

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34

Unravelling the genomic basis and evolution of the pea aphid male wing dimorphism

Binshuang Li^{1^}, Ryan D. Bickel^{1^}, Benjamin J. Parker¹, Neetha Nanoth Vellichirammal², Mary Grantham¹, Jean-Christophe Simon³, David L. Stern⁴, and Jennifer A. Brisson^{1*}

1. Department of Biology, University of Rochester, USA 14627, USA

2. University of Nebraska Medical Center, Omaha, NE 68198, USA

3. INRA/Agrocampus Ouest/Université Rennes 1, Le Rheu, Rennes 35653, France

4. Janelia Research Campus, Ashburn, VA 20147, USA

[^]Indicates equal contributions

* Correspondence to Jennifer.brisson@rochester.edu

Keywords: polymorphism, wing dimorphism, pea aphid, dispersal, genetic mapping

35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68

Summary

Wing dimorphisms have long served as models for examining the ecological and evolutionary tradeoffs associated with alternative morphologies [1], yet the mechanistic basis of morph determination remains largely unknown. Here we investigate the genetic basis of the pea aphid (*Acyrtosiphon pisum*) wing dimorphism, wherein males exhibit one of two alternative morphologies that differ dramatically in a set of correlated traits that included the presence or absence of wings [2-4]. Unlike the environmentally-induced asexual female aphid wing polyphenism [5], the male wing polymorphism is genetically determined by a single uncharacterized locus on the X chromosome called *aphicarus* (“aphid” plus “Icarus”, *api*) [6, 7]. Using recombination and association mapping, we localized *api* to a 130kb region of the pea aphid genome. No nonsynonymous variation in coding sequences strongly associated with the winged and wingless phenotypes, indicating that *api* is likely a regulatory change. Gene expression level profiling revealed an aphid-specific gene from the region expressed at higher levels in winged male embryos, coinciding with the expected stage of *api* action. Comparison of the *api* region across biotypes (pea aphid populations specialized to different host plants that began diverging ~16,000 years ago [8, 9]) revealed that the two alleles were likely present prior to biotype diversification. Moreover, we find evidence for a recent selective sweep of a wingless allele since the biotypes diversified. In sum, this study provides insight into how adaptive, complex traits evolve within and across natural populations.

69

Results and Discussion

70 **Mapping identifies the *api* locus.** Winged and wingless male pea aphids (Figure 1A) are
71 genetically determined by a single locus called *aphicarus* (*api*). This locus was previously
72 localized to a 10cM region on the X chromosome (Figure 1B, top) [6], a chromosome estimated
73 to contain a third of the genome [10]. To narrow down this region, we selfed F1 individuals from
74 the *api* mapping line produced by this original study to create F2 individuals. Using a panel of
75 448 of those F2 pea aphids, we simultaneously identified and scored single nucleotide
76 polymorphisms (SNPs) using multiplexed shotgun sequencing [11]. QTL analysis of these data
77 resulted in the identification of 19 scaffolds containing SNPs with LOD scores higher than the
78 1% significance level of 7.6 generated by 1000 permutations (Table S1).

79

80 Concurrently, to perform genome-wide association mapping, we sequenced the genomes of 44
81 pooled winged and 44 pooled wingless males collected from alfalfa (*Medicago sativa*) plants
82 across the U.S. (Table S2) to a total of 68X and 70X coverage, respectively. F_{ST} analysis
83 between the winged and wingless sequenced pools revealed two scaffolds with high levels of
84 differentiation (Figure 1C). The first was a scaffold identified from the QTL analysis (LOD=17.5),
85 while the second was not considered in the QTL analysis because it was smaller (42 kb) than
86 the minimum scaffold size used in QTL analysis (scaffolds>100kb).

87

88 We physically ordered the 19 genomic scaffolds containing SNPs with LOD scores greater than
89 7.6, plus the smaller scaffold with a high F_{ST} value. For ordering, we assayed a restriction
90 fragment length polymorphism (RFLP) marker for each scaffold using a panel of 40 F2
91 individuals that each carried a recombination event between the previous closest *api* flanking
92 markers identified by Braendle et al. [6]. 16 of the 19 scaffolds were localized within these
93 flanking markers, confirming the effectiveness of the QTL analysis. Two of these 16 scaffolds
94 contained RFLPs with perfect association with all 40 F2s (see the *api* region noted in Figure
95 1B); these were the same two scaffolds identified from the F_{ST} analysis. Thus, recombination
96 and association mapping each implicated the same two genomic scaffolds, a smaller scaffold
97 (~42kb, GL351389) and a larger scaffold (>350kb, GL349773; this scaffold is misassembled
98 after position ~350kb). We compared these scaffolds to their homologous regions from the
99 peach-potato aphid (*Myzus persicae*) and Russian wheat aphid (*Diuraphis noxia*) genomes [12,
100 13] and determined that these two scaffolds sit proximately, in opposing orientation (Figure S1).
101 Both of these species have a single genomic scaffold that spans the entire *api* region, with each
102 species' single scaffold containing homologous regions to the two pea aphid scaffolds. We

103 developed additional RFLP markers in this region, which narrowed down the *api* region to
104 between position 25kb on the smaller scaffold and position 107kb on the larger one, defining an
105 approximately 130kb *api* region spanning the two scaffolds.

106

107 **High associations between SNPs in the *api* region and the winged and wingless males.**

108 The winged and wingless male pooled sequence (pool-seq) data highlighted many SNPs across
109 the ~130kb *api* region that are strongly associated with the male wing phenotype (Figure 2A).

110 The wide distribution of associated SNPs may indicate that this is a region of low recombination

111 or that the *api* phenotype is generated by multiple SNPs that are maintained in linkage

112 disequilibrium. The pool-seq data were generated from pea aphids collected from alfalfa. We

113 wanted to determine whether a broader sample, across biotypes, would narrow down the

114 causative polymorphism. As noted above, the pea aphid is actually a species complex with as

115 many as 15 host plant adapted lineages, called biotypes, that began diverging 8,000-16,000

116 years ago [8, 9]. Biotypes have limited gene flow between them, and the ones with the least

117 genetic exchange have been described as incipient species [8, 9]. We used the complete

118 genomes of 23 genotypes from nine different biotypes (Table S3) to investigate patterns of

119 association in the *api* region: nine winged allele carrying genotypes from five biotypes, and 14

120 wingless carrying genotypes from six biotypes (two biotypes, pea and alfalfa, had individuals of

121 both allele classes). This data set thus overlapped with the alfalfa pool-seq data in that they

122 both contained alfalfa biotype aphids, but allowed us to further narrow the informative genomic

123 region. Because of the small sample size of this nine-biotype data set, here we focus

124 exclusively on the SNPs that perfectly segregated with the winged and wingless males. There

125 are 130 SNPs that meet that criteria. These 130 SNPs are indicated in orange, overlaid on the

126 pool-seq association data (Figure 2A).

