / 2\mu = \frac{0.0044}{2 \times 2.7 \times 10^{-10}} \approx 8\,148,000$  generations (95%CI 6 286,000 – 11 579,000). To convert the number of the generations into years, we assumed 15 generations per year: this results into a time of split of about 543,000 years (95% CI 419,000– 772,000). This estimate is conservative, as it includes the maximal number of asexual generations per year suggested for aphids: 14 (Loxdale and Balog 2018). We added one more generation to this estimate to calculate the total number of generations per year, as we assume that sexual generation is likely to have similar mutation rate. Even if the mutation rate is higher in sexual generations, it does not affect our estimate of the divergence time very strongly, as aphids have a single sexual generation per year, and the rest of the life cycle consists of asexual generations. If we assume that the pea aphid mutation rate during sexual generation is similar to other insects (i.e.  $3e-9$ , which is an average of the estimates obtained from (Keightley et al. 2014; Keightley et al. 2015; Yang et al. 2015; Liu et al. 2017; Oppold and Pfenninger 2017)), we can estimate the mutation rate which accounts for contribution of sexual reproduction every 15<sup>th</sup> generation of the life cycle as  $2.7e-10 \times (14/15) + 3e-9 \times (1/15) = 4.52e-10$ , and the age of the pea aphid divergence would be about 324,000 years.

Our estimate of the age of the pea aphid radiation contrasts with previous report of very recent and rapid diversification (Peccoud, Simon, et al. 2009), which was based on the mutation rate of maternally transmitted endosymbiont *Buchnera aphidicola* (Moran et al. 2009). However, the *B. aphidicola* genome behaves as a single gene, without recombination (Shigenobu et al. 2000) and exclusive maternal inheritance (Tóth 1933). Reconstruction of evolutionary histories from such single-locus datasets have been shown to be unreliable and strongly affected by processes such as selective sweeps or drift (e.g., Shaw 2002; Ballard and Whitlock 2004). Misleading phylogenies could affect the estimate of mutation rate, as well as subsequent estimates of age of divergence using this rate. Given our estimate is based on over 2 million sites sampled genome-wide, we expect our new estimate of the divergence age of the pea aphid to be much more reliable than previous estimates.

Our results contradict the Post-Pleistocene timing of speciation in the pea aphid complex and rule out the effect of anthropogenic Neolithic agriculture on their diversification. The older age of the radiation 543,000 years (95% CI 419,000– 772,000) raises the possibility that pea aphids experienced numerous Pleistocene habitat fragmentations (Hewitt 2004). They might have been trapped in separate refugia (Stewart et al. 2010) on novel host plants, and undergone adaptation

and divergence without the impeding effect of gene flow. This implies that patterns of gene flow present in pea aphids – especially for distantly related host races – can be more parsimoniously explained by allopatric isolation and secondary contact than divergence with gene flow. This has been suggested before by some authors (e.g., Futuyma 2008; Bierne et al. 2013; Harrison and Larson 2016) but remains largely unappreciated in the large body of work on pea aphid speciation.

Our results shed new light into the diversification of the pea aphid complex, and call for reassessment of current understanding within this study system. It will be especially important to understand how much geographic separation has contributed to the divergence of the closely related pea aphid host races (specialized on *Medicago sativa* and *Trifolium pratense*). Thus far, patterns of genomic differentiation between pea aphid host races have been interpreted in light of a scenario of speciation with gene flow. In particular, genome scans studies (e.g., Jaquiéry et al. 2012; Nouhaud et al. 2018) have identified highly differentiated genomic regions between host races and interpreted those as regions involved in adaptation to host plants, because under speciation with gene flow genes responsible for adaptation are expected to show elevated differentiation. Furthermore, the identification of large genomic regions of high divergence around quantitative trait loci in pea aphids has inspired the development of the theory of divergence hitchhiking, that is the process through which initial selection on only a few loci can extend divergence to larger genomic regions (Via and West 2008). However, under a scenario of allopatric isolation, any genomic region has the potential to differentiate without necessarily being involved in reproductive isolation or adaptation (e.g., Feder et al. 2013). Thus, both genome scan results and divergence hitchhiking models need to be reassessed in light of the potential role of geographic isolation in divergence of pea aphid host races. Finally, our results raise warning for other study systems as well – especially for those which are assumed to have recently divergence with ongoing gene flow. Only lineage-specific estimates of mutation rates, together with genome-wide estimates of divergence and implementation of appropriately complex demographic models will allow accurate inference of evolutionary histories and a more complete understanding of the processes driving diversification in these systems.

