Coding sequence-independent homology search identifies highly divergent homopteran putative effector gene family

David L. Stern* & Clair Han

Affiliation:

Janelia Research Campus of the Howard Hughes Medical Institute, 19700 Helix Drive, Ashburn, VA 20147, USA

* Corresponding author

Abstract

Many genomes contain rapidly evolving and highly divergent genes whose homology to genes of known function often cannot be determined from sequence similarity alone. However, coding sequence-independent features of genes, such as intron-exon boundaries, often evolve more slowly than coding sequences and can provide complementary evidence for homology. We found that a linear logistic regression classifier using only structural features of rapidly evolving *bicycle* aphid effector genes identified many putative *bicycle* homologs in aphids, phylloxerids, and scale insects, whereas sequence similarity search methods yielded few homologs in most aphids and no homologs in phylloxerids and scale insects. Subsequent examination of sequence features and intron locations supported homology assignments. Differential expression studies of newly-identified *bicycle* homologs, together with prior proteomic studies, support the hypothesis that BICYCLE proteins act as plant effector proteins in many aphid species and perhaps also in phylloxerids and scale insects.

Introduction

One challenge that is faced by evolutionary and functional genetic studies is that many genes have been identified as “orphan” or lineage-specific genes because homologous sequences cannot be identified outside of a limited taxonomic range. Many lineage-specific genes may not be truly novel, however, since sequence divergence can cause homologs to become undetectable by sequence-search methods (Vakirlis et al., 2020; Weisman et al., 2020). Identifying such extremely divergent homologs remains a significant bioinformatic challenge and limits functional inferences derived from homology (Loewenstein et al., 2009).

Genes involved in host-parasite systems often evolve extremely divergent sequences as a result of genetic conflict (Eizaguirre et al., 2012; Paterson et al., 2010). Aphids and their host plants represent one such antagonistic pair. Aphids are small insects that feed by inserting their thin mouthparts (stylets) into the phloem vessels of plants to extract nutrients (Dixon, 1978). Like many herbivorous insects, aphids introduce effector molecules into plant tissues to manipulate the physiology and development of plants to the insects’ advantage (Elzinga and Jander, 2013; Hogenhout and Bos, 2011). For example, aphids introduce calcium binding proteins that prevent the plant’s ability to block phloem cell transport in response to phloem vessel damage (Will et al., 2007). It is believed that aphids introduce a wide range of effector proteins and that these molecules contribute to the debilitating effects of multiple aphid species on plants, which imposes significant financial damage on most major agricultural crops (Schaible et al., 2008).

In an extreme form of aphid manipulation of plant physiology and development, some aphid species induce novel plant organs, called galls, which provide the aphids with a ready food source and with protection from the elements and from natural enemies (Giron et al., 2016; Mani, 1964; Shorthouse et al., 2005). Galls can be significant nutrient sinks (Burstein et al., 1994), demonstrating that galling aphids induce both local and long-range changes to plant physiology.

We reported recently that the genome of one species of galling aphid encodes 476 related genes encoding diverse, novel, presumptive effector proteins and many of these proteins contain two cysteine-tyrosine-cysteine motifs (CYC). These genes were therefore named *bicycle* (bi-CYC-like) genes (Korgaonkar et al., 2021). A genome-wide association study revealed that variation in one *bicycle* gene, called *determinant of gall color* (*dgc*), was strongly associated with a red versus green gall color polymorphism and this genetic polymorphism was associated with strong downregulation of *dgc* in aphids and strong upregulation of a small number of genes involved in anthocyanin production in plant galls. In addition, the majority of *bicycle* genes are strongly upregulated in the salivary glands specifically of the aphid generation that induces galls (the fundatrix). Together with genetic evidence that *dgc* regulates the gall phenotype, the enrichment of *bicycle* genes expressed in fundatrix salivary glands suggests that many *bicycle* genes contribute to gall development or physiology.

The primary amino-acid sequences of BICYCLE proteins have evolved rapidly, apparently in response to extremely strong positive selection (Korgaonkar et al., 2021). In preliminary studies, we attempted to identify *bicycle* homologs in other insect species using sequence similarity algorithms such as *BLAST* (Altschul et al., 1990) and

hmmer (Eddy, 2011; Johnson et al., 2010), but identified few putative homologs. It was therefore unclear if *bicycle* genes are a recently evolved family of genes or an ancient family that has evolved highly divergent coding sequences.

Extremely divergent homologs have been identified previously using features of gene structures that evolve more slowly than the coding sequence, such as exon sizes and intron positions (Bazan, 1991; Betts, 2001; Brown et al., 1995), and these characteristics of genes have also proven to be valuable for phylogeny reconstruction (Rokas et al., 1999; Rokas and Holland, 2000; Telford and Copley, 2011; Venkatesh et al., 1999). We noted previously that *bicycle* genes have unusual gene structures containing many micro-exons (Korgaonkar et al., 2021) and we considered the possibility that *bicycle* gene structure may be more conserved than *bicycle* gene sequences. To explore the evolutionary history of *bicycle* genes, we therefore sought a method that would allow identification of highly divergent *bicycle* gene homologs that did not rely on sequence similarity. To accomplish this, we built a logistic regression classifier based only on structural features of *bicycle* genes. We found that this classifier was very accurate, despite the fact that it does not include any sequence information. This classifier identified many highly-divergent *bicycle* homologs in *H. cornu* that we had overlooked in our initial study, including proteins that do not include the previously canonical CYC motif. In addition, the classifier identified many highly divergent candidate *bicycle* homologs in all aphids, phylloxerans, and scale insects we studied. We did not detect any putative homologs in three progressively basal outgroups. Multiple sequence alignment of these putative homologs revealed that many contain N-terminal signal sequences and CYC motifs, consistent with their

assignment as *bicycle* homologs. In addition, gene-structure aware sequence alignment revealed multiple apparently shared intron boundaries between putative homologs that share little obvious sequence similarity. In addition, putative *bicycle* homologs are highly enriched in the salivary gland mRNA of two gall forming and one non-gall forming aphid. All of this evidence supports the hypothesis that *bicycle* gene homologs encode effector proteins in gall-forming and non-gall-forming aphids and also in phylloxerans and scale insects.

Results

A classifier based on only gene structural features can identify* bicycle *genes with high probability

Bicycle genes in *H. cornu* were identified, originally, as a subset of previously unannotated genes that were enriched in the salivary gland mRNA of the gall-inducing foundress (Korgaonkar et al., 2021). *Bicycle* genes were found to be the single largest group of such genes that shared sequence similarity. These 476 genes are extremely divergent from one another at the amino-acid sequence level and appear to have evolved rapidly due to positive natural selection. We noticed, however, that non-sequence specific features of *bicycle* genes appeared to be relatively well conserved. For example, *bicycle* genes contain a large number of unusually small internal exons (Korgaonkar et al., 2021), making them clear outliers to most genes in the *H. cornu* genome. We also observed that almost all internal exons of *bicycle* genes start with

121 codon position 2, that is they exhibit exon phase 2 (Figure 1 Supplement 1A), which is
122 an extremely different distribution of exon phases compared with other genes in the
123 genome (Figure 1 Supplement 1B). Genes containing many exons of similar size all
124 with the same phase are rare in most genomes (Ruvinsky and Watson, 2007).

125 We hypothesized that some aspects of *bicycle* gene structure may evolve
126 more slowly than the primary protein sequences and thus provide an evolutionary
127 signal to detect distantly related *bicycle* homologs (Aguiar et al., 2015). For example,
128 the pattern of exon phases can allow identification of highly-divergent homologous
129 genes (Roy and Gilbert, 2005; Ruvinsky and Watson, 2007). Development of a full
130 model to characterize gene structure evolution is beyond the scope of this paper, but
131 we explored whether a linear logistic regression classifier could identify *bicycle*
132 homologs based on the following structural features of each gene: length of gene in
133 genome, from start to stop codons; first and last exon length; internal exon mean
134 length; length of the sum of codons; and number of internal exons that start in phase
135 0, 1, or 2. Since almost all internal *bicycle* exons are of phase 2, there is no additional
136 information from the order of exon phases in *bicycle* genes, and we therefore simply
137 counted the number of exons of each phase. A randomly selected 70% of *bicycle* and
138 annotated genes were used for training, and 30% used for validation. (Non-annotated
139 genes were excluded from model training and validation, because we hypothesized
140 that this set may include additional *bicycle* homologs.) Characterized *bicycle* genes
141 have an average of approximately 17 and a minimum of 5 exons (Figure 1
142 Supplement 1C, D), which is strongly different from the genome-wide background,
143 and Bicycle protein lengths are well explained by total exon number (c.f. Figure 1

Supplement 1E and F). We excluded all genes with zero or one internal exons to accurately summarize internal exon mean lengths and number of internal exons that start in each of the three phases. Thus, if any *bicycle* genes have evolved to have three or fewer exons, then they could not be detected with this classifier.

The classifier exhibited extremely good performance with all data partitions (Figure 1 Supplement 1G) and, therefore, a full model was constructed using all *bicycle* and annotated genes. The classifier categorized the vast majority of genes as either non-*bicycle* or *bicycle* genes with high probability (Figure 1 Supplement 1H). To decide on a probability cutoff to identify putative *bicycle* homologs, we considered the trade-off between precision (the proportion of true positives among both true and false positives) and recall (the proportion of true positives among true positives and false negatives). We chose a model probability cutoff that equalized precision and recall (P/R = 1), which yielded precision and recall values of 0.99 (Figure 1A,B). To determine whether all gene features were required for accurate classification, we removed one factor at a time, retrained the model, and calculated precision and recall values using the same threshold as the full model. We found that no single factor was critical for model performance (Figure 1C), although removal of the number of mode 2 exons resulted in the strongest decrement in model performance, resulting in precision and recall values of 0.98 and 0.96, respectively. Surprisingly, given the fact that *bicycle* genes contain many microexons, the removal of average internal exon size had almost no effect on model performance (Figure 1C), suggesting that internal exon size is not a uniquely defining feature of *bicycle* genes.

When we applied this classifier to all genes in the *H. cornu* genome, we identified five false negative and 181 “false positive” genes. Since we chose a probability cutoff that should report an approximately equal number of false negatives and false positives, we hypothesized that the vast majority of “false positives” are newly discovered *bicycle* homologs. We clustered these genes together with the original *bicycle* genes based on sequence similarity (Figure 2A), and found that 528 represented a single group of related genes (Cluster 2, Figure 2A) that all share sequence similarity with, and included, the originally defined 476 *bicycle* genes (Figure 2C).

We also identified a second cluster of 129 genes that did not share strong sequence similarity with the originally defined Bicycle proteins and encode proteins with incomplete CYC motifs (Cluster 1, Figure 2A). Approximately 20 of these genes were strongly enriched in the salivary glands of gall-inducing foundresses, but many of these Cluster 1 genes were strongly enriched in salivary glands of other life stages (Figure 2D-F). In addition, approximately 20 *bicycle* genes containing the canonical CYC motif (Cluster 2) were weakly expressed in fundatrix salivary gland, but strongly expressed at other life stages (Figure 2E-G).

Like the original 476 *bicycle* genes, the new putative *bicycle* homologs likely experienced increased levels of positive selection. We found that the new putative *bicycle* homologs displayed a similar pattern to the original *bicycle* genes of elevated ratios of non-synonymous (d_N) to synonymous (d_S) substitutions between *H. cornu* and its closely related sister species *H. hamamelidis* compared to the genomic background (Figure 2 Supplement 1A-C). In addition, we found that signals of selective sweeps,

189 based on *H. cornu* population site frequencies, are enriched near both the original
190 *bicycle* genes and the newly discovered putative *bicycle* homologs (Figure 2
191 Supplement 1D).

192 Since some of these new putative *bicycle* homologs shared little apparent
193 sequence similarity with the originally defined *bicycle* genes, we sought additional
194 evidence for homology. Since these genes were identified by our gene-structure based
195 classifier, we hypothesized that other details of gene structure, such as intron positions
196 in the sequence, may provide additional evidence of shared ancestry. We therefore
197 performed gene-structure-aware multiple sequence alignment of the original *bicycle*
198 genes and the new putative *bicycle* genes identified by the classifier (Gotoh, 2021).

199 Initial attempts to align all 653 candidate proteins were uninformative and we
200 hypothesized that this was because these proteins display a large range of lengths and
201 exon numbers (Figure 3A, Figure 1 Supplement 1C). We therefore divided Bicycle
202 proteins into four groups based on protein length (blue dotted lines in Figure 3A). Logo
203 plots of these four groups revealed a striking pattern, proteins contain one, two, three,
204 or four CYC motifs (Figure 3B-E), and we therefore call these *unicycle, bicycle, tricycle,*
205 and *tetracycle* genes, respectively. This evolutionary comparison suggests that the
206 CYC motif is a functional unit, which can be multimerized within individual proteins.

207 We then divided all 653 potential *H. cornu bicycle* homologs into four length
208 categories and performed gene-structure aware protein sequence alignments to
209 search for shared intron position between the originally defined and newly discovered
210 *bicycle* homologs. We identified many evolutionarily shared intron positions between

211 the two classes of putative homologs (Figure 3 Supplement 1A-D), providing evidence,
212 independent of our classifier, that all of these genes shared a common ancestor.
213 These results provide evidence for several conclusions. First, measures of gene
214 structure alone can reliably identify divergent *bicycle* genes. Second, some *bicycle*
215 homologs have evolved highly divergent sequences, raising the possibility that highly
216 divergent homologs may be present in other species but are undetectable with
217 sequence-similarity search algorithms. Third, some *bicycle* homologs are highly
218 expressed in salivary glands of non-gall forming life stages, suggesting that some of
219 these proteins may contribute to the effector-protein repertoire outside of the context
220 of gall development. Given these observations, we next tested whether the classifier
221 could identify putative *bicycle* homologs in a divergent galling aphid species.
222

223 ***Most candidate gall effector genes in* Tetraneura ulmi *are highly divergent* bicycle**
224 ***genes***

225

226 There are two major clades of gall forming aphids, the Hormaphididae, to which
227 *H. cornu* belongs, and the Pemphigidae. These two aphid families are thought to have
228 shared a common ancestor that induced galls. Therefore, to determine whether *bicycle*
229 genes were present in the common ancestor of gall forming aphids, we assembled and
230 annotated the genome of *Tetraneura ulmi* (Figure 4 Supplement 1A), a gall forming
231 aphid belonging to the Pemphigidae (Álvarez et al., 2013). We annotated the *T. ulmi*
232 genome using mRNA sequencing reads generated from salivary glands and carcasses
233 of the fundatrix (G1) and G2 generations (Figure 4 Supplement 1B).

We applied the *H. cornu bicycle* gene classifier to predicted genes of the *T. ulmi* genome and identified 279 candidate *bicycle* homologs. In contrast, sequence-based search using BLAST and *hmmer* identified 1 and 53 candidate homologs at E-value < 0.01, respectively. We clustered the 279 *T. ulmi* genes discovered by the gene-structure based classifier based on their predicted amino acid sequences (Figure 4A) and identified three clusters that included proteins with clear evidence for CYC motifs (Figure 4B-D). One cluster of 94 genes does not exhibit clear CYC motifs (Figure 4E). To test whether these genes likely shared a common ancestor, we performed gene-structure aware alignments of proteins with and without CYC motifs and identified many shared introns between putative homologs with and without CYC motifs (Figure 4 Supplement 1C). Thus, the *T. ulmi* genome contains apparent *bicycle* genes that have become essentially unrecognizable as *H. cornu bicycle* homologs at the sequence level.

To explore whether the *T. ulmi bicycle* genes, and especially the divergent putative *bicycle* homologs, might act as effector proteins, we dissected salivary glands from fundatrices and compared mRNA expression levels in salivary glands versus carcasses. Almost all of the *bicycle* homologs detected by the classifier are strongly over-expressed in the fundatrix salivary glands (Figure 4F), suggesting that they may contribute to the effector protein repertoire in *T. ulmi*.

The discovery of *bicycle* genes in *T. ulmi* using the gene-structure based classifier provides evidence that the classifier can identify putative *bicycle* homologs even when the genes cannot be identified using sequence-based homology methods and that *bicycle* genes were likely present in the common ancestor of gall-forming

aphids. We therefore next tested whether the classifier can detect putative *bicycle* homologs in an aphid species that does not induce galls.

The pea aphid* (Acyrthosiphon pisum) *genome includes many* bicycle *homologs that are candidate effector proteins

The gene-structure based classifier detected 159 putative *bicycle* homologs in the genome of *Acyrthosiphon pisum*, a species that does not induce galls (Li et al., 2019; Richards et al., 2010). In contrast, sequence-based search using BLAST and *hmmer* identified 1 and 5 candidate hits at E-value < 0.01, respectively. Multiple sequence alignment of the 159 *A. pisum* putative *bicycle* homologs revealed that most of these genes encode proteins with N-terminal secretion signal sequences and a single CYC motif (Figure 5A). Gene structure aware alignment identified multiple apparently shared introns between *H. cornu bicycle* genes and *A. pisum* putative *bicycle* homologs (Figure 5 Supplement 1), providing independent confirmation that these genes shared a common ancestor.

