Title: A novel family of secreted proteins linked to plant gall development

Authors: Aishwarya Korgaonkar[1], Clair Han[1], Andrew L. Lemire[1], Igor Siwanowicz[1], Djawed Bennouna[2], Rachel Kopec[2,3], Peter Andolfatto[4], Shuji Shigenobu[5,6,7], David L. Stern[1,*]

Affiliations:

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

[2] Dept. of Human Sciences, Division of Human Nutrition, The Ohio State University, 262G Campbell Hall, 1787 Neil Ave., Columbus, OH 43210

[3] Ohio State University's Foods for Health Discovery Theme, The Ohio State University, 262G Campbell Hall, 1787 Neil Ave., Columbus, OH 43210

[4] Department of Biology, Columbia University, 600 Fairchild Center, New York, NY 10027, USA

[5] Laboratory of Evolutionary Genomics, Center for the Development of New Model Organism, National Institute for Basic Biology, Okazaki, 444-8585 Japan

[6] NIBB Research Core Facilities, National Institute for Basic Biology, Okazaki, 444-8585 Japan

[7] Department of Basic Biology, School of Life Science, SOKENDAI (The Graduate University for Advanced Studies), 38 Nishigonaka, Myodaiji, Okazaki, 444-8585 Japan

[*] Corresponding author

Abstract:

In an elaborate form of inter-species exploitation, many insects hijack plant development to induce galls—novel plant 'organs' resulting from dramatic reprograming of plant cell biology. Insects induce galls by introducing unidentified molecules into plants. Here we identify an aphid gene, *determinant of gall color* (*dgc*), that is genetically associated with variation in a gall phenotype, providing the first example of an insect gene implicated in gall development. *Dgc* encodes a BICYCLE protein, a new family of secreted proteins. *Bicycle* genes are strongly expressed specifically in galling aphid salivary glands, suggesting they control many aspects of gall development. *Bicycle* genes have experienced unusually strong diversifying selection, consistent with their potential role controlling gall development in a molecular arms race between aphids and their host plants.

One Sentence Summary:

Aphid *bicycle* genes, which encode diverse secreted proteins, contribute to plant gall development.

Main Text:

Organisms often exploit individuals of other species, for example through predation or parasitism. Parasites sometimes utilize molecular weapons against hosts, which themselves respond with molecular defenses, and the genes that encode or synthesize these molecular weapons may evolve rapidly in a continuous 'arms race' (Obbard et al., 2006; Papkou et al., 2019; Paterson et al., 2010). Some of the most elaborate molecular defenses—such as adaptive immune systems, restriction modification systems, and CRISPR—have resulted from such host-parasite conflicts. In many less well-studied systems, parasites not only extract nutrients from their hosts but they also alter host behavior, physiology, or development to the parasite's advantage (Heil, 2016). Insect galls represent one of the most extreme forms of such inter-species manipulation.

Insect-induced galls are intricately patterned homes that provide insects with protection from predators (Mani, 1964). Galls are also resource sinks, drawing nutrients from distant plant organs, and providing insects with abundant food (Larson and Whitham, 1991). Insect galls are atypical plant growths that do not result simply from unpatterned cellular over-proliferation, as observed for microbial galls like the crown gall induced by *Agrobacterium tumefaciens*. Instead, each galling insect species appears to induce a distinctive gall, even when related insect species attack the same plant, implying that each species provides unique instructions to re-program latent plant developmental networks (Cook and Gullan, 2008; Crespi and Worobey, 1998; Dodson, 1991; Leatherdale, 1955; Martin, 1938; Parr, 1940; Plumb, 1953; Stern, 1995; Stone and Cook, 1998).

At least some gall-inducing insects produce phytohormones (Dorchin et al., 2009; McCalla et al., 1962; Suzuki et al., 2014; Tanaka et al., 2013; Tooker and Helms, 2014; Yamaguchi et al., 2012), although it is not yet clear whether insects introduce these hormones into plants to support gall development. However, injection of phytohormones alone probably cannot generate the large diversity of species-specific insect galls. In addition, galling insects induce plant transcriptional changes independent of phytohormone activity (Bailey et al., 2015; Nabity et al., 2013; Schultz et al., 2019; Shih et al., 2018; Takeda et al., 2019). Thus, given the complex cellular changes required for gall development, insects probably introduce many molecules to induce galls.

Other than the potential role of phytohormones in promoting gall growth, none of the mechanisms that generate specific patterns of gall development have yet been identified. For example, while many herbivorous insects introduce effector molecules into plants to influence plant physiology (Elzinga and Jander, 2013; Hogenhout and Bos, 2011; Kaloshian and Walling, 2016; Stuart, 2015), there is no evidence that any previously described effectors contribute to gall development. Since there are currently no galling insect model systems that would facilitate a genetic approach to this problem, we turned to natural variation to identify insect genes that contribute to gall development.

We studied the aphid, *Hormaphis cornu*, which induces galls on the leaves of witch hazel, *Hamamelis virginiana*, in the Eastern United States (Fig. 1A-F). In early spring, each *H. cornu* gall foundress (fundatrix) probes an expanding leaf with her very thin mouthparts (stylets) (Fig. 1A, B, Video S1). She pierces individual mesophyll cells with her stylets (Fig. 1G and H) and injects fluid directly into each cell and perhaps into inter-cellular spaces (Lewis and

82 Walton, 1947, 1958). We found that plant cells near injection sites over-proliferate through
83 periclinal cell divisions (Figure 1H), a pattern of cytokinesis not otherwise observed in leaves at
84 this stage of development (Figure 1I), and cells elongate, causing thickening and apical
85 expansion of leaf tissue (Fig. 1D-G).

86 After several days, the basal side of the gall encloses the fundatrix and the gall
87 continues to grow apically and laterally, providing the fundatrix and her offspring with
88 protection and abundant food. After several weeks, the basal side of the gall opens to allow
89 aphids to remove excreta (honeydew) and molted nymphal skins from the gall and, eventually,
90 to allow winged migrants to depart. Continued gall growth requires the constant presence of
91 the fundatrix and gall tissue dies in her absence (Lewis and Walton, 1958; Rehill and Schultz,
92 2001), suggesting that the fundatrix must continuously inject material to overcome plant
93 defenses.

94

95 **A natural gall color polymorphism is linked to regulatory variation in a novel aphid gene,**
96 ***determinant of gall color***

97

98 Populations of *H. cornu* include approximately 4% red galls and 96% green galls (Fig.
99 1F). We inferred that this gall color polymorphism results from differences among aphids,
100 rather than from differences associated with leaves or the location of galls on leaves, because
101 red and green galls are located randomly on leaves and often adjacent to each other on a
102 single leaf (Fig. 1F). We performed a genome-wide association study (GWAS) on fundatrices
103 isolated from 43 green galls and 47 red galls by resequencing their genomes to approximately

104 3X coverage. There is no evidence for genome-wide differentiation of samples from red and
105 green galls, suggesting that individuals making red and green galls were sampled from a single
106 interbreeding population (Fig. S1A-D). We identified a cluster of SNPs near 40.5 Mbp on
107 Chromosome 1 that were strongly associated with gall color (Fig. 2A). We re-sequenced
108 approximately 800 kbp flanking these SNPs to approximately 60X coverage and identified 11
109 single-nucleotide polymorphisms (SNPs) within the introns and upstream of gene
110 *Horco_16073* that were strongly associated with gall color (Fig. 2B-D). There is no evidence
111 that large scale chromosomal aberrations are associated with gall color (Supplementary Text).
112 Since GWAS can sometimes produce spurious associations, we performed an
113 independent replication study and found that all 11 SNPs were highly significantly associated
114 with gall color in fundatrices isolated from 435 green and 431 red galls (LOD = 191 – 236; Fig.
115 2E). All fundatrices from green galls were homozygous for the ancestral allele at 9 or more of
116 these SNPs (Fig. 2E). In contrast, 98% of fundatrices from red galls were heterozygous or
117 homozygous for derived alleles at 9 or more SNPs (Fig. 2E). This pattern suggests that alleles
118 contributing to red gall color are genetically dominant to alleles that generate green galls. A
119 small number of fundatrices making red galls were homozygous for ancestral alleles at these
120 SNPs and likely carry genetic variants elsewhere in the genome that confer red color to galls
121 (Supplementary Text).
122 Based on these genetic associations and further evidence presented below, we
123 assigned the name *determinant of gall color* (*dgc*) to *Horco_16073*. *Dgc* encodes a predicted
124 protein of 23 kDa with an N-terminal secretion signal sequence (Fig. S4). The putatively

125 secreted portion of the protein shares no sequence homology with any previously reported
126 proteins.

