Title: A novel family of secreted insect 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 novel plant organs called galls that provide the insect with a source of nutrition and a temporary home. Galls result from dramatic reprogramming of plant cell biology driven by insect molecules, but the roles of specific insect molecules in gall development have not yet been determined. Here we study the aphid *Hormaphis cornu*, which makes distinctive "cone" galls on leaves of witch hazel *Hamamelis virginiana*. We found that derived genetic variants in the aphid gene *determinant of gall color* (*dgc*) are associated with strong downregulation of *dgc* transcription in aphid salivary glands, upregulation in galls of seven genes involved in anthocyanin synthesis, and deposition of two red anthocyanins in galls. We hypothesize that aphids inject DGC protein into galls, and that this results in differential expression of a small number of plant genes. *Dgc* is a member of a large, diverse family of novel predicted secreted proteins characterized by a pair of widely spaced cysteine-tyrosine-cysteine (CYC) residues, which we named BICYCLE proteins. *Bicycle* genes are most strongly expressed in the salivary glands specifically of galling aphid generations, suggesting that they may regulate many aspects of gall development. *Bicycle* genes have experienced unusually frequent 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 environmental vicissitude and from some predators and parasites (Bailey et al., 2009; Mani, 1964; Stone and Schönrogge, 2003). 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

67 and Worobey, 1998; Dodson, 1991; Hearn et al., 2019; Leatherdale, 1955; Martin, 1938;
68 Martinson et al., 2020; Parr, 1940; Plumb, 1953; Stern, 1995; Stone and Cook, 1998).
69 At least some gall-inducing insects produce phytohormones (Dorchin et al., 2009;
70 McCalla et al., 1962; Suzuki et al., 2014; Tanaka et al., 2013; Tooker and Helms, 2014;
71 Yamaguchi et al., 2012), although it is not yet clear whether insects introduce these hormones
72 into plants to support gall development. However, injection of phytohormones alone probably
73 cannot generate the large diversity of species-specific insect galls. In addition, galling insects
74 induce plant transcriptional changes independently of phytohormone activity (Bailey et al.,
75 2015; Hearn et al., 2019; Nabity et al., 2013; Papkou et al., 2019; Shih et al., 2018; Takeda et al.,
76 2019). Thus, given the complex cellular changes required for gall development, insects probably
77 introduce many molecules into plant tissue to induce galls.
78 In addition to the potential role of phytohormones in promoting gall growth, candidate
79 gall effectors have been identified in several gall-forming insects (Cambier et al., 2019; Eitle et
80 al., 2019; Zhao et al., 2015). However, none of these candidate effectors have yet been shown
81 to contribute to gall development or physiology. In addition, while many herbivorous insects
82 introduce effector molecules into plants to influence plant physiology (Elzinga and Jander,
83 2013; Hogenhout and Bos, 2011; Kaloshian and Walling, 2016; Stuart, 2015), there is no
84 evidence that any previously described effectors contribute to gall development. Since there
85 are currently no galling insect model systems that would facilitate a genetic approach to this
86 problem, we turned to natural variation to identify insect genes that contribute to gall
87 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 microscopic mouthparts (stylets) (Fig. 1A, B, Video S1) and pierces individual mesophyll cells with her stylets (Fig. 1G and H) (Lewis and Walton, 1947, 1958). We found that plant cells near injection sites, revealed by the persistent stylet sheaths, over-proliferate through periclinal cell divisions (Figure 1H). This pattern of cytokinesis is not otherwise observed in leaves at this stage of development (Figure 1I) and contributes to the thickening and expansion of leaf tissue that generates the gall (Fig. 1D-G). The increased proliferation of cells near the tips of stylet sheaths suggests that secreted effector molecules produced in the salivary glands are deposited into the plant via the stylets.

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

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

110 We found that populations of *H. cornu* include approximately 4% red galls and 96%
111 green galls (Fig. 1F). We inferred that this gall color polymorphism results from differences
112 among aphids, rather than from differences associated with leaves or the location of galls on
113 leaves, because red and green galls are located randomly on leaves and often adjacent to each
114 other on a single leaf (Fig. 1F). We sequenced and annotated the genome of *H. cornu*
115 (Supplemental Material) and performed a genome-wide association study (GWAS) on
116 fundatrices isolated from 43 green galls and 47 red galls by resequencing their genomes to
117 approximately 3X coverage. There is no evidence for genome-wide differentiation of samples
118 from red and green galls, suggesting that individuals making red and green galls were sampled
119 from a single interbreeding population (Fig. S1A-D). We identified SNPs near 40.5 Mbp on
120 Chromosome 1 that were strongly associated with gall color (Fig. 2A). We re-sequenced
121 approximately 800 kbp flanking these SNPs to approximately 60X coverage and identified 11
122 single-nucleotide polymorphisms (SNPs) within the introns and upstream of gene *Horco_16073*
123 that were strongly associated with gall color (Fig. 2B-D). There is no evidence that large scale
124 chromosomal aberrations are associated with gall color (Supplementary Text).
125 Since GWAS can sometimes produce spurious associations, we performed an
126 independent replication study and found that all 11 SNPs were highly significantly associated
127 with gall color in fundatrices isolated from 435 green and 431 red galls (LOD = 191 – 236; Fig.
128 2E). All fundatrices from green galls were homozygous for the ancestral allele at 9 or more of
129 these SNPs (Fig. 2E). In contrast, 98% of fundatrices from red galls were heterozygous or
130 homozygous for derived alleles at 9 or more SNPs (Fig. 2E). This pattern suggests that alleles
131 contributing to red gall color are genetically dominant to alleles that generate green galls. Two

132 percent of fundatrices that induce red galls were homozygous for ancestral alleles at these
133 SNPs and likely carry genetic variants elsewhere in the genome that confer red color to galls
134 (Supplementary Text).

135 Based on these genetic associations and further evidence presented below, we assigned
136 the name *determinant of gall color* (*dgc*) to *Horco_16073*. *Dgc* encodes a predicted protein of
137 23 kDa with an N-terminal secretion signal sequence (Fig. S4). The putatively secreted portion
138 of the protein shares no detectable sequence homology with any previously reported proteins.

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

149

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

151

152 Since all 11 *dgc* polymorphisms associated with gall color occur outside of *dgc* exons
153 (Fig. 2D), we tested whether these polymorphisms influence expression of *dgc* or of any other

genes in the genome. We first determined that *dgc* is expressed highly and specifically in fundatrix salivary glands and lowly or not at all in other tissues or other life cycle stages (Fig. S7A). We then performed RNAseq on salivary glands from fundatrices with *dgc*^{Green} /*dgc*^{Green} or *dgc*^{Red}/*dgc*^{Green} genotypes. *Dgc* stands out as the most strongly differentially expressed gene between these genotypes (Figure 3A and S7B). Since *dgc*^{Red} alleles appeared to be dominant to the *dgc*^{Green} alleles for gall color, we expected that *dgc* transcripts would be upregulated in animals with *dgc*^{Red} alleles. In contrast, *dgc* transcripts were almost absent in fundatrices carrying *dgc*^{Red} alleles (Fig. 3B). That is, red galls are associated with strongly reduced *dgc* expression in fundatrix salivary glands.

Dgc expression is reduced approximately 20-fold in fundatrix salivary glands with *dgc*^{Red}/*dgc*^{Green} (27 ± 22.6 CPM, mean ± SD) versus *dgc*^{Green}/*dgc*^{Green} genotypes (536 ± 352.3 CPM, mean ± SD). This result suggested that *dgc*^{Red} alleles downregulate both the *dgc*^{Red} and *dgc*^{Green} alleles in heterozygotes. To confirm whether the *dgc*^{Red} allele downregulates the *dgc*^{Green} allele in *trans*, we identified exonic SNPs that were specific to each allele and could be identified in the RNAseq data. We found that both *dgc*^{Red} and *dgc*^{Green} alleles were strongly downregulated in heterozygotes, confirming the *trans* activity of the *dgc*^{Red} allele (Fig. S7D). We observed no systematic transcriptional changes in neighboring genes (Figure 3B), most of which are *dgc* paralogs, indicating that *dgc*^{Red} alleles exhibit a perhaps unique example of locus-specific repressive transvection (Duncan, 2002).

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

176

177 Since red galls are associated with strong differential expression of only *dgc*, we
178 wondered how the plant responds to changes in this single putative effector. To examine this
179 question, we sequenced and annotated the genome of the host plant *Hamamelis virginiana* and
180 then performed whole-genome differential expression on plant mRNA isolated from galls
181 induced by aphids with dgc^{Red}/dgc^{Green} versus dgc^{Green}/dgc^{Green} genotypes (Supplementary Text).
182 Only eight plant genes were differentially expressed between green and red galls and all eight
183 genes were downregulated in green galls (Fig. 3C-D). That is, high levels of *dgc* are associated
184 with downregulation of only eight plant genes in galls.

185 Red pigmented galls could result from production of carotenoids (Smits and Peterson,
186 1942), anthocyanins (Blunden and Challen, 1965; Bomfim et al., 2019), or betacyanins.
187 However, in red galls induced by *H.cornu*, the seven most strongly downregulated plant genes
188 are all homologous to genes annotated as enzymes of the anthocyanin biosynthetic pathway
189 (Fig. 3E). One gene encodes an enzyme (ACCA) that irreversibly converts acetyl-CoA to malonyl
190 CoA, a biosynthetic precursor of multiple anthocyanins. Two genes encode anthocyanidin 3-0-
191 glucosyltransferases (UFGT and UGT75C1), which glycosylate unstable anthocyanidins to allow
192 their accumulation (Springob et al., 2003). Two genes encode flavonoid 3'-5'
193 methyltransferases (FAOMT-1, FAOMT-2), which methylate anthocyanin derivatives (Hugueney
194 et al., 2009). Finally, two genes encode phi class glutathione S-transferases (GSTF11, GSTF12),
195 which conjugate glutathione to anthocyanins, facilitating anthocyanin transport and stable
196 accumulation in vacuoles (Marrs et al., 1995).

197 Six of the enzymes upregulated in red galls are required for final steps of anthocyanin
198 production and deposition (Fig. 3E) and their upregulation in red galls may account for the
199 accumulation of pigments in red galls. To test this hypothesis, we extracted and analyzed
200 pigments from galls (Fig. 3F) and identified high levels of two pigments only in red galls (Fig.
201 3G), the anthocyanins malvidin-3,5-diglucoside and peonidin-3,5-diglucoside (Fig. 3G-H, S8).
202 Thus, the pigments present in red galls are products of enzymes in the anthocyanin biosynthetic
203 pathway, such as those encoded by genes that are upregulated in red galls. The two abundant
204 anthocyanins are produced from distinct intermediate precursor molecules (Fig. 3E), three of
205 which were also detected in red galls (Fig. 3H, S8), and synthesis of these two anthocyanins
206 likely requires activity of different methyltransferases and glucosyltransferases. The three pairs
207 of glucosyltransferases, methyltransferases, and glutathione transferases upregulated in red
208 galls may provide the specific activities required for production of these two anthocyanins.
209 Taken together, these observations suggest that *dgc* represses transcription of seven
210 anthocyanin biosynthetic enzymes. It is not clear how *dgc* induces specific transcriptional
211 changes in seven plant genes; it may act by altering activity of an upstream regulator of these
212 plant genes.
213
214 **Aphids induce widespread transcriptomic changes in galls**
215
216 Gall color represents only one aspect of the gall phenotype, apparently mediated by
217 changes in expression of seven plant genes, and the full complement of cell biological events
218 during gall development presumably requires changes in many more plant genes. To estimate

219 how many plant genes are differentially expressed during development of the *H. cornu* gall on
220 *H. virginiana*, we performed differential expression analysis of plant genes in galls versus the
221 surrounding leaf tissue. Approximately 31% of plant genes were upregulated and 34% were
222 downregulated at FDR = 0.05 in galls versus leaf tissue (Figure S9E; 27% up and 29% down in
223 gall at FDR = 0.01). Results of gene ontology analysis of up and down-regulated genes is
224 consistent with the extensive growth of gall tissue and down regulation of chloroplasts seen in
225 aphid galls (Fig. S9F; Supplementary Text), a pattern observed in other galling systems (Hearn et
226 al., 2019).

227 Thus, approximately 15,000 plant genes are differentially expressed in galls,
228 representing a system-wide re-programming of plant cell biology. If other aphid effector
229 molecules act in ways similar to *dgc*, which is associated with differential expression of only
230 eight plant genes, then gall development may require injection of hundreds or thousands of
231 effector molecules.

232

233 **_Dgc_ is a member of a large class of novel _bicycle_ genes expressed specifically in the salivary**
234 **glands of gall-inducing aphids**

235

236 To identify additional proteins that aphids may inject into plants to contribute to gall
237 development, we exploited the fact that only some individuals in the complex life cycle of *H.*
238 *cornu* induce galls (Figure 4A). Only the fundatrix generation induces galls and only her
239 immediate offspring live alongside her in the developing gall. In contrast, individuals of
240 generations that live on river birch (*Betula nigra*) through the summer and the sexual

241 generation that feed on *H. virginiana* leaves in the autumn do not induce any leaf
242 malformations. Thus, probably only the salivary glands of the generations that induce galls (the
243 fundatrix (G1) and possibly also her immediate offspring (G2)) produce gall-effector molecules.
244 We identified 3,048 genes upregulated in fundatrix salivary glands versus the fundatrix body
245 (Fig. S10A) and 3,427 genes upregulated in salivary glands of fundatrices, which induce galls,
246 versus sexuals, which do not induce galls although they feed on the same host plant (Fig. S10B).
247 Intersection of these gene sets identified 1,482 genes specifically enriched in the salivary glands
248 of fundatrices (Figure 4B).

249 Half of these genes (744) were homologous to previously identified genes, many of
250 which had functional annotations. Gene Ontology analysis of the "annotated" genes suggests
251 that they contribute mostly to the demands for high levels of protein secretion in fundatrix
252 salivary glands (Fig. S10C). Most do not encode proteins with secretion signals (671; Fig. S10D)
253 and are thus unlikely to be injected into plants. We searched for homologs of genes that have
254 been proposed as candidate gall-effector genes in other insects but found no evidence that
255 these classes of genes contribute to aphid gall development (Supplementary Text). We
256 therefore focused on the remaining 738 unannotated genes, which included 459 genes
257 encoding proteins with predicted secretion signals (Fig. S10D).

258 Hierarchical clustering of the unannotated genes by sequence similarity identified one
259 large (476 genes) and one small (43 genes) cluster of related genes, and 222 genes sharing few
260 or no homologs (Fig. 4C). Genes in both the large and small clusters encode proteins with N-
261 terminal secretion signals, as expected for effector proteins that might be injected into plants.
262 The small cluster encodes a divergent set of proteins containing several conserved cysteines (C)

263 and a well conserved tryptophan (W) and glycine (G), and we named these CWG genes (Fig.
264 S11A).

265 Proteins encoded by the large cluster display conservation mainly of a pair of widely
266 spaced cysteine-tyrosine-cysteine (CYC) motifs and spacing between the C, Y, and C residues of
267 each motif is not well conserved (Fig. 4F). This pair of CYC motifs led us to name these *bicycle*
268 (bi-CYC-like) genes. The *bicycle* genes were the most strongly upregulated class of genes in
269 fundatrix salivary glands (Fig. 4D, Fig. S11B-F) and were expressed specifically in the salivary
270 glands of the two generations associated with galls (G1 and G2) (Fig. 4E). Many *bicycle* genes
271 are found in paralog clusters throughout the *H. cornu* genome (Fig. S10E and F) and each *bicycle*
272 gene contains approximately 5-25 microexons (Fig. S10F)—a large excess relative to the
273 genomic background (Fig. S10G)—interrupted by long introns (Fig. S10F).

274 We found that *dgc* shares many features with *bicycle* genes—it is strongly expressed
275 specifically in fundatrix salivary glands, and it exhibits many microexons and a pair of CYC motifs
276 (Fig. S4)—and that it is evolutionarily related to other *bicycle* genes. Thus, *dgc* is a member of a
277 diverse family of genes encoding secreted proteins expressed specifically in the salivary glands
278 of gall forming generations. *Bicycle* genes are therefore good candidates to encode many of the
279 molecules required to generate the extensive transcriptional changes observed in galls.