To explore whether these *A. pisum bicycle* homologs may contribute to the effector gene repertoire of *A. pisum*, we performed RNA-seq of salivary glands and carcasses. We found that 55% of the putative *bicycle* genes are strongly enriched in salivary glands (Figure 5B). In addition, 19 proteins encoded by genes with multiple microexons were identified by a previously-published proteomic study of proteins secreted by *A. pisum* into an artificial food medium (Dommel et al., 2020), and we found that at least 13 of these proteins are *bicycle* homologs (Table S1).

In summary, the *bicycle* classifier, the presence of CYC motifs and of introns apparently shared with *H. cornu bicycle* genes, RNA expression, and proteomics all support the inference that the *A. pisum* genome encodes approximately 159 *bicycle* homologs that may contribute to the *A. pisum* effector protein repertoire.

Bicycle genes are present in genomes of aphids, phylloxerids, and coccids

The presence of *bicycle* homologs in three divergent aphid species, *H. cornu*, *T. ulmi*, and *A. pisum* implies that *bicycle* genes were present in the common ancestor of aphids, which lived approximately 280 MYA. To further explore the origins and evolution of *bicycle* genes, we downloaded genomes and RNA-seq data for nine additional aphid species and ten outgroup species from NCBI. We annotated predicted genes in all genomes using the same bioinformatic pipeline that we have used previously to discover *bicycle* genes in other species, applied the *bicycle* gene classifier to all predicted genes, and manually annotated all putative *bicycle* homologs guided by RNA-seq data. One caveat of this analysis is that many of these genomes are fragmented into many contigs and genes bridging contigs are often misannotated or not annotated, resulting in possible under-counting of *bicycle* homologs. In addition, we detected multiple putative *bicycle* homologs near the ends of contigs. It is therefore possible that fragments of single *bicycle* homologs are present on multiple contigs, which would lead to over-counting *bicycle* gene homologs. While these problems with current annotations are expected to increase variance in estimates of *bicycle* homologs, we did not detect a dependence of number of *bicycle* homologs detected

on genome assembly quality (Figure 6 Supplement 1A). In addition, we employed chromosome-level genome assemblies for two outgroup species where no predicted *bicycle* genes were found, *P. venusta* (Li et al., 2020) and *T. vaporariorum* (Xie et al., 2020), suggesting that the failure to identify *bicycle* genes in these species did not result from poor genome assemblies.

We estimated the phylogeny for these 22 species using whole-genome proteomic predictions (Emms, 2019), and this phylogeny is in general agreement with earlier studies based on a small number of genes (Johnson et al., 2018; Nováková et al., 2013; von Dohlen and N. A. Moran, 1995) (Figure 6). We detected putative *bicycle* homologs in all aphid species studied here, supporting the inference that *bicycle* genes were present in the common ancestor of aphids (Figure 6). We found extensive variation in the number of *bicycle* genes between species. In all aphid species with a sufficient number of *bicycle* homologs, multiple sequence alignment revealed relatively conserved CYC domains (Figure 6 Supplement 1B-E). In most Aphidini species (e.g. *M. persicae*, *A. pisum*, *P. nigronervosa*), most homologs appear to be *unicycle* genes (Figure 6 Supplement 1B-C). In *Cinara cedri*, many homologs appear to be *tetracycles* (Figure 6 Supplement 1E).

We found 69 putative *bicycle* homologs in *Daktulosphaira vitifoliae* indicating that *bicycle* genes were present in the common ancestor of the Aphidomorpha (Aphidoidea + Phylloxeridea). We also found 74, 26 and 2 putative *bicycle* homologs in the coccid species *Maconellicoccus hirsutus*, *Phenacoccus solenopsis*, and *Ericerus pela*, respectively. The putative *bicycle* homologs from these non-aphid species do not contain obvious CYC motifs (Figure 6 Supplement 2A-C). However, gene structure

326 aware alignment with the *H. cornu bicycle* genes revealed that the putative homologs
327 from *D. vitifoliae*, *M. hirsutus*, and *P. solenopis* share multiple introns in the same
328 locations with *H. cornu bicycle* genes (Figure 6 Supplement 2D-F), supporting the
329 inference that these are *bicycle* homologs.

330 We did not detect *bicycle* homologs in any of the species sampled from the
331 Psylloidea, Aleyrodoidea, or Fulgoroidea. These results indicate that *bicycle* genes
332 were present in the common ancestor of coccids and aphids and may have evolved in
333 this lineage. However, for at least two reasons we cannot rule out the possibility that
334 *bicycle* genes originated earlier. First, we observed extensive variation in the number of
335 *bicycle* homologs across lineages, so *bicycle* genes may have been lost in the specific
336 outgroup species studied here. Second, ancestral *bicycle* genes may not display the
337 specific gene structure detected by our classifier. Deeper taxonomic sampling and
338 other homology search approaches may reveal an older origin of *bicycle* genes.

339

340 **Aphid genomes contain a conserved *megacycle* gene**

341

342 In many of the aphid genomes we studied, our classifier identified a single
343 extremely large gene containing more than 100 microexons and encoding a protein of
344 approximately 2000 amino acids as a candidate *bicycle* homolog (Figure 7A). Since this
345 gene was so large and most existing aphid genomes are incompletely assembled, we
346 observed that many of these gene models were incomplete. We therefore performed
347 *de novo* transcript assembly for all species studied using TRINITY (Grabherr et al.,

348 2011), and assembled consensus transcripts of these long candidate *bicycle* genes
349 manually.

350 Since we had previously found that *bicycle* gene length is correlated with the
351 number of encoded CYC motifs (Figure 3), we explored whether these genes encoded
352 proteins with multiple repeating units. Using Rapid Automatic Detection and Alignment
353 of Repeats (Heger and Holm, 2000) via EMBL-EBI (Madeira et al., 2019), we found that
354 these large proteins contain many repeating motifs and each motif is approximately the
355 size of two CYC motifs. This structure was most obvious for the *Melanaphis sacchari*
356 gene. Sequence-based alignment of 15 predicted repeats of the *M. sacchari* protein,
357 which did not use information of intron positions, revealed multiple shared intron
358 locations across repeats (Figure 7B). Thus, this protein appears to have evolved
359 through multiple internal tandem duplications. This large protein exhibited little obvious
360 sequence similarity to *bicycle* genes, but gene-structure aware alignment revealed
361 multiple apparently conserved intron positions shared between the internal
362 duplications of the large gene from *M. sacchari* and *H. cornu bicycle* genes (Figure 7
363 Supplement 1A), suggesting that they share a common ancestor.

364 These large genes appear to be divergent *bicycle* proteins that have duplicated
365 the core motif many times. We therefore call these *megacycle* genes. Compared to
366 most *bicycle* genes, *megacycle* orthologs are relatively well conserved across aphids
367 and a phylogenetic tree reconstructed from the inferred Megacycle proteins is similar
368 to the aphid phylogeny estimated from the whole proteome (Figure 7 Supplement 1B).
369 Most aphid genomes we examined contain a single copy of this gene; we did not
370 detect this gene in *Diuraphis noxia*, and we found three paralogs in both *Cinara cedri*

371 and *Tetraneura ulmi*. No *megacycle* orthologs were found in any of the non-aphid
372 species.
373 For the three species for which we have good salivary gland transcriptomic
374 data, *H. cornu*, *T. ulmi,* and *A. pisum* the *megacycle* genes are strongly expressed in
375 salivary glands (Figure 7C-E). Thus, *megacycle* genes may contribute to the effector
376 protein repertoire in aphids. The relatively conservated lengths and sequences of
377 *megacycle* homologs suggests that they may share a relatively conserved role across
378 aphids.
379
380 **Discussion**
381
382 Some gene features, such as the structure of the encoded protein or the intron-
383 exon structure of the gene, can provide evidence for homology that is independent of
384 sequence. Recent improvements in machine learning predictions of protein structure
385 (Baek et al., 2021; Jumper et al., 2021) may, in principle, facilitate protein structure
386 based homology assignment. In addition, gene structural features provide homology
387 information independent of both sequence and protein structure (Bazan, 1991; Betts,
388 2001; Brown et al., 1995). This source of information is becoming increasingly available
389 with the recent availability of many well-assembled, annotated genomes. It is possible
390 to conceptualize a model that jointly utilizes both sequence and exon-intron structure
391 in homology search, but we realized that *bicycle* genes are encoded with extremely
392 unusual intron-exon structures and that a model incorporating only general features of
393 exons may allow detection of distant homologs.

We found that a linear classifier using information on gene length, exon sizes, exon numbers, and exon phases provided a highly accurate predictor of *bicycle* homologs that could not be identified using sequence similarity searches. This predictor allowed identification of *bicycle* homologs across aphids and in several outgroup taxa (Figure 6). Homology was supported by *post hoc* observation of CYC motifs and shared introns across many homologs. Thus, sequence independent features provide substantial evidence of homology and integration of even a few sequence independent features into existing sequence-based search methods may substantially increase power to detect ancient homology relationships between genes. All gene structure features that we used contributed to classifier performance, but no single feature was critically important for classifier performance. Therefore, in future work it would be ideal to incorporate multiple features of gene structure into predictive models.

For several reasons, our classifier may have underestimated the number of *bicycle* homologs in each species. First, this approach is sensitive to the quality of genome annotation, which is itself dependent on genome assembly quality. We found that most automatically annotated gene models of *bicycle* genes required manual re-annotation. Thus, we may have overlooked additional *bicycle* genes in these genomes because of genome fragmentation and inaccurate automated gene annotation. Second, *bicycle* gene annotation in most species is likely hampered by the lack of salivary gland RNA-seq samples. Finally, it is possible that some *bicycle* homologs have evolved both divergent sequences and novel gene structures, preventing identification with any existing method.

417

418 **bicycle *genes are not strictly associated with gall formation***

419

420 *Bicycle* genes were detected originally as putatively secreted proteins strongly
421 expressed in the salivary glands of the gall-forming aphid generation of *H. cornu* and
422 variation in one *bicycle* gene is genetically associated with gall physiology (Korgaonkar
423 et al., 2021). These observations led to the hypothesis that many *bicycle* genes
424 participate in gall development. Our observations provide a revised assessment of
425 potential Bicycle protein functions. Two of the aphids studied here, *H. cornu* and *T.*
426 *ulmi*, induce complex galls and their genomes include more *bicycle* genes than other
427 species we examined. However, another gall-forming aphid, *E. lanigerum* has only 27
428 *bicycle* genes. The phylloxerid studied here, *D. vitifoliae*, also induces complex galls,
429 and has only 70 *bicycle* genes. In contrast, some non-galling aphids have more *bicycle*
430 genes, such as *Myzus persicae* and *A. pisum*, with 129 and 158, respectively. Thus,
431 there does not appear to be a strong correlation between the number of *bicycle* genes
432 and the gall-forming habit. In addition, the fact that all aphids we studied, which
433 includes nine non-galling aphids, have *bicycle* genes suggests that *bicycle* genes may
434 perform functions in most aphid species that are distinct from gall induction. In
435 principle, *bicycle* genes may have similar molecular functions across species, but
436 different targets in different plant species. In some cases, the targets may induce
437 patterned cell proliferation, resulting in galls, and in others the targets may alter plant
438 physiology in more cryptic ways to confer benefits on aphids.

None of the three coccid species we studied here induce galls. However, many coccids induce depressions, swellings, or other changes in plant organs (Beardsley and Gonzalez, 1975), including the induction of elaborate galls (Cook and Gullan, 2004). Parr (Parr, 1940) reported a remarkable series of experiments that appear to demonstrate that extracts of salivary glands of a coccid are sufficient to induce pit galls on oak trees. It may be worth exploring the hypothesis that *bicycle* genes contribute to the effector gene repertoire of scale insects and especially of the gall-forming species.

An evolutionary hypothesis for the unusual structure of bicycle genes

Each *bicycle* gene is encoded by multiple small exons spread across a large genomic region, even though most *bicycle* transcripts encode relatively short proteins. While each *bicycle* gene includes relatively long 5' and 3' exons that vary from about 50 to 300 bp (Figure 1 Supplement 1F), *bicycle* protein lengths are best predicted by the number of internal microexons (Figure 1 Supplement 1E), which vary in length from about 9 to 90 bp. Internal microexons appear to be gained and lost frequently, resulting in the extensive diversity of *bicycle* protein lengths. Most of these internal exons are of a length that is a multiple of 3bp and of phase 2 (Figure 1 Supplement 1A), which facilitates their gain and loss during evolution without disrupting the reading frame. We did not find strong evidence for extensive alternative splicing of *bicycle* transcripts. Current evidence suggests that the microexons of *bicycle* genes do not exist to increase transcript (and protein) diversity, but they may have been maintained for other reasons.

In principle, *bicycle* genes could maintain and evolve microexons in smaller genomic regions than they actually occupy. The real evolutionary puzzle then is why these genes have

large introns. We suggest three possible hypotheses to explain this pattern. First, if new microexons evolve often through *de novo* origins, then large introns may be maintained because longer introns have a higher probability than shorter introns of evolving new exons, which must have both the correct phase and be flanked by splice acceptor and donor sites.

Second, if microexons are gained and lost through recombination processes, large introns may facilitate illegitimate recombination amongst paralogs. The gene-structure aware alignments we performed suggest that some introns are strongly conserved, which suggests that illegitimate recombination is not the major source of new exons in most *bicycle* genes. In addition, while the *megacycle* genes most likely evolved initially through repeated internal duplications, they do not appear to have experienced rampant internal deletions and duplications subsequently.

A third possibility, which is agnostic to the mechanism of microexon gain and loss, is that large introns may provide a population genetic advantage that facilitates rapid adaptive protein evolution. Adaptive evolution is sometimes limited by linkage between two adaptive substitutions (Betancourt and Presgraves, 2002; Birky and Walsh, 1988), where selection for one allele prevents fixation of a second linked allele. This interference is amplified when adaptive substitutions occur within stretches of DNA that tend to be co-inherited because recombination is insufficiently frequent to randomize the inheritance of alternative alleles at each site, resulting in linkage disequilibrium. If each *bicycle* gene were encoded in a short stretch of genomic DNA (on average, each gene requires only about 1kbp of DNA), then multiple adaptive substitutions in different parts of the gene may interfere with each other and reduce the potential rate of Bicycle protein evolution. In contrast, multiple adaptive polymorphisms in widely spaced microexons likely experience relatively independent evolutionary fates and could be driven to fixation by positive selection efficiently in populations. We observed, in this paper and previously (Korgaonkar et al., 2021), pervasive, repeated, and

apparently strong positive selection on *bicycle* genes and here we report that some *bicycle* genes have diverged so dramatically that they cannot be recognized as *bicycle* genes from sequence comparisons. The unusual gene structure of *bicycle* genes may have facilitated their rapid evolutionary dynamics and their extreme diversification.

Methods

Gene structure based bicycle *gene classifier*

The *bicycle* gene classifier was built as a generalized linear model (GLM) with the following predictor variables: total gene length, total length of coding exons, first coding exon length, last coding exon length, mean internal exon length, and number of internal exons in phase 0, 1, and 2. All length measurements are in base pairs. All genes with zero or one internal exon were removed from the training and test datasets because mean internal exon lengths and number of internal exons starting in each of the three phases cannot be accurately estimated, hence our classifier was unable to detect any potential *bicycle* genes with only one to three exons. The dependent variable from the GLM is the predictor of *bicycle* gene classification. The GLM was implemented with the *glm* function as *binomial* with *logit* link, and response variables for the whole-genome gene set were predicted using *predict*, both in the *R* base package *stats* v.3.6.1.

In the training set, the dependent variable of the GLM is whether the transcript was previously classified as a *bicycle* gene (Korgaonkar et al., 2021), where *bicycle* genes were coded as 1 and non-bicycle genes were coded as 0. We used only genes classified as either *bicycle* genes or previously annotated genes in the training set due to the possibility that unannotated genes include additional previously undetected *bicycle* genes. Previous annotation was defined as having at least one significant match at $E < 0.01$ after Bonferroni correction for multiple searches with *blastp* (*blast* v.2.9.0), *blastx* (*blast* v.2.9.0), or *hmmscan* (*HMMER* v3.2.1) against the Pfam database.