## **Materials and Methods**

### ***Collection and rearing of aphids***

We sampled overlapping populations of pea aphids from several wild Fabaceae host plants in summers of 2015-2016 around Oxford, UK (Table S6). We genotyped all aphids collected with 14 microsatellite loci (Peccoud, Ollivier, et al. 2009) to confirm the host race assignment (except the pea aphid line collected from *Lotus arenarius*), and reared them as parthenogenetic lines in the lab at 14±1°C with a 16L : 8D photoperiod, on leaves of *Vicia faba* (Sutton variety, replaced weekly) placed in 1% agarose gel in Petri dishes.

### ***Mutation accumulation experiment***

For the mutation accumulation experiment, in May 2017 we established twelve parthenogenetic lines: three from *Lathyrus pratensis*, three from *Lotus corniculatus*, three from *Lotus pedunculatus*, two from *Medicago lupulina* and one from *Lotus arenarius*. To establish each line, before starting the mutation accumulation experiment we isolated a single asexual female and waited until reproduction, at which point we randomly selected six offspring to establish the first generation and used the remaining offspring for DNA extraction. Each generation thereafter was established by randomly selecting six offspring, ensuring low population sizes throughout the experiment. We continued these transfers until July-September 2018, at which point all lines had gone through 28 generations. For the penultimate generation, we isolated a random single adult female and collected its offspring for DNA extraction.

#### **DNA extraction and sequencing**

For each aphid line and generations 1 and 28, we extracted DNA from pools of six individuals (sisters) to obtain sufficient quantity for preparation of DNA libraries for whole-genome resequencing. Aphids were frozen in -80°C, processed with TissueLyser (Qiagen), and DNA extracted following the Supplementary Qiagen protocol (Purification of total DNA from insects using the DNeasy® Blood & Tissue Kit). We treated DNA extractions with 2 µl of RNAase A (ThermoScientific) and purified them with Monarch PCR & DNA Cleanup Kit (New England Biolabs). We assessed DNA quantity and quality with Qubit 3.0 (ThermoFisher), Nanodrop (ThermoFisher) and gel electrophoresis (0.7% agarose gel, TAE buffer). Preparation and sequencing of DNA libraries were performed by Novogene (Hong Kong). For the mutation accumulation lines we used NEBNext Ultra library kits (New England Biolabs), and sequenced with 150 bp paired-end reads to > 30× coverage per line and generation. For the analysis of genetic differentiation, we used NEBNext Ultra library kits (New England Biolabs) and sequenced to >15× coverage per line. Raw sequencing data is available from NCBI SRA under BioProject accession numbers PRJNA607096 and PRJNA607313.

#### **Identification of candidate mutations, validation, estimation of mutation rate**

We quality-controlled raw sequencing data with fastqc (Babraham Bioinformatics) and trimmed sequencing adaptors and low quality reads with trim-galore ([https://www.bioinformatics.babraham.ac.uk/projects/trim\\_galore/](https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/)) using Cutadapt (Martin 2011); we discarded bases with quality below 20 and discarded read pairs if any was shorter than 36 bp after trimming. We mapped the trimmed reads to the reference genome version Acyr\_2.0 using bwa (Li and Durbin 2009), with default parameter values and the -M option (which marks shorter split hits as secondary), and removed duplicate reads with samtools rmdup (Li et al. 2009). We realigned regions around indels using the GATK (McKenna et al. 2010) indel realignment tool. At each step of data processing, summary files were inspected in multiqc (Ewels et al. 2016).