280

281 ***bicycle* genes experienced intense diversifying selection, consistent with a potential arms race**
282 **between aphids and plants**

283

284 *Bicycle* genes are extremely diverse at the amino acid sequence level, as has been
285 observed for other candidate insect gall effector genes (Chen et al., 2010). Each *bicycle* protein
286 has accumulated approximately two substitutions per amino acid site since divergence from
287 paralogs (Fig. S10H). To explore whether this diversity resulted from natural selection rather
288 than genetic drift, we compared rates of non-synonymous (d_N) versus synonymous (d_S)
289 substitutions between the sister species *H. cornu* and *H. hamamelidis* (which also induces galls)
290 in *bicycle* versus non-*bicycle* genes, because d_N/d_S values greater than one provide evidence for
291 positive selection (Yang and Bielawski, 2000). To calculate polymorphism of orthologous genes
292 in each species, we mapped sequencing reads from individuals of each species to the *H. cornu*
293 genome. For divergence estimates, we estimated the *H. hamamelidis* genome by mapping *H.*
294 *hamamelidis* sequencing reads to the *H. cornu* genome (Supplemental Methods) and compared
295 this genome with the original *H. cornu* genome. A large excess of *bicycle* genes displayed d_N/d_S
296 significantly greater than 1 relative to the genomic background (Fig. S12, Table S1; $P < 2.2e-16$),
297 revealing recurrent adaptive amino acid substitution at many *bicycle* genes since these species
298 diverged.
299 We then quantified the frequency of adaptive amino acid substitutions at *bicycle* genes
300 and other categories of genes overexpressed in fundatrix salivary glands by calculating α, the
301 proportion of non-synonymous substitutions fixed by positive selection (Fay et al., 2001; Smith
302 and Eyre-Walker, 2002). We estimate that for all *bicycle* genes $\alpha = 0.33$ (95%CI 0.24 - 0.41) and
303 that for the subset of *bicycle* genes displaying d_N/d_S significantly greater than 1, $\alpha = 0.62$ (95%CI
304 0.45-0.73, Table S2). Other categories of genes overexpressed in fundatrix salivary glands
305 display values of α in the range of 0.27-0.38, however, *bicycle* genes display a considerably

306 higher ratio of non-synonymous to synonymous substitutions than other categories of genes
307 (Table S2). Most strikingly, as a fraction of protein length, *bicycle* genes display a considerably
308 greater fraction of adaptive amino acid substitutions than other categories of genes (Fig. S12C).
309 Since speciation between *H. cornu* and *H. hamamelidis*, positive selection has resulted, on
310 average, in fixation of approximately 2-3 substitutions in each *bicycle* gene, and approximately
311 10 substitutions in the most rapidly evolving *bicycle* genes. This represents a considerable
312 fraction of the average length of these proteins (~200 residues) and reveals intense selection on
313 *bicycle* genes, presumably for novel functions that require multiple amino acid substitutions.
314 The previous tests cannot detect selection on non-coding regions and do not
315 discriminate between selection acting in the deep past versus more recently. To search for
316 recent selection in *bicycle* gene regions, we examined patterns of polymorphism and
317 divergence within and between *H. cornu* and *H. hamamelidis*. Polymorphism was strongly
318 reduced relative to divergence in *bicycle* gene regions compared with the genomic background
319 (Fig. S13, S14) and patterns of reduced polymorphism were strikingly similar in both species
320 (Fig. S13, S14). This pattern is suggestive of recent selective sweeps in *bicycle* gene regions in
321 both species, which we tested by performing genome-wide scans for positive selection (Pavlidis
322 et al., 2013). Genome-wide signals of selective sweeps were enriched near *bicycle* genes and
323 multiple signals fell within *bicycle* gene clusters in both species (Fig. S14 and S15). Thus, in
324 addition to long-term adaptive protein evolution of bicycle proteins, it appears that strong
325 positive selection has acted recently and presumably frequently near many *bicycle* genes
326 throughout the genome.

327 In summary, we find evidence for widespread, strong, recent, and frequent positive
328 selection on *bicycle* genes. Since *bicycle* genes are likely secreted from salivary glands
329 specifically in gall-forming aphids, these observations are consistent with the hypothesis that
330 *bicycle* genes encode proteins that are intimately involved in reciprocal molecular evolution
331 between the aphid and their host plant.
332
333 **DISCUSSION**
334
335 We presented multiple lines of evidence that suggest that *bicycle* gene products provide
336 instructive molecules required for aphid gall development. The strongest evidence is that
337 eleven derived regulatory polymorphisms at *dgc* are associated with red galls, with almost
338 complete silencing of *dgc* in aphid salivary glands (Fig. 3A-B, S7B), and with upregulation of
339 seven anthocyanin biosynthetic genes and two red-purple anthocyanins in galls (Fig. 3C-H, S8).
340 Gall color is one small, but convenient, aspect of the panoply of cell biological changes required
341 for gall development. We hypothesize that the product of each *bicycle* gene has its own unique
342 set of targets in the plant and that the combined action of all *bicycle* gene products regulates
343 many aspects of gall development. Testing this hypothesis will require the development of new
344 methods to explore and manipulate this aphid-plant system.
345
346 **BICYCLE protein functions**
347

348 *Dgc* likely encodes a protein that is deposited by aphids into gall tissue, and current
349 evidence suggests that this protein specifically and dramatically results in the downregulation
350 of seven anthocyanin biosynthetic genes. The mechanisms by which this novel aphid protein
351 could alter plant transcription remains to be determined. The primary sequences of DGC and
352 other BICYCLE proteins provide few clues to their molecular mode of action. Outside of the N-
353 terminal secretion signal, BICYCLE proteins possess no similarity with previously reported
354 proteins and display no conserved domains that might guide functional studies. The relatively
355 well-conserved C-Y-C motif appears to define a pair of ~50-80 aa domains in each protein and
356 the paired cysteines may form disulfide bonds, which is commonly observed for secreted
357 proteins. Secondary structure prediction methods provide little evidence for structural
358 conservation across BICYCLE proteins and the extensive variation in the spacing between
359 conserved cysteines further suggests that BICYCLE proteins may display structural
360 heterogeneity. Identification of their molecular mode of action will require identification of
361 BICYCLE protein binding targets.
362
363 **Evolution of *bicycle* genes**
364
365 We were unable to detect any sequence homology between *bicycle* genes and
366 previously reported genes, which is one reason that it is difficult to infer the molecular function
367 of *bicycle* genes from sequence alone. It is possible that *bicycle* genes evolved *de novo* in an
368 ancestor of gall-forming aphids, perhaps through capture of a 5' exon encoding an N-terminal
369 signal sequence. However, if *bicycle* genes experienced strong diversifying selection since their

origin, perhaps in an ancestor of gall forming aphids about 280 MYA, then the rate of amino-acid substitution that we detected between two closely-related species would likely be sufficient to have eliminated sequence similarity that could be detected by homology-detection algorithms. Identifying the evolutionary antecedents of *bicycle* genes will likely require tracing their evolutionary history across genomes of related species. It may also be more fruitful to use the unusual gene structure of *bicycle* genes (Fig. S10F, G) to search for the antecedents of *bicycle* genes.

References:

Bailey, R., Schönrogge, K., Cook, J.M., Melika, G., Csóka, G., Thuróczy, C., and Stone, G.N. (2009). Host Niches and Defensive Extended Phenotypes Structure Parasitoid Wasp Communities. PLOS Biology *7*, e1000179.

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

Blunden, G., and Challen, S.B. (1965). Red pigment in leaf galls of Salix fragilis L. Nature *208*, 388–389.

Bomfim, P.M.S., Cardoso, J.C.F., Rezende, U.C., Martini, V.C., and Oliveira, D.C. (2019). Red galls: the different stories of two gall types on the same host. Plant Biology *21*, 284–291.

Cambier, S., Ginis, O., Moreau, S.J.M., Gayral, P., Hearn, J., Stone, G.N., Giron, D., Huguet, E., and Drezen, J.-M. (2019). Gall Wasp Transcriptomes Unravel Potential Effectors Involved in Molecular Dialogues With Oak and Rose. Front. Physiol. *10*, 926.

Chen, M.-S., Liu, X., Yang, Z., Zhao, H., Shukle, R.H., Stuart, J.J., and Hulbert, S. (2010). Unusual conservation among genes encoding small secreted salivary gland proteins from a gall midge. BMC Evolutionary Biology *10*, 296–296.

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

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

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

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

405 Eitle, M.W., Carolan, J.C., Griesser, M., and Forneck, A. (2019). The salivary gland proteome of
406 root-galling grape phylloxera (Daktulosphaira vitifoliae Fitch) feeding on Vitis spp. PLoS ONE *14*,
407 e0225881.

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

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

412 Hearn, J., Blaxter, M., Schönrogge, K., Nieves-Aldrey, J.-L., Pujade-Villar, J., Huguet, E., Drezen,
413 J.-M., Shorthouse, J.D., and Stone, G.N. (2019). Genomic dissection of an extended phenotype:
414 Oak galling by a cynipid gall wasp. PLOS Genetics *15*, e1008398.

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

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

419 Hugueney, P., Provenzano, S., Verriès, C., Ferrandino, A., Meudec, E., Batelli, G., Merdinoglu, D.,
420 Cheynier, V., Schubert, A., and Ageorges, A. (2009). A novel cation-dependent o-
421 methyltransferase involved in anthocyanin methylation in grapevine. Plant Physiology *150*,
422 2057–2070.

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

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

427 Leatherdale, D. (1955). Plant hyperplasia induced with a cell-free insect extract. Nature *175*,
428 553–554.

429 Lewis, I.F., and Walton, L. (1947). Initiation of the cone gall of witch hazel. Science *106*, 419–
430 420.

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

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

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

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

Martinson, E., Werren, J., and Egan, S. (2020). Tissue-specific gene expression shows cynipid wasps repurpose host gene networks to create complex and novel parasite-specific organs on oaks (Preprints).

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

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

Obbard, D.J., Jiggins, F.M., Halligan, D.L., and Little, T.J. (2006). Natural selection drives 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.

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

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

468 Smith, N.G., and Eyre-Walker, A. (2002). Adaptive protein evolution in Drosophila. Nature *415*,
469 1022–1024.

470 Smits, B.L., and Peterson, W.J. (1942). Carotenoids of Telial Galls of Gymnosporangium Juniperi-
471 Virginianae Lk. Science *96*, 210–211.

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

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

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

478 Stone, G.N., and Schönrogge, K. (2003). The adaptive significance of insect gall morphology.
479 Trends in Ecology and Evolution *18*, 512–522.

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

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

485 Takeda, S., Yoza, M., Amano, T., Ohshima, I., Hirano, T., Sato, M.H., Sakamoto, T., and Kimura, S.
486 (2019). Comparative transcriptome analysis of galls from four different host plants suggests the
487 molecular mechanism of gall development. PloS One *14*, e0223686–e0223686.

488 Tanaka, Y., Okada, K., Asami, T., and Suzuki, Y. (2013). Phytohormones in Japanese Mugwort
489 Gall Induction by a Gall-Inducing Gall Midge. Bioscience, Biotechnology and Biochemistry *77*,
490 1942–1948.

491 Tooker, J.F., and Helms, A.M. (2014). Phytohormone Dynamics Associated with Gall Insects, and
492 their Potential Role in the Evolution of the Gall-Inducing Habit. Journal of Chemical Ecology *40*,
493 742–753.

494 Yamaguchi, H., Tanaka, H., Hasegawa, M., Tokuda, M., Asami, T., and Suzuki, Y. (2012).
495 Phytohormones and willow gall induction by a gall-inducing sawfly. New Phytologist *196*, 586–
496 595.

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

Zhao, C., Escalante, L.N., Chen, H., Benatti, T.R., Qu, J., Chellapilla, S., Waterhouse, R.M., Wheeler, D., Andersson, M.N., Bao, R., et al. (2015). A Massive Expansion of Effector Genes Underlies Gall-Formation in the Wheat Pest Mayetiola destructor. Current Biology *25*, 613–620.

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:

HHMI has filed a provisional patent, number 63/092,942, for the inventors AK and DLS covering unique aphid polypeptides for use in modifying plants.

Figure Legends

Fig. 1. *Hormaphis cornu* aphids drive patterned over-proliferation of plant cells to produce galls of on leaves *Hamamelis virginiana*

521 (A-C) Photographs of the abaxial surfaces of *H. virginiana* leaves being attacked by first-instar
522 fundatrices of *H. cornu*. Nymphs gather on the unopened leaf buds (A) and soon after bud
523 break the fundatrices gather near leaf veins (B) and inject material that begins to transform the
524 leaf into a gall (C). Magnified region in blue rectangle of panel (A) shows three fundatrices
525 waiting on unopened bud (A').
526 (D-F) Photographs of the adaxial leaves of *H. virginiana*, showing galls at early (D), middle (E),
527 and late (F) growth stages.
528 (G-I) Confocal micrographs of sections through a *H. cornu* gall (G,H) and un-galled *H. virginiana*
529 leaf (I) stained with Congo red and calcofluor white. Extensive hypertropy is observed in the
530 mesophyll (yellow arrows) at a considerable distance from the location of the aphid's stylets
531 (blue arrows) (G). The tips of aphid stylets (pink arrow) can be observed within cells of *H.*
532 *virginiana* and plant tissue shows evidence of hyperproliferation and periclinal divisions (grey
533 arrows) close to the tips of stylets and the termini of stylet sheaths (H). Periclinal divisions are
534 observed in both spongy and palisade mesophyll cells during gall development, but never in
535 ungalled leaf tissue (I). Plant tissue is presented in the aphid's frame of reference, with abaxial
536 leaf surface up.
537
538 **Figure 2. A genome-wide association study (GWAS) identifies variation within a novel aphid**
539 **gene associated with gall color**
540 (A) A GWAS of fundatrices isolated from 43 green and 47 red galls identified a small region on
541 chromosome 1 of the *H. cornu* genome that is strongly associated with gall color. Red line
542 indicates FDR = 0.05. Colors of points on chromosomes are arbitrary.

543 (B-D) Resequencing 800 kbp of Chromosome 1 centered on the most significant SNPs from the
544 original GWAS to approximately 60X coverage identified 11 spatially clustered SNPs significantly
545 associated with gall color located within the introns and upstream of *Horco_g16073*, which was
546 named *dgc* (D). (Some SNPs are closely adjacent and cannot be differentiated at this scale.)
547 Significant SNPs are indicated with orange vertical lines in (D).
548 (E) Genotypes of all 11 SNPs associated with gall color from an independent sample of aphids
549 from 435 green and 431 red galls. Color of gall for aphid samples shown on left and genotype at
550 each SNP is shown adjacent in green (0/0, homozygous ancestral state), yellow (0/1,
551 heterozygous), or red (1/1, homozygous derived state). LOD scores for association with gall
552 color shown for each SNP at bottom of genotype plot. Histogram of frequencies of each
553 multilocus genotype ordered by frequency and collected within gall color is shown on the right.
554 All 11 SNPs are strongly associated with gall color ($P < 10^{-192}$), and a cluster of 5 SNPs in a 61bp
555 region are most strongly associated with gall color. Individuals homozygous for ancestral alleles
556 at all or most loci and making red galls likely carry variants at other loci that influence gall color
557 (Supplementary Text).
558
559 **Figure 3. *Dgc* is the most strongly differentially expressed gene in salivary glands of**
560 **fundatrices collected from red versus green galls**
561 (A) Genome-wide differential expression analysis of *H. cornu* fundatrix salivary gland transcripts
562 from individuals heterozygous for dgc^{Red}/dgc^{Green} (Red; N = 15) versus homozygous for dgc^{Green}
563 (Green; N = 7) illustrated as a volcano plot. D*gc* is strongly downregulated in dgc^{Red}/dgc^{Green}
564 fundatrices.