We evaluated the performance of the classifier using a precision-recall curve where $precision = \frac{TP}{TP+FP}$ and $recall = \frac{TP}{TP+FN}$, where TP, FP, and FN are the number of true positives, false positives, and false negatives, respectively. To determine a cutoff value on the dependent variable of the GLM for *bicycle* gene classification, we tested all possible cutoffs from 0 to 0.98 in increments of 0.02 to find the cutoff resulting in a precision to recall ratio closest to 1. We first validated our model by repeatedly sampling and training on 70% of the data and testing on the remaining 30% of the data for 100 replicates and found no substantial change to the trained classifier or the resulting cutoff value (mean 0.7338, s.d. 0.1578) amongst replicates. We therefore trained the final classifier on all *H. cornu bicycle* genes and annotated non-*bicycle* genes without subsampling and determined a cutoff value of 0.72 (precision = 0.9925, recall = 0.9925). The coefficients to the independent variables in the GLM are as follows: intercept = 6.26, total exon length = -0.0206, total gene length = -0.0000642, first exon length = -0.00632, last exon length = 0.0103, mean exon length = -0.0655, number of exons in phase 0 = -3.32, phase 1 = -1.12, and phase 2 = 1.21. We evaluated the contribution of each independent variable by training the *H. cornu* dataset using GLMs with each term removed one at a time. We applied the GLM trained by the *H. cornu* dataset to twenty-one other species belonging to the families of Aphidoidae, Phylloxeroidae, Coccoidea, Psylloidea, Aleyrodoidea, and Fulgoroidea.

Aphid collections and dissections

Aphids of *Tetraneura ulmi* were collected from galls collected on *Ulmus americana* on the grounds of Janelia Research Campus (Table S2-3). Aphids of *A. pisum* LSR1 were kindly provided by Greg Davis and grown on broad bean plants in the laboratory. Salivary glands

were dissected out of multiple instars of fundatrices and her offspring for *T. ulmi* and virginoparae for *A. pisum* and stored in 3 uL of Smart-seq2 lysis buffer (0.2% Triton X-100, 0.1 U/mL RNasin Ribonuclease Inhibitor) at -80°C for later mRNA isolation. Carcasses of both species were ground in PicoPure extraction buffer and mRNA was prepared using the PicoPure mRNA extraction kit. RNAseq libraries were prepared as described previously (Korgaonkar et al., 2021).

Genome sequencing, assembly, and annotation of Tetraneura ulmi

HMW gDNA was prepared from aphids isolated from a single gall following the Salting Out Method provided by 10X Genomics (https://support.10xgenomics.com/genome-exome/sample-prep/doc/demonstrated-protocol-salting-out-method-for-dna-extraction-from-cells). 10X linked-read libraries and Chromium sequencing was performed by HudsonAlpha Institute for Biotechnology. Genomes were assembled with *Supernova* v.2.1.0 commands *run* and *mkoutput* (Weisenfeld et al., 2017).

We used 133,280,000 reads for an estimated raw genome coverage of 56.18X. The genome size of the assembled scaffolds was 344.312 Mb and the scaffold N50 was 10 Mb. Using BUSCO completeness analysis with *gVolante* version 1.2.0 and BUSCO_v2/v3 (Nishimura et al., 2017), we found that the genome contains 1035 of 1066 (97.1%) completely detected core arthropod genes and 1043 of 1066 (97.8%) partially detected core genes. The genome was annotated for protein-coding genes using BRAKER (Hoff et al., 2019) with multiple RNA-seq libraries (Table S3).

This genome has been deposited to GenBank under the accession JAIUCT000000000.

Differential expression analysis

Candidate *bicycle* genes from *H. cornu*, *T. ulmi*, and *A. pisum* identified by the classifier were manually re-annotated in APOLLO (Dunn et al., 2019) guided by RNAseq data viewed in IGV (Thorvaldsdóttir et al., 2013). Samples used for analysis are described in Table S3-5. Differential expression analyses were performed as described previously (Korgaonkar et al., 2021) and full analysis scripts are provided at https://doi/figshare_spaceholder.

hmmer and BLAST searches

The phylogeny of all species studied here was estimated using *Orthofinder* v 2.3.1 (Emms and Kelly, 2018) on the complete gene sets predicted by BRAKER for all genomes using the following settings: *-M msa -S diamond -A mafft -T fasttree*. The phylogeny was plotted using *ggtree* v. 2.0.4 in *R*.

We searched for *bicycle* homologs in all aphid species and outgroups using Position-Specific Interactive Basic Local Alignment Search Tool (PSI-BLAST) (Altschul, 1997) and *hmmer* (Eddy, 2011). We ran *PSI-BLAST* v. 2.9.0+ using *-max_target_seqs 5* and all other parameters as default. We used a cutoff of $E < 0.01$, Bonferroni corrected for 476 genes used as query searches. We ran profile-based *hmmsearch* v.3.2.1 using all default parameters and cutoff of $E < 0.01$ without multiple testing correction.

Selection tests

DnDs between *H. cornu* and *H. hamamelidis* for all transcripts in the transcriptome was calculated using the *codeml* function from the *PAML* package v4.9j (Yang, 2007). Genome-

589 wide selection scan for *H. cornu* was done done using *SweeD-P* v.3.1 (Pavlidis et al., 2013).
590 Details of both analysis can be found in the Methods section of Korgaonkar et al., 2021.
591
592 *Gene-structure aware sequence alignment*
593
594 We performed gene structure aware alignment using *prrn5* v 5.2.0 (Gotoh, 2021) with
595 default weighting parameters. We prepared gene-structure-informed extended fasta files that
596 were suitable as input to *prrn5* using the *anno2gsiseq.pl* utility provided at
597 https://github.com/ogotoh/prrn_aln/blob/master/perl/anno2gsiseq.pl. We summarized the
598 results of these multiple sequence alignments by generating histograms with a bin size of 1 to
599 illustrate the fraction of sequences in the alignment that had an intron at each position in the
600 alignment.
601
602 *Re-annotation of previously sequenced genomes*
603
604 The *A. pisum* and *P. venusta* genomes are approximately full-chromosome genomes
605 that were kindly provided to us by Yiyuan Li and Nancy Moran (Li et al., 2020, 2019) prior to
606 publication. All other genomes, except for *H. cornu* and *T. ulmi*, were downloaded from NCBI
607 (Table S5). We downloaded RNAseq data from NCBI (Table S4) and predicted coding genes
608 using BRAKER (Hoff et al., 2019) with the same workflow we had used previously that had
609 allowed discovery of *bicycle* genes (Korgaonkar et al., 2021).
610
611 **Availability of Data and Materials**
612

All new sequence data generated during this study is available at NCBI at the accession numbers provided in Tables S2-4. All analysis scripts, which allow reproduction of all analyses and reproduction of all figures, are provided as Supplementary Material.

References

Aguiar ERGR, Olmo RP, Paro S, Ferreira FV, de Faria IJ da S, Todjro YMH, Lobo FP, Kroon EG, Meignin C, Gatherer D, Imler J-L, Marques JT. 2015. Sequence-independent characterization of viruses based on the pattern of viral small RNAs produced by the host. *Nucleic Acids Research* **43**:6191–6206. doi:10.1093/nar/gkv587

Altschul S. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**:3389–3402. doi:10.1093/nar/25.17.3389

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**:403–410. doi:10.1016/S0022-2836(05)80360-2

Álvarez RA, González-Sierra S, Candelas A, Martinez J-JI. 2013. Histological study of galls induced by aphids on leaves of Ulmus minor : Tetraneura ulmi induces globose galls and Eriosoma ulmi induces pseudogalls. *Arthropod-Plant Interactions* **7**:643–650. doi:10.1007/s11829-013-9278-8

Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, Millán C, Park H, Adams C, Glassman CR, DeGiovanni A, Pereira JH, Rodrigues AV, Dijk AA van, Ebrecht AC, Opperman DJ, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy MK, Dalwadi U, Yip CK, Burke JE, Garcia KC, Grishin NV, Adams PD, Read RJ, Baker D. 2021. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*. doi:10.1126/science.abj8754

Bazan JF. 1991. Neuropoietic cytokines in the hematopoietic fold. *Neuron* **7**:197–208. doi:10.1016/0896-6273(91)90258-2

Beardsley JW, Gonzalez RH. 1975. The Biology and Ecology of Armored Scales. *Annu Rev Entomol* **20**:47–73. doi:10.1146/annurev.en.20.010175.000403

Betancourt AJ, Presgraves DC. 2002. Linkage limits the power of natural selection in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America* **99**:13616–13620. doi:10.1073/pnas.212277199

Betts MJ. 2001. Exon structure conservation despite low sequence similarity: a relic of dramatic events in evolution? *The EMBO Journal* **20**:5354–5360. doi:10.1093/emboj/20.19.5354

Birky CW, Walsh JB. 1988. Effects of linkage on rates of molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America* **85**:6414–6418. doi:10.1073/pnas.85.17.6414

Brown NP, Whittaker AJ, Newell WR, Rawlings CJ, Beck S. 1995. Identification and analysis of multigene families by comparison of exon fingerprints. *Journal of Molecular Biology* **249**:342–359. doi:10.1006/jmbi.1995.0301

Burstein M, Wool D, Eshel A. 1994. Sink strength and clone size of sympatric, gall-forming aphids. *European Journal of Entomology* **91**:57–61.
Caillaud CM, Boutin M, Braendle C, Simon JC. 2002. A sex-linked locus controls wing polymorphism in males of the pea aphid, Acyrthosiphon pisum (Harris). *Heredity* **89**:346–352.
Cook L g., Gullan P j. 2004. The gall-inducing habit has evolved multiple times among the eriococcid scale insects (Sternorrhyncha: Coccoidea: Eriococcidae). *Biological Journal of the Linnean Society* **83**:441–452. doi:10.1111/j.1095-8312.2004.00396.x
Dixon AFG. 1978. Biology of aphids, Studies in Biology. London: Edward Arnold.
Dommel M, Oh J, Huguet-Tapia JC, Guy E, Boulain H, Sugio A, Murugan M, Legeai F, Heck M, Smith CM, White FF. 2020. Big Genes, Small Effectors: Pea Aphid Cassette Effector Families Composed From Miniature Exons. *Front Plant Sci* **11**. doi:10.3389/fpls.2020.01230
Dunn NA, Unni DR, Diesh C, Munoz-Torres M, Harris NL, Yao E, Rasche H, Holmes IH, Elsik CG, Lewis SE. 2019. Apollo: Democratizing genome annotation. *PLoS Comput Biol* **15**:e1006790. doi:10.1371/journal.pcbi.1006790
Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Computational Biology* **7**. doi:10.1371/journal.pcbi.1002195
Eizaguirre C, Lenz TL, Kalbe M, Milinski M. 2012. Rapid and adaptive evolution of MHC genes under parasite selection in experimental vertebrate populations. *Nature Communications* **3**:1–6. doi:10.1038/ncomms1632
Elzinga DA, Jander G. 2013. The role of protein effectors in plant-aphid interactions. *Current Opinion in Plant Biology* **16**:451–456. doi:10.1016/j.pbi.2013.06.018
Emms DM. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics 14.
Emms DM, Kelly S. 2018. STAG: Species Tree Inference from All Genes. *bioRxiv* 1–29. doi:https://doi.org/10.1101/267914
Giron D, Huguet E, Stone GN, Body MJA. 2016. Insect-induced effects on plants and possible effectors used by galling and leaf-mining insects to manipulate their host-plant. *Journal of Insect Physiology* **84**:70–89. doi:10.1016/j.jinsphys.2015.12.009
Gotoh O. 2021. Cooperation of Spaln and Prrn5 for Construction of Gene-Structure-Aware Multiple Sequence Alignment In: Katoh K, editor. Multiple Sequence Alignment, Methods in Molecular Biology. New York, NY: Springer US. pp. 71–88. doi:10.1007/978-1-0716-1036-7_5
Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**:644–652. doi:10.1038/nbt.1883
Heger A, Holm L. 2000. Rapid automatic detection and alignment of repeats in protein sequences. *Proteins: Structure, Function, and Bioinformatics* **41**:224–237. doi:10.1002/1097-0134(20001101)41:2<224::AID-PROT70>3.0.CO;2-Z
Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-genome annotation with BRAKER, Methods in Molecular Biology. doi:10.1007/978-1-4939-9173-0_5
Hogenhout SA, Bos JIB. 2011. Effector proteins that modulate plant-insect interactions. *Current Opinion in Plant Biology* **14**:422–428. doi:10.1016/j.pbi.2011.05.003

Johnson KP, Dietrich CH, Friedrich F, Beutel RG, Wipfler B, Peters RS, Allen JM, Petersen M, Donath A, Walden KKO, Kozlov AM, Podsiadlowski L, Mayer C, Meusemann K, Vasilikopoulos A, Waterhouse RM, Cameron SL, Weirauch C, Swanson DR, Percy DM, Hardy NB, Terry I, Liu S, Zhou X, Misof B, Robertson HM, Yoshizawa K. 2018. Phylogenomics and the evolution of hemipteroid insects. *Proceedings of the National Academy of Sciences* 201815820–201815820. doi:10.1073/pnas.1815820115

Johnson LS, Eddy SR, Portugaly E. 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**. doi:10.1186/1471-2105-11-431

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 1–11. doi:10.1038/s41586-021-03819-2

Korgaonkar A, Han C, Lemire AL, Siwanowicz I, Bennouna D, Kopec RE, Andolfatto P, Shigenobu S, Stern DL. 2021. A novel family of secreted insect proteins linked to plant gall development. *Current Biology* **31**:1836-1849.e12. doi:10.1016/j.cub.2021.01.104

Li Y, Park H, Smith TE, Moran NA, Singh N. 2019. Gene Family Evolution in the Pea Aphid Based on Chromosome-level Genome Assembly. *Molecular Biology and Evolution* **36**:2143–2156. doi:10.1093/molbev/msz138

Li Y, Zhang B, Moran NA. 2020. The Aphid X Chromosome Is a Dangerous Place for Functionally Important Genes: Diverse Evolution of Hemipteran Genomes Based on Chromosome-Level Assemblies. *Molecular Biology and Evolution* **37**:2357–2368. doi:10.1093/molbev/msaa095

Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, Orengo C, Thornton J, Tramontano A. 2009. Protein function annotation by homology-based inference. *Genome Biol* **10**:207. doi:10.1186/gb-2009-10-2-207

Madeira F, Park Y mi, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research* **47**:W636–W641. doi:10.1093/nar/gkz268

Mani MS. 1964. Ecology of Plant Galls. The Hague, The Netherlands: Springer-Science+Business Media, B.V. doi:10.2307/3756723

Nishimura O, Hara Y, Kuraku S. 2017. GVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics* **33**:3635–3637. doi:10.1093/bioinformatics/btx445

Nováková E, Hypša V, Klein J, Foottit RG, von Dohlen CD, Moran NA. 2013. Reconstructing the phylogeny of aphids (Hemiptera: Aphididae) using DNA of the obligate symbiont Buchnera aphidicola. *Molecular Phylogenetics and Evolution* **68**:42–54. doi:10.1016/j.ympev.2013.03.016

Parr T. 1940. Asterolecanium Variolosum Ratzeburg, A Gall-Forming Coccid, and its Effect Upon the Host Trees. *YALE SCHOOL OF THE ENVIRONMENT BULLETIN SERIES* 90.

Paterson S, Vogwill T, Buckling A, Benmayor R, Spiers AJ, Thomson NR, Quail M, Smith F, Walker D, Libberton B, Fenton A, Hall N, Brockhurst MA. 2010. Antagonistic coevolution accelerates molecular evolution. *Nature* **464**:275–278. doi:10.1038/nature08798

Pavlidis P, Zˇivkovic D, Stamatakis A, Alachiotis N. 2013. SweeD: Likelihood-Based Detection of Selective Sweeps in Thousands of Genomes. *Molecular Biology & Evolution* **30**:2224–2234.