We performed variant calling with samtools and bcftools (Li et al. 2009) with the multi-allele caller (-m option) and discarding reads with mapping quality below 20 and bases with base quality below 20. For each sample, we filtered variants with bcftools, excluding SNPs with low quality (<20) and with too low (<20) or too high (> 2.5 times average depth for each sample) depth. We further filtered heterozygous SNPs with fewer than 2 reads supporting each allele, SNPs near indels (< 3 bp), indels separated by 10 or fewer base pairs and SNPs supported by fewer than 2 reads in each direction (for heterozygous genotype calls that include the reference allele, i.e. GT=0/1). Versions of the software are provided in Table S7.

In order to obtain a first set of candidate *de novo* mutations in each line, we identified SNPs covered by at least 20 reads in each generation and with different genotype calls in generations 0 and 28 (start and end of mutation accumulation lines) considering only base substitutions (i.e. ignoring indels). Candidate mutations were further filtered out if reads supporting both alleles are present in both lines (generations 0 and 28) and if genotypes of the two samples imply more two different alleles. These steps were implemented in custom scripts (<https://github.com/brunonevado/MutationAccumulation>) that use the mpileup files produced by samtools for generations 0 and 28.

We further filtered this set of candidate mutations, by manually inspecting the alignment files in IGV (Thorvaldsdottir et al. 2013) for the following artefacts suggesting read mapping errors:

1. Candidate mutations do not show consistent linkage to other polymorphic sites on the same read (Long et al. 2016), as expected due to apomictic parthenogenesis.
2. Reads supporting candidate mutations appear both in Generation 1 and Generation 28, irrespective of read, base or mapping quality (i.e. considering even the low quality reads/bases that were filtered at the genotype calling step).
3. Reads supporting putative mutations have many substitutions, which are false positives according to the filtering criterion 2.
4. SNPs supported by fewer than two reads in each direction, only for heterozygotes genotype calls that do not include the reference allele, i.e. GT=1/2 (because these SNPs were not filtered on previous steps).
5. Candidate mutation can be resolved by indel realignment (Figure S3 in Keightley et al. 2014).

With these filters we obtained two sets of candidate *de novo* mutations. High confidence mutations passed all filters and were expected to be true *de novo* mutations. Low confidence mutations passed the filtering criteria, but were considered unlikely either because (1) reads supporting the candidate mutation were clipped or had many indels/substitutions (Figure S1 in Keightley et al. 2014), (2) more than one candidate mutation occurred within 50 bp region, or (3) candidate mutation implied a change from heterozygous to homozygous state. We also included

in the low confidence set one candidate (Line: PT1, Scaffold415, Position 393952) for which a single read supporting the new allele was also found in the Generation 1, to test if the filtering criterion 2 was too stringent. Candidate mutations after each step of filtering are shown in Table S8.

For all candidate mutations (low and high confidence) we retrieved FASTA files (based on the reference genome *Acyr*. 2.0) with 600 bp flanking the candidate mutations and attempted to design primers for these regions, using Primer3Plus (Untergasser et al. 2007). Primers were excluded if they would map to other scaffolds of the genome, as inferred with Primer-BLAST (Ye et al. 2012). This resulted in 25 primer pairs, 18 targeting regions around high confidence mutations, and 7 around low confidence mutations (for validation of 8 mutations, as two of them were located very close to each other: Line 16.107a, Scaffold 1165, Positions 78180 and 78222), which were used for PCR amplification of both generations (Table S9). PEG clean-up and Sanger sequencing of the PCR product (BigDye 3.1, Applied Biosystems capillary 3730XL DNA Analyzer) were performed by the Sequencing Core of the Department of Zoology, University of Oxford.

We estimated the mutation rate per haploid genome using the number of high confidence mutations and the number of callable sites (total number of sites with adequate depth of sequencing – at least 20 reads in in both generation 1 and generation 28 for each line) for each line. In addition, we estimated mutation rate for both X and Autosomes, using the assignment from (Jaquierey et al. 2019).