565 (B) Salivary gland expression of genes in the paralog cluster that includes *dgc*^{Red}/*dgc*^{Green}. Gene
566 models of the paralogous genes are shown at the top and expression levels normalized by
567 mean expression across all paralogs is shown below. Only *dgc* is strongly downregulated in this
568 gene cluster between individuals carrying *dgc*^{Red}/*dgc*^{Green} (Red) versus *dgc*^{Green}/*dgc*^{Green} (Green)
569 genotypes.
570 (C) Genome-wide differential expression analysis of *H. virginiana* transcripts isolated from galls
571 made by fundatrices heterozygous for *dgc*^{Red}/*dgc*^{Green} (Red; N = 17) versus homozygous for
572 *dgc*^{Green} (Green; N = 23) illustrated as a volcano plot. Only eight genes are differentially
573 expressed at FDR < 0.05, and all are overexpressed in red galls. The seven most strongly
574 differentially expressed genes that encode anthocyanin biosynthetic enzymes (FAOMT-1 =
575 Hamvi_23591; FAOMT-2 = Hamvi_7147 ; GSTF11 = Hamvi_134919; GSTF12 = Hamvi_109682;
576 UGT75C1 = Hamvi_14194; UFGT = Hamvi_22774; ACCA = Hamvi_97071).
577 (D) Expression levels, in counts per million reads, of the seven anthocyanin biosynthetic genes
578 overexpressed in red galls, in green (green) and red (red) galls and ungalled leaves (blue). Each
579 data point is from a separate genome-wide RNA-seq sample.
580 (E) Simplified diagram of the anthocyanin biosynthetic pathway. Enzyme classes upregulated in
581 red galls are shown in purple font. The two terminal anthocyanins that generate the red color in
582 galls, peonidin-3,5-diglucoside and malvidin-3,5-diglucoside, are shown in bold font, and three
583 precursor molecules found in red galls are shown in bold italic font. Anthocyanin names are
584 abbreviated (dg = 3,5-diglucoside).
585 (F) Photos of cross-sections of green (top left) and red (top right) and the pigments extracted
586 from green and red galls (below).

587 (G) UHPLC-DAD chromatograms at 520 nm of extract from red (red line) and green (green line)
588 galls.
589 (H) Overlaid UHPLC-MS chromatograms of green (top) and red (middle) gall extracts and
590 authentic standards (bottom). Each pigment is indicated with a different color: green =
591 delphinidin-3,5-diglucoside (*m/z* = 627.1551); black = cyanidin-3,5-diglucoside (*m/z* = 611.1602);
592 blue = petunidin-3,5-diglucoside (*m/z* = 641.1709); purple = peonidin-3,5-diglucoside (*m/z* =
593 625.1768); and red = malvidin-3,5-diglucoside (*m/z* = 655.1870). Anthocyanin names are
594 abbreviated (dg = 3,5-diglucoside). Peonidin-3,5-diglucoside and malvidin-3,5-diglucoside
595 together account for 87% of pigment detected in red galls.
596
597 **Figure 4. *bicycle* genes are salivary gland enriched transcripts of gall-associated *H. cornu***
598 **generations**
599 (A) Diagram of life cycle of *H. cornu*. *H. cornu* migrates annually between *H. virginiana* (blue
600 line) and *Betula nigra* (pink line) and the gall is produced only in the spring on *H. virginiana*
601 (brown line). Each nymph of the first generation (G1, the fundatrix) hatches from an over-
602 wintering egg and initiates development of a single gall. Her offspring (G2) feed and grow
603 within the gall and develop with wings, which allows them to fly to *B. nigra* in late spring. For
604 three subsequent generations (G3-G5) the aphids develop as small, coccidiform morphs on *B.*
605 *nigra*. In the fall, aphids develop with wings (G6), fly to *H. virginiana* plants, and deposit male
606 and female sexuals (G7), the only generation possessing meiotic cells. The sexuals feed and
607 complete development on the senescing leaves of *H. virginiana*. As adults they mate and the
608 females deposit eggs that overwinter and give rise to fundatrices the following spring.

609 (B) Differential expression of fundatrix versus sexual salivary glands with only genes significantly
610 upregulated in fundatrix salivary glands marked with colors, shown as a volcano plot. Genes
611 with and without homologs in public databases are labeled as “Annotated” (blue) and “Not
612 Annotated” (brown), respectively.
613 (C) Hierarchical clustering of unannotated salivary-gland specific genes reveals one large cluster
614 of *bicycle* genes, one small cluster of *CWG* genes, and many largely unique, unclustered genes.
615 (D) *Bicycle* (green) and remaining unannotated (brown) genes labelled on a differential
616 expression volcano plot illustrates that, on average, *bicycle* genes are the most strongly
617 differentially expressed genes expressed specifically in fundatrix salivary glands.
618 (E) *Bicycle* genes are expressed at highest levels in the salivary glands of the fundatrix and
619 mostly decline in expression during subsequent generations. (Sample sizes: G1 (N = 20); G2 (N =
620 4); G5 (N = 6); and G7 (N = 6).
621 (F) Amino-acid logo for predicted BICYCLE proteins. Sequence alignment used for logo was
622 filtered to highlight conserved positions. Most genes encode proteins with an N-terminal signal
623 sequence and a pair of conserved cysteine-tyrosine-cysteine motifs (CYC).
624

Figure 1

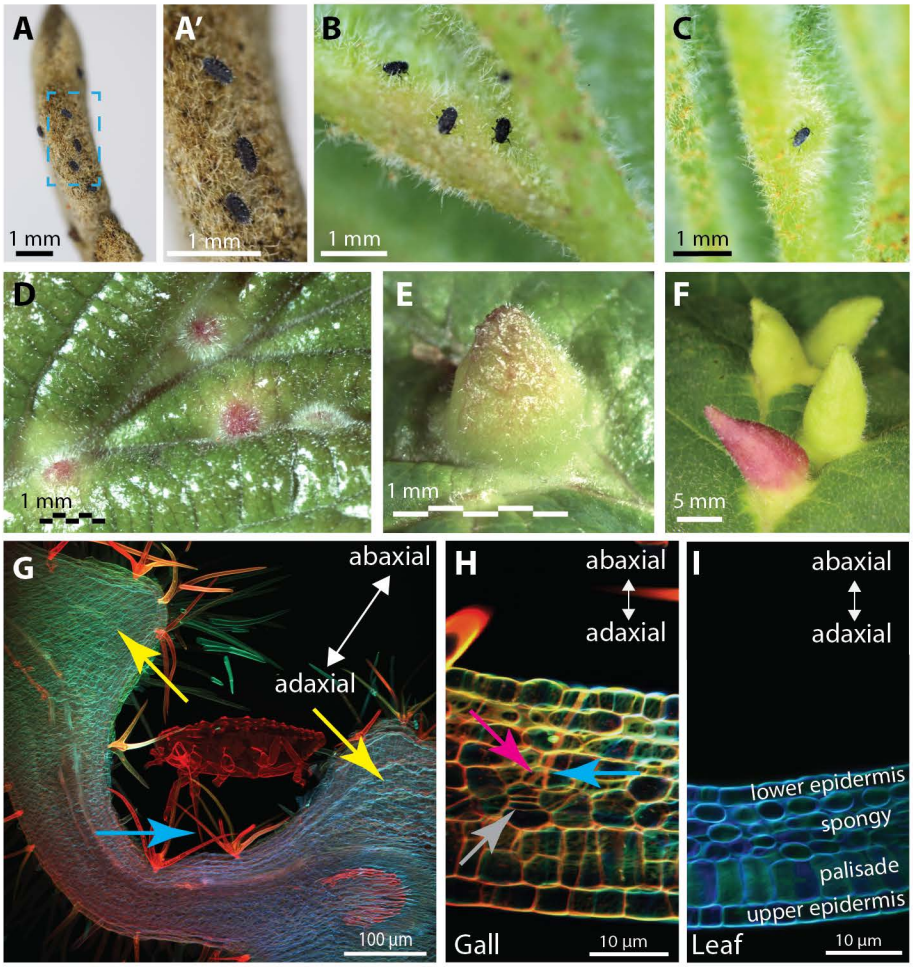

A
1 mm
A′
1 mm
B
1 mm
C
1 mm
D
1 mm
E
1 mm
F
5 mm
G
abaxial
adaxial
100 µm
H
abaxial
adaxial
Gall
10 µm
I
abaxial
adaxial
lower epidermis
spongy
palisade
upper epidermis
Leaf
10 µm

Figure 2

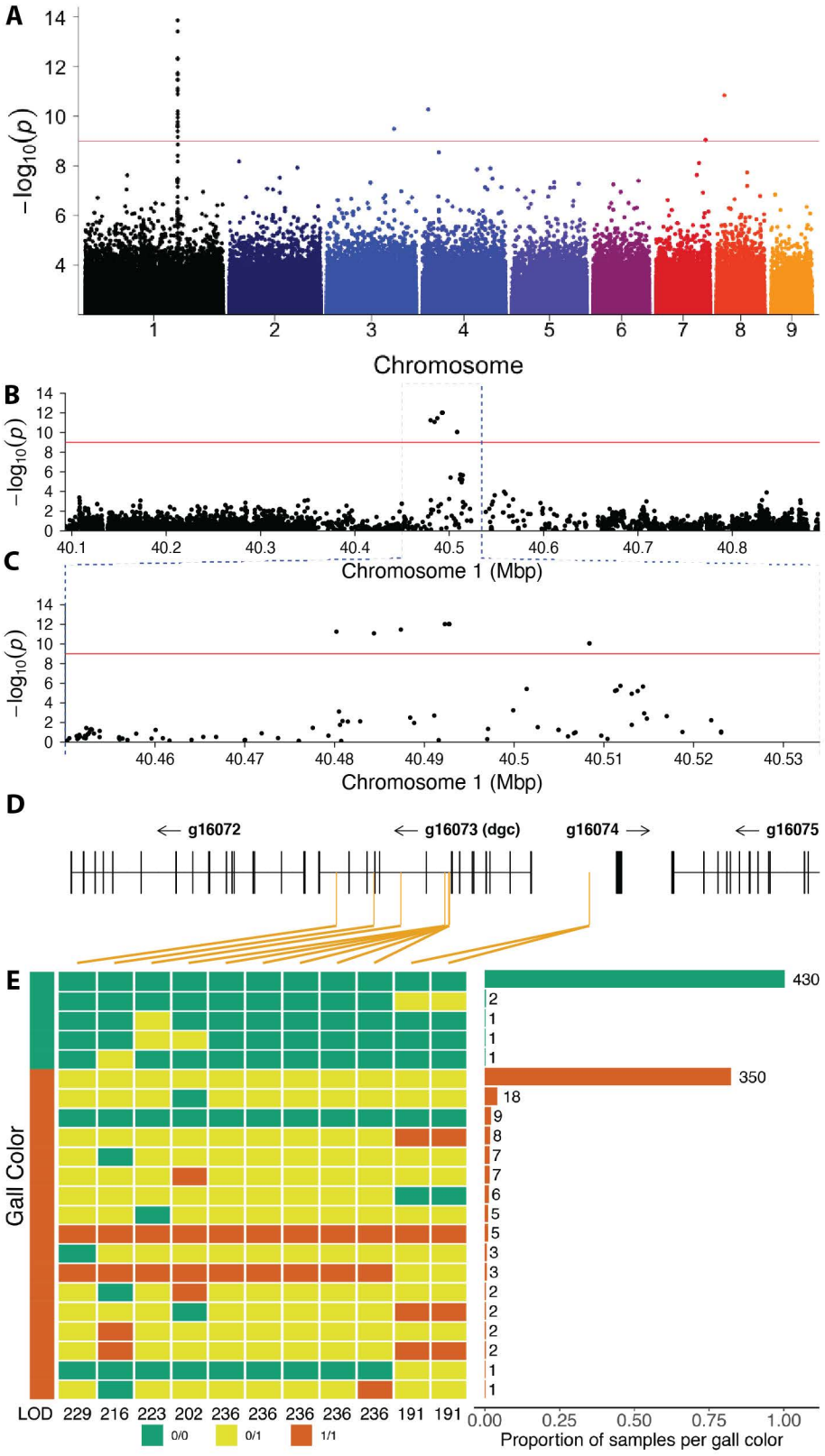

A
−log10(p)
Chromosome
B
−log10(p)
Chromosome 1 (Mbp)
C
−log10(p)
Chromosome 1 (Mbp)
D
g16072
g16073 (dgc)
g16074
g16075
E
Gall Color
LOD 229 216 223 202 236 236 236 236 236 191 191
0/0
0/1
1/1
Proportion of samples per gall color

Figure 3

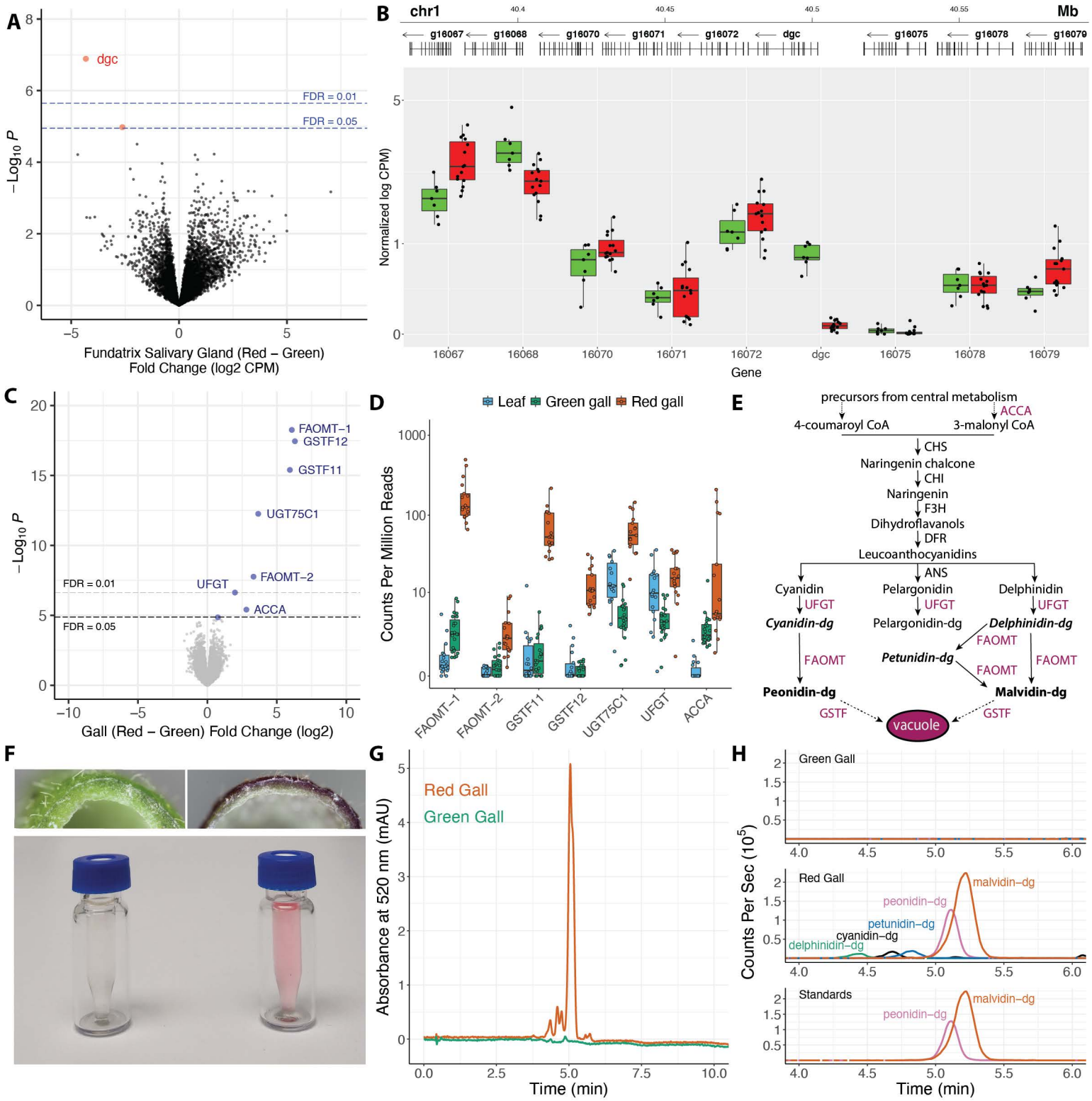
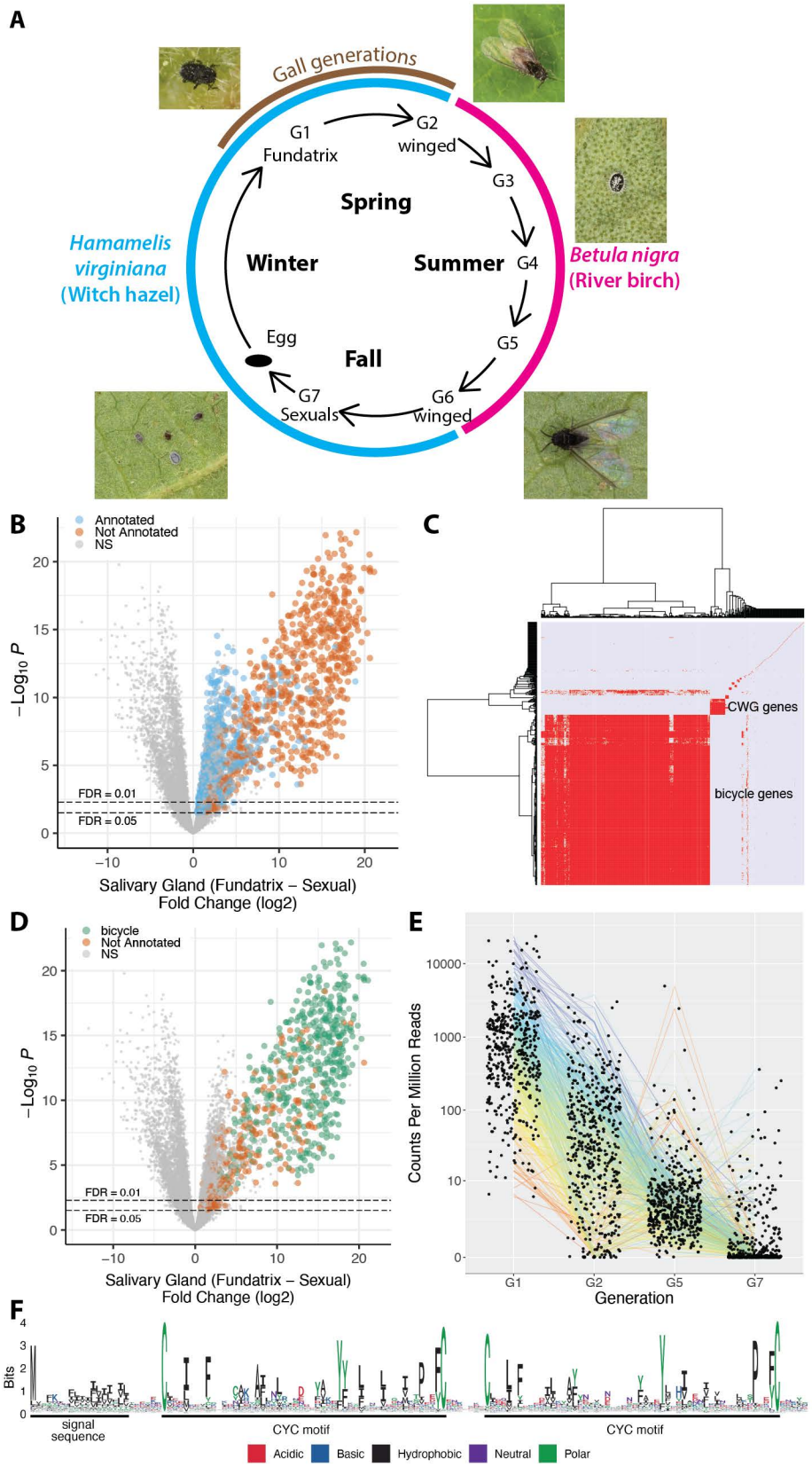