127

128 These two association studies, combined, indicate that the region that most likely contains *api* is

129 the ~95kb region that includes the ~10kb end of GL351389 (the smaller *api* scaffold), and the

130 first 85kb on GL349773 (the larger scaffold) (Figure 2A). Within this 95kb region, there are

131 multiple regions with considerable sequence divergence such that the Illumina sequence reads

132 from winged alleles cannot be aligned to the wingless reference genome, appearing as regions

133 with low coverage from the winged pool (Figure 2B). These regions, from position ~5kb to

134 ~33kb and from position ~65kb to ~75kb, could potentially be the causative variation, or contain

135 the causative SNP(s). Across the whole genome, the winged and wingless pools were

136 sequenced with near-equal effort (68X and 70x respectively), so sequence differences between

137 the alleles are driving this pattern.

138

139 **A candidate gene at the *api* locus.** The ~130kb *api* region contains 13 annotated genes in the
140 pea aphid annotation, v.2.1 (Figure 2D; Table S4). The pea aphid genome annotation was
141 primarily derived from gene prediction algorithms with aid from sequence information from
142 female RNA libraries [14], leaving the possibility that male-specific genes in the region were
143 missed during annotation efforts. We therefore sequenced four RNA-Seq libraries constructed
144 from different stages of winged and wingless male embryonic cDNA, but did not detect any
145 unannotated genes in the region. Of the 13 annotated genes, seven have transposable
146 element-related annotations (Table S4). The remaining six genes code for three aphid-specific
147 proteins with no conserved domains (*as1*, *as2*, and *as3*), a mitochondrial sorting and assembly
148 gene (*sam50*), a fibroblast growth factor receptor substrate (*frs2*), and a chromodomain-
149 encoding gene (*cdg*) (Figure 2D). The seven TE-related genes, along with *as1*, have no
150 discernable gene expression in our male embryo RNA-seq data or the 38 RNA-Seq libraries (36
151 from females of different ages, two from adult males) publicly available on Aphidbase.com. *as2*,
152 *sam50*, *frs2*, and *cdg* all exhibited evidence of expression in our male-specific RNA-Seq
153 libraries (Table S4); *as3* was not expressed in our male RNA-seq libraries, but was expressed in
154 female libraries on Aphidbase [15]. We conclude these five genes (*as2*, *as3*, *sam50*, *frs2*, and
155 *cdg*) are the only annotated genes that are transcribed and thus the only functional genes in the
156 region.

157

158 We found no nonsynonymous changes that very strongly associated with the *api* phenotype
159 across the pool-seq and biotype data; all nonsynonymous sites contained multiple reads in the
160 pool-seq data that contradicted the association. We thus inferred that the mutation(s) that
161 differentiates winged and wingless males must be regulatory. We measured the expression
162 levels of the five genes (*as2*, *as3*, *sam50*, *frs2*, and *cdg*) that had evidence of transcription,
163 using qRT-PCR. Winged and wingless males are morphologically different by the second
164 nymphal instar [16] and wing morph determination in the environmentally induced wing
165 polyphenism in pea aphid females occurs embryonically [17]. We thus reasoned that the action
166 of *api* would occur embryonically, but that potentially the first nymphal instar may be important,
167 too. Therefore, we focused on two developmental stages: embryos and first instar nymphs.

168

169 We observed that *as3* is not expressed in males at these stages, consistent with the RNA-Seq
170 results. Among the four expressed genes (*as2*, *sam50*, *frs2*, and *cdg*), only *as2* was

171 differentially expressed between winged and wingless embryos (two-sided t-test, $P=0.01$; Figure
172 3), with two-fold higher expression in the winged embryos. No genes significantly differed in
173 expression as first instars. *as2* is found in all sequenced aphid genomes [12, 13, 18], but is not
174 found outside of aphids. *as2* is physically located in the center of our identified region near a
175 large number of linked markers. It is also the only gene in the region that shows differential
176 expression in the embryo stage, when *api* is likely to act. Thus, *as2* is the most likely candidate
177 for *api*, although this hypothesis requires further validation.

178

179 **Molecular evolution at the *api* locus across pea aphid biotypes.** To investigate the evolution
180 of the *api* region within the pea aphid complex, we used the 23 resequenced genomes from
181 nine biotypes. To determine the evolutionary history of the X chromosome in these lineages we
182 constructed a phylogeny based on DNA polymorphisms from scaffolds located across the X
183 chromosome, but not in the *api* region (Figure 4A). This analysis confirmed the genetic grouping
184 of individuals from the same biotype, and that there is a continuum of sequence divergence
185 across the complex of biotypes [9]. The winged and wingless phenotypes are scattered across
186 this phylogenetic tree. In contrast, when we constructed trees from 10kb windows across the *api*
187 region, we found complete separation of the winged and wingless genotypes (Figure 4B;
188 Figures S2-3), suggesting that the wing morph is determined by the same variation in the
189 different biotypes. Furthermore, the *api* region shows a different pattern of evolution than the
190 rest of the X chromosome. These data suggest that the winged and wingless alleles pre-existed
191 as standing variation before the biotypes split, and both alleles have been segregating in at
192 least some lineages since then.