127 Most SNPs associated with green or red galls were found in one of two predominant
128 haplotypes (Fig. 2E) and exhibited strong linkage disequilibrium (LD) (Fig. S5). LD can result
129 from suppressed recombination. However, these 11 SNPs are in linkage equilibrium with many
130 other intervening and flanking SNPs (Fig. S5, S6). Also, multiple observed genotypes are
131 consistent with recombination between these 11 SNPs (Fig. 2E) and we found no evidence for
132 chromosomal aberrations that could suppress recombination (Fig. S1E-I; Supplementary Text).
133 Thus, LD among the 11 *dgc* SNPs associated with gall color cannot be explained by suppressed
134 recombination. It is more likely that the non-random association of the 11 dgc^{Red} SNPs has
135 been maintained by natural selection, suggesting that the combined action of all 11 SNPs may
136 have a stronger effect on gene function than any single SNP alone.

137

138 **Regulatory variants at *dgc* dominantly silence *dgc* expression**

139

140 Since all 11 *dgc* polymorphisms associated with gall color occur outside of *dgc* exons
141 (Fig. 2D), we tested whether these polymorphisms influence expression of *dgc* or of any other
142 genes in the genome. We first determined that *dgc* is expressed highly and specifically in
143 fundatrix salivary glands and lowly or not at all in other tissues or other life cycle stages (Fig.
144 S7A). We then performed RNAseq on salivary glands from fundatrices with dgc^{Green} /dgc^{Green} or
145 dgc^{Red}/dgc^{Green} genotypes. *Dgc* is the only strongly differentially expressed gene between
146 these genotypes (Figure 3A and S7B). Since dgc^{Red} alleles appeared to be dominant to the

147 *dgc*^{Green} alleles for gall color, we expected that *dgc* transcripts would be upregulated in animals
148 with *dgc*^{Red} alleles. In contrast, *dgc* transcripts were almost absent in fundatrices carrying
149 *dgc*^{Red} alleles (Fig. 3B). That is, red galls are associated with strongly reduced *dgc* expression in
150 fundatrix salivary glands.

151 *Dgc* expression is reduced approximately 20-fold in fundatrix salivary glands with
152 *dgc*^{Red}/*dgc*^{Green} (27 ± 22.6 CPM, mean ± SD) versus *dgc*^{Green}/*dgc*^{Green} genotypes (536 ± 352.3
153 CPM, mean ± SD). Thus, *dgc*^{Red} alleles probably downregulate both the *dgc*^{Red} and *dgc*^{Green}
154 alleles in heterozygotes. By examining exonic SNPs in *dgc*, we confirmed that the *dgc*
155 transcript is strongly downregulated from both *dgc*^{Red} and *dgc*^{Green} alleles in heterozygotes
156 (Fig. S7D). We observed no systematic transcriptional changes in neighboring genes (Figure
157 3B), most of which are *dgc* paralogs, indicating that *dgc*^{Red} alleles exhibit a perhaps unique
158 example of locus-specific repressive transvection (Duncan, 2002).

159

160 **High levels of *dgc* transcription are associated with downregulation specifically of plant**
161 **anthocyanin genes and two anthocyanins**

162

163 Since red galls are associated with strong differential expression of only *dgc*, we
164 wondered how the plant responds to changes in this single putative signal. To examine this
165 question, we performed whole-genome differential expression on plant mRNA isolated from
166 galls induced by aphids with *dgc*^{Red}/*dgc*^{Green} versus *dgc*^{Green}/*dgc*^{Green} genotypes (Supplementary
167 Text). Only eight plant genes were differentially expressed between green and red galls and all

168 eight genes were downregulated in green galls (Fig. 3C-D). That is, high levels of *dgc* are
169 associated with downregulation of eight plant genes in galls.

170 The seven most strongly downregulated plant genes all encode enzymes of the
171 anthocyanin biosynthetic pathway (Fig. 3E). One gene encodes an enzyme (ACCA) that
172 irreversibly converts acetyl-CoA to malonyl CoA, a biosynthetic precursor of multiple
173 anthocyanins. Two genes encode anthocyanidin 3-0-glucosyltransferases (UFGT and
174 UGT75C1), which glycosylate unstable anthocyanidins to allow their accumulation (Springob et
175 al., 2003). Two genes encode flavonoid 3'-5' methyltransferases (FAOMT-1, FAOMT-2), which
176 methylate anthocyanin derivatives (Hugueney et al., 2009). Finally, two genes encode phi class
177 glutathione S-transferases (GSTF11, GSTF12), which conjugate glutathione to anthocyanins,
178 facilitating anthocyanin transport and stable accumulation in vacuoles (Marrs et al., 1995).

179 Six of these enzymes are required for final steps of anthocyanin production and
180 deposition (Fig. 3E) and their upregulation in red galls may account for the accumulation of
181 pigments in red galls. To test this hypothesis, we extracted and analyzed pigments from galls
182 (Fig. 3F) and identified high levels of two pigments only in red galls (Fig. 3G), the anthocyanins
183 malvidin-3,5-diglucoside and peonidin-3,5-diglucoside (Fig. 3G-H, S8). Thus, the pigments
184 present in red galls are products of enzymes in the anthocyanin biosynthetic pathway, such as
185 those that are upregulated in red galls. The two abundant anthocyanins are produced from
186 distinct intermediate precursor molecules (Fig. 3E), three of which were also detected in red
187 galls (Fig. 3H, S8), and synthesis of these two anthocyanins likely requires activity of different
188 methyltransferases and glucosyltransferases. The three pairs of glucosyltransferases,

189 methyltransferases, and glutathione transferases upregulated in red galls may provide the
190 specific activities required for production of these two anthocyanins.
191 Taken together, these observations suggest that *dgc* represses transcription of seven
192 anthocyanin biosynthetic enzymes. It is not clear how *dgc* induces specific transcriptional
193 changes in seven plant genes; it may act by altering activity of an upstream regulator of these
194 plant genes.
195
196 **Aphids induce widespread transcriptomic changes in galls**
197
198 The gall color polymorphism reflects only a small part of the gall phenotype. To
199 estimate how many plant genes are differentially expressed during gall development, we
200 performed differential expression analysis of plant genes in galls versus the surrounding leaf
201 tissue. Approximately 31% of plant genes were upregulated and 34% were downregulated at
202 FDR = 0.05 in galls versus leaf tissue (Figure S9A; 27% up and 29% down in gall at FDR = 0.01).
203 Results of gene ontology analysis of up and down-regulated genes is consistent with the
204 extensive growth of gall tissue and down regulation of chloroplasts seen in galls (Fig. S9B;
205 Supplementary Text).
206 Thus, approximately 15,000 plant genes are differentially expressed in galls,
207 representing a system-wide re-programming of plant cell biology. If other aphid effector
208 molecules act in ways similar to *dgc*, which is associated with differential expression of only
209 eight plant genes, then gall development may require injection of hundreds or thousands of
210 effector molecules.