A
dgc
FDR = 0.01
FDR = 0.05
−Log10 P
Fundatrix Salivary Gland (Red – Green)
Fold Change (log2 CPM)
B
chr1
Mb
g16067
g16068
g16070
g16071
g16072
dgc
g16075
g16078
g16079
Normalized log CPM)
Gene
C
FAOMT−1
GSTF12
GSTF11
UGT75C1
FAOMT−2
UFGT
ACCA
FDR = 0.01
FDR = 0.05
−Log10 P
Gall (Red – Green) Fold Change (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

STAR Methods

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, David L. Stern (sternd@hhmi.org)

Material Availability

This study did not generate new unique reagents.

Data and Code Availability

All sequencing data generated during this study is available at the NCBI Short Read Archive and the accession numbers are provided for each sample in the tables below.
The genomes are available at Genbank and the genomes and gene annotations (gff files) are available at FigShare.

All of the computer code we produced for this study is freely available at GitHub and links to each piece of code are provided in the detailed descriptions of methods.

647 **Table STAR Methods 1. Details of biological samples, sequence files, and SRA accession**
648 **numbers for genome sequencing.**
649

Sample #	Collection Date	Collection Location	Species	Library Prep Method	Sequencing Method	# Read pairs	Accession #
MP2	15-Jun-2016	Morven Park, Leesburg, VA	*Hormaphis cornu*	10X Chromium	Illumina PE151	408,093,791	SRR11397593
MP2	15-Jun-2016	Morven Park, Leesburg, VA	*Hormaphis cornu*	Chicago	Illumina PE151	234,423,934	SRR11397592
80	31-May-2017	Gap Road, Rt. 651, Leesburg, VA	*Hormaphis cornu*	HiC	Illumina PE151	281,999,702	SRR11397591
ML97Red	7-Jun-2018	Mountain Lakes Biological Station, Pembroke, VA	*Hormaphis hamamelidis*	10X Chromium	Illumina PE151	91,611,428	SRR13258507
Hamamelis virginiana nuclei	17-Apr-2019	Janelia Research Campus, Ashburn, VA	*Hamamelis virginiana*	10X Chromium	Illumina PE150	303,926,536	SRR11397589

650

651 **Table STAR Methods 2. Details of biological samples, sequence files, and SRA accession**
652 **numbers for *H. cornu* RNA-sequencing. WGB refers to Whole Gene Body library**
653 **preparation method described in Supplemental Methods. SS refers to the Smart-SCRB**
654 **library preparation methods described in (Cembrowski et al., 2018). JRC = Janelia**
655 **Research Campus, Ashburn, VA.**
656

Sample ID	Collection Location	Collection Date	Stage	Generation	Organ	Library Prep Method	Sequencing Method	# Read pairs	Accession #
1st_I_FundSG	JRC	6-4-2016	1	1	SG	WGB	Illumina PE100	12643792, 33584393	SRR11428787
1st_II_FundSG	JRC	6-4-2016	1	1	SG	WGB	Illumina PE100	12412280, 29986671	SRR11428786
1st_III_FundSG	JRC	6-4-2016	1	1	SG	WGB	Illumina PE100	9566012, 28830029	SRR11428769
early2nd_I_FundSG	JRC	6-4-2016	2	1	SG	WGB	Illumina PE100	11846803, 31258786	SRR11428758
early2nd_II_FundSG	JRC	6-4-2016	2	1	SG	WGB	Illumina PE100	14097211, 33244826	SRR11428747
early2nd_III_FundSG	JRC	6-4-2016	2	1	SG	WGB	Illumina PE100	10610077, 26423617	SRR11428736
late2nd_I_FundSG_rep1	JRC	6-4-2016	2	1	SG	WGB	Illumina PE100	11783417, 29867606	SRR11428725
late2nd_II_FundSG_rep1	JRC	6-4-2016	2	1	SG	WGB	Illumina PE100	14189826, 37094462	SRR11428714
late2nd_III_FundSG_rep1	JRC	6-4-2016	2	1	SG	WGB	Illumina PE100	12137125, 30412534	SRR11428779
late2nd_IV_FundSG	JRC	6-4-2016	2	1	SG	WGB	Illumina PE100	34021474	SRR11428778
late2nd_V_FundSG	JRC	6-4-2016	2	1	SG	WGB	Illumina PE100	33255973	SRR11428785
late2nd_VI_FundSG	JRC	6-4-2016	2	1	SG	WGB	Illumina PE100	29588804	SRR11428783
3rd_I_FundSG	JRC	29-4-2016	2	1	SG	WGB	Illumina PE100	30833668	SRR11428777
3rd_II_FundSG	JRC	29-4-2016	2	1	SG	WGB	Illumina PE100	33651454	SRR11428776
3rd_III_FundSG	JRC	29-4-2016	2	1	SG	WGB	Illumina PE100	40510300	SRR11428775
4th_I_FundSG	JRC	3-5-2016	2	1	SG	WGB	Illumina PE100	36267453	SRR11428774
4th_II_FundSG	JRC	3-5-2016	2	1	SG	WGB	Illumina PE100	37712709	SRR11428773
4th_III_FundSG	JRC	3-5-2016	2	1	SG	WGB	Illumina PE100	37200793	SRR11428772
adult_I_FundSG	JRC	3-5-2016	2	1	SG	WGB	Illumina PE100	37826115	SRR11428771
adult_II_FundSG	JRC	3-5-2016	2	1	SG	WGB	Illumina PE100	31970558	SRR11428770
adult_III_FundSG	JRC	4-5-2016	2	1	SG	WGB	Illumina PE100	29795163	SRR11428768

whole_1st_I	JRC	6-4-2016	1	1	Whole	WGB	Illumina PE100	36313971	SRR11428767
whole_early2nd_I_Fund	JRC	6-4-2016	2	1	Whole	WGB	Illumina PE100	44067290	SRR11428766
whole_late2nd_I_Fund	JRC	6-4-2016	2	1	Whole	WGB	Illumina PE100	32843185	SRR11428765
whole_v_late2nd_I_Fund	JRC	6-4-2016	2	1	Whole	WGB	Illumina PE100	35509383	SRR11428764
whole_3rd_I_Fund	JRC	29-4-2016	3	1	Whole	WGB	Illumina PE100	39418696	SRR11428763
whole_3rd_II_Fund	JRC	29-4-2016	3	1	Whole	WGB	Illumina PE100	40421752	SRR11428762
whole_4th_I_Fund	JRC	3-5-2016	4	1	Whole	WGB	Illumina PE100	38585734	SRR11428761
whole_adult_I_Fund	JRC	4-5-2016	Ad	1	Whole	WGB	Illumina PE100	42742613	SRR11428760
2nd_gen_1st inst_SG1	JRC	2-6-2018	1	2	SG	WGB	Illumina PE150	11405653	SRR11428759
2nd_gen_1st inst_SG2	JRC	2-6-2018	1	2	SG	WGB	Illumina PE150	10794548	SRR11428757
2nd_gen_1st inst_SG3	JRC	2-6-2018	1	2	SG	WGB	Illumina PE150	12574061	SRR11428756
2ndgen_3rdinst_sg_I_1	JRC	19-5-2016	3	2	SG	WGB	Illumina PE100	20922005	SRR11428755
2ndgen_3rdinst_sg_II_1	JRC	19-5-2016	3	2	SG	WGB	Illumina PE100	19238796	SRR11428754
2ndgen_3rdinst_sg_III_1	JRC	19-5-2016	3	2	SG	WGB	Illumina PE100	19740444	SRR11428753
2ndgen_3rdinst_whole_I_1	JRC	19-5-2016	3	2	Whole	WGB	Illumina PE100	18321283	SRR11428752
2ndgen_3rdinst_whole_II_1	JRC	19-5-2016	3	2	Whole	WGB	Illumina PE100	18971169	SRR11428751
2ndgen_3rdinst_whole_III_1	JRC	19-5-2016	3	2	Whole	WGB	Illumina PE100	19463547	SRR11428750
2nd_gen_Ad_SG4	JRC	2-6-2018	3	Ad	SG	WGB	Illumina PE150	12714848	SRR11428749
2nd_gen_Ad_SG5	JRC	2-6-2018	3	Ad	SG	WGB	Illumina PE150	10858738	SRR11428748
2nd_gen_Ad_SG6	JRC	2-6-2018	3	Ad	SG	WGB	Illumina PE150	13120645	SRR11428746
ex_bet_AdSG8	JRC	8-7-2018	Ad	Ex Betula	SG	WGB	Illumina PE150	4914246, 13175533	SRR11428745
ex_bet_4thSG9	JRC	8-7-2018	4	Ex Betula	SG	WGB	Illumina PE150	6611299, 18243185	SRR11428744
ex_bet_4thSG10	JRC	8-7-2018	4	Ex Betula	SG	WGB	Illumina PE150	6578419, 17834081	SRR11428743
ex_bet_AdSG11	JRC	8-7-2018	Ad	Ex Betula	SG	WGB	Illumina PE150	5627200, 10425494	SRR11428742
ex_bet_AdSG12	JRC	8-7-2018	Ad	Ex Betula	SG	WGB	Illumina PE150	6915842, 13798518	SRR11428741

ex_bet_AdSG13	JRC	8-7-2018	Ad	Ex Betula	SG	WGB	Illumina PE150	7365598, 14951729	SRR11428740
ex_bet_nymphs_15	JRC	8-7-2018	4	Ex Betula	SG	WGB	Illumina PE150	11106332	SRR11428739
ex_bet_nymphs_16	JRC	8-7-2018	4	Ex Betula	SG	WGB	Illumina PE150	12481764	SRR11428738
sex_SG_19_1	JRC	22-10-2019	Nymph	Sexual	SG	WGB	Illumina PE150	9662029	SRR11428737
sex_SG_19_2	JRC	22-10-2019	Nymph	Sexual	SG	WGB	Illumina PE150	12690319	SRR11428735
sex_SG_19_3	JRC	22-10-2019	Nymph	Sexual	SG	WGB	Illumina PE150	10875978	SRR11428734
sex_SG_19_4	JRC	22-10-2019	Nymph	Sexual	SG	WGB	Illumina PE150	9540662	SRR11428733
sex_SG_19_5	JRC	22-10-2019	Nymph	Sexual	SG	WGB	Illumina PE150	9632389	SRR11428732
D02	JRC	18-11-2018	Nymph	Sexual	3 pooled heads	SS	Illumina SE125	23182206	SRR11428731
E02	JRC	18-11-2018	Nymph	Sexual	3 pooled bodies	SS	Illumina SE125	23157666	SRR11428730
F02	JRC	18-11-2018	Nymph	Sexual	4 pooled salivary glands	SS	Illumina SE125	24353653	SRR11428729
G02	JRC	18-11-2018	Nymph	Sexual	3 pooled salivary glands	SS	Illumina SE125	21591985	SRR11428728
H02	JRC	18-11-2018	Nymph	Sexual	3 pooled salivary glands	SS	Illumina SE125	17705013	SRR11428727

659 **Table STAR Methods 3. Details of biological samples, sequence files, and SRA accession**
660 **numbers for *H. cornu* GWAS whole-genome re-sequencing.**
661

Sample #	Collection Location	Collection Date	Gall Color	Library Prep Method	Sequencing Method	# Read pairs	Accession #
B02L02-gr	Moore St & Guyot Ave., Princeton, NJ	10-May-2017	Green	Tn5	Illumina PE150	6366491	SRR11363255
B03L03	Moore St & Guyot Ave., Princeton, NJ	10-May-2017	Green	Tn5	Illumina PE150	4593475	SRR11363254
GCL74	Goose Creek Lane, Leesburg, VA	31-May-2017	Green	Tn5	Illumina PE150	4809727	SRR11363163
HW03L02	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	3884184	SRR11363152
HW03L07	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	8469155	SRR11363141
HW04L01	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	4589629	SRR11363130
HW04L07	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	6432189	SRR11363119
HW05L02	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	5734918	SRR11363108
HW06L01G01	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	5168885	SRR11363097
HW07L01	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	5796597	SRR11363086
HW08L03	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	5326728	SRR11363253
HW09L05	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	3783083	SRR11363210
HW10L01	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	3390296	SRR11363199
HW11L01	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	6969888	SRR11363188
HW02L05	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	4714807	SRR11363241