Richards S, Gibbs RA, Gerardo NM, Moran N, Nakabachi A, Stern D, Tagu D, Wilson ACC, Muzny D, Kovar C, Cree A, Chacko J, Chandrabose MN, Dao MD, Dinh HH, Gabisi RA, Hines S, Hume J, Jhangian SN, Joshi V, Lewis LR, Liu YS, Lopez J, Morgan MB, Nguyen NB, Okwuonu GO, Ruiz SJ, Santibanez J, Wright RA, Fowler GR, Hitchens ME, Lozado RJ, Moen C, Steffen D, Warren JT, Zhang J, Nazareth LV, Chavez D, Davis C, Lee SL, Patel BM, Pu LL, Bell SN, Johnson AJ, Vattathil S, Williams RL, Shigenobu S, Dang PM, Morioka M, Fukatsu T, Kudo T, Miyagishima SY, Jiang H, Worley KC, Legeai F, Gauthier JP, Collin O, Zhang L, Chen HC, Ermolaeva O, Hlavina W, Kapustin Y, Kiryutin B, Kitts P, Maglott D, Murphy T, Pruitt K, Sapojnikov V, Souvorov A, Thibaud-Nissen F, Câmara F, Guigó R, Stanke M, Solovyev V, Kosarev P, Gilbert D, Gabaldón T, Huerta-Cepas J, Marcet-Houben M, Pignatelli M, Moya A, Rispe C, Ollivier M, Quesneville H, Permal E, Llorens C, Futami R, Hedges D, Robertson HM, Alioto T, Mariotti M, Nikoh N, McCutcheon JP, Burke G, Kamins A, Latorre A, Ashton P, Calevro F, Charles H, Colella S, Douglas AE, Jander G, Jones DH, Febvay G, Kamphuis LG, Kushlan PF, Macdonald S, Ramsey J, Schwartz J, Seah S, Thomas G, Vellozo A, Cass B, Degnan P, Hurwitz B, Leonardo T, Koga R, Altincicek B, Anselme C, Atamian H, Barribeau SM, De Vos M, Duncan EJ, Evans J, Ghanim M, Heddi A, Kaloshian I, Vincent-Monegat C, Parker BJ, Pérez-Brocal V, Rahbé Y, Spragg CJ, Tamames J, Tamarit D, Tamborindeguy C, Vilcinskas A, Bickel RD, Brisson JA, Butts T, Chang CC, Christiaens O, Davis GK, Duncan EJ, Ferrier D, Iga M, Janssen R, Lu HL, McGregor A, Miura T, Smagghe G, Smith J, Van Der Zee M, Velarde R, Wilson M, Dearden P, Edwards OR, Gordon K, Hilgarth RS, Rider SD, Srinivasan D, Walsh TK, Ishikawa A, Jaubert-Possamai S, Fenton B, Huang W, Rizk G, Lavenier D, Nicolas J, Smadja C, Zhou JJ, Vieira FG, He XL, Liu R, Rozas J, Field LM, Campbell P, Carolan JC, Fitzroy CIJ, Reardon KT, Reeck GR, Singh K, Wilkinson TL, Huybrechts J, Abdel-Latief M, Robichon A, Veenstra JA, Hauser F, Cazzamali G, Schneider M, Williamson MMS, Stafflinger E, Hansen KK, Grimmelikhuijzen CJP, Price DRG, Caillaud M, Van Fleet E, Ren Q, Gatehouse JA, Brault V, Monsion B, Diaz J, Hunnicutt L, Ju HJ, Pechuan X, Aguilar J, Cortés T, Ortiz-Rivas B, Martínez-Torres D, Dombrovsky A, Dale RP, Davies TGE, Williamson MMS, Jones A, Sattelle D, Williamson S, Wolstenholme A, Cottret L, Sagot MF, Heckel DG, Hunter W, Consortium TIAG. 2010. Genome Sequence of the Pea Aphid Acyrthosiphon pisum. *PLoS Biol* **8**:e1000313–e1000313. doi:10.1371/journal.pbio.1000313

Rokas A, Holland PWH. 2000. Rare genomic changes as a tool for phylogenetics. *Trends in Ecology and Evolution* **15**:454–459.

Rokas A, Kathirithamby J, Holland PWH. 1999. Intron insertion as a phylogenetic character: the engrailed homeobox of Strepsiptera does not indicate affinity with Diptera. *Insect Molecular Biology* **8**:527–530.

Roy SW, Gilbert W. 2005. Rates of intron loss and gain: Implications for early eukaryotic evolution. *Proc Natl Acad Sci U S A* **102**:5773–5778. doi:10.1073/pnas.0500383102

Ruvinsky A, Watson C. 2007. Intron Phase Patterns in Genes: Preservation and Evolutionary Changes. *TOEVOLJ* **1**:1–14. doi:10.2174/1874404400701010001

Schaible G, Lee DJ, Kim CS, Schaible G, Garrett L, Lubowski R. 2008. Economic Impacts of the U. S. Soybean Aphid Infestation : A Multi-Regional Competitive Dynamic Analysis. *Agricultural and Resource Economics Review* **37**:227–242. doi:10.1017/S1068280500003026

Shorthouse JD, Wool D, Raman A. 2005. Gall-inducing insects - Nature's most sophisticated herbivores. *Basic and Applied Ecology* **6**:407–411. doi:10.1016/j.baae.2005.07.001

Telford MJ, Copley RR. 2011. Improving animal phylogenies with genomic data. *Trends in Genetics* **27**:186–195. doi:10.1016/j.tig.2011.02.003

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**:178–192. doi:10.1093/bib/bbs017

Vakirlis N, Carvunis A-R, McLysaght A. 2020. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife* **9**:e53500. doi:10.7554/eLife.53500

Venkatesh B, Ning Y, Brenner S. 1999. Late changes in spliceosomal introns define clades in vertebrate evolution. *Proceedings of the National Academy of Sciences* **96**:10267–10271. doi:10.1073/pnas.96.18.10267

von Dohlen and N. A. Moran CD. 1995. Molecular phylogeny of the Homoptera: a paraphyletic taxon. *J Mol Evol* **41**:211–223.

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. *Genome research* **27**:757–767. doi:10.1101/gr.235812.118

Weisman CM, Murray AW, Eddy SR. 2020. Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLoS Biol* **18**:e3000862. doi:10.1371/journal.pbio.3000862

Will T, Tjallingii WF, Thönnessen A, van Bel AJE. 2007. Molecular sabotage of plant defense by aphid saliva. *Proceedings of the National Academy of Sciences of the United States of America* **104**:10536–10541. doi:10.1073/pnas.0703535104

Xie W, He C, Fei Z, Zhang Y. 2020. Chromosome-level genome assembly of the greenhouse whitefly (Trialeurodes vaporariorum Westwood). *Molecular Ecology Resources* **20**:995–1006. doi:10.1111/1755-0998.13159

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **24**:1586–1591. doi:10.1093/molbev/msm088

Acknowledgements

We thank Andrew Lemire and the whole Quantitative Genomics team at Janelia Research Campus for generating and sequencing the RNA-seq samples, Tom Dolafi for maintaining our Apollo Annotation server, Zari Zavala-Ruiz and the Janelia Visitor Project for supporting CH, Fernando Bazan for recommending the use of RADAR software to detect internal repeats in *megacycle* genes, Gregory Davis for providing samples of *A. pisum*, and

members of the Stern lab for critical feedback on this work and this manuscript, especially Aishwarya Korgaonkar.

Funding

The authors are employees of Howard Hughes Medical Institute. The funders had no role in the design of the study or in data collection, analysis, and interpretation or in writing of the paper.

Contributions

DLS conceptualized the project, collected aphids, performed dissections, made DNA and most RNA samples, performed initial analyses, and wrote the paper. CH developed the bicycle gene classifier, independently verified all of the computational analyses, and assisted with preparations of figures and manuscript.

Ethics declarations

HHMI has filed a provisional patent, number 63/243,904, for the inventors DLS and CH covering a gene structure, coding sequence-independent method of identifying bicycle genes.

850 **Figure Legends**

851

852 Figure 1. Performance of a gene-structure based classifier of *bicycle* genes.

853 (A) Precision (the proportion of true positives among both true and false positives)
854 versus recall (the proportion of true positives among true positives and false
855 negatives) for the model trained on 476 previously identified *H. cornu bicycle*
856 genes. Differently colored dots represent the precision and recall for different
857 probability cutoffs of the model, shown in panel (C).

858 (B) The ratio of precision to recall for different cutoffs. Precision approximately
859 equals recall (P/R = 1) at a cutoff of 0.72.

860 (C) Precision and recall values for models where one variable at a time was removed from
861 the model. Removal of the number of exons with mode2 had the largest decrement on
862 performance, but all models performed well.

863

864 Figure 1 Supplement 1. Salient features of gene structure of *H. cornu bicycle* genes.

865 (A-B) Almost all internal exons of *bicycle* genes are of phase 2 (A), whereas the majority of
866 internal exons of other genes in the *H. cornu* genome are of phase 0(B).

867 (C-D) *Bicycle* genes have a wide range of exon numbers, with a distribution (C) distinctly
868 different from non-*bicycle* genes (D).

869 (E-F) The total length of *bicycle* proteins is better predicted by the number of exons in the
870 gene (E) than by the length of the first and last exon (F).

871 (G) Precision versus recall curves for replicate training runs of the classifier using 100
872 different subsets of the training data. The curves across all subsets are very similar,
873 indicating that the classifier was not unduly influenced by a specific subset of the data.

(H) Histogram of the number of transcripts with a particular classifier response shows a bimodal distribution. The vast majority of genes (>10,000) had a response close to 0. A threshold of 0.72 was chosen at a precision to recall ratio of 1.

Figure 2. The gene-structure based classifier identifies additional CYC and non-CYC motif containing genes in the *H. cornu* genome that may contribute to the effector protein repertoire of multiple stages of the *H. cornu* life cycle.

(A) Hierarchical clustering of candidate *H. cornu bicycle* homologs detected by the gene-structure based on amino-acid sequence similarity measure reveals two clusters of genes with similar sequences. Some of the genes in Cluster 1 appear to have some sequence similarity to Cluster 2 genes.

(B) Genes in Cluster 1 encode proteins with N-terminal signal sequences and weak similarity to *bicycle* genes.

(C) Genes in Cluster 2 encode proteins with sequence similarity to the previously described *bicycle* genes.

(D) Diagram of life cycle of *H. cornu*. Individuals from different generations exhibit phenotypes that are specialized for each stage of the complex life cycle. The fundatrix (G1) is the only generation that induces a gall.

(E) A volcano plot of the strength of evidence for differential expression (-log10(P value)) versus the fold-change differential expression of salivary glands isolated from fundatrices versus sexuals reveals that some of the Cluster1 (black) and 2 (orange) genes identified by the gene structure based classifier are as strongly enriched in fundatrix salivary glands as the originally described *bicycle* genes (blue). In addition, several Cluster 1 and 2 genes are more strongly enriched in salivary glands of sexuals.

(F-G) Plots of gene expression levels for Cluster 1 (F) and 2 (G) genes across four generations of the life cycle (Figure S2C) reveals that while most *bicycle* homologs are most strongly expressed in fundatrix salivary glands (e.g. purple lines), some are weakly expressed in fundatrix salivary glands and then strongly expressed in salivary glands of individuals from other stages of the life cycle (e.g. red lines).

Figure 2 Supplement 1. Performance measures of the *H. cornu* gene-structure based classifier and diagram of the life cycle of *H. cornu.*

(A-C) d_N/d_S ratio for the 476 original *bicycle* genes (C), the newly putative *bicycle* genes (D) and the non-*bicycle* genomic background (E) between *H. cornu* and *H, hamamelidis*. Both the original *bicycle* genes and the newly putative *bicycle* genes show elevated ratio of d_N/d_S above 1 compared to the background. Dashed vertical line indicates d_N/d_S = 1, the putative neutrality.

(D) Median distance to the closest sweep signal for original *bicycle* genes (red dashed line) and new putative *bicycle* genes (blue dashed line). Histogram shows median distance to closest sweep signal for 1000 permutations of gene labels. The original *bicycle* genes and the new putative *bicycle* genes are both closer to sweep signals than expected by chance.

Figure 3. Features of *H. cornu bicycle* genes

(A) The originally identified *bicycle* genes encode proteins that exhibit a wide size range, with representatives of approximately four size classes, indicated by the vertical dotted blue lines.

922 (B-E) Logo plots of the proteins from each of the four size classes of Bicycle proteins
923 identifies Unicycle (B), Bicycle (C), Tricycle (D), and Tetracycle (E) proteins. The CYC
924 motifs in each protein are identified by thick black lines below each Logo plot.
925
926 Figure 3 Supplement 1. Gene-structure aware alignments reveal extensive conservation of
927 intron locations between the originally-defined *bicycle* genes and putative *bicycle* homologs
928 identified by the gene-structure based classifier.
929 (A-D) Predicted proteins encoded by the original or newly identified *bicycle* homologs were
930 divided by sequence length into Unicycle (A), Bicycle (B), Tricycle (C), or Tetracycle (D)
931 proteins and aligned using gene-structure aware alignment (Gotoh, 2021). The
932 frequency of introns at each location in the alignments are plotted for the original
933 (magenta) and newly defined (cyan) *bicycle* homologs and positions without introns
934 were excluded from the plot. Intron positions are therefore relative to each other and do
935 not indicate absolute position in the proteins or the alignments. Coincidence of intron
936 positions is shown by purple bars.
937
938 Figure 4. The *T. ulmi bicycle* homologs detected by the gene-structure based classifier display
939 CYC motifs and highly divergent protein sequences.
940 (A) Hierarchical clustering of candidate *T. ulmi bicycle* homologs based on amino-acid
941 sequence similarity measure reveals four clusters of genes.
942 (B-E) The four clusters of *T. ulmi bicycle* homologs display two (D) or one (E) clear CYC
943 motifs, one clear and one degenerate CYC motif (B), or no clear CYC motifs (C). The
944 cluster identities are indicated in each subplot.

(E) Volcano plot of genes differentially expressed in fundatrix salivary glands versus carcass with *bicycle* genes labelled. Most of the *T. ulmi* candidate *bicycle* homologs are strongly over-expressed in the fundatrix salivary glands.

Figure 4 Supplement 1. *Tetraneura ulmi* life cycle and additional features of *T. ulmi bicycle* genes.

(A) Photo of mature galls of *T. ulmi* on leaf of *Ulmus americana*. Galls collected at Janelia Research Campus, Ashburn, VA, USA.

(B) Diagram of life cycle of *T. ulmi*. Samples for differential expression analysis were collected from generations G1 and G2.

(C) The *T. ulmi* putative *bicycle* homologs from Clusters 3 and 2, with and without CYC motifs, respectively, share multiple intron positions. Proteins from the two classes of genes were aligned using gene-structure aware alignment (Gotoh, 2021) and the frequency of introns at each location in the alignment was plotted in different colors for the two classes of genes and positions without introns were excluded from the plot. Intron positions are therefore relative to each other and do not indicate absolute position in the proteins or the alignments. Sites of coincident introns are indicated in purple.

Figure 5. *A. pisum bicycle* homologs contain single CYC motif and are strongly enriched in salivary glands.

(A) Logo plot of *A. pisum bicycle* homologs detected by gene-structure based classifier. Most *A. pisum* homologs contain a single CYC motif.

(B) Volcano plot of differential expression between salivary glands and carcass reveals that *A. pisum bicycle* homologs are strongly over-expressed in salivary glands.

970

971 Figure 5 Supplement 1. The *A. pisum* putative *bicycle* homologs share multiple intron positions
972 with *H. cornu bicycle* genes. Proteins were aligned using gene-structure aware alignment
973 (Gotoh, 2021) and the frequency of introns at each location in the alignment was plotted in
974 different colors for the two classes of genes and positions without introns were excluded from
975 the plot. Intron positions are therefore relative to each other and do not indicate absolute
976 position in the proteins or the alignments. Sites of coincident introns are indicated in purple.

977

978 Figure 6. The gene-structure based classifier detects *bicycle* genes in genomes of aphids,
979 phylloxerids, and coccids. A whole-proteome phylogeny of the species studied is shown on
980 the left. Family membership is shown in color beside the species names. The number of
981 candidate *bicycle* homologs detected by *PSI-BLAST*, *hmmer*, and the gene-structure based
982 classifier is shown on the right.

983

984 Figure 6 Supplement 1. Features of bicycle homologs detected in aphid species.

985 (A) A plot of the number of *bicycle* homologs detected by the gene-structure based
986 classifier versus genome N50, a measure of genome assembly quality, shows no
987 significant relationship between these variables (R^2 = 0.1, F = 1.6, P = 0.22).

988 (B-E) Logo plots of putative *bicycle* homologs detected in four aphid species, *M. persicae*
989 (B), *P. nigronervosa* (C), *R. maidis* (D), and *C. cedri* (E) reveals that they all contain CYC
990 motifs. The *M. persicae* and *P nigronervosa* genes are primarily *unicycles*, the *R. maidis*
991 genes are primarily *bicycles*, and the *C. cedri* genes are primarily *tetracycles*.

992

993 Figure 6 Supplement 2. Features of *bicycle* homologs detected in outgroups to aphids.

994 (A-C) Logo plots of proteins encoded by putative *bicycle* homologs detected in *D. vitilifoliae*
995 (A), *M. hirsutus* (B), and *P. solenopsis* (C). The logo plots from these putative homologs
996 show no obvious CYC motifs.
997 (D-F) Gene-structure aware alignments (Gotoh, 2021) reveal that multiple introns are shared
998 between these highly divergent putative *bicycle* homologs and *H. cornu bicycle* genes.
999 Positions in alignments without introns were excluded from the plot. Intron positions are
1000 therefore relative to each other and do not indicate absolute position in the proteins or
1001 the alignments.
1002
1003 Figure 7. *Megacycle* genes have a repeating protein structure and are overexpressed in
1004 salivary glands of multiple aphid species.
1005 (A) Gene structure of the *M. sacchari megacycle* gene (*g20694*). The gene includes 118
1006 exons encoded in approximately 100 kbp of DNA. The protein is 2092 AA long.
1007 (B) Rapid Automatic Detection and Alignment of Repeats (RADAR) analysis of the
1008 *M. sacchari megacycle* gene provides evidence for 15 repeating units within the
1009 protein encoded by *g20694*. Multiple sequence alignment of these repeat units
1010 (not using information on gene structure) reveals that many introns, marked in
1011 green, are located in identical or similar positions in the aligned sub-sequences.
1012 A logo plot of the aligned repeated sequences is shown above the alignment.
1013 There is no obvious evidence for CYC motifs, although the motif is
1014 approximately the same length as the proteins encoded by many *bicycle* genes.
1015 (C-E) Volcano plots of differential expression between salivary glands and carcasses for
1016 three species, *H. cornu* (C), *T. ulmi* (D), and *A. pisum* (E), reveals that *megacycle* genes
1017 are strongly over-expressed in salivary glands of all three species.