211

212 ***Dgc* is the founding member of a large class of novel *bicycle* genes expressed specifically in**

213 **the salivary glands of gall-inducing aphids**

214

215 To identify additional proteins that aphids may inject into plants to contribute to gall

216 development, we exploited the fact that only some individuals in the complex life cycle of *H.*

217 *cornu* induce galls (Figure 4A). We hypothesized that expression of genes required for gall

218 formation should be enriched specifically in the salivary glands of the generations that induce

219 galls: the fundatrix (G1) and possibly also her immediate offspring (G2). We identified 3,048

220 genes upregulated in fundatrix salivary glands versus the fundatrix body (Fig. S10A) and 3,427

221 genes upregulated in salivary glands of fundatrices, which induce galls, versus sexuals, which

222 do not induce galls although they feed on the same host plant (Fig. S10B). Intersection of

223 these gene sets identified 1,489 genes specifically enriched in the salivary glands of

224 fundatrices (Figure 4B).

225 Half of these genes (744) were homologous to previously identified genes, many of

226 which had functional annotations. Gene Ontology analysis of the "annotated" genes suggests

227 that they contribute mostly to the demands for high levels of protein secretion in fundatrix

228 salivary glands (Fig. S10C). Most do not encode proteins with secretion signals (671; Fig. S10D)

229 and are thus unlikely to be injected into plants. We searched for homologs of genes that have

230 been proposed as candidate gall-effector genes in other insects but found no evidence that

231 these classes of genes contribute to aphid gall development (Supplementary Text). We

232 therefore focused on the remaining 738 unannotated genes, which included 459 genes
233 encoding proteins with predicted secretion signals (Fig. S10D).
234 Hierarchical clustering of the unannotated genes by sequence similarity identified one
235 large (476 genes) and one small (43 genes) cluster of related genes, and 222 genes sharing few
236 or no homologs (Fig. 4C). Genes in both the large and small clusters encode proteins with N-
237 terminal secretion signals, as expected for effector proteins that might be injected into plants.
238 The small cluster encodes a divergent set of proteins containing several conserved cysteines
239 (C) and a well conserved tryptophan (W) and glycine (G) (Fig. S11A).
240 Proteins encoded by the large cluster display conservation mainly of a pair of widely
241 spaced cysteine-tyrosine-cysteine (CYC) motifs and spacing between the C, Y, and C residues of
242 each motif is not well conserved (Fig. 4F). This pair of CYC motifs led us to name these *bicycle*
243 (bi-CYC-like) genes. The *bicycle* genes were the most strongly upregulated class of genes in
244 fundatrix salivary glands (Fig. 4D, Fig. S11B-F) and were expressed specifically in the salivary
245 glands of the two generations associated with galls (G1 and G2) (Fig. 4E). Many *bicycle* genes
246 are found in paralog clusters throughout the *H. cornu* genome (Fig. S10E and F) and each
247 *bicycle* gene contains approximately 5-25 microexons (Fig. S10F)—a large excess relative to
248 the genomic background (Fig. S10G)—interrupted by long introns (Fig. S10F).
249 We found that *dgc* shares many features with *bicycle* genes—it is strongly expressed
250 specifically in fundatrix salivary glands, and it exhibits many microexons and a pair of CYC
251 motifs (Fig. S4)—and that it is evolutionarily related to other *bicycle* genes. Thus, *dgc*, which is
252 associated with gall color, is the founding member of a diverse family of genes encoding
253 secreted proteins expressed specifically in the salivary glands of gall forming generations.

254 *Bicycle* genes are therefore good candidates to encode many of the molecules required to
255 generate the extensive transcriptional changes observed in galls.
256
257 ***bicycle* genes experienced intense diversifying selection, consistent with a potential arms**
258 **race between aphids and plants**
259
260 *Bicycle* genes are extremely diverse at the amino acid sequence level. Each protein has
261 accumulated approximately two substitutions per amino acid site since divergence from
262 paralogs (Fig. S10H). To explore whether this diversity resulted from natural selection rather
263 than genetic drift, we compared rates of non-synonymous (d_N) versus synonymous (d_S)
264 substitutions between the sister species *H. cornu* and *H. hamamelidis* in *bicycle* versus non-
265 *bicycle* genes, because d_N/d_S values greater than one provide evidence for positive selection
266 (Yang and Bielawski, 2000). A large excess of *bicycle* genes displayed d_N/d_S significantly greater
267 than 1 relative to the genomic background (Fig. S12, Table S1; P < 2.2e-16), revealing recurrent
268 adaptive amino acid substitution at many *bicycle* genes since these species diverged,
269 approximately 1 MYA.
270 We then quantified the frequency of adaptive amino acid substitution at *bicycle* genes
271 and other categories of genes overexpressed in fundatrix salivary glands by calculating α, the
272 proportion of non-synonymous substitutions fixed by positive selection (Fay et al., 2001; Smith
273 and Eyre-Walker, 2002). We estimate that for all *bicycle* genes α = 0.33 and that for the subset
274 of *bicycle* genes displaying d_N/d_S significantly greater than 1, α = 0.62 (Table S2). Other
275 categories of genes overexpressed in fundatrix salivary glands also display values of α in the

range of 0.27-0.38 (Table S2), however, *bicycle* genes display a considerably higher ratio of non-synonymous to synonymous substitutions than other categories of genes (Table S2). Most strikingly, as a fraction of protein length, bicycle genes display a considerably greater fraction of adaptive non-synonymous substitutions than other categories of genes (Fig. S12C). That is, over the past 1 MY, positive selection has resulted on average in fixation of approximately 2-3 substitutions in each *bicycle* gene, and approximately 10 substitutions in the most rapidly evolving *bicycle* genes. This represents a considerable fraction of the total number of amino acids (~200) in each protein and reveals intense selection on these *bicycle* genes, presumably for novel functions that require multiple amino acid substitutions.

The previous tests cannot detect selection on non-coding regions and do not discriminate between selection acting in the deep past or more recently. To search for recent selection in *bicycle* gene regions, we examined patterns of polymorphism and divergence within and between *H. cornu* and *H. hamamelidis*. Polymorphism was strongly reduced relative to divergence in *bicycle* gene regions compared with the genomic background (Fig. S13, S14) and patterns of reduced polymorphism were strikingly similar in both species (Fig. S13, S14). This pattern is suggestive of recent selective sweeps in *bicycle* gene regions in both species, which we tested by performing genome-wide scans for positive selection (Pavlidis et al., 2013). Genome-wide signals of selective sweeps were enriched near *bicycle* genes and multiple signals fell within *bicycle* gene clusters in both species (Fig. S14 and S15). Thus, strong selection has acted not only in the past to fix non-synonymous substitutions in *bicycle* genes, but also recently and presumably frequently on many *bicycle* loci throughout the genome.