HWL04	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	5643363	SRR11363230
HWL07	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	4917264	SRR11363187
JRC03	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Green	Tn5	Illumina PE150	4863866	SRR11363176
JRC04	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Green	Tn5	Illumina PE150	4523085	SRR11363165
JRC10	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Green	Tn5	Illumina PE150	5182279	SRR11363164
JRC12	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Green	Tn5	Illumina PE150	4178879	SRR11363162
JRC13	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Green	Tn5	Illumina PE150	1985664	SRR11363161
JRC15	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Green	Tn5	Illumina PE150	4191677	SRR11363160
JRC17	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Green	Tn5	Illumina PE150	5230834	SRR11363159
JRC19	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Green	Tn5	Illumina PE150	4844552	SRR11363158
JRC20	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Green	Tn5	Illumina PE150	4750713	SRR11363157
JRC27	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Green	Tn5	Illumina PE150	4868486	SRR11363156
MLR36	Man Lu Ridge, Maryland	16-May-2017	Green	Tn5	Illumina PE150	3547891	SRR11363155
MLR37	Man Lu Ridge, Maryland	16-May-2017	Green	Tn5	Illumina PE150	3401601	SRR11363154
MP55	Morven Park, Leesburg, VA	20-May-2017	Green	Tn5	Illumina PE150	4030936	SRR11363153
MP56	Morven Park, Leesburg, VA	20-May-2017	Green	Tn5	Illumina PE150	5264763	SRR11363151

OWR61	Old Waterford Road, Leesburg, VA	20-May-2017	Green	Tn5	Illumina PE150	4227643	SRR11363150
OWR73	Old Waterford Road, Leesburg, VA	30-May-2017	Green	Tn5	Illumina PE150	4276012	SRR11363149
POR29	Point of Rocks, Maryland	16-May-2017	Green	Tn5	Illumina PE150	3962098	SRR11363148
SM01L01	St. Michael's Reserve, Princeton, NJ	1-Jun-2017	Green	Tn5	Illumina PE150	4883070	SRR11363147
SM02L01	St. Michael's Reserve, Princeton, NJ	1-Jun-2017	Green	Tn5	Illumina PE150	4381921	SRR11363146
SM03L01	St. Michael's Reserve, Princeton, NJ	1-Jun-2017	Green	Tn5	Illumina PE150	4311021	SRR11363145
SM05L01	St. Michael's Reserve, Princeton, NJ	1-Jun-2017	Green	Tn5	Illumina PE150	4506282	SRR11363144
SM05L05	St. Michael's Reserve, Princeton, NJ	1-Jun-2017	Green	Tn5	Illumina PE150	3972013	SRR11363143
W02L01G01	Woodfield Reservation, Princeton, NJ	26-May-2017	Green	Tn5	Illumina PE150	5347819	SRR11363142
W03L01	Woodfield Reservation, Princeton, NJ	26-May-2017	Green	Tn5	Illumina PE150	4536267	SRR11363140
W05L03G02	Woodfield Reservation, Princeton, NJ	26-May-2017	Green	Tn5	Illumina PE150	6397623	SRR11363139
W06L02	Woodfield Reservation, Princeton, NJ	26-May-2017	Green	Tn5	Illumina PE150	6315684	SRR11363138
B01L01	Moore St & Guyot Ave., Princeton, NJ	10-May-2017	Red	Tn5	Illumina PE150	8135750	SRR11363137
B01L03	Moore St & Guyot Ave., Princeton, NJ	10-May-2017	Red	Tn5	Illumina PE150	4429167	SRR11363136
B02L01	Moore St & Guyot Ave., Princeton, NJ	10-May-2017	Red	Tn5	Illumina PE150	4620277	SRR11363135
B02L02-red	Moore St & Guyot Ave., Princeton, NJ	10-May-2017	Red	Tn5	Illumina PE150	5625715	SRR11363134
B03L01	Moore St & Guyot Ave., Princeton, NJ	10-May-2017	Red	Tn5	Illumina PE150	10312327	SRR11363133
HW02L02	Herrontown Woods, Princeton, NJ	31-May-2017	Red	Tn5	Illumina PE150	4039449	SRR11363132

HW03L01	Herrontown Woods, Princeton, NJ	31-May-2017	Red	Tn5	Illumina PE150	4130302	SRR11363131
HW03L03	Herrontown Woods, Princeton, NJ	31-May-2017	Red	Tn5	Illumina PE150	3384712	SRR11363129
HW03L04	Herrontown Woods, Princeton, NJ	31-May-2017	Red	Tn5	Illumina PE150	3904224	SRR11363128
HW06L01G02	Herrontown Woods, Princeton, NJ	31-May-2017	Red	Tn5	Illumina PE150	5403059	SRR11363127
HW07L02	Herrontown Woods, Princeton, NJ	31-May-2017	Red	Tn5	Illumina PE150	8799194	SRR11363126
HW09L02	Herrontown Woods, Princeton, NJ	31-May-2017	Red	Tn5	Illumina PE150	3673337	SRR11363125
HW09L03	Herrontown Woods, Princeton, NJ	31-May-2017	Red	Tn5	Illumina PE150	5235539	SRR11363124
HW09L07	Herrontown Woods, Princeton, NJ	31-May-2017	Red	Tn5	Illumina PE150	8256988	SRR11363123
HW10L02	Herrontown Woods, Princeton, NJ	31-May-2017	Red	Tn5	Illumina PE150	5492996	SRR11363122
HWL03	Herrontown Woods, Princeton, NJ	31-May-2017	Red	Tn5	Illumina PE150	7814556	SRR11363121
JRC01	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Red	Tn5	Illumina PE150	4617289	SRR11363120
JRC06	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Red	Tn5	Illumina PE150	5709657	SRR11363118
JRC11	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Red	Tn5	Illumina PE150	4441152	SRR11363117
JRC14	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Red	Tn5	Illumina PE150	4255537	SRR11363116
JRC16	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Red	Tn5	Illumina PE150	7587374	SRR11363115
JRC18	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Red	Tn5	Illumina PE150	4944638	SRR11363114

JRC22	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Red	Tn5	Illumina PE150	6692470	SRR11363113
JRC23	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Red	Tn5	Illumina PE150	4475288	SRR11363112
JRC25	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Red	Tn5	Illumina PE150	5506970	SRR11363111
JRC84	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	7362956	SRR11363110
JRC85	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	7150172	SRR11363109
JRC86	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	6592107	SRR11363107
JRC87	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	6071483	SRR11363106
JRC88	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	3476862	SRR11363105
JRC89	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	6474867	SRR11363104
JRC90	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	7186265	SRR11363103
JRC91	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	6643361	SRR11363102
JRC92	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	4674510	SRR11363101
JRC93	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	5182706	SRR11363100
JRC94	Janelia Research	21-May-2017	Red	Tn5	Illumina PE150	4946068	SRR11363099

	Campus, Ashburn, VA						
JRC95	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	7026186	SRR11363098
JRC96	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	4946848	SRR11363096
JRC97	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	6115771	SRR11363095
OWR71	Old Waterford Road, Leesburg, VA	30-May-2017	Red	Tn5	Illumina PE150	5324101	SRR11363094
POR30	Point of Rocks, Maryland	16-May-2017	Red	Tn5	Illumina PE150	3709000	SRR11363093
POR31	Point of Rocks, Maryland	16-May-2017	Red	Tn5	Illumina PE150	6515377	SRR11363092
SM05L02	St. Michael's Reserve, Princeton, NJ	1-Jun-2017	Red	Tn5	Illumina PE150	5328410	SRR11363091
SM05L04	St. Michael's Reserve, Princeton, NJ	1-Jun-2017	Red	Tn5	Illumina PE150	4428657	SRR11363090
W02L02	Woodfield Reservation, Princeton, NJ	31-May-2017	Red	Tn5	Illumina PE150	4068904	SRR11363089
W02L06	Woodfield Reservation, Princeton, NJ	31-May-2017	Red	Tn5	Illumina PE150	3336967	SRR11363088
W05L05	Woodfield Reservation, Princeton, NJ	31-May-2017	Red	Tn5	Illumina PE150	4073188	SRR11363087

663 **Table STAR Methods 4. Details of biological samples, sequence files, and SRA accession**
664 **numbers for *H. cornu* GWAS targeted re-sequencing of approximately 800 kbp of**
665 **Chromosome 1.**
666

Sample #	Collection Location	Collection Date	Gall Color	Library Prep Method	Sequencing Method	# Read pairs	Accession #
B01L02	Moore St & Guyot Ave., Princeton, NJ	10-May-2017	Green	Tn5	Illumina PE150	214163	SRR11363085
B02L02-gr	Moore St & Guyot Ave., Princeton, NJ	10-May-2017	Green	Tn5	Illumina PE150	190639	SRR11363084
B03L03	Moore St & Guyot Ave., Princeton, NJ	10-May-2017	Green	Tn5	Illumina PE150	212907	SRR11363083
GCL74	Goose Creek Lane, Leesburg, VA	31-May-2017	Green	Tn5	Illumina PE150	306230	SRR11363082
HW03L02	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	291669	SRR11363081
HW03L07	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	306531	SRR11363080
HW04L07	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	270871	SRR11363079
HW05L02	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	262310	SRR11363078
HW06L01G01	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	269379	SRR11363077
HW07L01	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	237951	SRR11363076
HW08L03	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	257010	SRR11363252
HW09L05	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	211921	SRR11363219
HW10L01	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	241629	SRR11363218
HW11L01	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	237027	SRR11363217
HWL02L05	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	272737	SRR11363216

HWL04	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	249525	SRR11363215
HWL07	Herrontown Woods, Princeton, NJ	31-May-2017	Green	Tn5	Illumina PE150	304184	SRR11363214
JH01L04G01	Jockey Hollow, Morristown, NJ	12-Jun-2017	Green	Tn5	Illumina PE150	225593	SRR11363213
JH02L04G02	Jockey Hollow, Morristown, NJ	3-Apr-2017	Green	Tn5	Illumina PE150	241035	SRR11363212
JRC03	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Green	Tn5	Illumina PE150	270579	SRR11363211
JRC04	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Green	Tn5	Illumina PE150	291029	SRR11363209
JRC12	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Green	Tn5	Illumina PE150	291716	SRR11363208
JRC15	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Green	Tn5	Illumina PE150	278157	SRR11363207
JRC19	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Green	Tn5	Illumina PE150	296693	SRR11363206
JRC27	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Green	Tn5	Illumina PE150	289882	SRR11363205
MLR36	Man Lu Ridge Rd., Jefferson, MD	16-May-2017	Green	Tn5	Illumina PE150	237264	SRR11363204
MLR37	Man Lu Ridge Rd., Jefferson, MD	16-May-2017	Green	Tn5	Illumina PE150	215201	SRR11363203
MP55	Morven Park, Leesburg, VA	20-May-2017	Green	Tn5	Illumina PE150	233325	SRR11363202
MP56	Morven Park, Leesburg, VA	20-May-2017	Green	Tn5	Illumina PE150	201670	SRR11363201
OWR61	Old Waterford Road, Leesburg, VA	20-May-2017	Green	Tn5	Illumina PE150	175533	SRR11363200
OWR73	Old Waterford Road, Leesburg, VA	20-May-2017	Green	Tn5	Illumina PE150	132423	SRR11363198
POR29	Point of Rocks, MD	16-May-2017	Green	Tn5	Illumina PE150	236262	SRR11363197
SM01L01	St. Michael's Reserve, Princeton, NJ	1-Jun-2017	Green	Tn5	Illumina PE150	325832	SRR11363196
SM02L01	St. Michael's Reserve, Princeton, NJ	1-Jun-2017	Green	Tn5	Illumina PE150	313663	SRR11363195
SM03L01	St. Michael's Reserve, Princeton, NJ	1-Jun-2017	Green	Tn5	Illumina PE150	257124	SRR11363194

SM05L01	St. Michael's Reserve, Princeton, NJ	1-Jun-2017	Green	Tn5	Illumina PE150	292603	SRR11363193
SM05L05	St. Michael's Reserve, Princeton, NJ	1-Jun-2017	Green	Tn5	Illumina PE150	231169	SRR11363192
W01L01	Woodfield Reservation, Princeton, NJ	26-May-2017	Green	Tn5	Illumina PE150	300257	SRR11363191
W02L01G01	Woodfield Reservation, Princeton, NJ	26-May-2017	Green	Tn5	Illumina PE150	315314	SRR11363190
W03L01	Woodfield Reservation, Princeton, NJ	26-May-2017	Green	Tn5	Illumina PE150	313024	SRR11363189
W05L03G02	Woodfield Reservation, Princeton, NJ	26-May-2017	Green	Tn5	Illumina PE150	223027	SRR11363251
W06L02	Woodfield Reservation, Princeton, NJ	26-May-2017	Green	Tn5	Illumina PE150	303327	SRR11363250
B01L01	Moore St. & Guyot Ave., Princeton, NJ	10-May-2017	Red	Tn5	Illumina PE150	303803	SRR11363249
B01L03	Moore St & Guyot Ave., Princeton, NJ	10-May-2017	Red	Tn5	Illumina PE150	323593	SRR11363248
B02L01	Moore St & Guyot Ave., Princeton, NJ	10-May-2017	Red	Tn5	Illumina PE150	256759	SRR11363247
B02L02	Moore St & Guyot Ave., Princeton, NJ	10-May-2017	Red	Tn5	Illumina PE150	305206	SRR11363246
B03L01	Moore St & Guyot Ave., Princeton, NJ	10-May-2017	Red	Tn5	Illumina PE150	319941	SRR11363245
HW02L02	Herrontown woods, NJ	31-May-2017	Red	Tn5	Illumina PE150	333873	SRR11363244
HW03L01	Herrontown woods, NJ	31-May-2017	Red	Tn5	Illumina PE150	258120	SRR11363243
HW03L03	Herrontown woods, NJ	31-May-2017	Red	Tn5	Illumina PE150	275293	SRR11363242
HW03L04	Herrontown woods, NJ	31-May-2017	Red	Tn5	Illumina PE150	299356	SRR11363240
HW06L01G02	Herrontown woods, NJ	31-May-2017	Red	Tn5	Illumina PE150	260553	SRR11363239
HW06L02G03	Herrontown woods, NJ	31-May-2017	Red	Tn5	Illumina PE150	300112	SRR11363238
HW07L02	Herrontown woods, NJ	31-May-2017	Red	Tn5	Illumina PE150	259311	SRR11363237
HW09L02	Herrontown woods, NJ	31-May-2017	Red	Tn5	Illumina PE150	260464	SRR11363236
HW09L03	Herrontown woods, NJ	31-May-2017	Red	Tn5	Illumina PE150	247150	SRR11363235

HW09L07	Herrontown woods, NJ	31-May-2017	Red	Tn5	Illumina PE150	252787	SRR11363234
HW10L02	Herrontown woods, NJ	31-May-2017	Red	Tn5	Illumina PE150	254752	SRR11363233
HWL03	Herrontown woods, NJ	31-May-2017	Red	Tn5	Illumina PE150	254968	SRR11363232
JRC01	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Red	Tn5	Illumina PE150	252369	SRR11363231
JRC06	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Red	Tn5	Illumina PE150	317542	SRR11363229
JRC11	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Red	Tn5	Illumina PE150	304971	SRR11363228
JRC14	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Red	Tn5	Illumina PE150	327441	SRR11363227
JRC16	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Red	Tn5	Illumina PE150	253464	SRR11363226
JRC18	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Red	Tn5	Illumina PE150	362357	SRR11363225
JRC22	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Red	Tn5	Illumina PE150	226260	SRR11363224
JRC23	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Red	Tn5	Illumina PE150	338219	SRR11363223
JRC25	Janelia Research Campus, Ashburn, VA	3-Apr-2017	Red	Tn5	Illumina PE150	235476	SRR11363222
JRC84	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	244087	SRR11363110
JRC85	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	242809	SRR11363220
JRC86	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	277938	SRR11363186
JRC87	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	294687	SRR11363185
JRC88	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	283813	SRR11363184
JRC89	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	293816	SRR11363183
JRC90	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	307410	SRR11363182