Figure 7 Supplement 1. *Megacycle* genes share intron locations with *bicycle* genes and are found across many aphid species.

(A) Gene-structure aware alignments (Gotoh, 2021) reveal that multiple introns are shared between *megacycle* genes and *H. cornu bicycle* genes. Positions in alignments without introns were excluded from the plot. Intron positions are therefore relative to each other and do not indicate absolute position in the proteins or the alignments. Shared introns are shown in purple.

(B) Maximum likelihood phylogeny of the protein sequences encoded by *megacycle* genes found in aphid species. *Megacycle* genes were not detected outside of aphids. The phylogeny has a topology similar to the aphid portion of the whole-proteome phylogeny shown in Figure 6. Only one *megacycle* gene was detected in most aphid species, but three paralogs were detected in *C. cedri* and *T. ulmi*. Values at nodes are bootstrap support values. Scale bar is 1 substitution per residue.

Supplementary Tables

Table S1. Many microexon genes producing secreted putative effector proteins in *A. pisum* (Dommel et al., 2020) are identified as *bicycle* genes by the classifier.

Gene Name (Dommel)	Transcript ID	Gene Locus	Augustus gene	Peptide sequence	*bicycle* gene according to classifier	Notes
KQY1a	NM_001162442	LOC100158789	g18916.t1	MIFFKQYSMMITFIVIAVWVMPAITSE	Yes	
KQY1b	AK339882	LOC100158789	g18916.t1	MIFFKQYSMMITFIVIAVWVMPAITSE	Yes	
KQY2a	AK342599	LOC100302371	No perfect match. 3'	MVFYKQYLLTITCIVITAWVIPTSA	Yes	

			match to g100001.1			
KQY2b	AK342927	LOC100302371	No perfect match. 3' match to g100001.1	MVFYKQYLLTITCIVITAWVIPTSA	Yes	
KQY3	AK341661	LOC100301916	g17857.t1	MVFFKQYLITLTCIVISVWITPVNT	Yes	
KQY4a	AK342948	LOC100302370	g39706.t1	MIFFKQYLIILTFIVIAVLVMPVTP	Yes	
KQY4b	AK342242	LOC100302370	g39706.t1	MIFFKQYLIILTFIVIAVLVMPVTP	Yes	
KQY4c	AK342690	LOC100302370	g39706.t1	MIFFKQYLIILTFIVIAVLVMPVTP	Yes	
KQY4d	AK342678	LOC100302370	g39706.t1	MIFFKQYLIILTFIVIAVLVMPVTP	Yes	
KQY5a	AK342406	LOC100302375	g41375.t1	MVFFKQYLLTLTCIVIVVQVMPASA	Yes	
KQY5b	AK342378	LOC100302376	g41375.t1	MVFFKQYLLTLTCIVIVVQVMPASA	Yes	
KQY6	AK340126	LOC100302481	g39707.t1	MIFFKQYLIMLTFIIIAVWVMPANT	Yes	
KQY7	AK340563	LOC100302485	g100003.t1	MSFFKQYLTLTFIVISVWNMSEA	No	Single exon of only signal peptide
KQY8	AK342473	LOC100302439	g23349.t1	MVFFKQFLITLTVIIITEA	No	No obvious signal peptide. Gene on small contig and annotation likely truncated
KQY9	AK342683	LOC100302403	g41375.t1	MVFFKLYLLTLTCIVIAVWVMPVSA	Yes	
KQY10	AK342808	LOC100302381	g13591.t1	MFNVLIILSLISYTFEPSYTLYKFKMVFFKQDLLMLTCITIAVWIMPPSASTN	Yes	
KQY11	AK340121	LOC100302480	g23349 (best match)	MVFFRQFLITLSVILITEA	No	Cannot find exact match
KQYMp	XM_022322515	LOC111039170	g15805.t1	MHFFKHYLIVLTYIVISFWFMPSASL	Yes	Myzus persicae
KHI1a	AK339863	LOC100159750	g19277.t1	MFKHIIVLVLCFMAYFVGNLDA	Yes	
KHI1b	AK339862	LOC100159750	g19277.t1	MFKHIIVLVLCFMAYFVGNLDA	Yes	
KHI2a	AK341162	LOC100302383	g1192.t1	MDKHIIMLALCLMVYIIGNIDA	No	On small contig, likely partial annotation
KHI2b	AK341161	LOC100302383	g1192.t1	MDKHIIMLALCLMVYIIGNIDA	No	On small contig, likely partial annotation
KHI2c	AK342769	LOC100302383	g1192.t1	MDKHIIMLALCLMVYIIGNIDA	No	On small contig, likely partial annotation

KHI3a	AK340197	LOC100166702	g34738	MLKHIIVLALYLMAYIIGNIDA	Yes	
KHI3b	AK342603	LOC100166702	g34738	MLKHIIVLALYLMAYIIGNIDA	Yes	
KHI4	AK341077	LOC100570519	g100004	MLKHILLALCFMAYIIENIG	Yes	
KHI5	AK341390	LOC100534636	g100004 (best match)	MLKHILLALCFMAYIIENIGA	Yes	Cannot find exact match
KHI6	AK340760	LOC100571631	g100005	MLKHIIVLVLCFMPYIIG	No	2 exons on small contig; not searched by classifier
C002Ap	XM_001948323	LOC100167863	g43317	MGSYKLYVAVMAIAIAVVQEVRC	No	2 exon gene; not searched by classifier
Hypothetical protein ACYPI003695	NM_001293542.1	LOC100162547	g2264		Yes	

Table S2—Details of *Tetraneura ulmi* sample, sequence files, and SRA accession numbers for genome sequencing. JRC refers to Janelia Research Campus, Ashburn VA, USA. These sequences can be found under the SRA BioProject PRJNA759586.

Sample ID	Collection Date	Collection Location	Species	Library Prep Method	Sequencing Method	# Read pairs	Accession #
		JRC	*Tetraneura ulmi*	10X Chromium	Illumina PE151	441,975,717	SRR15704183

Table S3 - Details of biological samples, sequence files, and SRA accession numbers for *Tetraneua ulmi* RNA-sequencing. JRC refers to Janelia Research Campus, Ashburn VA. These sequences can be found under the SRA BioProject PRJNA762862.

Sample ID	Collection Date	Collection Location	Generation	Organ	Library Prep Method	Sequencing Method	# Reads	Accession #
Tulm_Fund_SG_2	4/22/19	JRC	1	SG	WGB	Illumina PE150	8475508, 5308849	SRR15871517
Tulm_Fund_SG_4	4/22/19	JRC	1	SG	WGB	Illumina PE150	6087930, 6250323	SRR15871516
Tulm_Fund_SG_5	4/22/19	JRC	1	SG	WGB	Illumina PE150	7474183, 7147731	SRR15871515
Tulm_Fund_Carc_2	4/22/19	JRC	1	Carcass	WGB	Illumina PE150	6208251, 6194859	SRR15871514
Tulm_Fund_Carc_4	4/22/19	JRC	1	Carcass	WGB	Illumina PE150	7453030, 6095776	SRR15871513
Tulm_Fund_Carc_5	4/22/19	JRC	1	Carcass	WGB	Illumina PE150	8399343, 7332616	SRR15871512
Tulm_2ndGen_SG_2_1	4/22/19	JRC	2	SG	WGB	Illumina PE150	6902051, 7037142	SRR15871511

Tulm_2ndGen_SG_2_2	4/22/19	JRC	2	SG	WGB	Illumina PE150	6888742, 6997123	SRR15871510
Tulm_2ndGen_SG_2_3	4/22/19	JRC	2	SG	WGB	Illumina PE150	6049169, 6696157	SRR15871509

Table S4 - Details of biological samples, sequence files, and SRA accession numbers for *Acyrthosiphon pisum* RNA-sequencing. The aphid strain used was LSR1 (Caillaud et al., 2002), originally collected in the vicinity of Ithaca, NY, USA in summer 1998. Aphids were grown on broad bean plants in the laboratory at Janelia Reseach Campus and dissected for these samples on 21 February 2021. Sequences can be found under SRA BioProject PRJNA762703.

Sample ID	Collection Date	Collection Location	Generation	Organ	Library Prep Method	Sequencing Method	# Reads	Accession #
SG1	1998	Ithaca, NY, USA	1	SG	WGB	Illumina PE150	1273082, 5724222	SRR15862168
SG2	1998	Ithaca, NY, USA	1	SG	WGB	Illumina PE150	1557673, 6606339	SRR15862167
SG3	1998	Ithaca, NY, USA	1	SG	WGB	Illumina PE150	1199959, 5221105	SRR15862166
SG4	1998	Ithaca, NY, USA	1	SG	WGB	Illumina PE150	1159521, 5169430	SRR15862165
carc1	1998	Ithaca, NY, USA	1	Carcass	WGB	Illumina PE150	998879, 4363153	SRR15862164
carc2	1998	Ithaca, NY, USA	1	Carcass	WGB	Illumina PE150	1185062, 5433749	SRR15862163
carc3	1998	Ithaca, NY, USA	1	Carcass	WGB	Illumina PE150	1045749, 4901408	SRR15862162

Table S5 – SRA accession numbers for Genomes and RNAseq samples downloaded from NCBI.

Species	**Original genomes Genbank assembly accession or alternative location**	**Manually corrected BRAKER Annotation**	**TRINITY assembly**
Acyrthosiphon pisum	GCF_005508785.1	Apis.Augustus.updated_w_ annots.7vii21.gff3	apis.Trinity.fasta.gz
Aphis gossypi	GCF_004010815.1	Agos.Augustus.updated_w_ annots.24vii21.gff3	agos.Trinity.fasta.gz
Bemisia tabaci	GCA_001854935.1	Btab.Augustus.updated_w_ annots.7vii21.gff3	
Cinara cedri	GCA_902439185.1	Cced.Augustus.updated_w_ annots.7vii21.gff3	cced.Trinity.fasta.gz
Daktulosphaira vitifoliae	https://bipaa.genouest.org/sp/daktulosphaira_vitifoliae/ download/genome/v3.1/ Dv_mitochondrial_genome.fa	Dvit.Augustus.updated_w_ annots.7vii21.gff3	
Diaphorina citri	GCA_000475195.1	Dcit.Augustus.updated_w_ annots.7vii21.gff3	
Diuraphis noxia	GCA_001186385.1	Dnox.Augustus.updated_w_ annots.7vii21.gff3	dnox.Trinity.fasta.gz
Ericerus pela	GCA_011428145.1	Epel.Augustus.updated_w_ annots.7vii21.gff3	
Eriosoma lanigerum	GCA_013282895.1	Elan.Augustus.updated_w_ annots.7vii21.gff3	elan.Trinity.fasta.gz

Hormaphis cornu	GCA_017140985.1	Augustus.updated_w_ annots.19viii21_final.gff3	hcor.Trinity.fasta.gz
Laodelphax striatella	GCA_003335185.2	Lstr.Augustus.updated_w_ annots.7vii21.gff3	
Maconellicoccus hirsutus	GCA_003261595.1	Mhir.Augustus.updated_w_ annots.7vii21.gff3	
Melanaphis sacchari	GCA_002803265.2	Msac.Augustus.updated_w_ annots.7vii21.gff3	msac.Trinity.fasta.gz
Myzus persicae	GCA_001856785.1	Mper.Augustus.updated_w_ annots.7vii21.gff3	mper.Trinity.fasta.gz
Nilaparvata lugens	GCA_000757685.1	Nlug.Augustus.updated_w_ annots.7vii21.gff3	
Pachypsylla venusta	GCA_012654025.1	Pven.Augustus.updated_w_ annots.7vii21.gff3	
Pentalonia nigronervosa	GCA_014851325.1	Pnig.Augustus.updated_w_ annots.7vii21.gff3	pnig.Trinity.fasta.gz
Phenacoccus solenopsis	GCA_009761765.1	Psol.Augustus.updated_w_ annots.7vii21.gff3	
Rhopalosiphum maidis	GCF_003676215.2	Rmai.Augustus.updated_w_ annots.7vii21.gff3	rmai.Trinity.fasta.gz
Sipha flava	GCA_003268045.1	Sfla.Augustus.updated_w_ annots.7vii21.gff3	sfla.Trinity.fasta.gz
Tetraneura ulmi	to be uploaded	Tulm.Augustus.updated_w_ annots.7vii21.gff3	tulm.Trinity.fasta.gz
Trialeurodes vaporariorum	GCA_011764245.1	Tvap.Augustus.updated_w_ annots.7vii21.gff3	

1067

Figure 1

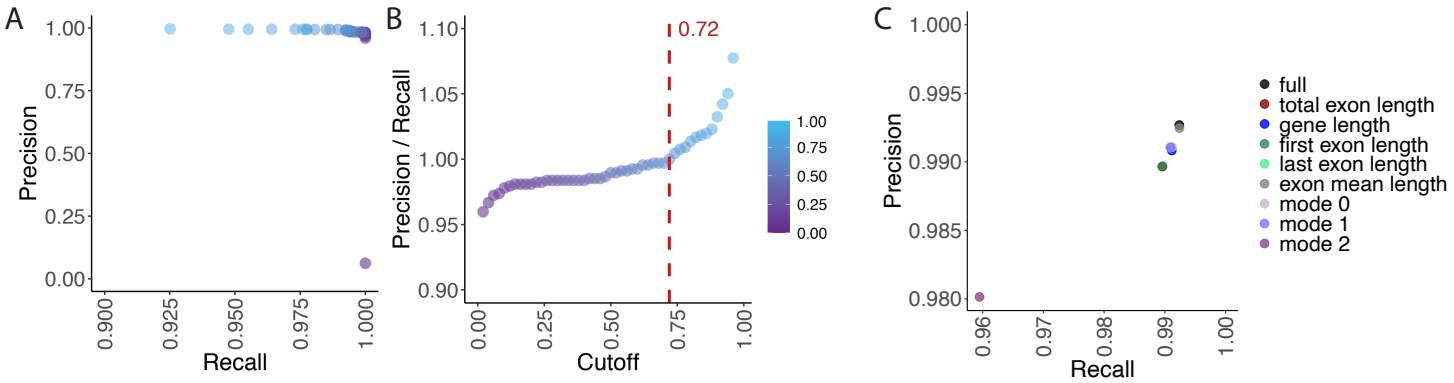

A
Precision
Recall
B
Precision / Recall
Cutoff
0.72
C
Precision
Recall
full
total exon length
gene length
first exon length
last exon length
exon mean length
mode 0
mode 1
mode 2