193

194 **A recent selective sweep of the wingless allele.** The wingless alleles exhibit less genetic
195 differentiation than the winged alleles in the *api* region (short branch lengths in Figure 4B;
196 Figures S2-S3). To further explore this observation, we examined patterns of sequence
197 variation in the alfalfa biotype using the pool-seq data. Since the two alleles do not seem to be
198 freely recombining with one another, we examined Tajima's D [19] separately in the two alleles
199 to understand the evolutionary history of each. We found markedly lower Tajima's D values in
200 wingless males between positions ~35kb and ~65kb (Figure 2C), suggesting a selective sweep
201 in the wingless allele. This signature of a selective sweep is located just upstream of *as2*, our
202 *api* candidate gene. Within the *api* biotype trees (Figure 4B; Figures S2-S3), there is a long
203 branch leading to three individuals of the *Lathyrus* (Lap1-3) biotype, a biotype that does not
204 seem to hybridize with the other biotypes [9]. All examined biotypes currently carry both the

205 winged and wingless *api* alleles ([20], Li et al., unpublished data). While the winged and
206 wingless alleles arose before diversification of biotypes, it is not clear if both alleles have been
207 maintained by natural selection in all biotypes since then. There is still limited gene flow
208 between the biotypes [8] which could allow a biotype to recover an allele if it were lost. The
209 reduced variation and the pattern of Tajima's D in the wingless allele suggest that one wingless
210 allele variant has recently swept through the alfalfa biotype (Figure 2C) and the other biotypes,
211 except the reproductively isolated *Lathyrus* biotype (Figure 4). It is not clear if this is a new
212 evolutionarily advantageous wingless allele (or a linked favorable allele), or the reintroduction of
213 a wingless allele that has been lost from a large number of biotypes and recently become
214 favorable. In addition, the selective advantage of this wingless allele is unclear but decreased
215 dispersal due to the wingless phenotype could reinforce ecological speciation in the pea aphid
216 complex by increasing mating within biotypes [21].

217

218

Conclusions

219 Our study shows that the pea aphid provides a robust model for investigating the molecular
220 basis of morphological variation. We have identified a single ~130kb-region of the pea aphid X
221 chromosome that causes the differences between winged and wingless males. *as2*, an aphid-
222 specific gene in the region, is expressed at higher levels in winged embryos relative to wingless
223 embryos, making it a strong candidate for regulating this wing dimorphism. This study is the first
224 demonstration of the genetic basis of a dispersal dimorphism, which have evolved repeatedly
225 across insects because of the respective benefits of having winged and wingless morphs [1].
226 Moreover, we have demonstrated that a complex whole body phenotype can be regulated by
227 what is likely a single gene. Finally, our results form the foundation for future comparative
228 studies aimed at (1) discovering how male wing dimorphisms have been lost and gained
229 repeatedly during the evolution of aphids [22, 23], (2) better understanding the factors promoting
230 the maintenance of the male wing polymorphism, and (3) determining how the male wing
231 polymorphism relates to the environmentally determined female wing polyphenism [24], which is
232 present in most extant aphid taxa [3].

233

234

Figure Legends

235

236 **Figure 1. Linkage and association mapping of the *api* region.** (A) Winged (top) and wingless
237 (bottom) males. (B) The *api* region was initially localized to a 10cM region on the X chromosome
238 (green box, top [6]). Seven ordered scaffolds in this region are shown in green. Further

239 refinement of the linkage map narrowed the location of *api* to a ~130kb region spanning two
240 genomic scaffolds (black box). The inferred recombination breakpoints of eight recombinant F2
241 individuals using RFLP markers are shown below the scaffolds. Black indicates sequence from
242 the winged parent and grey from the wingless parent, and phenotype is indicated to the right
243 (W:winged, WL:wingless). The recombination breakpoints are approximate (see methods for
244 details). (C) Genome-wide view of genetic differentiation between winged and wingless males
245 using F_{st} values calculated from 20kb windows across all scaffolds greater than 20kb (1,896
246 scaffolds). Scaffolds are ordered by their scaffold number, which is roughly from largest to
247 smallest. F_{st} values from the 19 scaffolds identified from the recombination mapping analysis
248 are indicated with green points.

249

250 **Figure 2. Population genetics statistics in the 130kb *api* region defined by recombination**

251 **mapping.** (A) Points show the large number of highly associated SNPs in the *api* region,
252 illustrated as the $-\log(P\text{-value})$ from a Fisher's exact test between SNPs and the male wing
253 phenotype using the alfalfa biotype pool-seq data. Orange points are SNPs that are perfectly
254 segregating with the phenotype across the biotype data. 0 kb is the breakpoint between
255 scaffolds GL351389, the smaller *api* scaffold is indicated by negative values, and GL349773,
256 indicated with positive values. (B) The read count per site for pool-seq of winged and wingless
257 males is shown on the y-axis, with a maximum of 200. The lower read depth in winged males
258 despite near-equal sequence effort indicates sequence divergence. (C) The y-axis shows
259 Tajima's D values calculated across 10kb windows in 5kb steps for winged and wingless males
260 separately. Only windows with greater than 30% of sites having a read count of at least 20 are
261 presented. Some windows have insufficient data to calculate Tajima's D. The low Tajima's D
262 values in the wingless males indicate a selective sweep. (D) Annotated genes in the *api* region.
263 Grey indicates genes not expressed in any publicly available RNA-Seq data set, while black
264 indicates expressed genes. *as1*, *as2*, *as3*=three aphid-specific genes, *sam50*=mitochondrial
265 sorting and assembly gene, *frs2*=fibroblast growth factor receptor substrate,
266 *cdg*=chromodomain-containing gene.

267

268 **Figure 3. Gene expression levels for the four expressed genes in the *api* region.** Each
269 gene was measured in five biological replicates collected from winged and wingless male
270 embryos and first instar nymphs. Each point represents a replicate. Y-axes show the ΔC_T values
271 for each sample subtracted from the average ΔC_T value of wingless embryos ($-\Delta\Delta C_T$): higher

272 values therefore represent stronger relative gene expression. Short grey horizontal lines show
273 means. * indicates $P < 0.05$.

274

275 **Figure 4. Relationships among winged and wingless genotypes from a range of pea**

276 **aphid biotypes.** (A) The evolutionary history of the X chromosome based on 27.6 Mb of
277 sequence from across the X chromosome. (B) Phylogenetic tree based on positions 50-60kb in
278 scaffold GL349773 in the center of the *api* region. Winged (W) lines are in blue, wingless (WL)
279 in red. Trees are inferred using maximum likelihood, and numbers on the tree branches are
280 bootstrap values; bootstrap values below 75 are not shown. Abbreviations: Mo: *Melilotus*
281 *officinalis*, Tp: *Trifolium pratense*, Lap: *Lathyrus pratensis*, Ps: *Pisum sativum*, Ms: *Medicago*
282 *sativa*, Os: *Ononis spinosa*, Mes: *Melilotus suaveolens*, Vc: *Vicia cracca*, Cs: *Cytisus scoparius*.