JRC91	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	301578	SRR11363181
JRC92	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	274484	SRR11363180
JRC93	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	251875	SRR11363179
JRC94	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	246812	SRR11363178
JRC95	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	255648	SRR11363177
JRC96	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	228158	SRR11363175
JRC97	Janelia Research Campus, Ashburn, VA	21-May-2017	Red	Tn5	Illumina PE150	242492	SRR11363174
OWR71	Old Waterford Road, Leesburg, VA	30-May-2017	Red	Tn5	Illumina PE150	271725	SRR11363173
POR30	Point of Rocks, MD	16-May-2017	Red	Tn5	Illumina PE150	264484	SRR11363172
POR31	Point of Rocks, MD	16-May-2017	Red	Tn5	Illumina PE150	132066	SRR11363171
SM05L02	St. Michael's preserve, NJ	1-Jun-2017	Red	Tn5	Illumina PE150	317520	SRR11363170
SM05L04	St. Michael's preserve, NJ	1-Jun-2017	Red	Tn5	Illumina PE150	327054	SRR11363169
W02L02	Woodfield reservation, NJ	31-May-2017	Red	Tn5	Illumina PE150	272076	SRR11363168
W02L06	Woodfield reservation, NJ	31-May-2017	Red	Tn5	Illumina PE150	266667	SRR11363167
W05L05	Woodfield reservation, NJ	31-May-2017	Red	Tn5	Illumina PE150	233816	SRR11363166

668 **Table STAR Methods 5. Details of biological samples, sequence files, and SRA accession**
669 **numbers for *H. hamamelidis* whole-genome re-sequencing. MLBS = Mountain Lakes**
670 **Biological Station, Pembroke, Virginia.**
671

Sample #	Collection Location	Collection Date	Library Prep Method	Sequencing Method	# Read pairs	Accession #
GCC77	Goose Creek Lane, Leesburg, VA	31-May-2017	Tn5	Illumina PE150	7327275	SRR13258155
ML10_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	7333060	SRR13258154
ML14_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	7868181	SRR13258160
ML15_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	6656988	SRR13258022
ML17_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	7010451	SRR13257979
ML18_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	6481897	SRR13258033
ML2_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	7344032	SRR13258044
ML21_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	7166893	SRR13257983
ML23_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	7503826	SRR13258124
ML24_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	4641363	SRR13257994
ML27_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	6152152	SRR13258153
ML28_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	7763943	SRR13258063
ML30_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	6873459	SRR13258068
ML33_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	5378458	SRR13258073
ML34_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	6891338	SRR13258084
ML36_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	228271	SRR13258094
ML38_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	7212647	SRR13258102
ML40_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	6162598	SRR13258113
ML41_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	6170559	SRR13258002
ML43_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	7580980	SRR13258012
ML44_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	7912117	SRR13258161
ML48_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	5620965	SRR13258162
ML49_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	4744505	SRR13257971

ML50_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	7383271	SRR13258015
ML51_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	7607244	SRR13258016
ML53_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	5317613	SRR13258017
ML55_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	7580879	SRR13258018
ML57_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	1605322	SRR13258019
ML6_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	7702032	SRR13258020
ML60_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	7457247	SRR13258021
ML61_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	7359768	SRR13258023
ML62_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	7278081	SRR13258024
ML65_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	7448093	SRR13258025
ML70_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	6912997	SRR13257972
ML71_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	8049317	SRR13257973
ML73_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	7124219	SRR13257974
ML77_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	8407556	SRR13257975
ML79_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	8352011	SRR13257976
ML82_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	6974944	SRR13257977
MLR38	Man Lu Ridge Rd., Maryland	16-May-2017	Tn5	Illumina PE150	7732998	SRR13257978
MLR40	Man Lu Ridge Rd., Maryland	16-May-2017	Tn5	Illumina PE150	8121428	SRR13257980
MLR50	Man Lu Ridge Rd., Maryland	16-May-2017	Tn5	Illumina PE150	6605756	SRR13257981
MLR51	Man Lu Ridge Rd., Maryland	16-May-2017	Tn5	Illumina PE150	7442782	SRR13257982
MLR52	Man Lu Ridge Rd., Maryland	16-May-2017	Tn5	Illumina PE150	7346231	SRR13258026
MLR53	Man Lu Ridge Rd., Maryland	16-May-2017	Tn5	Illumina PE150	7679601	SRR13258027
MLR54	Man Lu Ridge Rd., Maryland	16-May-2017	Tn5	Illumina PE150	6710199	SRR13258028
OWR57	Old Waterford Road, Leesburg, VA	22-May-2017	Tn5	Illumina PE150	7307248	SRR13258029
OWR70	Old Waterford Road, Leesburg, VA	30-May-2017	Tn5	Illumina PE150	6821266	SRR13258030
ML10_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	7157138	SRR13258031

ML12_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	135505	SRR13258032
ML13_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	8118063	SRR13258034
ML16_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	9041960	SRR13258035
ML18_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	8826653	SRR13258036
ML19_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	8143654	SRR13258037
ML2_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	7175142	SRR13258038
ML20_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	8225262	SRR13258039
ML22_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	8664565	SRR13258040
ML23_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	8102497	SRR13258041
ML24_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	7884498	SRR13258042
ML25_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	8547869	SRR13258043
ML27_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	7831540	SRR13258045
ML29_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	7930016	SRR13258129
ML3_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	7532536	SRR13258128
ML32_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	7411318	SRR13258046
ML35_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	7957062	SRR13258047
ML36_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	8051736	SRR13258048
ML37_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	8015006	SRR13258049
ML38_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	8549973	SRR13258050
ML39_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	7014367	SRR13258051
ML41_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	7008886	SRR13258052
ML42_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	6255046	SRR13257984
ML44_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	8196265	SRR13257985
ML45_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	7370732	SRR13257986
ML46_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	7101764	SRR13258053
ML47_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	7935806	SRR13258054
ML52_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	8092411	SRR13258055

ML54_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	5607371	SRR13258056
ML56_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	7892227	SRR13258121
ML57_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	6630514	SRR13258122
ML58_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	7561525	SRR13258123
ML59_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	7637387	SRR13258125
ML62_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	7912262	SRR13258126
ML63_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	6922716	SRR13258127
ML64_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	7470732	SRR13257987
ML65_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	7195503	SRR13257988
ML66_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	7596370	SRR13257989
ML67_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	7351721	SRR13257990
ML68_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	6448176	SRR13257991
ML69_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	7633376	SRR13257992
ML70_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	8179649	SRR13257993
ML9_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	6874469	SRR13257995
MLR32	Man Lu Ridge Rd., Maryland	16-May-2017	Tn5	Illumina PE150	5102731	SRR13257996
MLR34	Man Lu Ridge Rd., Maryland	16-May-2017	Tn5	Illumina PE150	6204711	SRR13257997
MLR43	Man Lu Ridge Rd., Maryland	16-May-2017	Tn5	Illumina PE150	7370217	SRR13257998
MLR49	Man Lu Ridge Rd., Maryland	16-May-2017	Tn5	Illumina PE150	6120506	SRR13258057
OWR58	Old Waterford Road, Leesburg, VA	22-May-2017	Tn5	Illumina PE150	7305751	SRR13258058

Table STAR Methods 6. Details of biological samples, sequence files, and SRA accession numbers for *H. hamamelidis* targeted genome re-sequencing of approximately 800 kbp of Chromosome 1. MLBS = Mountain Lakes Biological Station, Pembroke, Virginia.

Sample #	Collection Location	Collection Date	Library Prep Method	Sequencing Method	# Read pairs	Accession #
GCC77	Goose Creek Lane, Leesburg, Virginia	31-May-2017	Tn5	Illumina PE150	314257	SRR13258059
ML10_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	334744	SRR13258060
ML14_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	325866	SRR13258061

ML15_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	298479	SRR13258062
ML17_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	313916	SRR13258152
ML18_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	290694	SRR13258151
ML2_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	331929	SRR13258150
ML21_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	329166	SRR13258149
ML23_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	327831	SRR13258148
ML24_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	178226	SRR13258147
ML27_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	275723	SRR13258146
ML28_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	346212	SRR13258145
ML30_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	300476	SRR13258144
ML33_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	235769	SRR13258143
ML34_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	313858	SRR13258064
ML36_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	6195	SRR13258142
ML38_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	312165	SRR13258141
ML40_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	262504	SRR13258140
ML41_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	266728	SRR13258065
ML43_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	334740	SRR13258066
ML44_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	355990	SRR13258139
ML48_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	76870	SRR13258138
ML49_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	219732	SRR13258137
ML50_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	319064	SRR13258067
ML51_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	335990	SRR13258069
ML53_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	240432	SRR13258136
ML55_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	334902	SRR13258135
ML57_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	89944	SRR13258134
ML6_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	340415	SRR13258070
ML60_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	283159	SRR13258071

ML61_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	317924	SRR13258133
ML62_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	312738	SRR13258132
ML65_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	332910	SRR13258131
ML70_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	313079	SRR13258072
ML71_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	353342	SRR13258074
ML73_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	309115	SRR13258075
ML77_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	374346	SRR13258076
ML79_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	367693	SRR13258077
ML82_Gr	MLBS	7-Jul-2018	Tn5	Illumina PE150	280482	SRR13258078
MLR38	Man Lu Ridge Rd., Maryland	16-May-2017	Tn5	Illumina PE150	331238	SRR13258079
MLR40	Man Lu Ridge Rd., Maryland	16-May-2017	Tn5	Illumina PE150	356281	SRR13258080
MLR50	Man Lu Ridge Rd., Maryland	16-May-2017	Tn5	Illumina PE150	290972	SRR13258081
MLR51	Man Lu Ridge Rd., Maryland	16-May-2017	Tn5	Illumina PE150	327369	SRR13258082
MLR52	Man Lu Ridge Rd., Maryland	16-May-2017	Tn5	Illumina PE150	327149	SRR13258083
MLR53	Man Lu Ridge Rd., Maryland	16-May-2017	Tn5	Illumina PE150	323702	SRR13258085
MLR54	Man Lu Ridge Rd., Maryland	16-May-2017	Tn5	Illumina PE150	285992	SRR13258086
OWR57	Old Waterford Road, Leesburg, VA	22-May-2017	Tn5	Illumina PE150	306545	SRR13258087
OWR70	Old Waterford Road, Leesburg, VA	30-May-2017	Tn5	Illumina PE150	302774	SRR13258088
ML10_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	329939	SRR13258089
ML12_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	3091	SRR13258090
ML13_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	365121	SRR13258091
ML16_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	402478	SRR13258092
ML18_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	395314	SRR13258093
ML19_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	369502	SRR13258156
ML2_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	323168	SRR13258095
ML20_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	360596	SRR13258096

ML22_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	380225	SRR13258097
ML23_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	355840	SRR13258098
ML24_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	360261	SRR13258157
ML25_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	381256	SRR13258158
ML27_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	354745	SRR13258099
ML29_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	360523	SRR13258100
ML3_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	341628	SRR13258101
ML32_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	332324	SRR13258159
ML35_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	353534	SRR13258103
ML36_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	369572	SRR13258104
ML37_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	349811	SRR13258105
ML38_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	393709	SRR13258106
ML39_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	304153	SRR13258107
ML41_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	312756	SRR13258108
ML42_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	281446	SRR13258109
ML44_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	360792	SRR13258110
ML45_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	331967	SRR13258111
ML46_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	306892	SRR13258112
ML47_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	312341	SRR13258114
ML52_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	355827	SRR13258115
ML54_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	220263	SRR13258116
ML56_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	341262	SRR13258117
ML57_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	287334	SRR13258118
ML58_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	337310	SRR13258119
ML59_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	342148	SRR13258120
ML62_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	360053	SRR13257999
ML63_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	303988	SRR13258000

ML64_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	326164	SRR13258001
ML65_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	308072	SRR13258003
ML66_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	334547	SRR13258004
ML67_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	313769	SRR13258005
ML68_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	285472	SRR13258006
ML69_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	335029	SRR13258007
ML70_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	368484	SRR13258008
ML9_Red	MLBS	7-Jul-2018	Tn5	Illumina PE150	311326	SRR13258009
MLR32	Man Lu Ridge Rd., Maryland	16-May-2017	Tn5	Illumina PE150	206644	SRR13258010
MLR34	Man Lu Ridge Rd., Maryland	16-May-2017	Tn5	Illumina PE150	271798	SRR13258011
MLR43	Man Lu Ridge Rd., Maryland	16-May-2017	Tn5	Illumina PE150	320317	SRR13258130
MLR49	Man Lu Ridge Rd., Maryland	16-May-2017	Tn5	Illumina PE150	256744	SRR13258013
OWR58	Old Waterford Road, Leesburg, VA	22-May-2017	Tn5	Illumina PE150	320996	SRR13258014

675

676 **Table STAR Methods 7. Details of biological samples, sequence files, and SRA accession**
677 **numbers for *Hormaphis cornu* salivary gland RNA-sequencing from fundatrices from red**
678 **and green galls. JRC = Janelia Research Campus.**
679

Sample ID	Collection Location	Collection Date	Gall color	Library Prep Method	Sequencing Method	# Read pairs	Accession #
AKG_4	JRC	28-Apr-2019	Green	WGB	Illumina PE 150	11607393	SRR11428726
DSG_6	JRC	28-Apr-2019	Green	WGB	Illumina PE 150	12179326	SRR11428724
DSG_7	JRC	30-Apr-2019	Green	WGB	Illumina PE 150	13321660	SRR11428723
DSG_8	JRC	30-Apr-2019	Green	WGB	Illumina PE 150	10433415	SRR11428722
DSG_9	JRC	30-Apr-2019	Green	WGB	Illumina PE 150	14776468	SRR11428721
DSG_10	JRC	30-Apr-2019	Green	WGB	Illumina PE 150	13105789	SRR11428720
DSG_11	JRC	01-May-2019	Green	WGB	Illumina PE 150	12867296	SRR11428719
DSG_13	JRC	01-May-2019	Green	WGB	Illumina PE 150	10772840	SRR11428718
DSR_11	JRC	01-May-2019	Red	WGB	Illumina PE 150	15042624	SRR11428717
DSR_12	JRC	01-May-2019	Red	WGB	Illumina PE 150	11051413	SRR11428716
DSR_13	JRC	01-May-2019	Red	WGB	Illumina PE 150	15396180	SRR11428715
DSR_15	JRC	01-May-2019	Red	WGB	Illumina PE 150	18975841	SRR11428784
DSR_16	JRC	01-May-2019	Red	WGB	Illumina PE 150	22465588	SRR11428782
DSR_17	JRC	01-May-2019	Red	WGB	Illumina PE 150	15618318	SRR11428781
DSR_18	JRC	01-May-2019	Red	WGB	Illumina PE 150	16229009	SRR11428780
AKR1	JRC	28-Apr-2019	Red	WGB	Illumina PE 150	14353107	SRR13258428
AKR2	JRC	28-Apr-2019	Red	WGB	Illumina PE 150	17302776	SRR13258427
AKR3	JRC	28-Apr-2019	Red	WGB	Illumina PE 150	24214993	SRR13258426
AKR4	JRC	28-Apr-2019	Red	WGB	Illumina PE 150	12914802	SRR13258425
DSR5	JRC	28-Apr-2019	Red	WGB	Illumina PE 150	16153051	SRR13258424
DSR6	JRC	28-Apr-2019	Red	WGB	Illumina PE 150	10998399	SRR13258423
DSR7	JRC	28-Apr-2019	Red	WGB	Illumina PE 150	10264369	SRR13258422
DSR_8	JRC	28-Apr-2019	Red	WGB	Illumina PE 150	19769187	SRR13258421

DSR_9	JRC	28-Apr-2019	Red	WGB	Illumina PE 150	18487599	SRR13258420
DSR	JRC	28-Apr-2019	Red	WGB	Illumina PE 150	9649227	SRR13258419

680

681 **Table STAR Methods 8. Details of biological samples, sequence files, and SRA accession**
682 **numbers for *Hamamelis virginiana* RNA-sequencing from galls and leaf around galls. JRC**
683 **= Janelia Research Campus, Ashburn, VA.**
684

Sample ID	Collection Location	Collection Date	Library prep method	Sequencing method	# Read pairs	Accession #
Gall_9	JRC	11-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	14453904	SRR11362686
Gall_11	JRC	11-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	7224527	SRR11362685
Early_gall_18	JRC	12-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	25279196	SRR11362642
Gall_20	JRC	12-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	11692754	SRR11362631
Gall_22	JRC	12-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	10990345	SRR11362620
Gall_26	JRC	14-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	6419861	SRR13258463
Gall_27	JRC	14-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	7086413	SRR13258462
Gall_28	JRC	14-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	7618914	SRR13258461
Gall_29	JRC	14-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	5633325	SRR13258460
Gall_31	JRC	14-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	12324002	SRR11362672
Early_gall_33	JRC	14-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	31823857	SRR11362661
gall_35	JRC	14-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	11104190	SRR11362652
Gall_42	JRC	15-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	6199468	SRR13258459
Gall_44	JRC	15-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	11104190	SRR11362683
Gall_46	JRC	15-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	11965583	SRR11362684