### **Estimation of time of divergence**

For the analysis of genomic differentiation, we chose distantly related host races according to the phylogeny in (Fazalova et al. 2018), and included 13 individuals of *L. pratensis* and 12 individuals of *V. cracca*. Quality control, trimming, removal of duplicate reads, mapping, indel realignment and variant calling were performed in the same way as for the mutation accumulation analysis (see above). Filtering was less stringent compared to the mutation accumulation pipeline due to lower coverage: obtained SNPs were filtered to remove low quality SNP calls (< 15); SNPs with low depth (< 8 reads); heterozygous SNPs with less than 2 reads supporting each allele; SNPs near indels (< 3 bp); indels separated by 10 or fewer base pairs; and SNPs with high depth (more than 2.5x the average depth of coverage of each sample). We used the option to report homozygous-reference blocks for each individual with a minimum depth of 8 reads. Versions of the software are provided in the Table S7.

In order to obtain an estimate of divergence that is least affected by selection, for analysis of divergence we used only 4-fold degenerate sites from each gene. We used vcf2fas (available from <https://github.com/brunonevado/vcf2fas>) to obtain FASTA files for each scaffold, and extracted coding sequences (CDS) according to the available annotation of the pea aphid genome obtained from GenBank (GCF\_000142985.2\_*Acyr*\_2.0\_genomic.gff). We assigned CDS to X-

chromosomes and Autosomes according to (Jaquiere et al. 2019). For each gene, we extracted and concatenated 4-fold degenerate sites from all exons using custom scripts (<https://github.com/brunonevado/MutationAccumulation>).

In order to calculate genetic differentiation between *L. pratensis* and *V. cracca*, while taking into account ancestral polymorphism, we calculated population polymorphism (Watterson theta (Watterson 1975)) and absolute divergence ( $d_{XY}$ , (Nei and Li 1979)) using Popgenome (Pfeifer et al. 2014), and used the formula of Nei and Li (Nei and Li 1979)  $d_a = d_{XY} - (\theta_x + \theta_y)/2$ , where  $d_{XY}$  is absolute divergence between two populations, and  $\theta_x$  and  $\theta_y$  is polymorphism within *L. pratensis* and *V. cracca* respectively. To convert our estimate of genetic differentiation into absolute age of divergence (in generation) we used then used the formula  $T = d_a/2\mu$ , where  $\mu$  is the mutation rate (per site per generation).

## Acknowledgments

We are grateful to Charles Godfray for useful comments on the manuscript and to James Clark for proofreading the manuscript. VF is grateful to Ciara Mann for assistance in collecting and maintaining aphids, to Edgar Wong for lab assistance, to Jason Hogg for maintenance of controlled temperature rooms, and to the department of Plant Sciences of the University of Oxford for hosting after the closure of the Tinbergen building. The authors acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work. <http://dx.doi.org/10.5281/zenodo.22558>. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 659847 (to VF), and from the John Fell Fund grant no. 0005545 from the University of Oxford (to VF).

## Author Contributions

VF designed the study, collected the data and performed the mutation accumulation experiment. VF and BN performed the analysis. VF wrote the manuscript with input from BN. Both authors agree with the content of the manuscript.

## Competing interests

The authors declare no competing interests.

## References

- Ballard JWO, Whitlock MC. 2004. The incomplete natural history of mitochondria. *Mol. Ecol.* 13:729–744.
- Bierne N, Gagnaire P, David P. 2013. The geography of introgression in a patchy environment and the thorn in the side of ecological speciation. *Curr. Zool.* 59:72–86.
- Bromham L. 2009. Why do species vary in their rate of molecular evolution? *Biol. Lett.* 5:401–

351 404.

352 Coyne JA, Orr HA. 1997. "Patterns of speciation in *Drosophila* " Revisited. *Evolution* (N. Y).  
353 51:295–303.

354 Denver DR, Wilhelm LJ, Howe DK, Gafner K, Dolan PC, Baer CF. 2012. Variation in base-  
355 substitution mutation in experimental and natural lineages of *Caenorhabditis* nematodes.  
356 *Genome Biol. Evol.* 4:513–522.

357 Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for  
358 multiple tools and samples in a single report. *Bioinformatics* 32:3047–3048.

359 Fazalova V, Nevado B, McLean A, Godfray HCJ. 2018. Intrinsic pre-zygotic reproductive isolation  
360 of distantly related pea aphid host races. *Biol. Lett.* 14:20180332.