Figure 1 Supplement 1

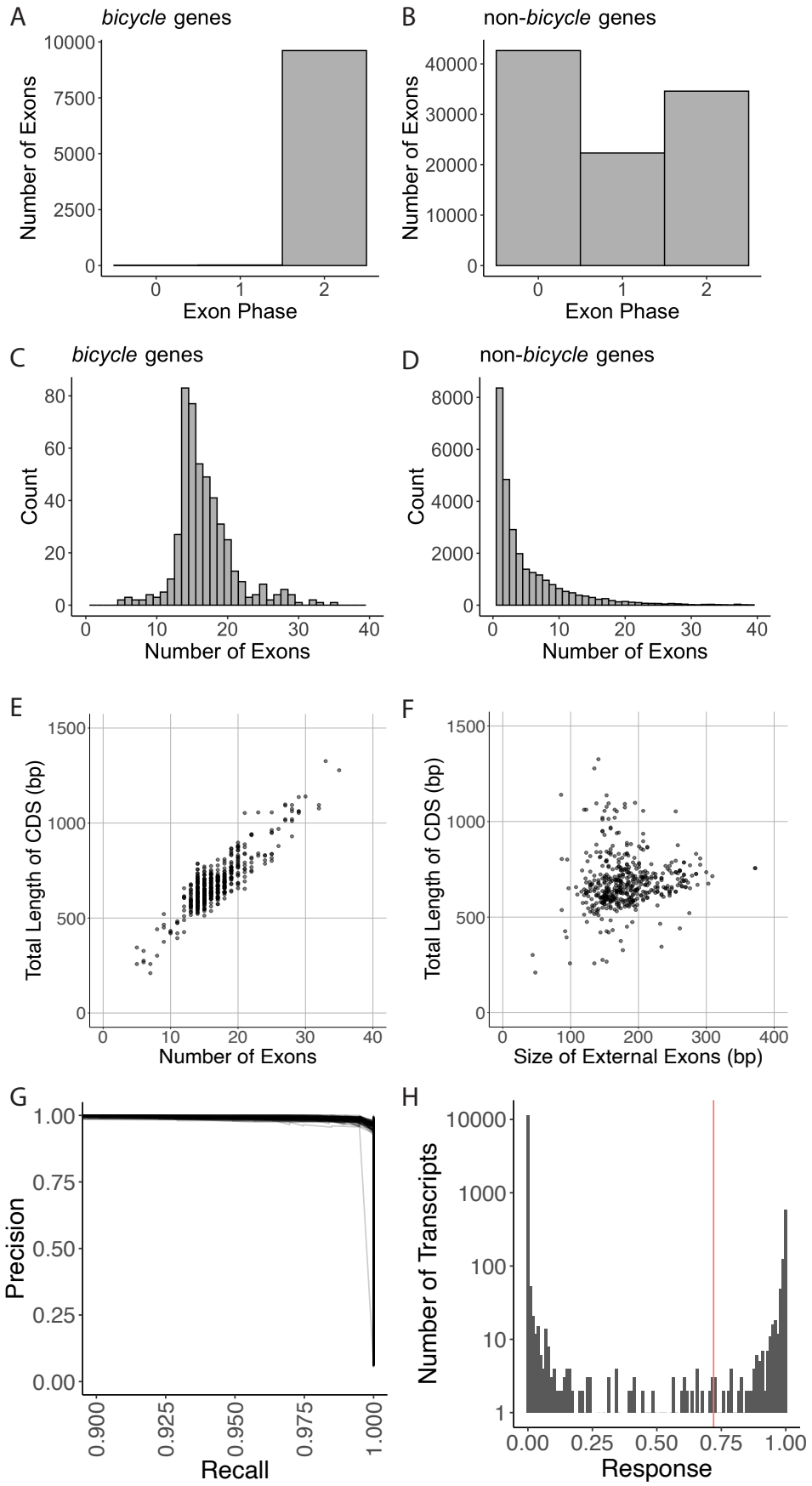


Figure 2

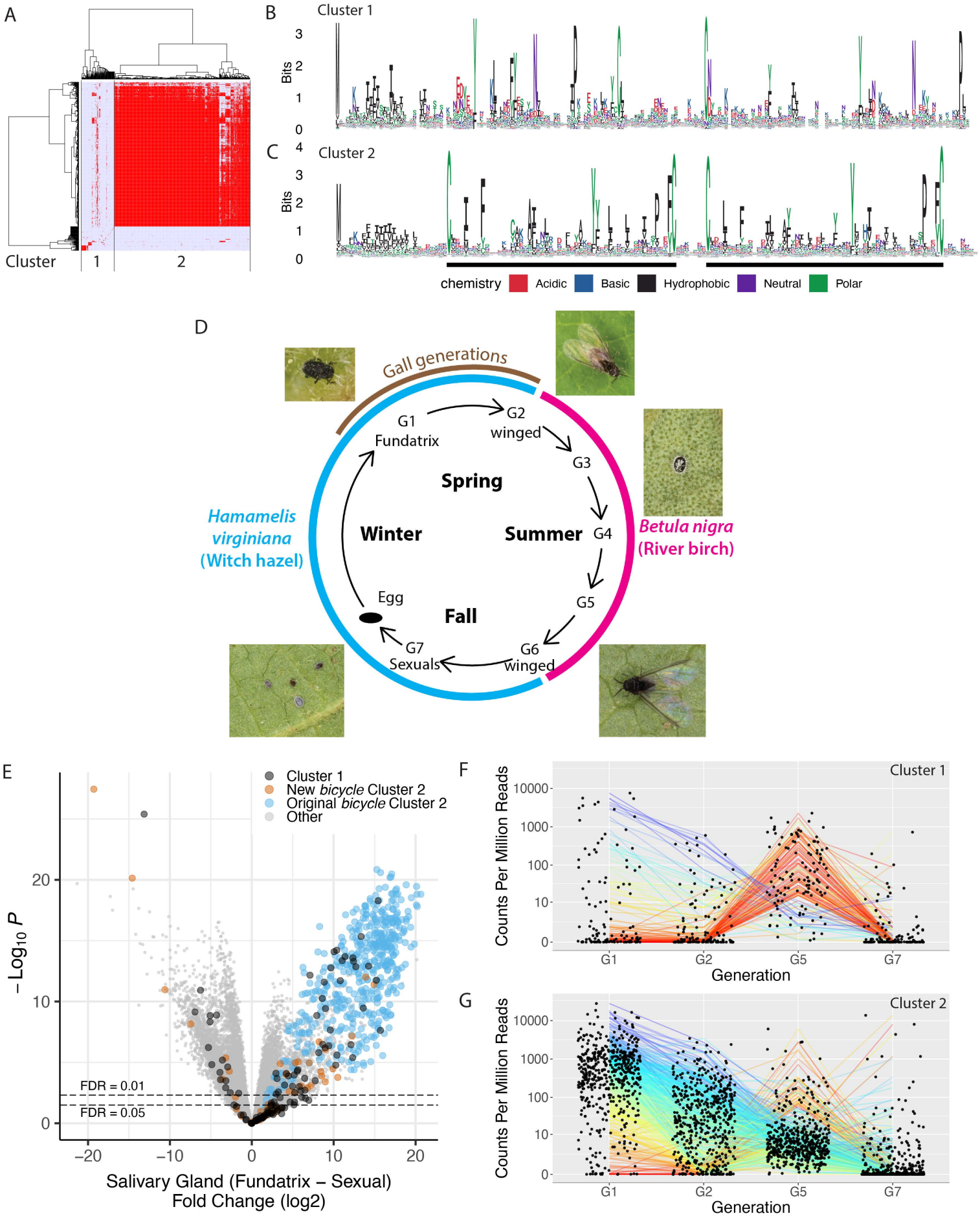

A
Cluster
1
2
B
Cluster 1
Bits
C
Cluster 2
chemistry
Acidic
Basic
Hydrophobic
Neutral
Polar
D
Gall generations
G1
Fundatrix
G2
winged
G3
G4
G5
G6
winged
G7
Sexuals
Egg
Spring
Summer
Fall
Winter
Hamamelis virginiana
(Witch hazel)
Betula nigra
(River birch)
E
Cluster 1
New bicycle Cluster 2
Original bicycle Cluster 2
Other
FDR = 0.01
FDR = 0.05
– Log10 P
Salivary Gland (Fundatrix – Sexual)
Fold Change (log2)
F
Cluster 1
Counts Per Million Reads
Generation
G
Cluster 2
Counts Per Million Reads
Generation

Figure 2 Supplement 1

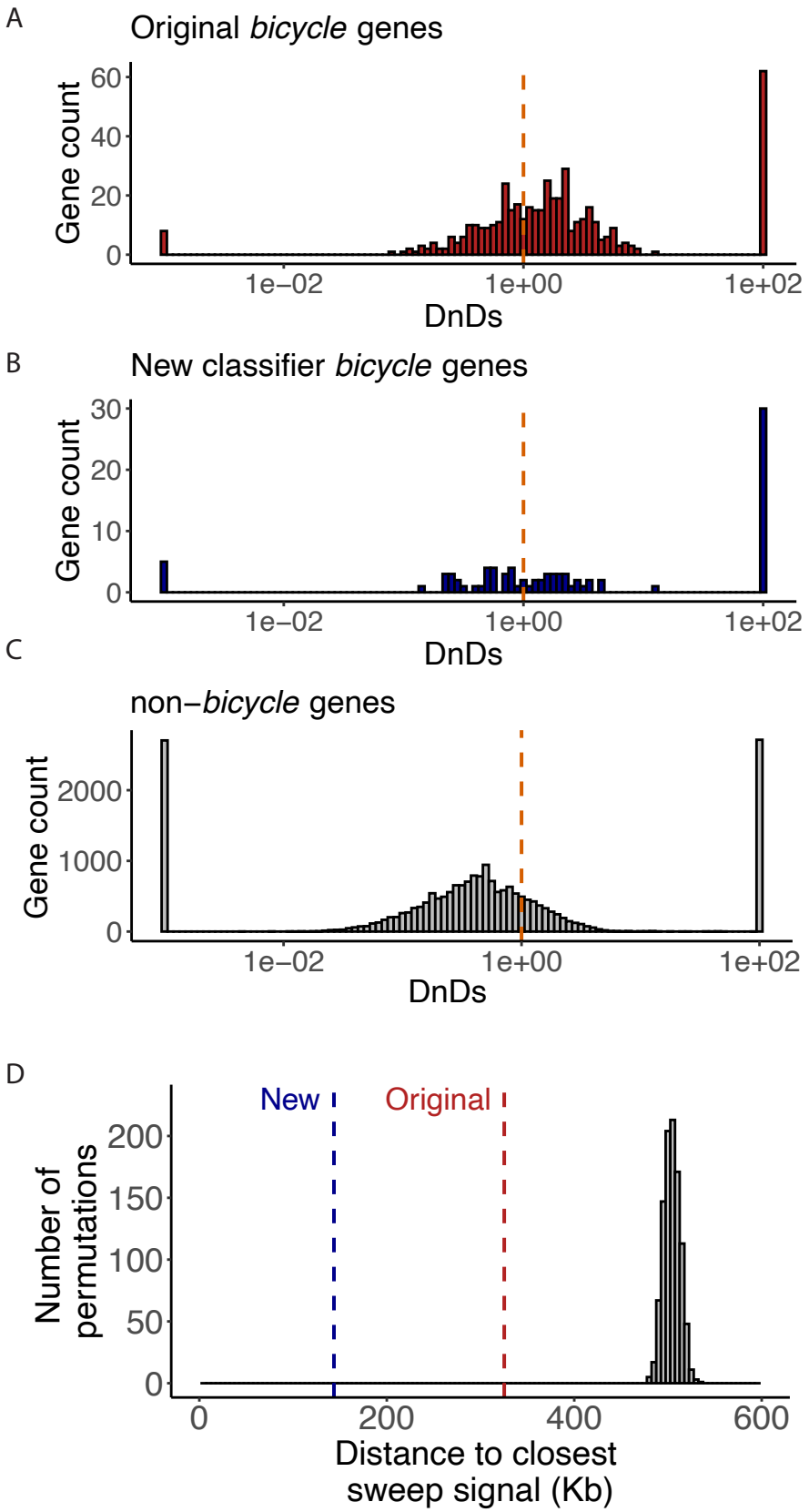


Figure 3

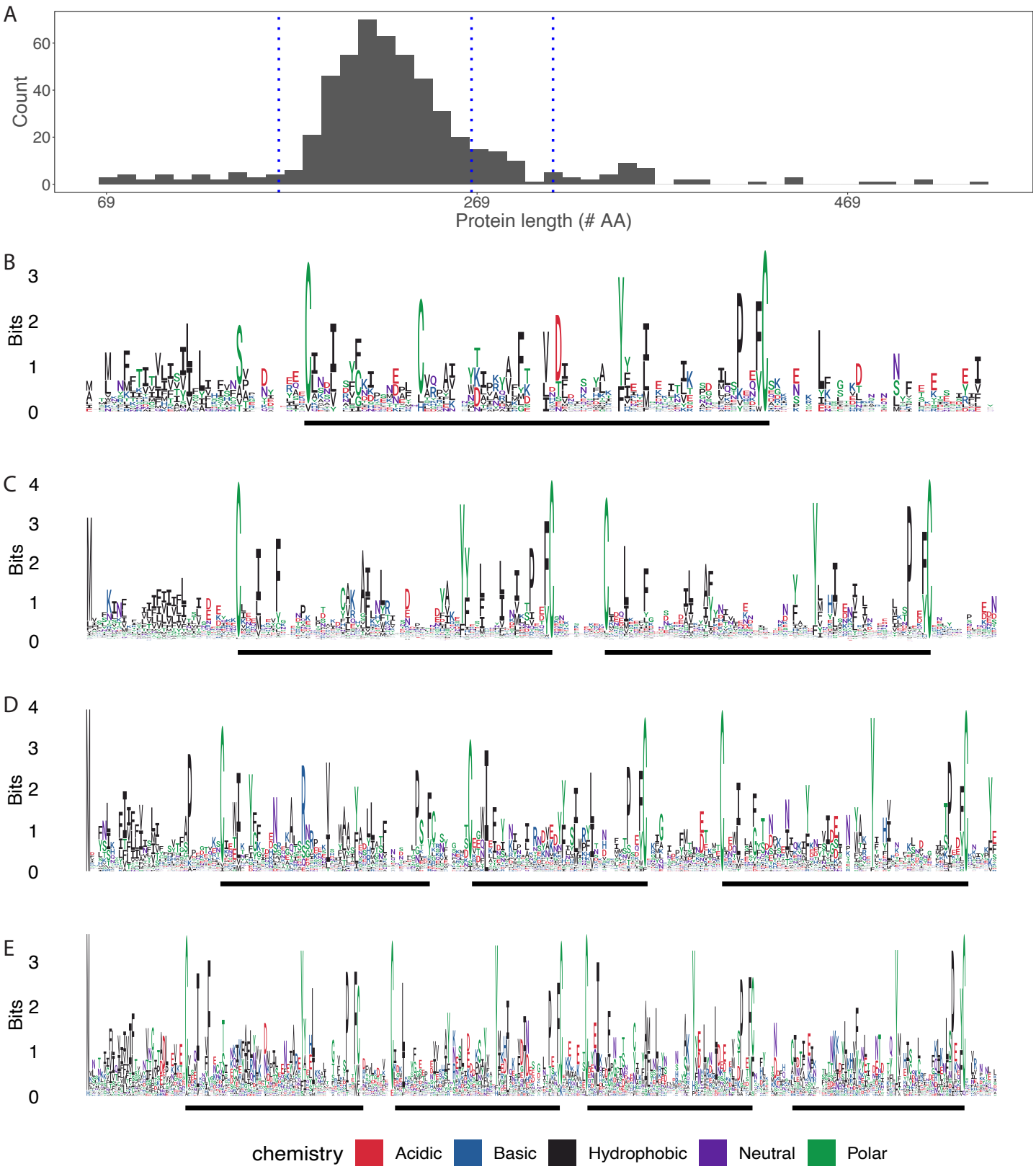

A
Count
60
40
20
0
69
269
469
Protein length (# AA)
B
Bits
3
2
1
0
C
Bits
4
3
2
1
0
D
Bits
4
3
2
1
0
E
Bits
3
2
1
0
chemistry
Acidic
Basic
Hydrophobic
Neutral
Polar