Gall_47	JRC	15-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	18509404	SRR11362682
Bottom_gall_51	JRC	17-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	13291659	SRR11362650
Top_gall_52	JRC	17-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	11679869	SRR11362649
Bottom_gall_53	JRC	17-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	11661968	SRR11362648
Top_gall_54	JRC	17-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	12277949	SRR11362647
Bottom_gall_55	JRC	17-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	10705594	SRR11362646
Top_gall_56	JRC	17-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	12245151	SRR11362645
Bottom_gall_57	JRC	17-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	5688881	SRR11362644
Top_gall_58	JRC	17-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	8195240	SRR11362643
Gall_60	JRC	17-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	7075731	SRR13258458
Gall_61	JRC	17-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	5226728	SRR13258472
Gall_62	JRC	17-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	3985629	SRR13258471
Gall_63	JRC	17-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	6194899	SRR13258470
Gall_65	JRC	17-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	10541245	SRR13258469
Gall_66	JRC	17-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	5324437	SRR13258468
Apical_gall_118	JRC	22-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	8305160	SRR13258474
Apical_gall_122	JRC	22-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	3971025	SRR13258473
Basal_gall_116	JRC	22-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	10405803	SRR13258465

Basal_gall_120	JRC	22-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	4887393	SRR13258464
Medial_gall_117	JRC	22-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	3452858	SRR13258467
Medial_gall_121	JRC	22-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	5465679	SRR13258466
Leaf_around_gall_10	JRC	11-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	3055859	SRR11362625
Leaf_around_gall_12	JRC	11-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	2924715	SRR11362624
Leaf_around_gall_19	JRC	11-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	39720279	SRR11362623
Leaf_around_gall_21	JRC	12-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	4346238	SRR11362622
Leaf_around_gall_23	JRC	12-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	9527411	SRR11362621
Leaf_around_gall_30	JRC	14-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	10162178	SRR11362619
Leaf_around_gall_34	JRC	14-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	34938887	SRR11362681
Leaf_around_gall_38	JRC	14-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	5195676	SRR11362680
Leaf_around_gall_43	JRC	15-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	11911791	SRR11362679
Leaf_around_gall_45	JRC	15-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	9012995	SRR11362678
Leaf_around_gall_49	JRC	15-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	2405452	SRR11362677
Leaf_around_gall_50	JRC	17-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	4591989	SRR11362676
Leaf_around_gall_59	JRC	17-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	6517483	SRR11362675
Leaf_around_gall_64	JRC	17-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	4871569	SRR11362674
Leaf_around_gall_67	JRC	17-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	4352513	SRR11362673

Leaf_around_gall_119	JRC	22-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	4505696	SRR11362671
Leaf_around_gall_123	JRC	22-Apr-2017	Nugen Universal Plus mRNA-seq	Illumina PE150	2655968	SRR11362670

685

686 **Table STAR Methods 9. Details of biological samples, sequence files, and SRA accession**
687 **numbers for *Hamamelis virginiana* RNA-sequencing from green and red galls.**
688

Sample ID	Collection Location	Collection Date	Leaf Sample	Gall color	Library Prep Method	Sequencing Method	# Read pairs	Accession #
20JRC105R	JRC	08-May-2020	1	Red	WGB	Illumina SE75	5813855	SRR13258378
20JRC104G	JRC	08-May-2020	1	Green	WGB	Illumina SE75	6618273	SRR13258369
20JRC107R	JRC	08-May-2020	2	Red	WGB	Illumina SE75	7330505	SRR13258344
20JRC106G	JRC	08-May-2020	2	Green	WGB	Illumina SE75	6999471	SRR13258368
20JRC108R	JRC	08-May-2020	3	Red	WGB	Illumina SE75	7444078	SRR13258376
20JRC109G	JRC	08-May-2020	3	Green	WGB	Illumina SE75	7711014	SRR13258350
20JRC110R	JRC	08-May-2020	4	Red	WGB	Illumina SE75	6961512	SRR13258372
20JRC111G	JRC	09-May-2020	4	Green	WGB	Illumina SE75	6765543	SRR13258346
20JRC113R	JRC	08-May-2020	5	Red	WGB	Illumina SE75	6707106	SRR13258371
20JRC112G	JRC	09-May-2020	5	Green	WGB	Illumina SE75	7274181	SRR13258345
20JRC116R	JRC	08-May-2020	7	Red	WGB	Illumina SE75	12806892	SRR13258370
20JRC1076R	JRC	08-May-2020	8	Red	WGB	Illumina SE75	9172629	SRR13258377
20JRC1077R	JRC	08-May-2020	8	Red	WGB	Illumina SE75	9280201	SRR13258366
20JRC1074G	JRC	08-May-2020	8	Green	WGB	Illumina SE75	9200050	SRR13258367
20JRC1075G	JRC	08-May-2020	8	Green	WGB	Illumina SE75	9001911	SRR13258365
20JRC1079R	JRC	08-May-2020	9	Red	WGB	Illumina SE75	8765890	SRR13258355
20JRC1078G	JRC	08-May-2020	9	Green	WGB	Illumina SE75	8534470	SRR13258364
20JRC1081R	JRC	08-May-2020	10	Red	WGB	Illumina SE75	7555292	SRR13258343
20JRC1080G	JRC	08-May-2020	10	Green	WGB	Illumina SE75	8840709	SRR13258363
20JRC1083R	JRC	08-May-2020	11	Red	WGB	Illumina SE75	9657276	SRR13258342
20JRC1082G	JRC	08-May-2020	11	Green	WGB	Illumina SE75	8984115	SRR13258362

20JRC1085R	JRC	08-May-2020	12	Red	WGB	Illumina SE75	9542403	SRR13258341
20JRC1086R	JRC	08-May-2020	13	Red	WGB	Illumina SE75	9508095	SRR13258340
20JRC1087G	JRC	08-May-2020	13	Green	WGB	Illumina SE75	9552857	SRR13258361
20JRC1089R	JRC	08-May-2020	14	Red	WGB	Illumina SE75	8580538	SRR13258339
20JRC1088G	JRC	08-May-2020	14	Green	WGB	Illumina SE75	9348175	SRR13258360
20JRC1091R	JRC	08-May-2020	15	Red	WGB	Illumina SE75	10193238	SRR13258375
20JRC1090G	JRC	09-May-2020	15	Green	WGB	Illumina SE75	9782456	SRR13258359
20JRC1092G	JRC	09-May-2020	15	Green	WGB	Illumina SE75	10354726	SRR13258358
20JRC1093R	JRC	08-May-2020	16	Red	WGB	Illumina SE75	10632771	SRR13258374
20JRC1094G	JRC	09-May-2020	16	Green	WGB	Illumina SE75	9767489	SRR13258357
20JRC1095G	JRC	09-May-2020	16	Green	WGB	Illumina SE75	9957641	SRR13258356
20JRC1096G	JRC	09-May-2020	16	Green	WGB	Illumina SE75	9938437	SRR13258354
20JRC1097G	JRC	09-May-2020	16	Green	WGB	Illumina SE75	8222414	SRR13258353
20JRC1100R	JRC	08-May-2020	17	Red	WGB	Illumina SE75	5354681	SRR13258373
20JRC1098G	JRC	09-May-2020	17	Green	WGB	Illumina SE75	5890682	SRR13258352
20JRC1099G	JRC	09-May-2020	17	Green	WGB	Illumina SE75	6057780	SRR13258351
20JRC1101G	JRC	09-May-2020	17	Green	WGB	Illumina SE75	7237058	SRR13258349
20JRC1102G	JRC	08-May-2020	18	Green	WGB	Illumina SE75	4943324	SRR13258348
20JRC1103G	JRC	08-May-2020	18	Green	WGB	Illumina SE75	7407432	SRR13258347

693 **Method Details**

694

695 Imaging of leaves and fundatrices inside developing galls

696

697 Young *Hamamelis virginiana* (witch hazel) leaves or leaves with early stage galls of
698 *Hormaphis cornu* were fixed in Phosphate Buffered Saline (PBS: 137 mM NaCl, 2.7 mM KCl,
699 10 mM Na2HPO4, 1.8 mM KH2PO4, pH 7.4) containing 0.1% Triton X-100, 2%
700 paraformaldehyde and 0.5% glutaraldehyde (paraformaldehyde and glutaraldehyde were EM
701 grade from Electron Microscopy Services) at room temperature for two hours without agitation
702 to prevent the disruption of the aphid stylet inserted into leaf tissue. Fixed leaves or galls were
703 washed in PBS containing 0.1% Triton X-100, hand cut into small sections (~10 mm2), and
704 embedded in 7% agarose for subsequent sectioning into 0.3 mm thick sections using a Leica
705 Vibratome (VT1000s). Sectioned plant tissue was stained with Calcofluor White (0.1 mg/mL,
706 Sigma-Aldrich, F3543) and Congo Red (0.25 mg/mL, Sigma-Aldrich, C6767) in PBS
707 containing 0.1% Triton X-100 with 0.5% DMSO and Escin (0.05 mg/mL, Sigma-Aldrich,
708 E1378) at room temperature with gentle agitation for 2 days. Stained sections were washed with
709 PBS containing 0.1% Triton X-100. Soft tissues were digested and cleared to reduce light
710 scattering during subsequent imaging using a mixture of 0.25 mg/mL collagenase/dispase
711 (Roche #10269638001) and 0.25 mg/mL hyaluronidase (Sigma Aldrich #H3884) in PBS
712 containing 0.1% Triton X-100 for 5 hours at 37°C. To avoid artifacts and warping caused by
713 osmotic shrinkage of soft tissue and agarose, samples were gradually dehydrated in glycerol
714 (2% to 80%) and then ethanol (20% to 100 %) (Ott 2008) and mounted in methyl salicylate
715 (Sigma-Aldrich, M6752) for imaging. Serial optical sections were obtained at 2 µm intervals on

716 a Zeiss 880 confocal microscope with a Plan-Apochromat 10x/0.45 NA objective, at 1 µm with
717 a LD-LCI 25x/0.8 NA objective or at 0.5 µm with a Plan-Apochromat 40x/0.8 NA objective.
718 Maximum projections of confocal stacks or rotation of images were carried out using FIJI
719 (Schindelin et al. 2012).
720
721 *Hormaphis cornu* genome sequencing, assembly, and annotation
722
723 We collected *H. cornu* aphids from a single gall for genome sequencing (Table STAR
724 Methods 1). All aphids within the gall were presumed to be clonal offspring of a single
725 fundatrix, because all *H. cornu* galls we have ever examined contain only a single fundatrix and
726 the ostiole of the galls was closed at the time we collected this gall, so there is little chance of
727 inter-gall migration. High molecular weight (HMW) DNA was prepared by gently grinding
728 aphids with a plastic pestle against the inside wall of a 2 mL Eppendorf tube in 1 mL of 0.5%
729 SDS, 200 mM Tris-HCl pH 8, 25 mM EDTA, 250 mM NaCl with 10 uL of 1 mg/mL RNAse
730 A. Sample was incubated at 37°C for 1 hour and then 30uL of 10 mg/mL ProteinaseK was
731 added and the sample was incubated for an additional 1 hr at 50°C with gentle agitation at 300
732 RPM. 1 mL of Phenol:Chloroform:Isoamyl alcohol (25:24:1) was added and the sample was
733 centrifuged at 16,000 RCF for 10 min. The supernatant was removed to a new 2mL Eppendorf
734 tube and the Phenol:Chloroform:Isoamyl alcohol extraction was repeated. The supernatant was
735 removed to a new 2 mL tube and 2.5 X volumes of absolute ethanol were added. The sample
736 was centrifuged at 16,000 RCF for 15 min and then washed with fresh 70% ethanol. All ethanol
737 was removed with a pipette and the sample was air dried for approximately 15 minutes and

738 DNA was resuspended in 50 uL TE. This sample was sent to HudsonAlpha Institute for
739 Biotechnology for genome sequencing.
740 DNA quality control, library preparation, and Chromium 10X linked read sequencing
741 were performed by HudsonAlpha Institute for Biotechnology. Most of the mass of the HMW
742 DNA appeared greater than approximately 50 kb on a pulsed field gel and paired end
743 sequencing on an Illumina HiSeq X10 yielded 816M reads. The genome was assembled using
744 *Supernova* (Weisenfeld et al. 2017) using 175M reads, which generated the best genome N50 of
745 a range of values tested. This 10X genome consisted of 21,072 scaffolds of total length 319.352
746 MB. The genome scaffold N50 was 839.101 KB and the maximum scaffold length was 3.495
747 MB.
748 We then contracted with Dovetail Genomics to apply Chicago (*in vitro* proximity
749 ligation) and HiC (*in vivo* proximity ligation) to generate larger scaffolds
750 (https://dovetailgenomics.com/ga_tech_overview/). We submitted HMW gDNA from the same
751 sample used for 10X genome sequencing for Chicago and a separate sample of frozen aphids
752 for HiC (Table STAR Methods 1). The Dovetail genome consisted of 11,244 scaffolds of total
753 length 320.34 MB with a scaffold N50 of 36.084 Mb. This genome, named
754 hormaphis_cornu_26Sep2017_PoQx8, contains 9 main scaffolds, each longer than 17.934 Mb,
755 which appear to represent the expected 9 chromosomes of *H. cornu* (Blackman and Eastop
756 1994). This assembly also includes the circular genome of the bacterial endosymbiont
757 *Buchnera aphidicola* of 643,259 bp. Assembly and *BUSCO* analysis statistics (Simao et al.
758 2015) using the *gVolante* web interface (Nishimura et al., 2017) with the Arthropod gene set are
759 shown below.
760

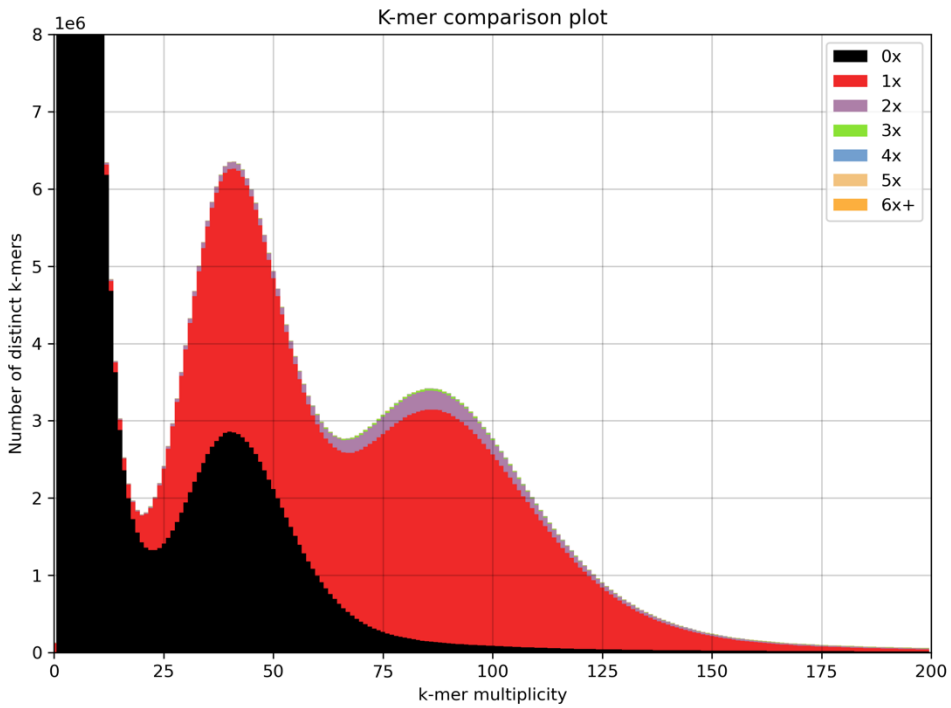
Assembly statistics	**hormaphis_cornu_26Sep2017_PoQx8**
# of scaffolds:	11242
Total length:	320,336,030
Longest sequence:	60,222,264
N50 sequence length:	36,083,769
Sum length of sequences > 1M	297,831,669 (93.0% of total length)
Sum length of sequences > 10M	294,347,208 (91.9% of total length)