283

284 **Supplemental Figure 1. Comparison of scaffolds from three aphid species.** The pea aphid,
285 *Acyrtosiphon pisum*, and *Myzus persicae* diverged from one another around 43 MYA [25];
286 *Diuraphis noxia* is more distantly related from both [26]. Lines are genomic scaffolds, with
287 arrows indicating the orientation of a scaffold. Whole scaffolds from *D. noxia* and *M. persicae*
288 are shown. The entire scaffolds of GL350308 and GL351389 and the homologous region of
289 GL349773 from *A. pisum* are presented and to scale (GL349773 is misassembled above
290 position ~350,000 and is not shown entirely).

291

292 **Supplemental Figures 2-3. Relationships among 9 winged and 14 wingless genomes from**
293 **across nine biotypes.** Unrooted trees are constructed from nucleotide variation from 10kb
294 windows in the *api* region. Information below each tree indicates the scaffold, the 10kb window
295 on the scaffold used, and the number of variable sites in that window. Trees for intervals
296 containing less than 150 variable sites are not shown. Individuals containing *api* winged alleles
297 are colored in blue and wingless alleles in red. Numbers on the tree branches are bootstrap
298 values. Abbreviations: Mo: *Melilotus officinalis*, Tp: *Trifolium pratense*, Lap: *Lathyrus*
299 *pratensis*, Ps: *Pisum sativum*, Ms: *Medicago sativa*, Os: *Ononis spinosa*, Mes: *Melilotus*
300 *suaveolens*, Vc: *Vicia cracca*, Cs: *Cytisus scoparius*;

301

302

303

304

305

Materials and Methods

306

307

308 **Linkage mapping.** Braendle et al. [6] previously established an *api* linkage mapping population.
309 We used the F1 line from that population to generate additional F2 recombinants. Specifically,
310 F1 asexual females were placed on *Vicia faba* plants in an incubator at 16°C and a photoperiod
311 of 12h light and 12h dark. After two generations of asexual reproduction, these conditions
312 induce the production of sexual females and males. We crossed F1 females to F1 males,
313 collected fertilized eggs, sterilized them in 1% calcium propionate on Whatman paper, and
314 placed them in an incubator that alternated between 4°C for 12h light and 0°C for 12hr dark.
315 After 90 days, eggs were removed from this incubator and placed in a 19°C incubator that
316 alternated between 16h light and 8h dark. F2 hatchlings were transferred to individual plants for
317 asexual reproduction to establish a line of that F2 individual.

318

319 Genomic DNA of 448 F2 females and 22 F2 males were isolated, quantified, and diluted to
320 2ngs/μl. Multiplexed shotgun sequencing of all 470 F2 individuals was carried out according to
321 [27] with minor modifications: individual F2 DNA quantity for Mse1 digestion was increased to
322 10ngs and the unique barcoded adapter concentration was reduced to 2.5nmols. These
323 changes were made to increase ligation efficiency and to reduce the formation of ligated linker-
324 dimers. The final amplified libraries were sequenced on an Illumina Genome Analyzer.

325

326 We identified genomic scaffolds that exhibited linkage with *api* using Rqtl [28]. For each scaffold
327 of interest, we developed a diagnostic restriction fragment length polymorphism (RFLP) marker.
328 RFLP markers were tested on a panel of 10 F2 individuals to confirm that the scaffold was
329 indeed X-linked and linked to *api*. The scaffolds were ordered relative to one another using a
330 panel of up to 40 F2 individuals. Recombination breakpoints for the two scaffolds in the *api*
331 region were localized to within ~10kb (left side defined by markers at positions 15,767 and
332 25,460 on GL351389 and right side by markers at positions 97,258 and 107,073 on GL349773).
333 The others scaffolds contained only one RFLP marker each and thus breakpoints in Figure 1C
334 are approximated by representing them in the middle of the adjacent scaffolds.

335

336 **Pool-seq.** We used 44 winged and 44 wingless male pea aphids induced from females
337 collected from Nebraska, New York, California and Massachusetts (Table S2). Genomic DNA
338 (gDNA) from each male was extracted using the Qiagen DNeasy Blood & Tissue Kit. 50ng of
339 gDNA from each male was pooled together with males of the same phenotype. Paired-end

340 libraries were prepared with the TruSeq DNA PCR-Free Library Preparation Kit at the University
341 of Rochester Genomics Research Center and sequenced on an Illumina HiSeq2500 Sequencer
342 with paired 125nt reads. Reads were mapped to the pea aphid reference genome v.2 using bwa
343 (v.0.7.9a-r786) [29] using default parameters. Reads were filtered for a mapping quality of 20
344 and BAM files were sorted by coordinates using samtools (Version: 0.1.19-44428cd). The
345 coverage of both libraries was calculated using the samtools depth function for the mapped
346 reads.

347
348 **Association, F_{ST} , and Tajima's D analyses.** An mpileup file was created from the winged and
349 wingless male BAM files using the samtools mpileup function with the -B option to disable BAQ
350 computation. This was further processed by the mpileup2sync.jar script in PoPoolation2 v.1.201
351 [30] to generate a synchronized mpileup file (sync file), with fastq type set to sanger and
352 minimum quality set to 20. For association analyses, Fisher's exact tests were performed using
353 the fisher-test.pl script included in PoPoolation2 with a window-size of 1, step-size of 1,
354 minimum coverage of 2, and minimum allele count of 2. F_{ST} values were calculated with the fst-
355 sliding.pl script included in PoPoolation2 with a window size set to 20kb, a step size set to 10kb,
356 minimum coverage set to 30, a maximum coverage set to 200, pool size set to 44. The
357 variance-sliding.pl script was used to calculate Tajima's D, setting a minimum allele count of 2,
358 a minimum base quality of 20, a minimum coverage of 10, a maximum coverage of 400, a pool
359 size of 44, a window size of 10,000, and a step size of 5,000; fastq type was set to sanger. Data
360 are shown for windows with a 30% minimum covered sequence fraction.