361 Feder JL, Flaxman SM, Egan SP, Comeault AA, Nosil P. 2013. Geographic mode of speciation  
362 and genomic divergence. *Annu. Rev. Ecol. Evol. Syst.* 44:73–97.

363 Ferrari J, Godfray HCJ, Faulconbridge AS, Prior K, Via S. 2006. Population differentiation and  
364 genetic variation in host choice among pea aphids from eight host plant genera. *Evolution*  
365 60:1574–1584.

366 Ferrari J, West JA, Via S, Godfray HCJ, Al ET. 2012. Population genetic structure and secondary  
367 symbionts in host-associated populations of the pea aphid complex. *Evolution* (N. Y).  
368 66:375–390.

369 Flynn JM, Chain FJJ, Schoen DJ, Cristescu ME. 2017. Spontaneous mutation accumulation in  
370 *Daphnia pulex* in selection-free vs. competitive environments. *Mol. Biol. Evol.* 34:160–173.

371 Futuyma D. 2008. Sympatric speciation: norm or exception? In: Tilmon KJ, editor. Specialization,  
372 speciation, and radiation: the evolutionary biology of herbivorous insects. Vol. 91. 1st ed.  
373 University of California Press. p. 136–148.

374 Harrison RG, Larson EL. 2016. Heterogeneous genome divergence, differential introgression,  
375 and the origin and structure of hybrid zones. *Mol. Ecol.* 25:2454–2466.

376 Hewitt GM. 2004. Genetic consequences of climatic oscillations in the Quaternary. Willis KJ,  
377 Bennett KD, Walker D, editors. *Philos. Trans. R. Soc. London. Ser. B Biol. Sci.* 359:183–  
378 195.

379 Jaquiere J, Peccoud J, Ouisse T, Legeai F, Prunier-Leterme N, Gouin A, Nouhaud P, Brisson JA,

380 Bickel R, Purandare S, et al. 2019. Disentangling the causes for faster-X evolution in  
381 aphids. *Genome Biol. Evol.* 10:520.

382 Jaquiéry J, Stoeckel S, Nouhaud P, Mieuze L, Mahéo F, Legeai F, Bernard N, Bonvoisin A,  
383 Vitalis R, Simon J-C. 2012. Genome scans reveal candidate regions involved in the  
384 adaptation to host plant in the pea aphid complex. *Mol. Ecol.* 21:5251–5264.

385 Keightley PD, Ness RW, Halligan DL, Haddrill PR. 2014. Estimation of the spontaneous mutation  
386 rate per nucleotide site in a *Drosophila melanogaster* full-sib family. *Genetics* 196:313–320.

387 Keightley PD, Pinharanda A, Ness RW, Simpson F, Dasmahapatra KK, Mallet J, Davey JW,  
388 Jiggins CD. 2015. Estimation of the spontaneous mutation rate in *Heliconius melpomene*.  
389 *Mol. Biol. Evol.* 32:239–243.

390 Kozak KM, Wahlberg N, Neild AFE, Dasmahapatra KK, Mallet J, Jiggins CD. 2015. Multilocus  
391 species trees show the recent adaptive radiation of the mimetic *Heliconius* butterflies. *Syst.*  
392 *Biol.* 64:505–524.

393 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.  
394 *Bioinformatics* 25:1754–1760.

395 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.  
396 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.

397 Liu H, Jia Y, Sun X, Tian D, Hurst LD, Yang S. 2017. Direct determination of the mutation rate in  
398 the bumblebee reveals evidence for weak recombination-associated mutation and an  
399 approximate rate constancy in insects. *Mol. Biol. Evol.* 34:119–130.

400 Long H, Winter DJ, Chang AY-C, Sung W, Wu SH, Balboa M, Azevedo RBR, Cartwright RA,  
401 Lynch M, Zufall RA. 2016. Low base-substitution mutation rate in the germline genome of  
402 the ciliate *Tetrahymena thermophila*. *Genome Biol. Evol.* 8:3629–3639.

403 Loxdale HD, Balog A. 2018. Aphid specialism as an example of ecological-evolutionary  
404 divergence. *Biol. Rev.* 93:642–657.

405 Mandrioli M, Zamboni G, Manicardi GC. 2017. Comparative gene mapping as a tool to  
406 understand the evolution of pest crop insect chromosomes. *Int. J. Mol. Sci.* 18:1919.