297 In summary, we find evidence for widespread, strong, recent, and frequent positive
298 selection on *bicycle* genes. Since *bicycle* genes are likely secreted from salivary glands
299 specifically in gall-forming aphids, these observations are consistent with the hypothesis that
300 *bicycle* genes encode proteins that are intimately involved in reciprocal molecular evolution
301 between the aphid and their host plant.
302
303 **DISCUSSION**
304
305 We presented multiple lines of evidence that suggest that *bicycle* gene products
306 provide instructive molecules required for aphid gall development. The strongest evidence is
307 that eleven derived regulatory polymorphisms at *dgc* are associated with almost complete
308 silencing of *dgc* in aphid salivary glands (Fig. 3A-B, S7B) and upregulation of seven anthocyanin
309 biosynthetic genes and two red-purple anthocyanins in galls (Fig. 3C-H, S8). Gall color is simply
310 one small, but convenient, aspect of the panoply of cell biological changes required for gall
311 development. We hypothesize that the product of each *bicycle* gene has its own unique set of
312 targets in the plant and that the combined action of all *bicycle* gene products regulates many
313 aspects of gall development. Testing this hypothesis will require the development of new
314 methods to explore and manipulate this aphid-plant system.
315
316 **BICYCLE protein functions**
317

318 *Dgc* likely encodes a protein that is deposited by aphids into gall tissue, but the
319 mechanism by which this novel aphid protein specifically and dramatically downregulates
320 seven anthocyanin biosynthetic genes remains to be determined. The primary sequences of
321 DGC and other BICYCLE proteins provide few clues to their molecular mode of action. Outside
322 of the N-terminal secretion signal, BICYCLE proteins possess no similarity with previously
323 reported proteins and display no conserved domains that might guide functional studies. The
324 relatively well-conserved C-Y-C motif appears to define a pair of ~50-80 aa domains in each
325 protein and the paired cysteines may form disulfide bonds, which is commonly observed for
326 secreted proteins. Secondary structure prediction methods provide little evidence for
327 structural conservation across BICYCLE proteins and the extensive variation in the spacing
328 between conserved cysteines further suggests that BICYCLE proteins may display structural
329 heterogeneity. Identification of their molecular mode of action will require identification of
330 BICYCLE protein binding targets.
331
332 **Evolution of *bicycle* genes**
333
334 We were unable to detect any sequence homology between *bicycle* genes and
335 previously reported genes, which is one reason that it is difficult to infer the molecular
336 function of *bicycle* genes from sequence alone. It is possible that *bicycle* genes evolved *de*
337 *novo* in an ancestor of gall-forming aphids, perhaps through capture of a 5' exon encoding an
338 N-terminal signal sequence. However, if *bicycle* genes experienced strong diversifying
339 selection since their origin, perhaps in an ancestor of gall forming aphids about 280 MYA, then

340 the rate of amino-acid substitution that we detected between two closely-related species
341 would likely be sufficient to have eliminated sequence similarity that could be detected by
342 homology-detection algorithms. Identifying the evolutionary antecedents of *bicycle* genes will
343 likely require tracing their evolutionary history across genomes of related species. It may also
344 be more fruitful to use the unusual gene structure of *bicycle* genes (Fig. S10F, G) to search for
345 the antecedents of *bicycle* genes.
346

347 **References:**

348 Bailey, S., Percy, D.M., Hefer, C.A., and Cronk, Q.C.B. (2015). The transcriptional landscape of
349 insect galls: psyllid (Hemiptera) gall formation in Hawaiian Metrosideros polymorpha
350 (Myrtaceae). BMC Genomics *16*, 943–943.

351 Bilder, D., and Irvine, K.D. (2017). Taking Stock of the Drosophila Research Ecosystem. *206*,
352 1227–1236.

353 Cook, L.G., and Gullan, P.J. (2008). Insect, not plant, determines gall morphology in the
354 Apiomorpha pharetrata species-group (Hemiptera: Coccoidea). Australian Journal of
355 Entomology *47*, 51–57.

356 Crespi, B., and Worobey, M. (1998). Comparative analysis of gall morphology in Australian gall
357 thrips: The evolution of extended phenotypes. Evolution *52*, 1686–1696.

358 Dodson, G.N. (1991). Control of gall morphology: tephritid gallformers (Aciurina spp.) on
359 rabbitbrush (Chrysothamnus). Ecol. Entomol *16*, 177–181.

360 Dorchin, N., Hoffmann, J.H., Stirk, W.A., NovÁk, O., Strnad, M., and Van Staden, J. (2009).
361 Sexually dimorphic gall structures correspond to differential phytohormone contents in male
362 and female wasp larvae. Physiological Entomology *34*, 359–369.

363 Elzinga, D.A., and Jander, G. (2013). The role of protein effectors in plant-aphid interactions.
364 Current Opinion in Plant Biology *16*, 451–456.

365 Fay, J.C., Wyckoff, G.J., and Wu, C.-I. (2001). Positive and Negative Selection on the Human
366 Genome. Genetics *158*, 1227+1234.

367 Heil, M. (2016). Host Manipulation by Parasites: Cases, Patterns, and Remaining Doubts. Front.
368 Ecol. Evol. *4*.

369 Hogenhout, S.A., and Bos, J.I.B. (2011). Effector proteins that modulate plant-insect
370 interactions. Current Opinion in Plant Biology *14*, 422–428.

371 Hugueney, P., Provenzano, S., Verriès, C., Ferrandino, A., Meudec, E., Batelli, G., Merdinoglu,
372 D., Cheynier, V., Schubert, A., and Ageorges, A. (2009). A novel cation-dependent o-
373 methyltransferase involved in anthocyanin methylation in grapevine. Plant Physiology *150*,
374 2057–2070.

375 Kaloshian, I., and Walling, L.L. (2016). Hemipteran and dipteran pests: Effectors and plant host
376 immune regulators. Journal of Integrative Plant Biology *58*, 350–361.

377 Larson, K.C., and Whitham, T.G. (1991). Manipulation of food resources by a gall-forming
378 aphid: the physiology of sink-source interactions. Oecologia *88*, 15–21.

379 Leatherdale, D. (1955). Plant hyperplasia induced with a cell-free insect extract. Nature *175*,
380 553–554.

381 Lewis, I.F., and Walton, L. (1947). Initiation of the cone gall of witch hazel. Science *106*, 419–
382 420.

383 Lewis, I.F., and Walton, L. (1958). Gall-formation on Hamamelis virginiana resulting from
384 material injected by the aphid Hormaphis hamamelidis. Trans. Am. Microscop. Soc. *77*, 146–
385 200.

386 Mani, M.S. (1964). Ecology of Plant Galls (The Hague, The Netherlands: Springer-
387 Science+Business Media, B.V.).

388 Marrs, K.A., Alfenito, M.R., Lloyd, A.M., and Walbot, V. (1995). A glutathione S-transferase
389 involved in vacuolar transfer encoded by the maize gene Bronze-2. Nature *375*, 397–400.

390 Martin, J.P. (1938). Stem galls of sugar cane induced with an insect extract. Hawaiian Planters'
391 Record *42*, 129–134.

392 McCalla, D.R., Genthe, M.K., and Hovanitz, W. (1962). Chemical Nature of an Insect Gall
393 Growth-Factor. Plant Physiol *37*, 98–103.

394 McDonald, J.H., and Kreitman, M. (1991). Adaptive Protein Evolution at the Adh Locus in
395 Drosophila. Nature *351*, 652–654.

396 Nabity, P.D., Haus, M.J., Berenbaum, M.R., and DeLucia, E.H. (2013). Leaf-galling phylloxera on
397 grapes reprograms host metabolism and morphology. Proceedings of the National Academy of
398 Sciences of the United States of America *110*, 16663–16668.

399 Obbard, D.J., Jiggins, F.M., Halligan, D.L., and Little, T.J. (2006). Natural selection drives
400 extremely rapid evolution in antiviral RNAi genes. Current Biology *16*, 580–585.

Papkou, A., Guzella, T., Yang, W., Koepper, S., Pees, B., Schalkowski, R., Barg, M.C., Rosenstiel, P.C., Teotónio, H., and Schulenburg, H. (2019). The genomic basis of red queen dynamics during rapid reciprocal host–pathogen coevolution. Proceedings of the National Academy of Sciences of the United States of America *116*, 923–928.