361
362 **Fully re-sequenced genomes.** In addition to the reference pea aphid genome which is *api*
363 wingless homozygous, we obtained the sequence of 22 additional pea aphid genomes: 9
364 carrying only the winged allele and 13 the wingless allele (Table S3). These genotypes have
365 been collected in the wild as parthenogenetic females, on distinct legume species. Biotypes of
366 these genotypes have been assigned based on their microsatellite profiles [9] and
367 representatives have been then selected for genome resequencing [31]. Lines were sequenced
368 to 17X to 30X coverage with Illumina 100nt paired-end reads. The reads of each sample were
369 aligned to the pea aphid reference genome with default settings in bowtie2 (version 2.2.1) [32].
370 The consensus sequence of each sample was acquired using the recommended pipeline in
371 samtools (version 1.3.1) and bcftools (version 1.3).

372
373 **Phylogenetic trees.** Trees were built using Raxml v. 8.2.9 [33] using the substitution model

374 GTRGAMMA. One thousand bootstraps were run on distinct starting trees. For the neutral tree,
375 we used a concatenated data set of genomic scaffolds from the pea aphid genome build v.2
376 with a probability of being on the X above 90% (111 scaffolds, total of 27.6 Mb [34]). The *api*
377 trees were constructed using the 9 winged and 14 wingless biotype sequences in 10kb
378 intervals. The trees were graphed using Figtree (v 1.4.0)[] and rooted by the included midpoint
379 method.

380
381 **Male embryo RNA-seq.** Stage 18 embryos [35] and a mixed sample of embryos younger than
382 stage 18 were obtained separately for F2 lines homozygous for the *api* winged or wingless allele
383 (four libraries total). Total RNA was isolated using TRIzol® (Thermo Fisher Scientific, Waltham,
384 MA). Library construction was performed using the TruSeq Stranded mRNA Sample
385 Preparation Kit (Illumina, San Diego, CA). The four libraries were constructed and sequenced
386 with single end 100 nt reads in one lane of an Illumina HiSeq 2500 sequencer at the University
387 of Rochester Genomics Research Center. The reads were aligned to the reference genome
388 using bowtie2 [32] and processed into bam files using samtools with default settings. Bam files
389 were visualized with IGV (v.2.3.72) [36]. The coverage of mapped reads was reported using
390 samtools and compared to the gene annotations. No coverage of more than six reads was
391 discovered outside of annotated exon regions within the *api* region.

392
393 **Quantitative reverse-transcriptase PCR (qRT-PCR).** We also used the two *api* homozygous
394 F2 lines to collect male embryos and first instar nymphs. To produce males, we transferred
395 asexual female adults into an incubator at 16°C with a photoperiod of 12h light and 12h dark.
396 Two generations in this environment resulted in asexual females whose offspring would be
397 males and sexual females. We dissected stage 18 [35] embryos from the females or collected
398 first instar nymphs. We confirmed the sex of embryo or nymph using an RFLP on the X
399 chromosomes (forward primer: ATCGATGCTTTTGAATTGTTTTAC; reverse primer:
400 TGTAGGGTCTCTTGAAGTTGTTTG; restriction enzyme: TaqαI; double bands are females
401 while single bands are males) while simultaneously collecting tissue for RNA as in [37]. Five
402 biological replicates were included. Each embryo replicate contained 20 embryos from 6 to 10
403 females, while each first instar nymph replicate contained 10 individuals produced by 3 to 5
404 females. Quantitative PCR was performed on a Bio-Rad CFX-96 Real-Time System using 12μL
405 reactions of 40ng cDNA, 1X PCR buffer, 2nM Mg²⁺, 0.2nM dNTPs, 1X EvaGreen, and 0.025
406 units/μL Invitrogen Taq with the following conditions: 95°C 3 min, 40x (95°C 10s, 55°C 30s).
407 Primer concentrations were optimized to 100+/-5% reaction efficiency with an R² value of > 0.99

408 [G3PDH (ACYPI009769): 400nM Forward primer, 350nM Reverse primer; NADH
409 (ACYPI009382): 350nM F, 300nM R primer; 2281: 175nM; 25525: 150nM; 25532: 250nM; and
410 25533: 150nM]. Each of the five biological replicates was run on a single plate, with three
411 technical replicates of each reaction. ΔC_T values were calculated by subtracting the average C_T
412 value of the two endogenous controls (G3PDH and NADH) from the C_T values of each target
413 gene. For each pair of winged and wingless samples, ΔC_T values were analyzed using two-
414 sided t-tests after checking for normality.

415

416

Acknowledgements

417 We gratefully thank Jen Keister for technical assistance. This research was supported by award
418 R00 and R01GM116867 from the National Institute of General Medical Sciences to J.A.B. RNA-
419 Seq and pool-seq sequencing was performed at the University of Rochester Genomics
420 Research Center. J.C.S. was supported by the Agence Nationale de la Recherche (grant ANR-
421 11-BSV7-007). Sequencing of the YR2 line was performed at the Genoscope (project 62 AAP
422 2009/2010 to J.C.S.). We thank Prof. Shin-Ichi Akimoto who provided sample and *api*
423 phenotype for pea aphid lines 09003A, Sap05VC7 and Iwamizawa.

424

425

Author Contributions

426 Conceptualization, J.A.B., R.D.B., and D.L.S.; Methodology, J.A.B., R.D.B., and D.L.S.; Formal
427 Analysis: B.L., R.D.B., and D.L.S.; Investigation, B.L., R.D.B., B.J.P., N.N.V., M.G., D.L.S., and
428 J.A.B.; Writing – Original Draft, J.A.B., R.D.B., and B. L.; Writing – Review & Editing, D.L.S. and
429 J.-C.S.; Funding Acquisition, J.A.B.; Resources, J.-C.; Supervision, J.A.B.

430

431

432

References Cited

433

- 434 1. Zera, A.J. and R.F. Denno. 1997. Physiology and ecology of dispersal polymorphism in
435 insects. *Annu Rev Entomol* 42.
- 436 2. Dixon, A.F.G. and M.T. Howard, Dispersal in aphids, a problem in resource allocation, in
437 *Insect Flight: Dispersal and Migration*, W. Danthanarayana, Editor. 1986, Springer-
438 Verlag: Berlin. p. 145-151.
- 439 3. Braendle, C., G.K. Davis, J.A. Brisson and D.L. Stern. 2006. Wing dimorphism in aphids.
440 *Heredity* 97: 192-199.
- 441 4. Ogawa, K., A. Ishikawa, T. Kanbe, S.-i. Akimoto and T. Miura. 2012. Male-specific flight
442 apparatus development in *Acyrtosiphon pisum* (Aphididae, Hemiptera, Insecta):
443 comparison with female wing polyphenism. *Zoomorphology* 131(3): 197-207.
- 444 5. Dixon, A.F.G., *Biology of Aphids*. 1973, London: Edward Arnold Ltd.