Figure 3 Supplement 1

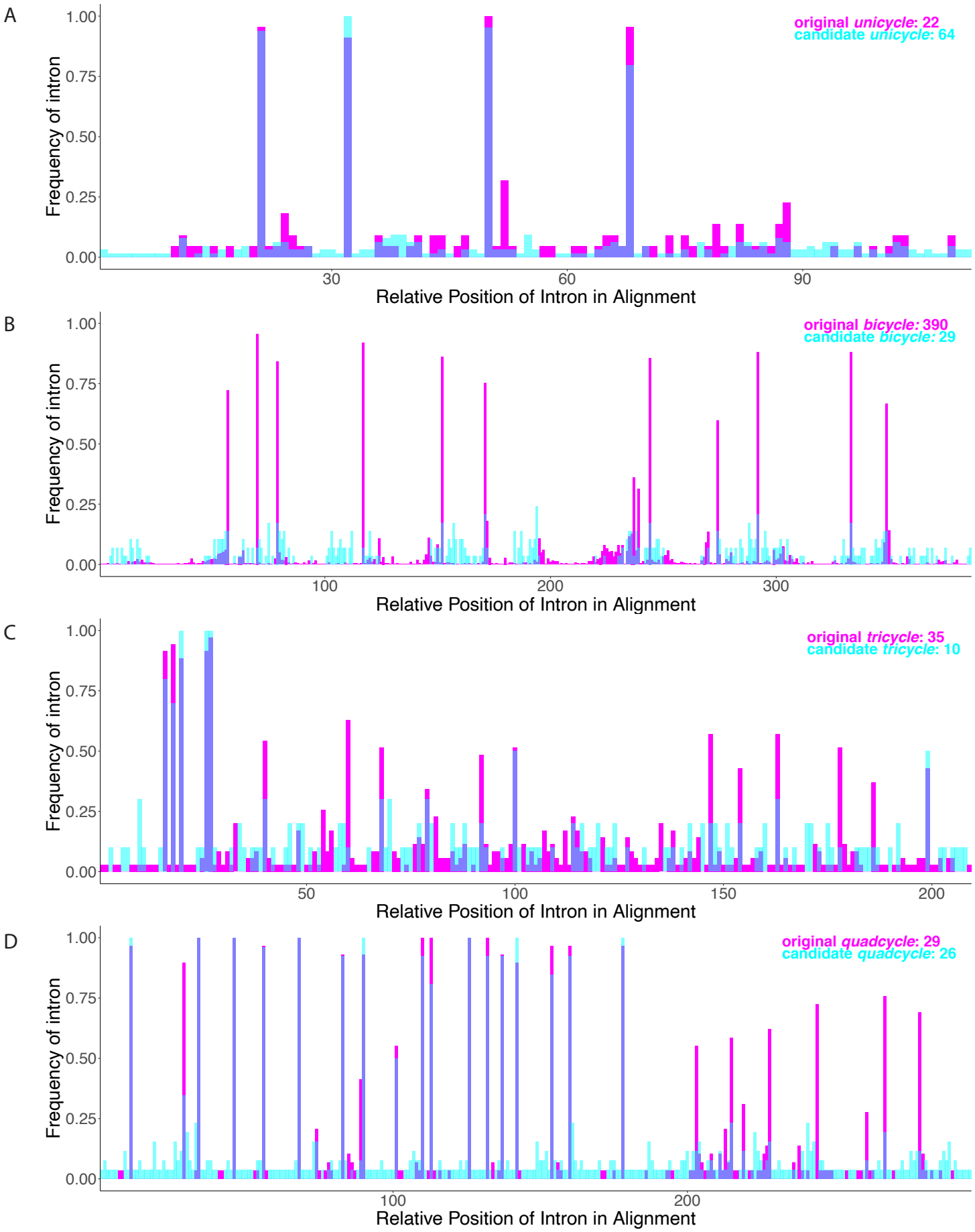


Figure 4

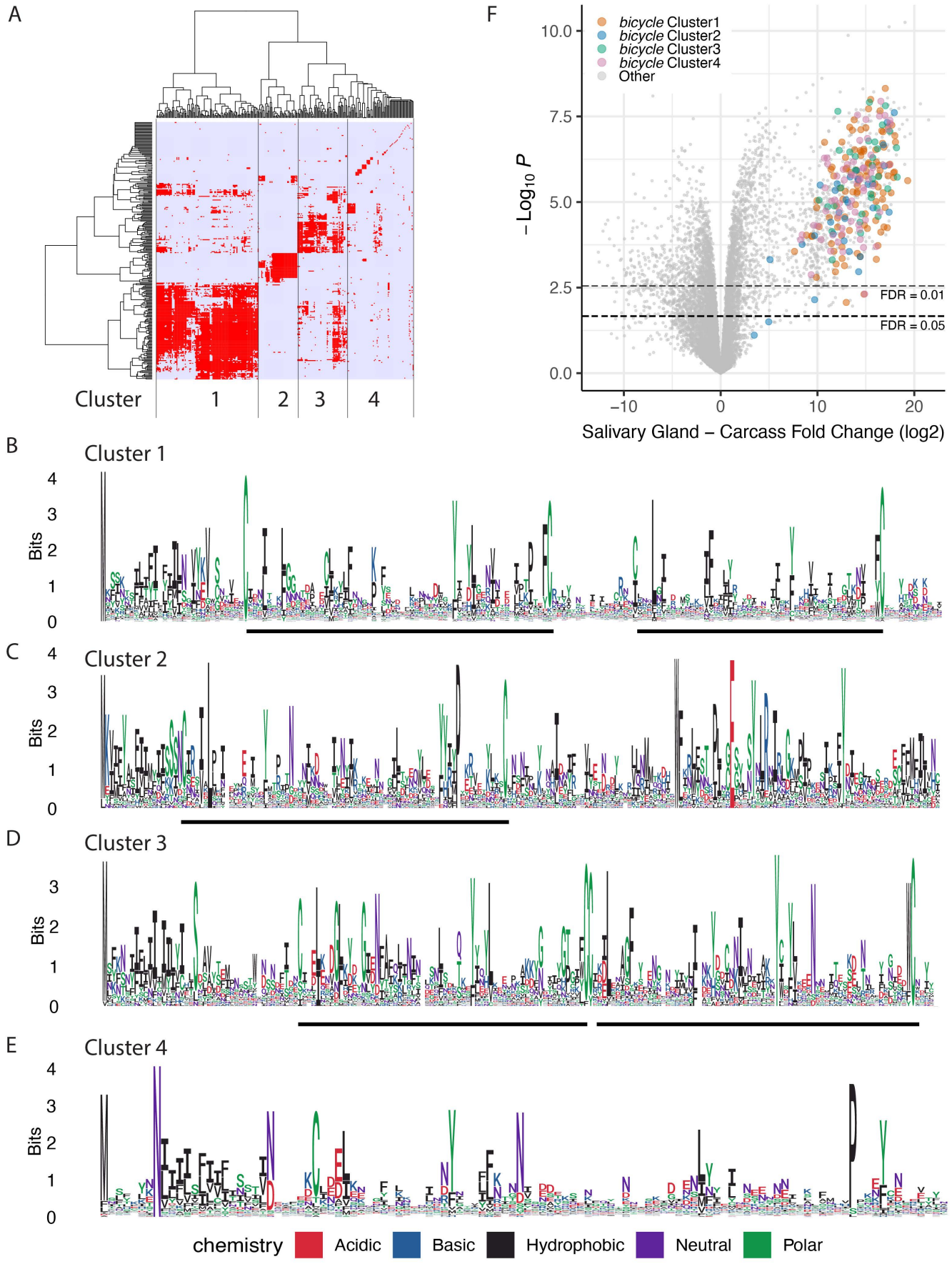

A
Cluster
1
2
3
4
F
bicycle Cluster1
bicycle Cluster2
bicycle Cluster3
bicycle Cluster4
Other
−Log10 P
FDR = 0.01
FDR = 0.05
Salivary Gland − Carcass Fold Change (log2)
B
Cluster 1
C
Cluster 2
D
Cluster 3
E
Cluster 4
Bits
chemistry
Acidic
Basic
Hydrophobic
Neutral
Polar

Figure 4 Supplement 1

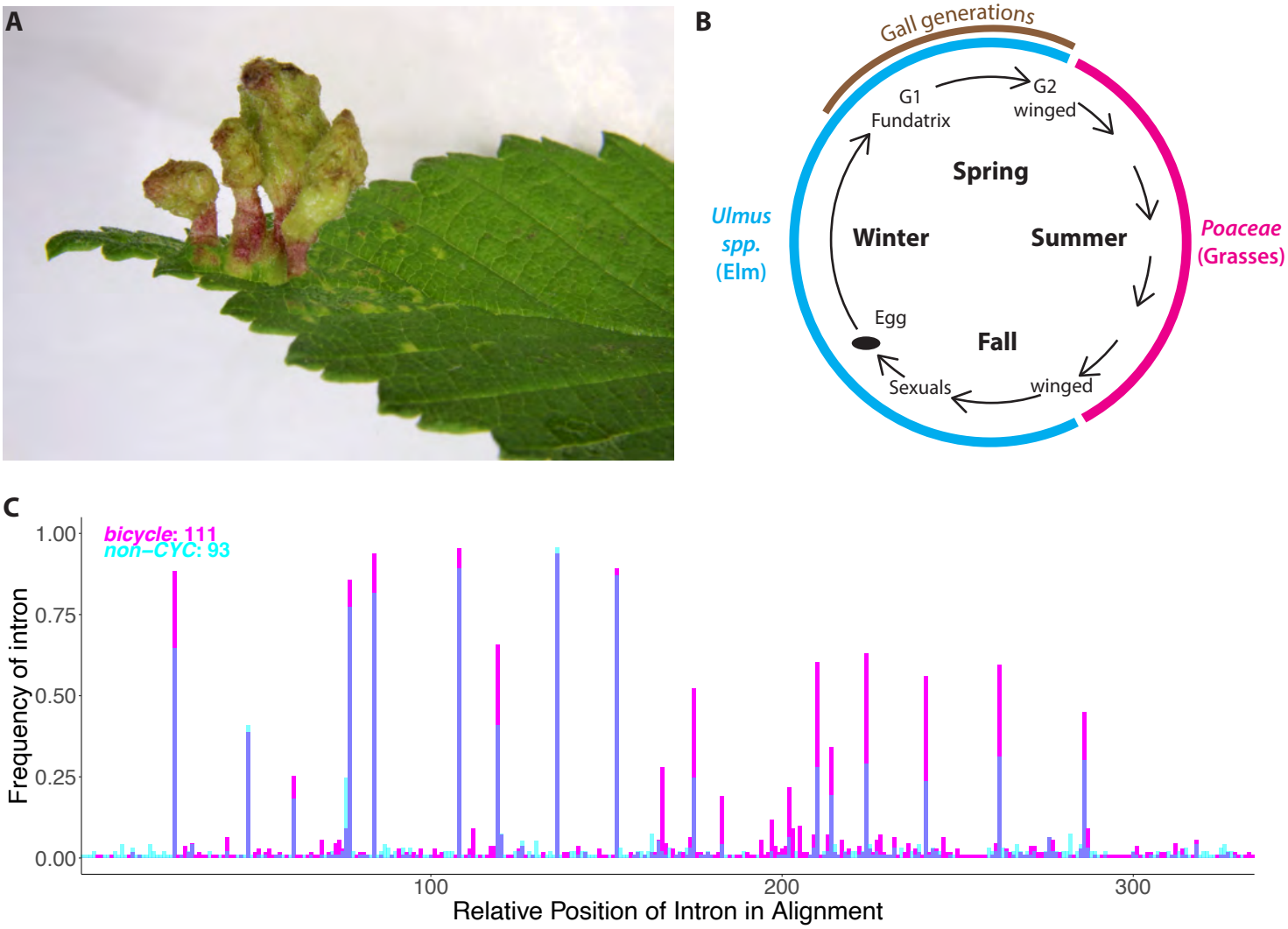
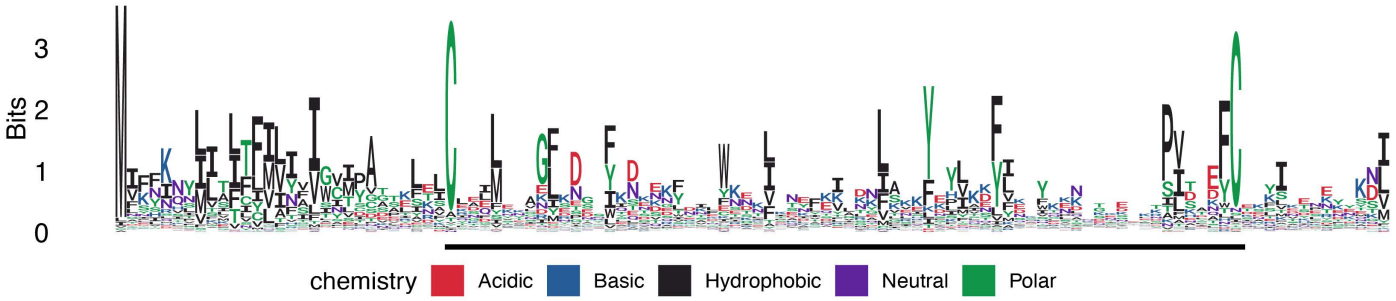


Figure 5

A


Bits
3
2
1
0
chemistry
Acidic
Basic
Hydrophobic
Neutral
Polar

B

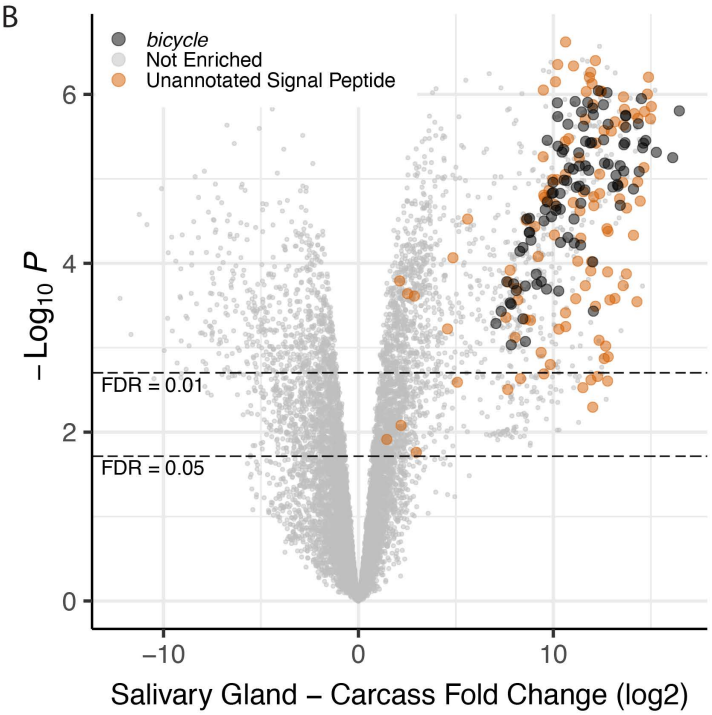

bicycle
Not Enriched
Unannotated Signal Peptide
FDR = 0.01
FDR = 0.05
$-\mathrm{Log}_{10}\,P$
6
4
2
0
−10
0
10
Salivary Gland − Carcass Fold Change (log2)

Figure 5 Supplement 1

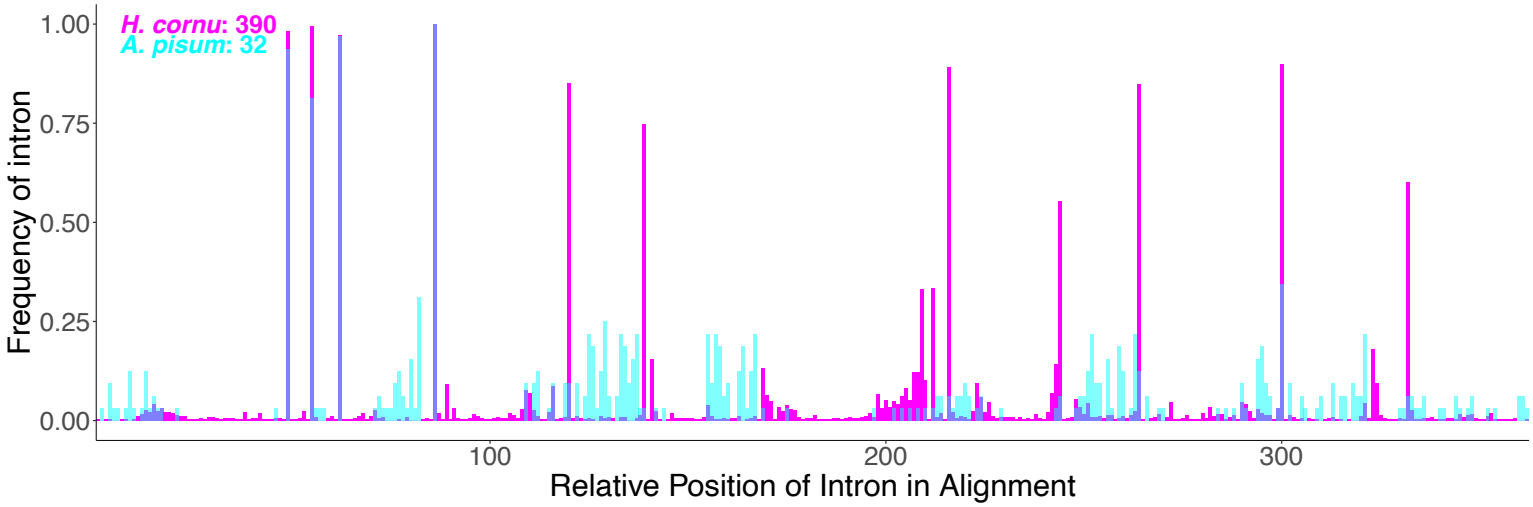
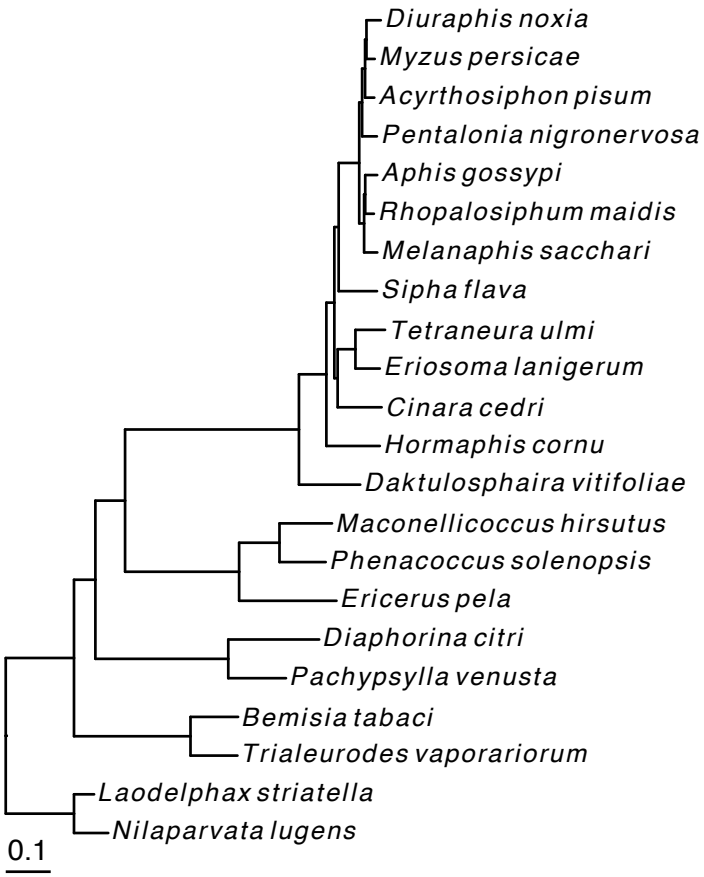