407 Manicardi GC, Mandrioli M, Blackman RL. 2015. The cytogenetic architecture of the aphid  
408 genome. *Biol. Rev.* 90:112–125.

- 409 Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads.  
410 *EMBnet.journal* 17:10.
- 411 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler  
412 D, Gabriel S, Daly M, et al. 2010. The genome analysis toolkit: A MapReduce framework for  
413 analyzing next-generation DNA sequencing data. *Genome Res.* 20:1297–1303.
- 414 Moran NA, McLaughlin HJ, Sorek R. 2009. The dynamics and time scale of ongoing genomic  
415 erosion in symbiotic bacteria. *Science* 323:379–382.
- 416 Müller FP. 1962. Biotypen und Unterarten der „Erbsenlaus“ *Acyrtosiphon pisum* (Harris).  
417 *Zeitschrift für Pflanzenkrankheiten und Pflanzenschutz* 69:129–136.
- 418 Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction  
419 endonucleases. *Proc. Natl. Acad. Sci. U. S. A.* 76:5269–5273.
- 420 Nouhaud P, Gautier M, Gouin A, Jaquière J, Peccoud J, Legeai F, Mieuzet L, Smadja CM,  
421 Lemaitre C, Vitalis R, et al. 2018. Identifying genomic hotspots of differentiation and  
422 candidate genes involved in the adaptive divergence of pea aphid host races. *Mol. Ecol.*  
423 27:3287–3300.
- 424 Oppold A-M, Pfenninger M. 2017. Direct estimation of the spontaneous mutation rate by short-  
425 term mutation accumulation lines in *Chironomus riparius*. *Evol. Lett.* 1:86–92.
- 426 Peccoud J, Mahéo F, de la Huerta M, Laurence C, Simon J-C. 2015. Genetic characterisation of  
427 new host-specialised biotypes and novel associations with bacterial symbionts in the pea  
428 aphid complex. Leather SR, editor. *Insect Conserv. Divers.* 8:484–492.
- 429 Peccoud J, Ollivier A, Plantegenest M, Simon J-C. 2009. A continuum of genetic divergence from  
430 sympatric host races to species in the pea aphid complex. *Proc. Natl. Acad. Sci. U. S. A.*  
431 106:7495–7500.
- 432 Peccoud J, Simon J-C. 2010. The pea aphid complex as a model of ecological speciation. *Ecol.*  
433 *Entomol.* 35:119–130.
- 434 Peccoud J, Simon J-C, McLaughlin HJ, Moran NA. 2009. Post-Pleistocene radiation of the pea  
435 aphid complex revealed by rapidly evolving endosymbionts. *Proc. Natl. Acad. Sci. U. S. A.*  
436 106:16315–16320.
- 437 Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ. 2014. PopGenome: an efficient Swiss  
438 army knife for population genomic analyses in R. *Mol. Biol. Evol.* 31:1929–1936.

439 Pool JE. 2015. The mosaic ancestry of the *Drosophila* genetic reference panel and the *D.*  
440 *melanogaster* reference genome reveals a network of epistatic fitness Interactions. *Mol.*  
441 *Biol. Evol.* 32:3236–3251.

442 Presgraves DC. 2018. Evaluating genomic signatures of “the large X-effect” during complex  
443 speciation. *Mol. Ecol.* 27:3822–3830.

444 Shaw KL. 2002. Conflict between nuclear and mitochondrial DNA phylogenies of a recent species  
445 radiation: what mtDNA reveals and conceals about modes of speciation in Hawaiian  
446 crickets. *Proc. Natl. Acad. Sci. U. S. A.* 99:16122–16127.

447 Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H. 2000. Genome sequence of the  
448 endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* 407:81–86.

449 Stewart JR, Lister AM, Barnes I, Dalén L. 2010. Refugia revisited: individualistic responses of  
450 species in space and time. *Proc. R. Soc. B Biol. Sci.* 277:661–671.

451 Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-  
452 performance genomics data visualization and exploration. *Brief. Bioinform.* 14:178–192.