Parr, T. (1940). Asterolecanium variolosum Ratzeburg, a gall-forming coccid, and its effect upon the host trees. Yale University School of Forestry Bulletin *46*, 1–49.

Paterson, S., Vogwill, T., Buckling, A., Benmayor, R., Spiers, A.J., Thomson, N.R., Quail, M., Smith, F., Walker, D., Libberton, B., et al. (2010). Antagonistic coevolution accelerates molecular evolution. Nature *464*, 275–278.

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

Plumb, G.H. (1953). The formation and development of the Norway Spruce gall caused by Adelges abietis L. Bulletin of the Connecticut Experiment Station *566*, 1–77.

Rehill, B.J., and Schultz, J.C. (2001). Hormaphis hamamelidis and gall size: a test of the plant vigor hypothesis. Oikos *95*, 94–104.

Schultz, J.C., Edger, P.P., Body, M.J.A., and Appel, H.M. (2019). A galling insect activates plant reproductive programs during gall development. Scientific Reports *9*, 1833–1833.

Shih, T.H., Lin, S.H., Huang, M.Y., Sun, C.W., and Yang, C.M. (2018). Transcriptome profile of cup-shaped galls in Litsea acuminata leaves. PLoS ONE *13*, 1–14.

Smith, N.G., and Eyre-Walker, A. (2002). Adaptive protein evolution in Drosophila. Nature *415*, 1022–1024.

Springob, K., Nakajima, J.-I., Yamazaki, M., and Saito, K. (2003). Recent advances in the biosynthesis and accumulation of anthocyanins.

Stern, D.L. (1995). Phylogenetic evidence that aphids rather than plants, determine gall morphology. Proceedings of Royal Society London B *260*, 85–89.

Stone, G.N., and Cook, J.M. (1998). The structure of cynipid oak galls: patterns in the evolution of an extended phenotype. Proceedings of the Royal Society, London Series B *265*, 979–988.

Stuart, J. (2015). Insect effectors and gene-for-gene interactions with host plants. Current Opinion in Insect Science *9*, 56–61.

Suzuki, H., Yokokura, J., Ito, T., Arai, R., Yokoyama, C., Toshima, H., Nagata, S., Asami, T., and Suzuki, Y. (2014). Biosynthetic pathway of the phytohormone auxin in insects and screening of its inhibitors. Insect Biochemistry and Molecular Biology *53*, 66–72.

Takeda, S., Yoza, M., Amano, T., Ohshima, I., Hirano, T., Sato, M.H., Sakamoto, T., and Kimura, S. (2019). Comparative transcriptome analysis of galls from four different host plants suggests the molecular mechanism of gall development. PloS One *14*, e0223686–e0223686.

Tanaka, Y., Okada, K., Asami, T., and Suzuki, Y. (2013). Phytohormones in Japanese Mugwort Gall Induction by a Gall-Inducing Gall Midge. Bioscience, Biotechnology and Biochemistry *77*, 1942–1948.

Tooker, J.F., and Helms, A.M. (2014). Phytohormone Dynamics Associated with Gall Insects, and their Potential Role in the Evolution of the Gall-Inducing Habit. Journal of Chemical Ecology *40*, 742–753.

Yamaguchi, H., Tanaka, H., Hasegawa, M., Tokuda, M., Asami, T., and Suzuki, Y. (2012). Phytohormones and willow gall induction by a gall-inducing sawfly. New Phytologist *196*, 586–595.

Yang, Z., and Bielawski, J.P. (2000). Statistical methods for detecting molecular adaptation. Trends in Ecology & Evolution *15*, 496–503.

Acknowledgements:

We thank Jim Truman for performing the initial aphid salivary gland dissections and for teaching us the tricks, which enabled the entire project, Erika Gajda and Henry Horn for help collecting galls, Patrick Reilly for his genotyping pipeline and helpful conversations, Goran Ceric for getting all software packages to run on the Janelia compute cluster, Tom Dolafi for maintaining the Apollo genome annotation service, Mountain Lakes Biological Station for permission to collect specimens, and Jim Truman, Nicolas Frankel, Richard Mann, Brett Mensh, and Vanessa Ruta for helpful comments on the manuscript.

Competing Interests:

460 HHMI has filed a provisional patent, number 63/092,942, for the inventors AK and DLS
461 covering unique aphid polypeptides for use in modifying plants.
462
463 **Figure Legends**
464
465 **Fig. 1. *Hormaphis cornu* aphids drive patterned over-proliferation of plant cells to produce**
466 **galls of on leaves *Hamamelis virginiana***
467 (A-C) Photographs of the abaxial surfaces of *H. virginiana* leaves being attacked by first-instar
468 fundatrices of *H. cornu*. Nymphs gather on the unopened leaf buds (A) and soon after bud
469 break the fundatrices gather near leaf veins (B) and inject material that begins to transform
470 the leaf into a gall (C). Magnified region in blue rectangle of panel (A) shows three fundatrices
471 waiting on unopened bud (A').
472 (D-F) Photographs of the adaxial leaves of *H. virginiana*, showing galls at early (D), middle (E),
473 and late (F) growth stages.
474 (G-I) Confocal micrographs of sections through a *H. cornu* gall (G,H) and un-galled *H. virginiana*
475 leaf (I) stained with Congo red and calcofluor white. Extensive hypertropy is observed in the
476 mesophyll (yellow arrows) at a considerable distance from the location of the aphid's stylets
477 (blue arrows) (G). The tips of aphid stylets (pink arrow) can be observed within cells of *H.*
478 *virginiana* and plant tissue shows evidence of hyperproliferation and periclinal divisions (grey
479 arrows) close to the tips of stylets and the termini of stylet sheaths (H). Periclinal divisions are
480 observed in both spongy and palisade mesophyll cells during gall development, but never in

481 ungalled leaf tissue (I). Plant tissue is presented in the aphid's frame of reference, with abaxial
482 leaf surface up.
483
484 **Figure 2. A genome-wide association study (GWAS) identifies variation within a novel aphid**
485 **gene associated with gall color**
486 (A) A GWAS of fundatrices isolated from 43 green and 47 red galls identified a small region on
487 chromosome 1 of the *H. cornu* genome that is strongly associated with gall color. Red line
488 indicates FDR = 0.05. Colors of points on chromosomes are arbitrary.
489 (B-D) Resequencing 800 kbp of Chromosome 1 centered on the most significant SNPs from the
490 original GWAS to approximately 60X coverage identified 11 spatially clustered SNPs
491 significantly associated with gall color located within the introns and upstream of
492 *Horco_g16073*, which was named *dgc* (D). (Some SNPs are closely adjacent and cannot be
493 differentiated at this scale.) Significant SNPs are indicated with orange vertical lines in (D).
494 (E) Genotypes of all 11 SNPs associated with gall color from an independent sample of aphids
495 from 435 green and 431 red galls. Color of gall for aphid samples shown on left and genotype
496 at each SNP is shown adjacent in green (0/0, homozygous ancestral state), yellow (0/1,
497 heterozygous), or red (1/1, homozygous derived state). LOD scores for association with gall
498 color shown for each SNP at bottom of genotype plot. Histogram of frequencies of each
499 multilocus genotype ordered by frequency and collected within gall color is shown on the
500 right. All 11 SNPs are strongly associated with gall color ($P < 10^{-192}$), and a cluster of 5 SNPs in a
501 61bp region are most strongly associated with gall color. Individuals homozygous for ancestral