Figure 6



		BLAST	HMMER	Classifier
Diuraphis noxia	Aphidoidae	0	0	15
Myzus persicae	Aphidoidae	2	2	130
Acyrthosiphon pisum	Aphidoidae	1	5	159
Pentalonia nigronervosa	Aphidoidae	1	0	64
Aphis gossypi	Aphidoidae	3	2	29
Rhopalosiphum maidis	Aphidoidae	2	3	65
Melanaphis sacchari	Aphidoidae	1	0	27
Sipha flava	Aphidoidae	0	1	3
Tetraneura ulmi	Aphidoidae	1	53	279
Eriosoma lanigerum	Aphidoidae	0	3	27
Cinara cedri	Aphidoidae	4	15	70
Hormaphis cornu	Aphidoidae	510	458	653
Daktulosphaira vitifoliae	Phylloxeroidae	0	0	69
Maconellicoccus hirsutus	Coccoidea	0	0	74
Phenacoccus solenopsis	Coccoidea	0	0	26
Ericerus pela	Coccoidea	0	0	2
Diaphorina citri	Psylloidea	0	0	0
Pachypsylla venusta	Psylloidea	0	0	0
Bemisia tabaci	Aleyrodoidea	0	0	0
Trialeurodes vaporariorum	Aleyrodoidea	0	0	0
Laodelphax striatella	Fulgoroidea	0	0	0
Nilaparvata lugens	Fulgoroidea	0	0	0

Figure 6 Supplement 1

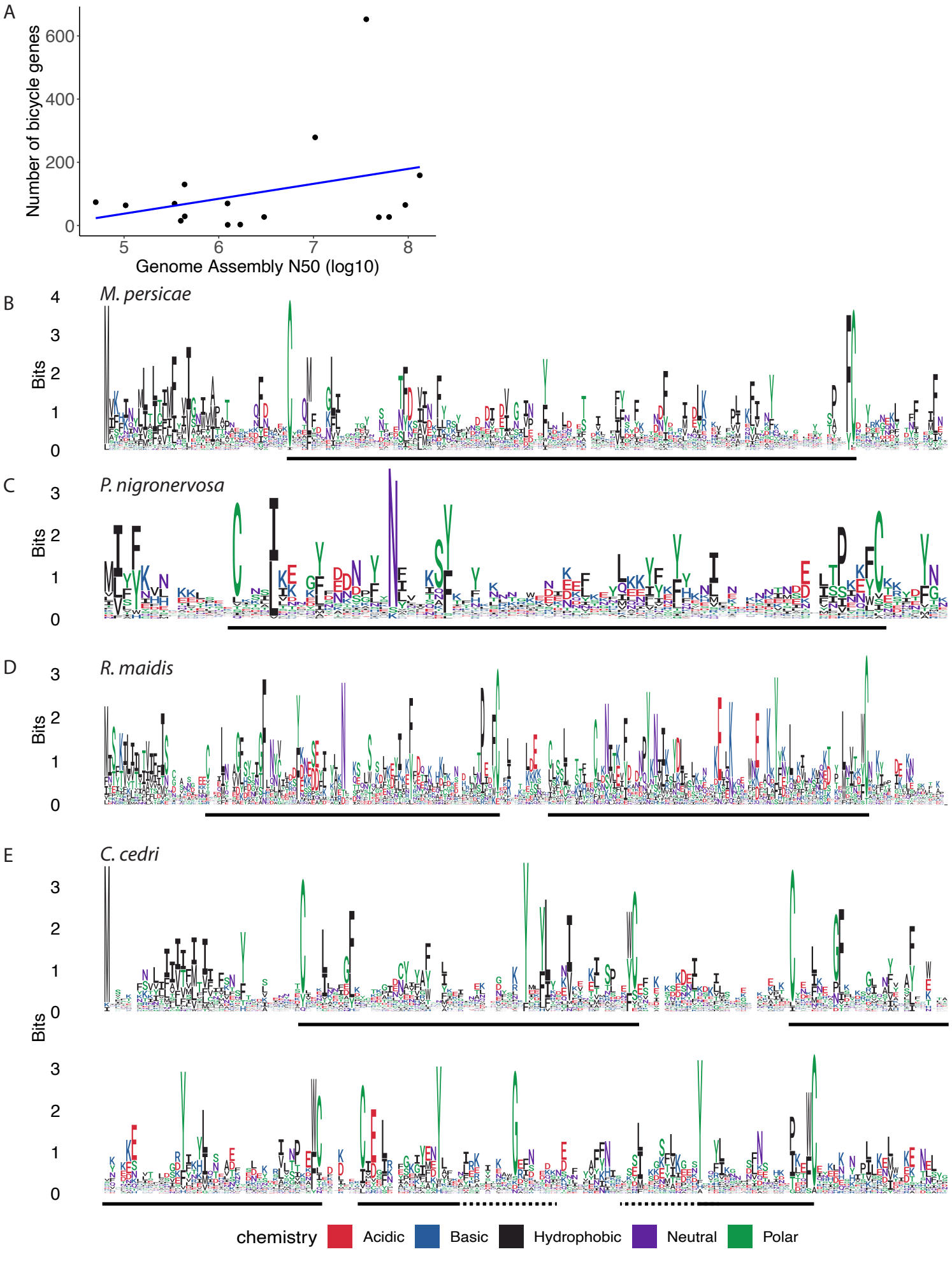


Figure 6 Supplement 2

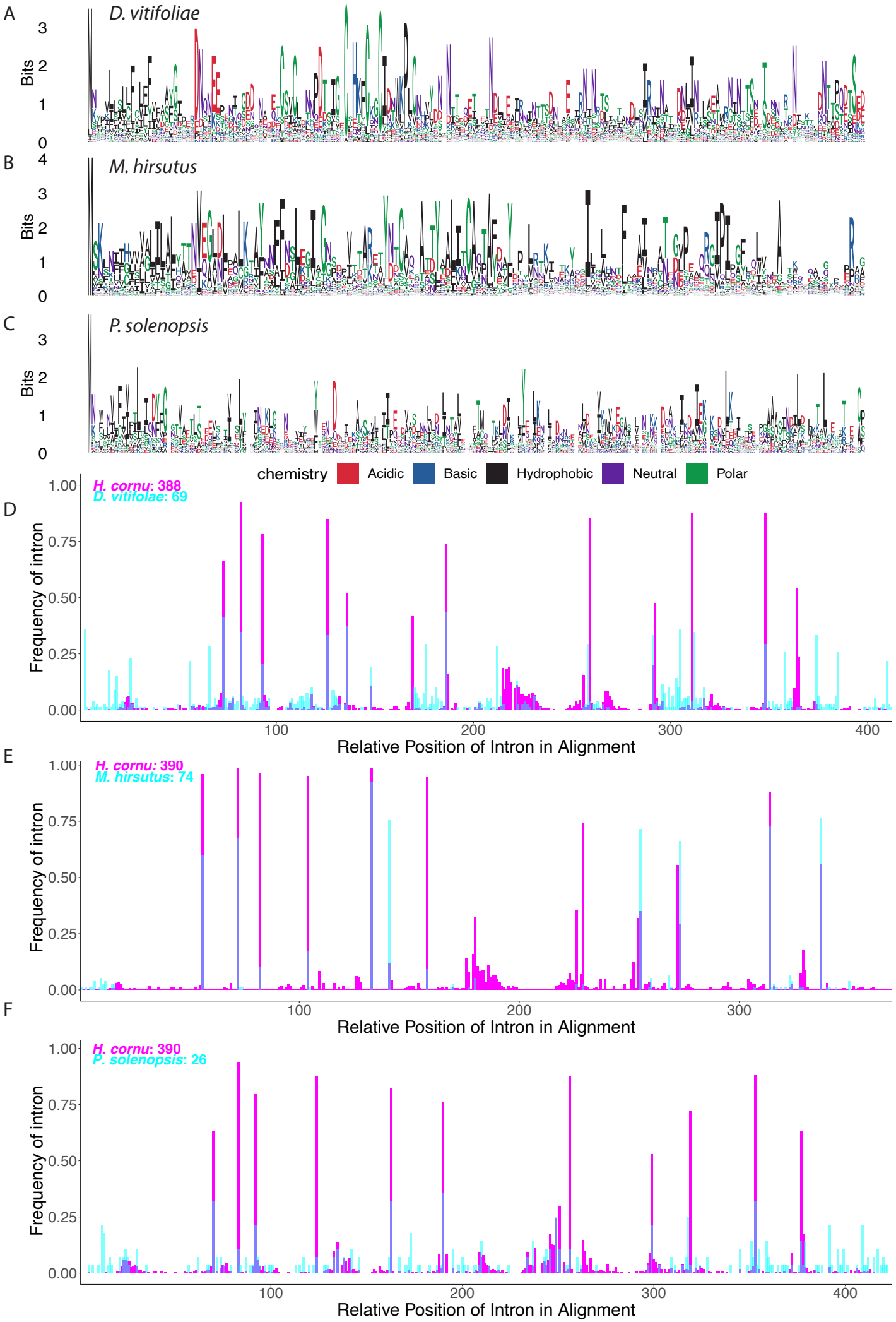
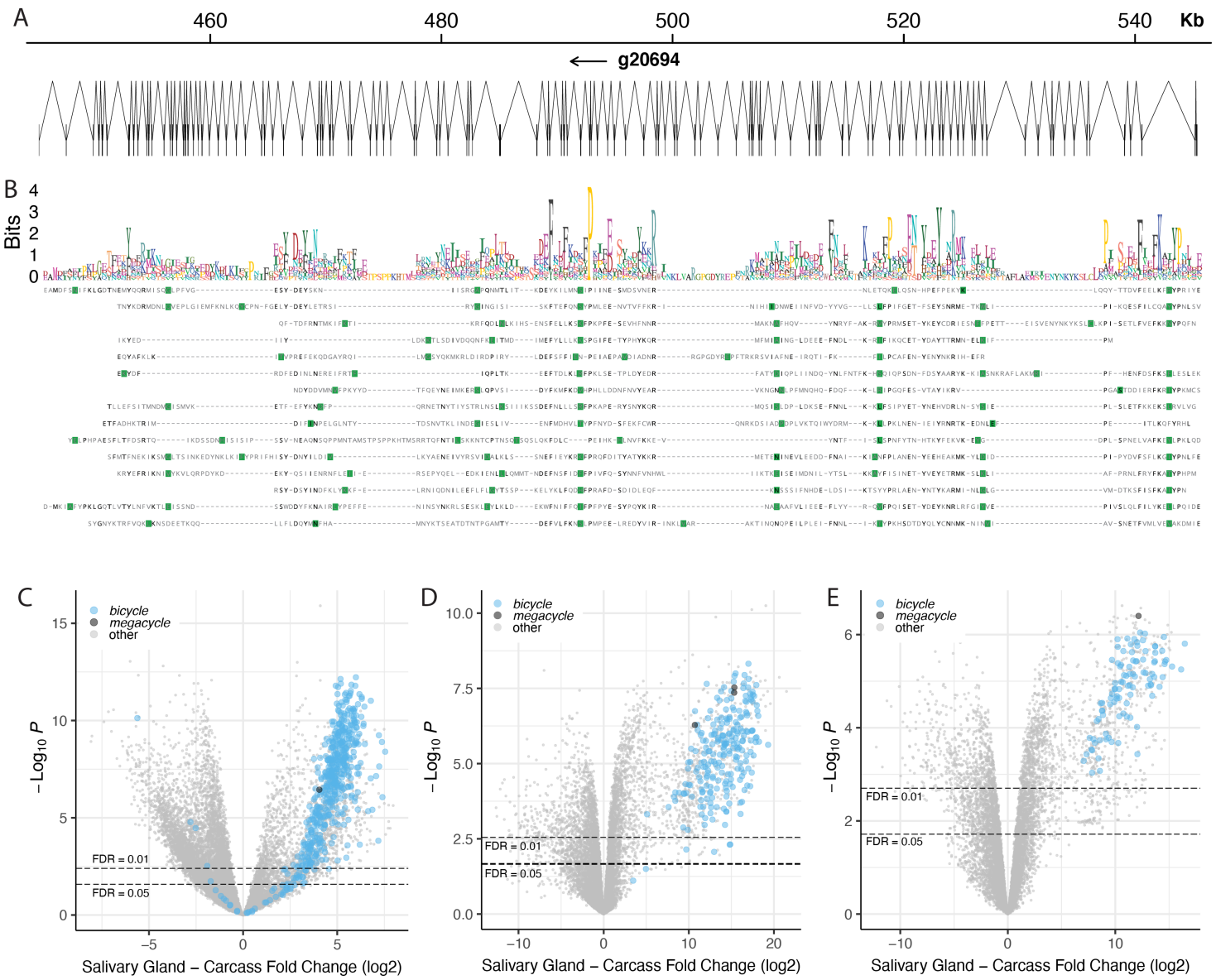


Figure 7


A
460
480
500
520
540
Kb
g20694
B
Bits
C
bicycle
megacycle
other
FDR = 0.01
FDR = 0.05
−Log10 P
Salivary Gland – Carcass Fold Change (log2)
D
bicycle
megacycle
other
FDR = 0.01
FDR = 0.05
−Log10 P
Salivary Gland – Carcass Fold Change (log2)
E
bicycle
megacycle
other
FDR = 0.01
FDR = 0.05
−Log10 P
Salivary Gland – Carcass Fold Change (log2)

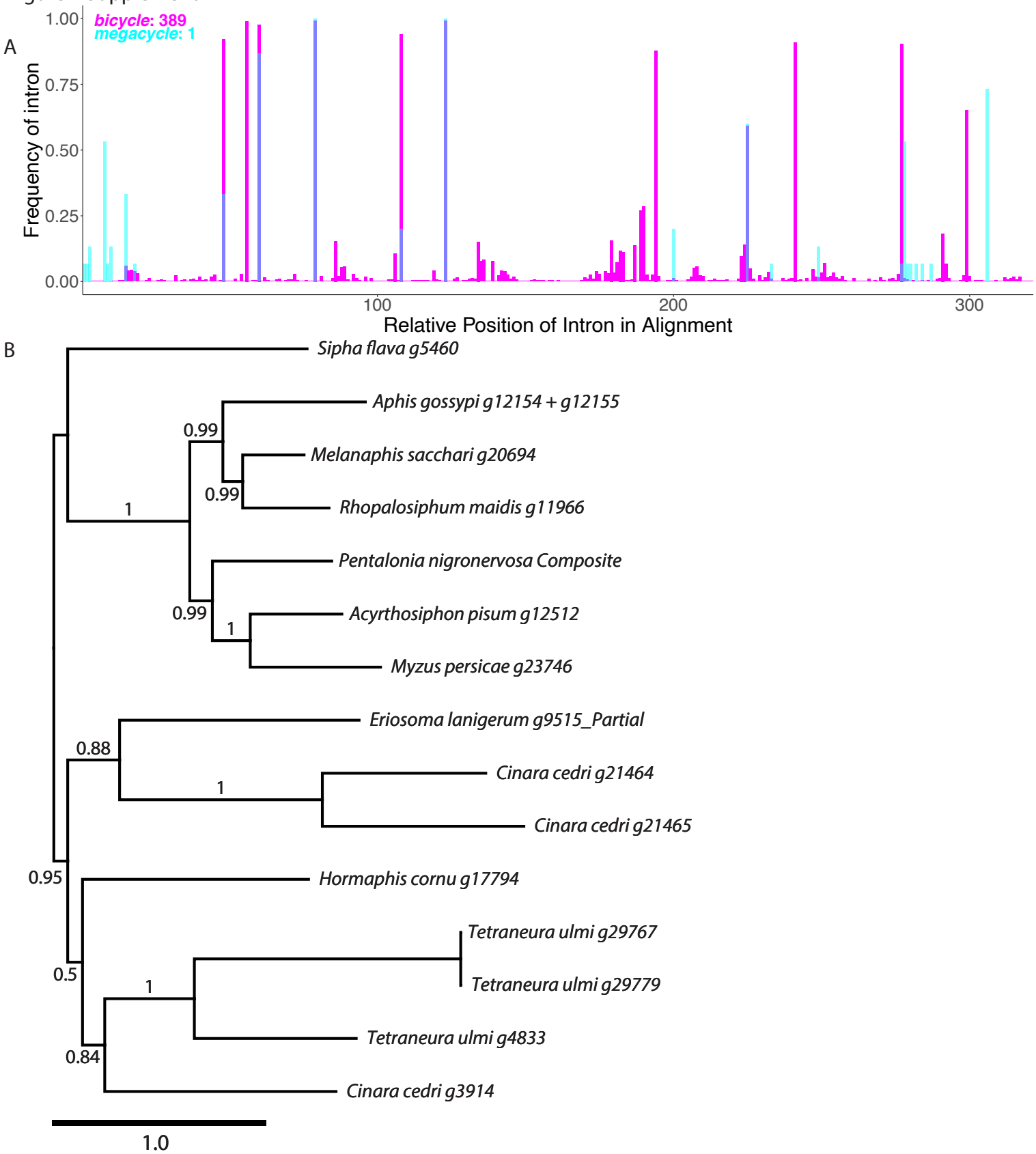



Figure 7 Supplement 1