453 Tóth L. 1933. Über die frühembryonale Entwicklung der viviparen Aphiden. *Zeitschrift für*  
454 *Morphol. und Ökologie der Tiere* 4:692–731.

455 Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM. 2007. Primer3Plus, an  
456 enhanced web interface to Primer3. *Nucleic Acids Res.* 35:W71–W74.

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

459 Via S, Conte G, Mason-Foley C, Mills K. 2012. Localizing  $F_{ST}$  outliers on a QTL map reveals  
460 evidence for large genomic regions of reduced gene exchange during speciation-with-gene-  
461 flow. *Mol. Ecol.* 21:5546–5560.

462 Via S, West J. 2008. The genetic mosaic suggests a new role for hitchhiking in ecological  
463 speciation. *Mol. Ecol.* 17:4334–4345.

464 Watterson GA. 1975. On the number of segregating sites in genetical models without  
465 recombination. *Theor. Popul. Biol.* 7:256–276.

466 Xu S, Stapley J, Gablenz S, Boyer J, Appenroth KJ, Sree KS, Gershenzon J, Widmer A, Huber  
467 M. 2019. Low genetic variation is associated with low mutation rate in the giant duckweed.

468       *Nat. Commun.* 10:1243.

469   Yang S, Wang L, Huang J, Zhang X, Yuan Y, Chen J-Q, Hurst LD, Tian D. 2015. Parent–progeny  
470       sequencing indicates higher mutation rates in heterozygotes. *Nature* 523:463–467.

471   Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. 2012. Primer-BLAST: a tool  
472       to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* 13:134.

473

**Table 1.** Number of callable sites, number of mutations and mutation rate (per site per haploid genome per generation) for each asexual lineage.

| Line          | Host race              | Number of callable sites | Number of mutations | Mutation rate (uncorrected) | Mutation rate (corrected for false positives) |
|---------------|------------------------|--------------------------|---------------------|-----------------------------|---|
| 05b           | <i>M. lupulina</i>     | 175,281,226              | 4                   | $4.2 \times 10^{-10}$       | $3.5 \times 10^{-10}$                         |
| 09b           | <i>L. corniculatus</i> | 185,010,536              | 5                   | $5.0 \times 10^{-10}$       | $4.2 \times 10^{-10}$                         |
| 107a          | <i>Lat. pratensis</i>  | 196,589,460              | 2                   | $1.9 \times 10^{-10}$       | $1.6 \times 10^{-10}$                         |
| 118a          | <i>L. corniculatus</i> | 280,177,299              | 4                   | $2.6 \times 10^{-10}$       | $2.2 \times 10^{-10}$                         |
| 120a          | <i>Lat. pratensis</i>  | 189,053,054              | 8                   | $7.8 \times 10^{-10}$       | $6.5 \times 10^{-10}$                         |
| 122a          | <i>L. corniculatus</i> | 258,621,867              | 1                   | $7.2 \times 10^{-11}$       | $6.0 \times 10^{-11}$                         |
| 29d           | <i>Lat. pratensis</i>  | 155,191,035              | 2                   | $2.4 \times 10^{-10}$       | $2.0 \times 10^{-10}$                         |
| 37a           | <i>L. pedunculatus</i> | 166,587,314              | 3                   | $3.3 \times 10^{-10}$       | $2.8 \times 10^{-10}$                         |
| 38a           | <i>L. pedunculatus</i> | 185,499,166              | 6                   | $6.0 \times 10^{-10}$       | $5.0 \times 10^{-10}$                         |
| 54a           | <i>M. lupulina</i>     | 258,825,042              | 4                   | $2.9 \times 10^{-10}$       | $2.4 \times 10^{-10}$                         |
| 81a           | <i>L. pedunculatus</i> | 204,485,061              | 2                   | $1.8 \times 10^{-10}$       | $1.5 \times 10^{-10}$                         |
| PT            | <i>L. arenarius?</i>   | 213,058,132              | 2                   | $1.7 \times 10^{-10}$       | $1.4 \times 10^{-10}$                         |
| <b>Total:</b> |                        | 2,468,379,192            | 43                  | $3.2 \times 10^{-10}$       | $2.7 \times 10^{-10}$                         |