502 alleles at all or most loci and making red galls likely carry variants at other loci that influence
503 gall color (Supplementary Text).
504
505 **Figure 3. *Dgc* is the only strongly differentially expressed gene in salivary glands of**
506 **fundatrices collected from red versus green galls**
507 (A) Genome-wide differential expression analysis of *H. cornu* fundatrix salivary gland
508 transcripts from individuals heterozygous for dgc^{Red}/dgc^{Green} (Red; N = 15) versus homozygous
509 for dgc^{Green} (Green; N = 7) illustrated as a volcano plot. D*gc* is strongly downregulated in
510 dgc^{Red}/dgc^{Green} fundatrices.
511 (B) Salivary gland expression of genes in the paralog cluster that includes dgc^{Red}/dgc^{Green}. Gene
512 models of the paralogous genes are shown at the top and expression levels normalized by
513 mean expression across all paralogs is shown below. Only *dgc* is strongly downregulated in this
514 gene cluster between individuals carrying dgc^{Red}/dgc^{Green} (Red) versus dgc^{Green}/dgc^{Green} (Green)
515 genotypes.
516 (C) Genome-wide differential expression analysis of *H. virginiana* transcripts isolated from galls
517 made by fundatrices heterozygous for dgc^{Red}/dgc^{Green} (Red; N = 17) versus homozygous for
518 dgc^{Green} (Green; N = 22) illustrated as a volcano plot. Only eight genes are differentially
519 expressed at FDR < 0.05, and all are overexpressed in red galls. The seven most strongly
520 differentially expressed genes that encode anthocyanin biosynthetic enzymes (FAOMT-1 =
521 Hamvi_23591; FAOMT-2 = Hamvi_7147 ; GSTF11 = Hamvi_134919; GSTF12 = Hamvi_109682;
522 UGT75C1 = Hamvi_14194; UFGT = Hamvi_22774; ACCA = Hamvi_97071).

523 (D) Expression levels, in counts per million reads, of the seven anthocyanin biosynthetic genes
524 overexpressed in red galls, in green (green) and red (red) galls and ungalled leaves (blue). Each
525 data point is from a separate genome-wide RNA-seq sample.
526 (E) Simplified diagram of the anthocyanin biosynthetic pathway. Enzyme classes upregulated
527 in red galls are shown in purple font. The two terminal anthocyanins that generate the red
528 color in galls, peonidin-3,5-diglucoside and malvidin-3,5-diglucoside, are shown in bold font,
529 and three precursor molecules found in red galls are shown in bold italic font. Anthocyanin
530 names are abbreviated (dg = 3,5-diglucoside).
531 (F) Photos of cross-sections of green (top left) and red (top right) and the pigments extracted
532 from green and red galls (below).
533 (G) UHPLC-DAD chromatograms at 520 nm of extract from red (red line) and green (green line)
534 galls.
535 (H) Overlaid UHPLC-MS chromatograms of green (top) and red (middle) gall extracts and
536 authentic standards (bottom). Each pigment is indicated with a different color: green =
537 delphinidin-3,5-diglucoside (*m/z* = 627.1551); black = cyanidin-3,5-diglucoside (*m/z* =
538 611.1602); blue = petunidin-3,5-diglucoside (*m/z* = 641.1709); purple = peonidin-3,5-
539 diglucoside (*m/z* = 625.1768); and red = malvidin-3,5-diglucoside (*m/z* = 655.1870).
540 Anthocyanin names are abbreviated (dg = 3,5-diglucoside). Peonidin-3,5-diglucoside and
541 malvidin-3,5-diglucoside together account for 87% of pigment detected in red galls.
542
543 **Figure 4. *bicycle* genes are salivary gland enriched transcripts of gall-associated *H. cornu***
544 **generations**

545 (A) Diagram of life cycle of *H. cornu*. *H. cornu* migrates annually between *H. virginiana* (blue
546 line) and *Betula nigra* (pink line) and the gall is produced only in the spring on *H. virginiana*
547 (brown line). Each nymph of the first generation (G1, the fundatrix) hatches from an over-
548 wintering egg and initiates development of a single gall. Her offspring (G2) feed and grow
549 within the gall and develop with wings, which allows them to fly to *B. nigra* in late spring. For
550 three subsequent generations (G3-G5) the aphids develop as small, coccidiform morphs on *B.*
551 *nigra*. In the fall, aphids develop with wings (G6), fly to *H. virginiana* plants, and deposit male
552 and female sexuals (G7), the only generation possessing meiotic cells. The sexuals feed and
553 complete development on the senescing leaves of *H. virginiana*. As adults they mate and the
554 females deposit eggs that overwinter and give rise to fundatrices the following spring.
555 (B) Differential expression of fundatrix versus sexual salivary glands with only genes
556 significantly upregulated in fundatrix salivary glands marked with colors, shown as a volcano
557 plot. Genes with and without homologs in public databases are labeled as "Annotated" (blue)
558 and "Not Annotated" (brown), respectively.
559 (C) Hierarchical clustering of unannotated salivary-gland specific genes reveals one large
560 cluster of *bicycle* genes, one small cluster of *CWG* genes, and many largely unique, unclustered
561 genes.
562 (D) *Bicycle* (green) and remaining unannotated (brown) genes labelled on a differential
563 expression volcano plot illustrates that, on average, *bicycle* genes are the most strongly
564 differentially expressed genes expressed specifically in fundatrix salivary glands.

565 (E) *Bicycle* genes are expressed at highest levels in the salivary glands of the fundatrix and
566 mostly decline in expression during subsequent generations. (Sample sizes: G1 (N = 20); G2 (N
567 = 4); G5 (N = 6); and G7 (N = 6).
568 (F) Amino-acid logo for predicted BICYCLE proteins. Sequence alignment used for logo was
569 filtered to highlight conserved positions. Most genes encode proteins with an N-terminal
570 signal sequence and a pair of conserved cysteine-tyrosine-cysteine motifs (CYC).
571