- 445 6. Braendle, C., M.C. Caillaud and D.L. Stern. 2005. Genetic mapping of *aphicarus* - a sex-
446 linked locus controlling a wing polymorphism in the pea aphid (*Acyrtosiphon pisum*).
447 Heredity 94: 435-442.
- 448 7. Caillaud, M.C., M. Boutin, C. Braendle and J.-C. Simon. 2002. A sex-linked locus
449 controls wing polymorphism in males of the pea aphid, *Acyrtosiphon pisum* (Harris).
450 Heredity 89: 346-352.
- 451 8. Peccoud, J., A. Ollivier, M. Plantagenest and J.-C. Simon. 2009. A continuum of genetic
452 divergence from sympatric host races to species in the pea aphid complex. Proc. Nat.
453 Acad. Sci.
- 454 9. Peccoud, J., J.-C. Simon, H.J. McLaughlin and N.A. Moran. 2009. Post-Pleistocene
455 radiation of the pea aphid complex revealed by rapidly evolving endosymbionts.
456 Proceedings of the National Academy of Sciences of the United States of America
457 106(38): 16315-16320.
- 458 10. Jaquiere, J., J. Peccoud, T. Ouisse, F. Legeai, N. Prunier-Leterme, A. Gouin, P.
459 Nouhaud, J.A. Brisson, R.D. Bickel, S. Purandare, J. Poulain, C. Battail, C. Lemaitre, L.
460 Mieuzet, G. Le Trionnaire, J.C. Simon, and C. Rispe. In review. Disentangling the
461 causes for faster-X evolution in aphids. bioRxiv 125310.
- 462 11. Andolfatto, P. and M. Przeworski. 2000. A genome-wide departure from the standard
463 neutral model in natural populations of *Drosophila*. Genetics 156: 257-268.
- 464 12. Mathers, T.C., Y. Chen, G. Kaithakottil, F. Legeai, S.T. Mugford, P. Baa-Puyoulet, A.
465 Bretaudeau, B. Clavijo, S. Colella, O. Collin, T. Dalmay, T. Derrien, H. Feng, T.
466 Gabaldon, A. Jordan, I. Julca, G.J. Kettles, K. Kowitzanich, D. Lavenier, P. Lenzi, S.
467 Lopez-Gomollon, D. Loska, D. Mapleson, F. Maumus, S. Moxon, D.R. Price, A. Sugio, M.
468 van Munster, M. Uzest, D. Waite, G. Jander, D. Tagu, A.C. Wilson, C. van Oosterhout, D.
469 Swarbreck, and S.A. Hogenhout. 2017. Rapid transcriptional plasticity of duplicated
470 gene clusters enables a clonally reproducing aphid to colonise diverse plant species.
471 Genome Biol 18(1): 27.
- 472 13. Nicholson, S.J., M.L. Nickerson, M. Dean, Y. Song, P.R. Hoyt, H. Rhee, C. Kim, and G.J.
473 Puterka. 2015. The genome of *Diuraphis noxia*, a global aphid pest of small grains. BMC
474 Genomics 16: 429.
- 475 14. Li, H. and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-
476 Wheeler transform. Bioinformatics 25(14): 1754-60.
- 477 15. IAGC. 2010. Genome sequence of the pea aphid *Acyrtosiphon pisum*. PLoS Biol. 8:
478 e1000313.
- 479 16. Legeai, F., S. Shigenobu, J.-P. Gauthier, J. Colbourne, C. Rispe, O. Collin, S. Richards,
480 A.C.C. Wilson, and D. Tagu. 2010. AphidBase: A centralized bioinformatic resource for
481 annotation of the pea aphid genome. Insect molecular biology 19(0 2): 5-12.
- 482 17. Ogawa, K. and T. Miura. 2013. Two developmental switch points for the wing
483 polymorphisms in the pea aphid *Acyrtosiphon pisum*. EvoDevo 4(1): 30.
- 484 18. Sutherland, O.R.W. 1969. The role of crowding in the production of winged forms by two
485 strains of the pea aphid, *Acyrtosiphon pisum*. J Insect Physiol 15.
- 486 19. Wenger, J.A., B.J. Cassone, F. Legeai, J.S. Johnston, R. Bansal, A.D. Yates, B.S.
487 Coates, V.A. Pavinato, and A. Michel. 2017. Whole genome sequence of the soybean
488 aphid, *Aphis glycines*. Insect Biochem Mol Biol.
- 489 20. Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA
490 polymorphism. Genetics 123(3): 585-595.
- 491 21. Frantz, A., M. Plantegenest and J.-C. Simon. 2010. Host races of the pea aphid
492 *Acyrtosiphon pisum* differ in male wing phenotypes. Bull. Ent. Res. 100: 59-66.
- 493 22. Peccoud, J. and J.C. Simon. 2010. The pea aphid complex as a model of ecological
494 speciation. Ecological Entomology 35: 119-130.