Figure 1

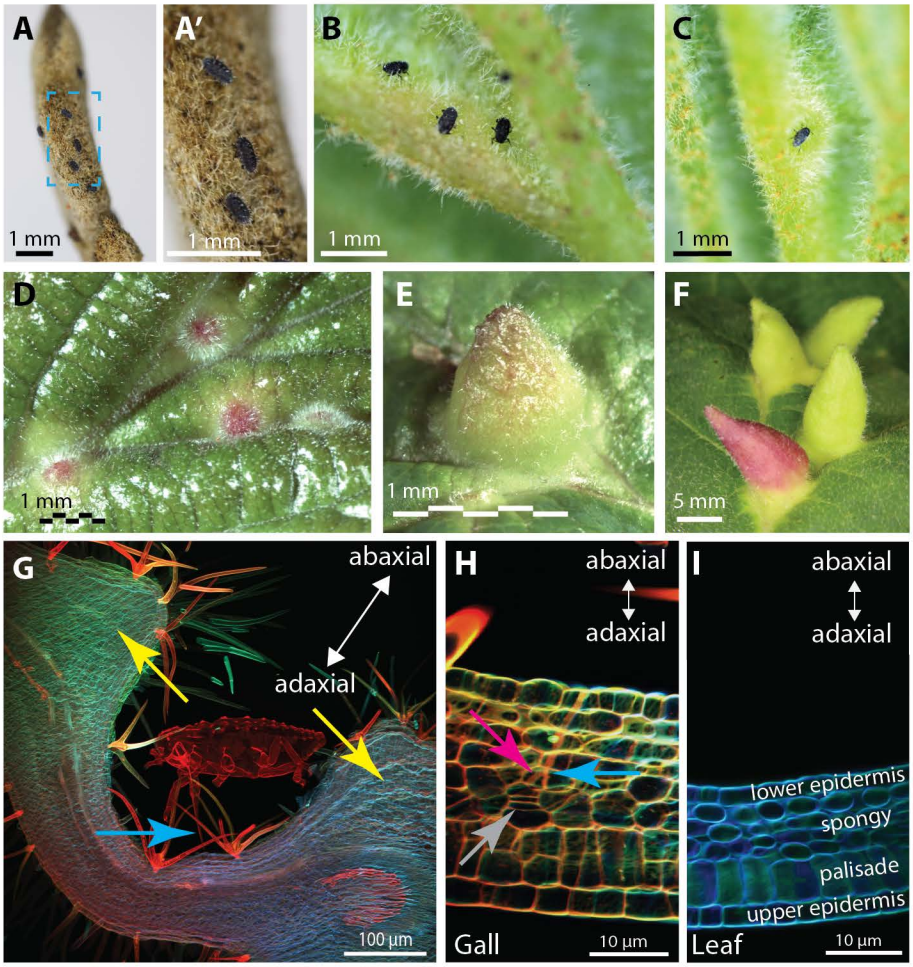

A
1 mm
A′
1 mm
B
1 mm
C
1 mm
D
1 mm
E
1 mm
F
5 mm
G
abaxial
adaxial
100 µm
H
abaxial
adaxial
Gall
10 µm
I
abaxial
adaxial
lower epidermis
spongy
palisade
upper epidermis
Leaf
10 µm

Figure 2

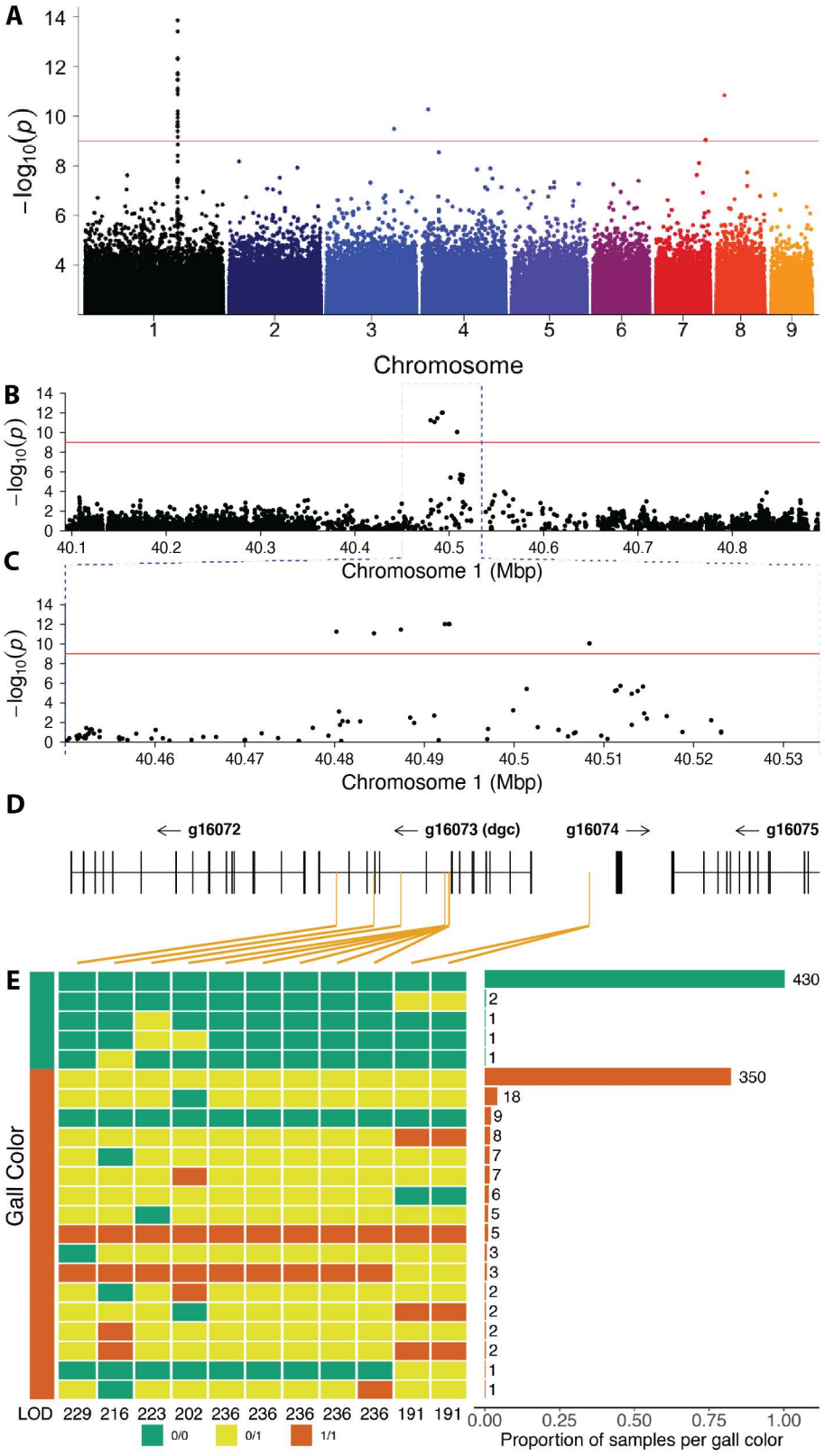

A
−log10(p)
Chromosome
B
−log10(p)
Chromosome 1 (Mbp)
C
−log10(p)
Chromosome 1 (Mbp)
D
g16072
g16073 (dgc)
g16074
g16075
E
Gall Color
LOD 229 216 223 202 236 236 236 236 236 191 191
0/0
0/1
1/1
Proportion of samples per gall color

Figure 3

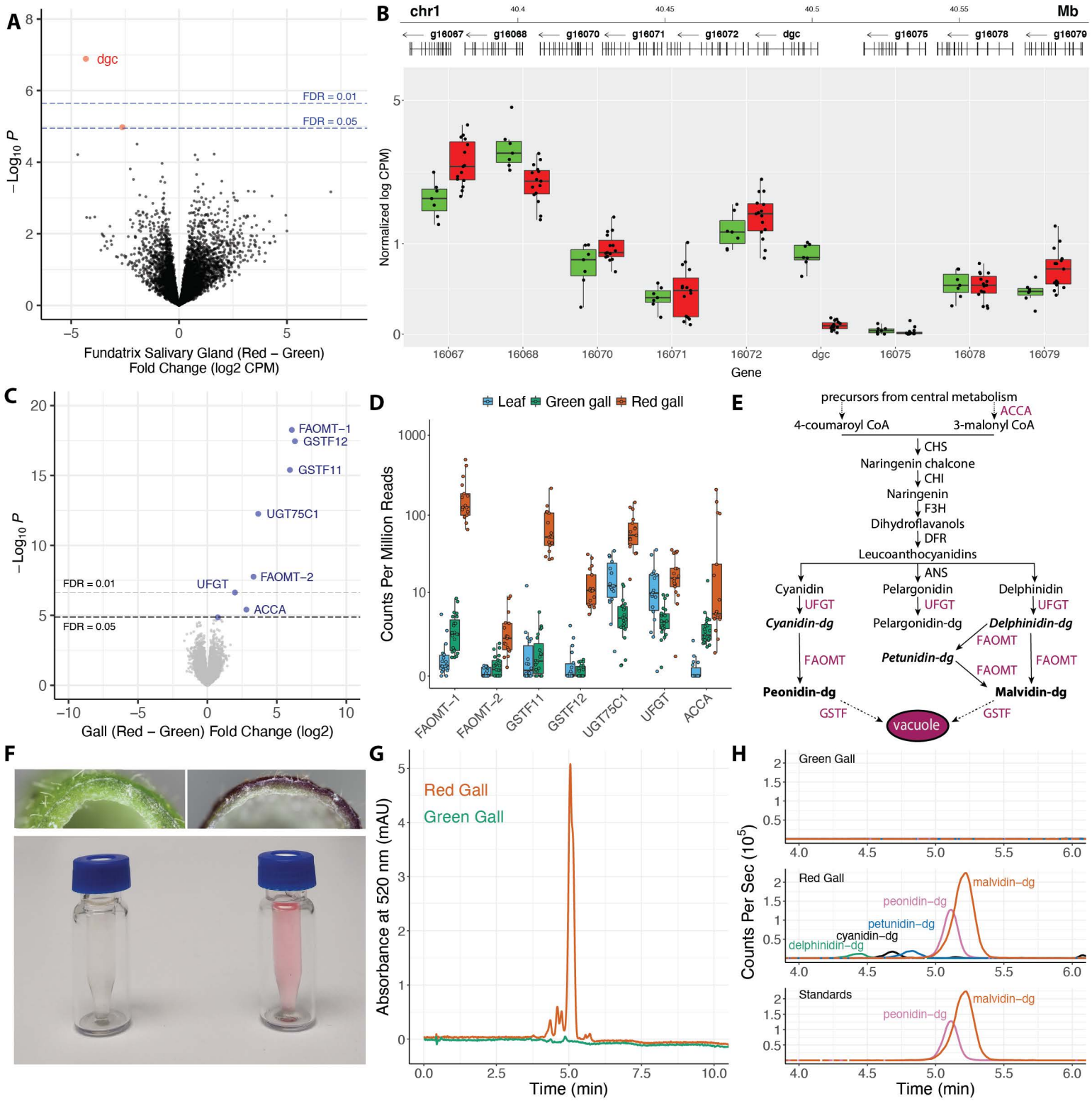
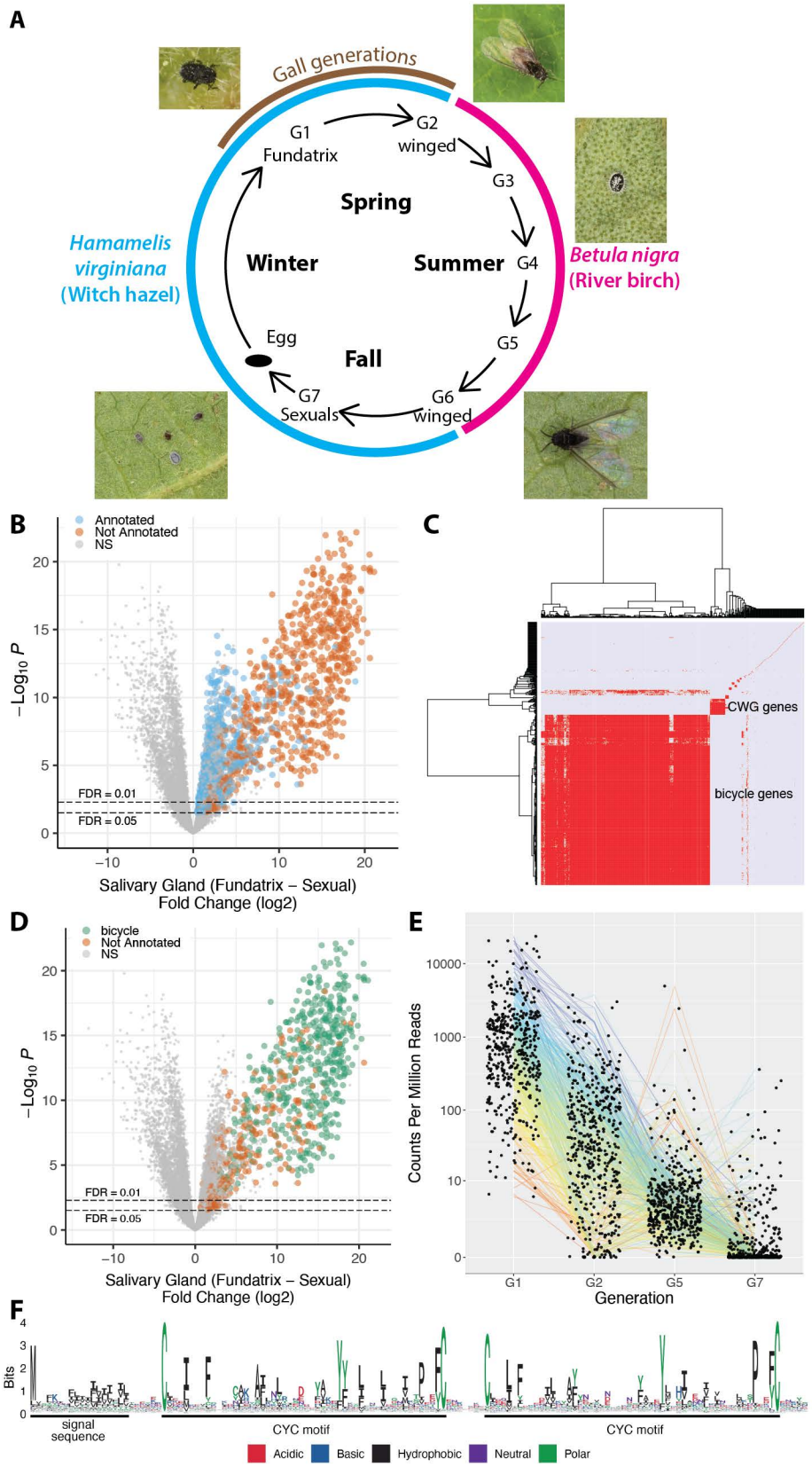

A
dgc
FDR = 0.01
FDR = 0.05
−Log10 P
Fundatrix Salivary Gland (Red – Green)
Fold Change (log2 CPM)
B
chr1
Mb
g16067
g16068
g16070
g16071
g16072
dgc
g16075
g16078
g16079
Normalized log CPM)
Gene
C
FAOMT−1
GSTF12
GSTF11
UGT75C1
FAOMT−2
UFGT
ACCA
FDR = 0.01
FDR = 0.05
−Log10 P
Gall (Red – Green) Fold Change (log2)
D
Leaf
Green gall
Red gall
Counts Per Million Reads
FAOMT−1
FAOMT−2
GSTF11
GSTF12
UGT75C1
UFGT
ACCA
E
precursors from central metabolism
ACCA
4-coumaroyl CoA
3-malonyl CoA
CHS
Naringenin chalcone
CHI
Naringenin
F3H
Dihydroflavanols
DFR
Leucoanthocyanidins
ANS
Cyanidin
Pelargonidin
Delphinidin
UFGT
Cyanidin-dg
Pelargonidin-dg
Delphinidin-dg
FAOMT
Petunidin-dg
Peonidin-dg
Malvidin-dg
GSTF
vacuole
F
G
Red Gall
Green Gall
Absorbance at 520 nm (mAU)
Time (min)
H
Green Gall
Red Gall
Standards
malvidin−dg
peonidin−dg
petunidin−dg
cyanidin−dg
delphinidin−dg
Counts Per Sec (10^5)
Time (min)

Figure 4


A
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)
B
Annotated
Not Annotated
NS
−Log10 P
FDR = 0.01
FDR = 0.05
Salivary Gland (Fundatrix − Sexual)
Fold Change (log2)
C
CWG genes
bicycle genes
D
bicycle
Not Annotated
NS
−Log10 P
FDR = 0.01
FDR = 0.05
Salivary Gland (Fundatrix − Sexual)
Fold Change (log2)
E
Counts Per Million Reads
Generation
G1
G2
G5
G7
F
Bits
signal sequence
CYC motif
CYC motif
Acidic
Basic
Hydrophobic
Neutral
Polar