- 495 23. Smith, M.A.H. and P.A. MacKay. 1989. Genetic variation in male alary dimorphism in
496 populations of pea aphid, *Acyrtosiphon pisum*. *Entomol Exp Appl* 51.
- 497 24. Brisson, J.A. 2010. Aphid wing dimorphisms: linking polyphenism and polymorphism.
498 *Phil. Trans. R. Soc. London B* 365: 605-616.
- 499 25. Braendle, C., I. Friebe, M.C. Caillaud and D.L. Stern. 2005. Genetic variation for an
500 aphid wing polyphenism is genetically linked to a naturally occurring wing polymorphism.
501 *Proc R Soc B* 272.
- 502 26. Kim, H., S. Lee and Y. Jang. 2011. Macroevolutionary patterns in the Aphidini aphids
503 (Hemiptera: Aphididae): Diversification, host association, and biogeographic origins.
504 *PLoS ONE* 6(9): e24749.
- 505 27. Novakova, E., V. Hypsa, J. Klein, R.G. Footitt, C.D. von Dohlen and N.A. Moran. 2013.
506 Reconstructing the phylogeny of aphids (Hemiptera: Aphididae) using DNA of the
507 obligate symbiont *Buchnera aphidicola*. *Mol Phylogenet Evol* 68(1): 42-54.
- 508 28. Andolfatto, P., D. Davison, D. Erezyilmaz, T.T. Hu, J. Mast, T. Sunayama-Morita and D.L.
509 Stern. 2011. Multiplexed shotgun genotyping for rapid and efficient genetic mapping.
510 *Genome Res.* 21: 610-617.
- 511 29. Broman, K.W., H. Wu, S. Sen and G.A. Churchill. 2003. R/qtl: QTL mapping in
512 experimental crosses. *Bioinformatics* 19(7): 889-90.
- 513 30. Kofler, R., R.V. Pandey and C. Schlotterer. 2011. PoPoolation2: identifying
514 differentiation between populations using sequencing of pooled DNA samples (Pool-
515 Seq). *Bioinformatics* 27(24): 3435-6.
- 516 31. Gouin, A., F. Legeai, P. Nouhaud, A. Whibley, J.C. Simon and C. Lemaitre. 2015.
517 Whole-genome re-sequencing of non-model organisms: lessons from unmapped reads.
518 *Heredity* 114(5): 494-501.
- 519 32. Langmead, B., C. Trapnell, M. Pop and S.L. Salzberg. 2009. Ultrafast and memory-
520 efficient alignment of short DNA sequences to the human genome. *Genome Biology*
521 10(3).
- 522 33. Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-
523 analysis of large phylogenies. *Bioinformatics* 30(9): 1312-1313.
- 524 34. Bickel, R.D., J.P. Dunham and J.A. Brisson. 2013. Widespread selection across coding
525 and noncoding DNA in the pea aphid genome. *G3* 3: 993-1001.
- 526 35. Miura, T., C. Braendle, A. Shingleton, G. Sisk, S. Kambhampati and D.L. Stern. 2003. A
527 comparison of parthenogenetic and sexual embryogenesis of the pea aphid
528 *Acyrtosiphon pisum* (Hemiptera : Aphidoidea). *J. Exp. Zool. B. Mol. Dev. Evol.* 295B(1):
529 59-81.
- 530 36. Thorvaldsdottir, H., J.T. Robinson and J.P. Mesirov. 2013. Integrative Genomics Viewer
531 (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*
532 14(2): 178-92.
- 533 37. Ghanim, M. and K.P. White. 2006. Genotyping method to screen individual *Drosophila*
534 embryos prior to RNA extraction. *Biotechniques* 41: 414-418.
- 535

Figure 1

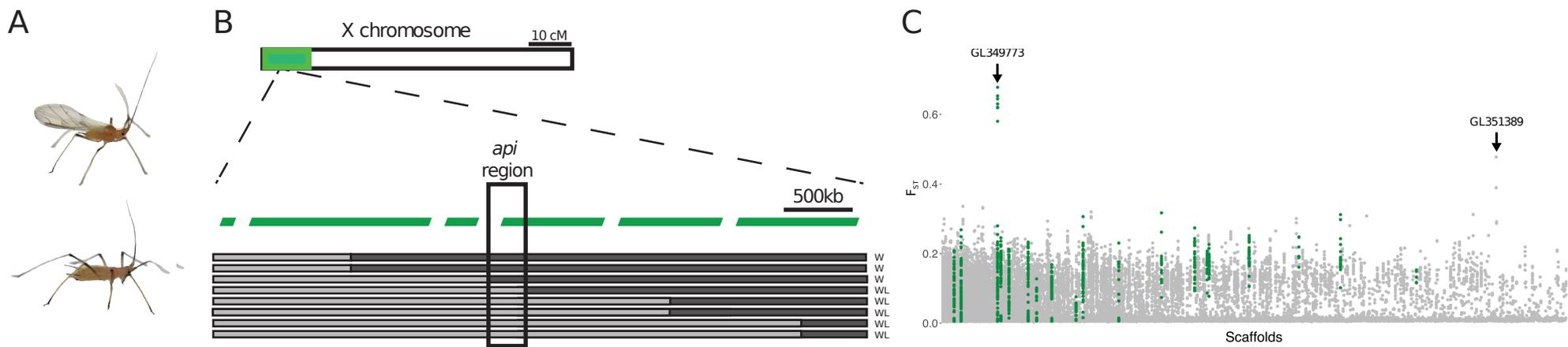


Figure 2

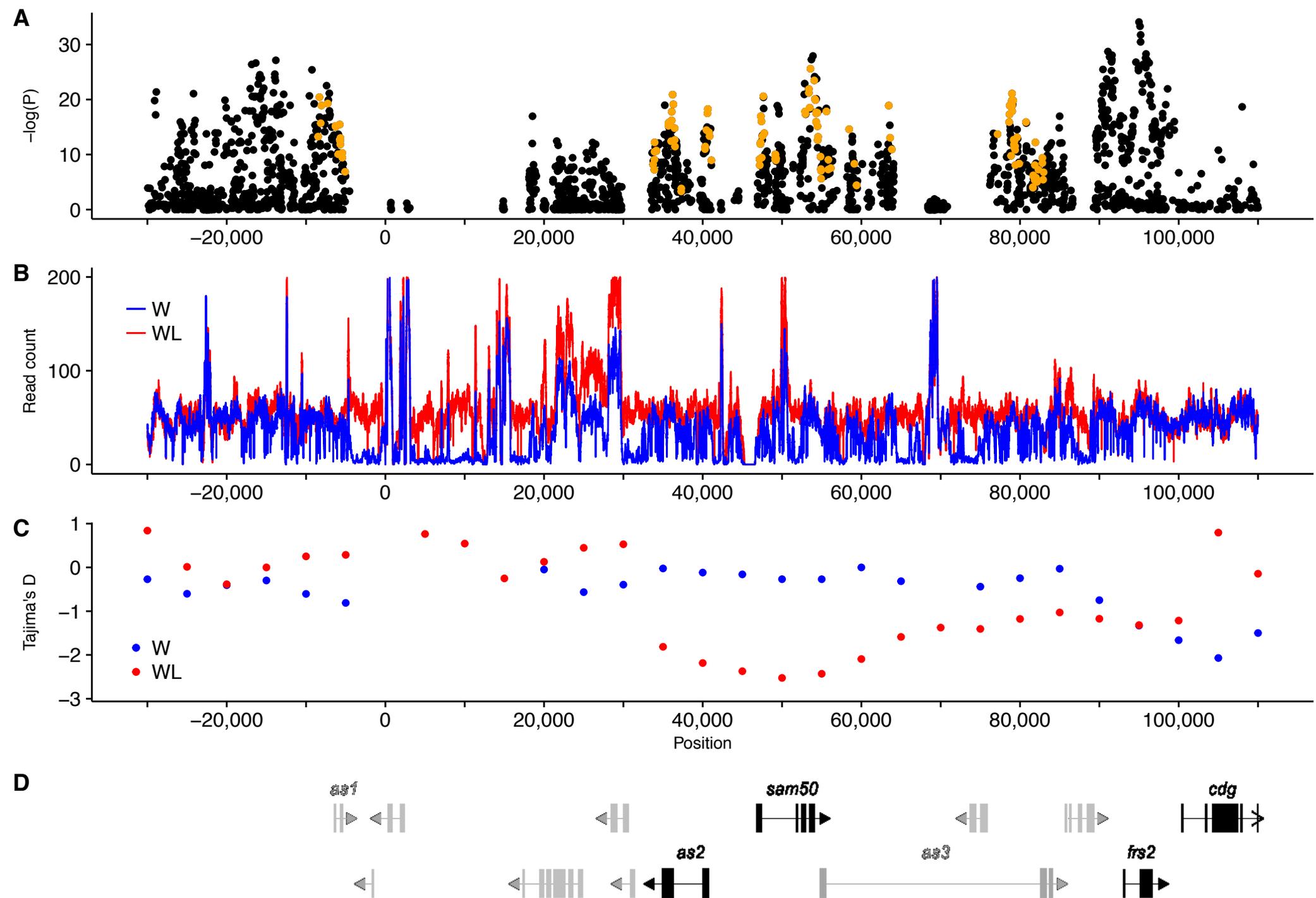


Figure 3

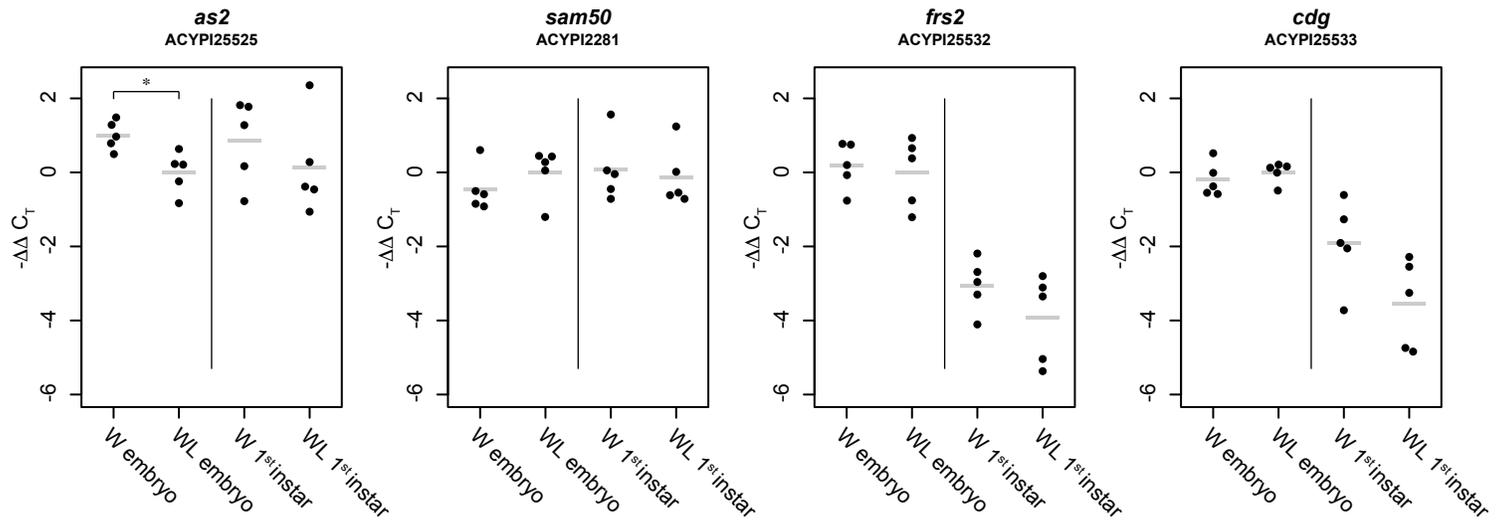
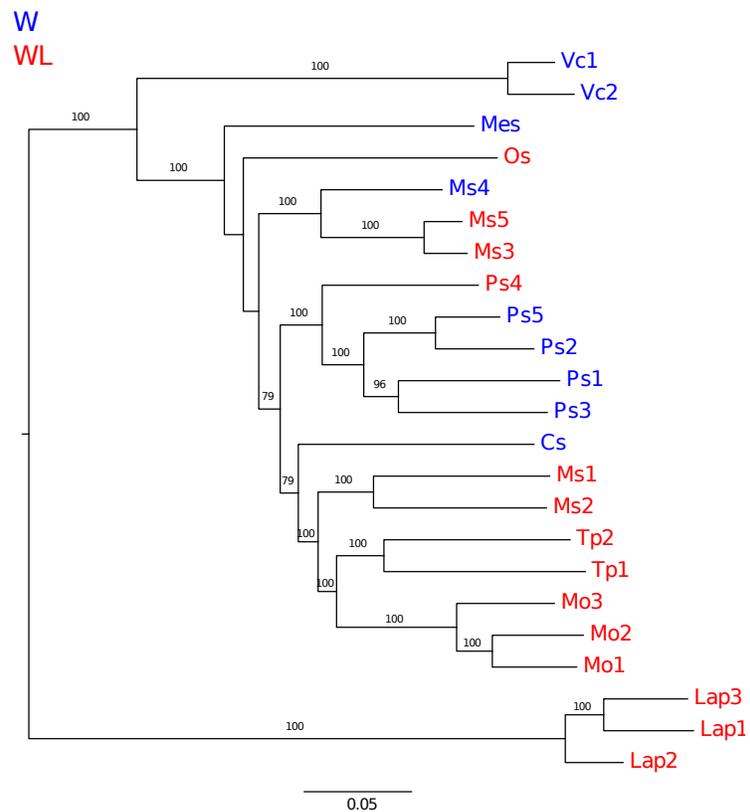


Figure 4

A. X chromosome



B. *api* scaffold GL349773 (50kb - 60kb)