761

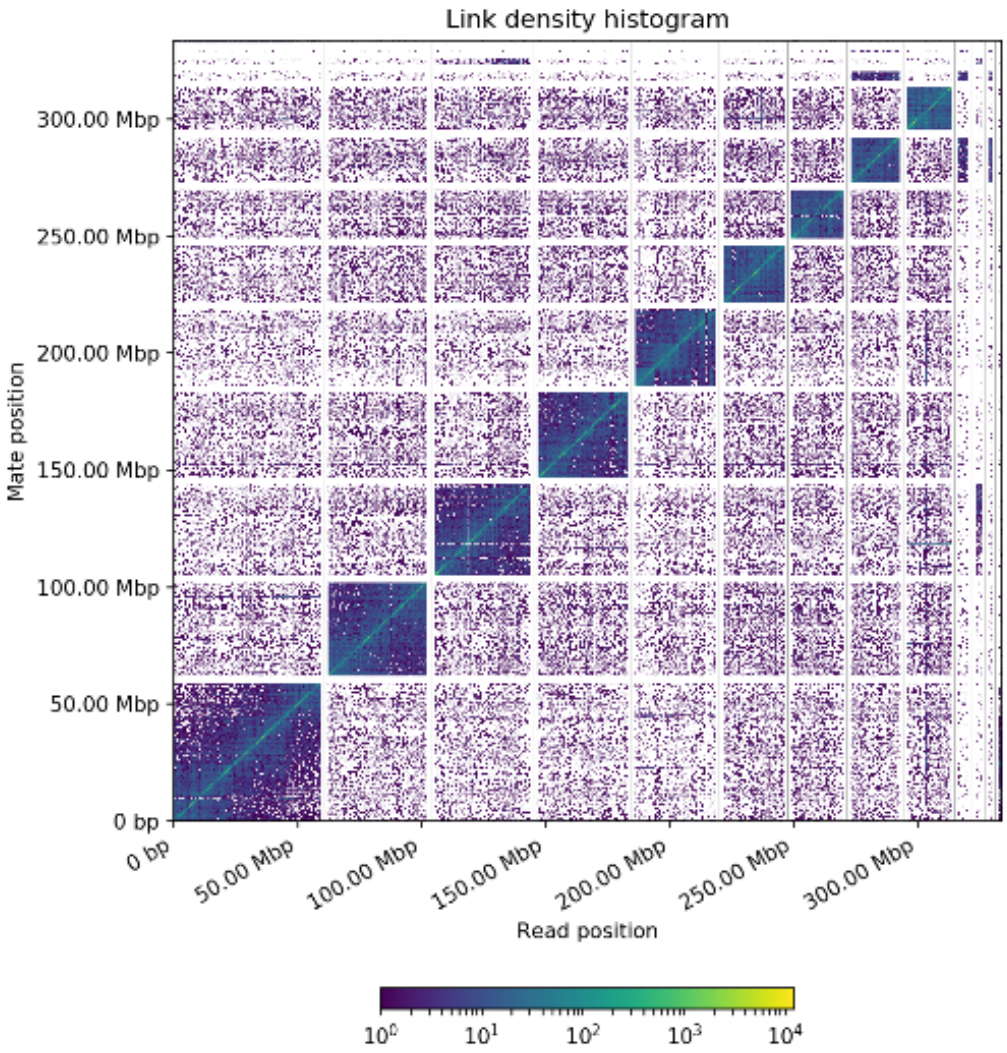
BUSCO Analysis	**hormaphis_cornu_26Sep2017_PoQx8**
Total # of core genes queried	1066
# of complete core genes detected	1026 (96.25%)
# of complete and partial core genes detected	1038 (97.37%)
# of missing core genes:	28 (2.63%)
Average # of orthologs per core genes:	1.02
% of detected core genes that have more than 1 ortholog:	2.34

762

763 As further checks on the quality of this genome assembly, we examined the K-mer
764 spectra and examined the HiC contact map. The K-mer spectra comparing kmers from the 10X
765 genome sequencing and the Dovetail assembled genome shown below was calculated using
766 KAT (Mapleson et al. 2017). As shown below, there are two peaks for multiplicity 1X, as
767 expected when sequencing a heterozygous individual, and the vast majority of sequence reads
768 map uniquely to the assembled genome 1X, revealing few duplications.



The HiC contact map, shown below, reveals that the great majority of read pairs map to approximately the same location within each of the nine major scaffolds. The X and Y axes indicate the mapping position for the first and second read of each read pair and the color indicates the number of read pairs mapping to each bin. Only scaffolds greater than 1 Mbp are shown.



We annotated this genome for protein-coding genes using RNA-seq data collected from salivary glands and carcasses of many stages of the *H. cornu* life cycle (Table STAR Methods 2) using *BRAKER* (Altschul et al. 1990; Barnett et al. 2011; Camacho et al. 2009; Hoff et al. 2016, 2019; Li 2013; Lomsadze, Burns, and Borodovsky 2014; Stanke et al. 2006). To increase the efficiency of mapping RNA-seq reads for differential expression analysis, we predicted 3' UTRs using *UTRme* (Radío et al. 2018). We found that *UTRme* sometimes predicted UTRs within introns. We therefore applied a custom R script to remove such predicted UTRs. Later,

783 after discovering the *bicycle* genes, we manually annotated all predicted *bicycle* genes,
784 including 5' and 3' UTRs, in *APOLLO* (Lewis et al. 2002) by examining evidence from
785 RNAseq reads aligned to the genome with the *Integrative Genomics Viewer* (Robinson et al.
786 2011; Thorvaldsdóttir, Robinson, and Mesirov 2013). We found that the start sites of many
787 *bicycle* genes were incorrectly annotated by *BRAKER* at a downstream methionine,
788 inadvertently excluding predicted putative signal peptides from these genes. RNAseq evidence
789 often supported transcription start sites that preceded an upstream methionine and these exons
790 were corrected in *APOLLO*. The combined collection of 18,895 automated and 687 manually
791 curated gene models (19,582 total) were used for all subsequent analyses of *H. cornu* genomic
792 data.
793
794 Genome-wide association study of aphids inducing red and green galls
795
796 Galls produced by *H. cornu* were collected in the early summer (Table STAR Methods
797 3) and dissected by making a single vertical cut down the side of each gall with a razor blade to
798 expose the aphids inside. DNA was extracted using the Zymo ZR-96 Quick gDNA kit from the
799 foundress of each gall, because foundresses do not appear to travel between galls. We
800 performed tagmentation of genomic DNA derived from 47 individuals from red galls and 43
801 from green galls using barcoded adaptors compatible with the Illumina sequencing platform
802 (Hennig et al. 2018). Tagmented samples were pooled without normalization, PCR amplified
803 for 14 cycles, and sequenced on an Illumina NextSeq 500 to generated paired end 150 bp reads
804 to an average depth of 2.9X genomic coverage. The average sequencing depth before filtering
805 was calculated by multiplying the number of read pairs generated by *SAMtools flagstat* version

806 1.3 (Li et al. 2009) by the read length of 150bp, then dividing by the total genome size
807 (323,839,449bp).
808 We performed principle component analysis on the genome-wide polymorphism data to
809 detect any potential population structure that might confound GWAS study. Reads were
810 mapped using *bwa mem* version 0.7.17-r1188 (Li 2013) and joint genotyped using *SAMtools*
811 *mpileup* version 1.3, with the flag *-ugf,* followed by *BCFtools call* version 1.9 (Li 2011), with
812 the flag *-m*. Genotype calls were then filtered for quality and missingness using *BCFtools filter*
813 and *view* version 1.9, where only SNPs with MAF > 0.05, QUAL > 20, and genotyped in at
814 least 80% of the individuals were kept. To limit the number of SNPs for computational
815 efficiency, the SNPs were additionally thinned using *VCFtools –thin* version 0.1.15 to exclude
816 any SNPs within 1000bp of each other. PCA was performed using the *snpgdsPCA* function
817 from the R package *SNPRelate* version 1.20.1 in *R* version 3.6.1 (Zheng et al. 2012).
818 We performed genome-wide association mapping with these low coverage data by
819 mapping reads with *bwa mem* version 0.7.17-r1188, and then calculated the likelihood of
820 association with gall color with *SAMtools mpileup* version 0.1.19 and *BCFtools view -vcs*
821 version 0.1.19 using BAM files as the input. Association for each SNP was measured by the
822 likelihood-ratio test (LRT) value in the INFO field of the output VCF file, which is a one-
823 degree association test p-value. This method calculates association likelihoods using genotype
824 likelihoods rather than hard genotype calls, ameliorating the issue of low-confidence genotype
825 calls resulting from low-coverage data (Li 2011). The false discovery rate was set as the
826 Bonferroni corrected value for 0.05, which was calculated as 0.05 / 50,957,130 (the total
827 number of SNPs in the genome-wide association mapping).
828

829 Enrichment and sequencing of the genomic region containing highly significant GWAS hits

830

831 The low coverage GWAS identified multiple linked SNPS on chromosome 1 that were
832 strongly associated with gall color (Figure 2A). To identify all candidate SNPs in this genomic
833 region and to generate higher-confidence GWAS calls, we enriched this genomic region from a
834 library of pooled tagmented samples of fundatrix DNA from red and green galls using custom
835 designed Arbor Bioscience *MyBaits* for a 800,290 bp region on chromosome 1 spanning the
836 highest scoring GWAS SNP (40,092,625 - 40,892,915 bp). This enriched library was sequenced
837 on an Illumina NextSeq 550 generating paired-end 150 bp reads and resulted in usable re-
838 sequencing data for 48 red gall-producing individuals and 42 green gall-producing individuals,
839 with average pre-filtered sequencing depth of 58.2X.

840 We mapped reads with *bwa mem* version 0.7.17-r1188 and sorted bam files with
841 *SAMtools sort* version 1.7, marked duplicate reads with *Picard MarkDuplicates* version 2.18.0
842 (http://broadinstitute.github.io/picard/), re-aligned indels using *GATK IndelRealigner* version
843 3.4 (McKenna et al. 2010), and called variants using *SAMtools mpileup version* 1.7 and
844 *BCFtools call* version 1.7 (https://github.com/SAMtools/bcftools). This genotyping pipeline is
845 available at https://github.com/YourePrettyGood/PseudoreferencePipeline (thereafter referred
846 to as *PseudoreferencePipeline*). SNPs were quality filtered from the VCF file using *BCFtools*
847 *view* version 1.7 at DP > 10 and MQ > 40 and merged using *BCFtools merge*.

848 For PCA analysis, the joint genotype calls were filtered for quality and missingness
849 using *BCFtools filter* and *view* version 1.9, where only SNPs with MAF > 0.05 and genotyped
850 in at least 80% of the individuals were retained. PCA was performed using the *snpgdsPCA*
851 function from the R package *SNPRelate* version 1.20.1 in *Rstudio* version 3.6.1.

852 Association testing was performed using *PLINK* version 1.90 (Purcell et al. 2007) with
853 minor allele frequency filtered at MAF > 0.2. We did not apply a Hardy-Weinberg equilibrium
854 filter because the samples were not randomly collected from nature. Red galls are rare in our
855 population and we oversampled fundatrices from red galls to roughly match the number of
856 fundatrices sampled from green galls. Results of the GWAS were plotted using the
857 *plotManhattan* function of *Sushi* version 1.24.0 (Phanstiel et al. 2014).

858 To calculate LD for the 45kbp region surrounding the 11 GWAS SNPs in Fig. S5B, we
859 extracted positions chromosome1:40,475,000 – 40,520,000 from the merged VCF and retained
860 SNPs that both exhibited MAF > 0.05 and were genotyped in at least 80% of the samples using
861 *BCFtools view* and *filter* version 1.9.

862 To plot LD for the entire target enrichment in Fig. S6D, we filtered the VCF for only
863 SNPs with MAF>0.2 and genotyped in at least 80% of the samples using *bcftools view* and
864 *filter* version 1.9, and thinned the resulting SNP set using *VCFtools–thin* version 0.1.15 to
865 exclude any SNPs within 500bp of each other. We further merged back the 11 significant
866 GWAS SNPs using *bcftools concat*, since the thinning process could have removed one or more
867 of these SNPs. We also removed SNPs in regions where the *H. cornu* reference genome did not
868 align with the genome of the sister species *H. hamamelidis* using *BEDTools intersect* version
869 2.29.2 (Quinlan and Hall 2010).

870 The LD heatmaps were generated using the R packages *vcfR* version 1.10.0 (Knaus and
871 Grünwald 2017), *snpStats* version 1.36.0 (Clayton 2019) and *LDheatmap* version 0.99.8 (Shin
872 et al. 2006) in *Rstudio* version 3.6.1. The R code used to generate the LD heatmap figure was
873 adapted from code provided at sfustatgen.github.io/LDheatmap/articles/vcfOnLDheatmap.html.

874 The gene models were plotted using the *plotGenes* function from the R package *Sushi* version
875 1.24.0.
876
877 Lack of evidence for chromosomal aberrations
878
879 To identify possible chromosomal rearrangements or transposable elements that might
880 be linked to the GWAS SNPs, we first trimmed adapters from the *H. cornu* target enrichment
881 data using *Trim Galore!* version 0.6.5 and *cutadapt* version 2.7 (Martin 2011). The trimming
882 pipeline is available at github.com/YourePrettyGood/ParsingPipeline. We then mapped the
883 reads to the *H. cornu* reference genome with *bwa mem* version 0.7.17, sorted BAM files with
884 *SAMtools sort* version 1.9, and marked duplicate reads with *Picard MarkDuplicates* version
885 2.22.7 (http://broadinstitute.github.io/picard/), all done with the *MAP* function of the
886 *PseudoreferencePipeline*. The analysis includes 43 high coverage red individuals and 42 high
887 coverage green individuals. The five individuals isolated from red galls that did not carry the
888 associated GWAS SNPs in *dgc* were excluded since the genetic basis for their gall coloration is
889 unknown.
890 We subset the BAM file for each individual to contain only the target enrichment region
891 on chromosome 1 from 40,092,625 to 40,892,915 bp using *SAMtools view* version 1.9 and
892 generated merged BAM file for each color using *SAMtools merge*. The discordant reads were
893 then extracted from each BAM file using *SAMtools view* with flag *-F 1286* and the percentage
894 of discordant reads was calculated as the ratio of the number of discordant reads over the total
895 number of mapped reads for each 5000 bp window.

896 To further explore the possibility that chromosomal aberrations near the GWAS signal
897 might differ between red- and green-gall producing individuals, we plotted the mapping
898 locations of discordant reads in the 100 kbp region near the 11 GWAS hits (40,450,000 -
899 40,550,000 bp) for red individuals, since the *H. cornu* reference was made from a green
900 individual. We obtained the read ID for all the discordant reads within the 100kbp region and
901 extracted all occurrences of these reads from the whole genome BAM file, regardless of their
902 mapping location. We then extracted the paired-end reads from the discordant reads BAM file
903 using *bedtools bamtofastq* version 2.29.2 and used *bwa mem* to map these reads as single-end
904 reads for read 1 and read 2 separately to a merged reference containing the *H. cornu* genome
905 and the 343 *Acyrthosiphon pisum* transposable elements annotated in *RepBase* (Jurka 1998).
906 We then removed duplicates and sorted the BAM file using *SAMtools rmdup* and *sort* and
907 determined the mapping location of all discordant reads using *SAMtools view*. We masked the
908 windows from chromosome 1: 40,400,000 - 40,499,999 bp and 40,500,000 - 40,599,999 bp in
909 the genome-wide scatter plot of discordant reads mapping because the majority of the
910 discordant reads are expected to map to these regions and displaying their counts would obscure
911 potential signals in the rest of the genome.
912
913 <u>Large scale survey of 11 *dgc* SNPs associated with gall color</u>
914
915 Aphids were collected from red and green galls as described above for the GWAS study
916 directly into Zymo DNA extraction buffer and ground with a plastic pestle. DNA was prepared
917 using the ZR-96 Quick gDNA kit. We developed qPCR assays and amplicon-seq assays to

genotype all individuals at all 11 SNPs (primers, etc below). PCR amplicon products were barcoded and samples were pooled for sequencing on an Illumina platform.

Adaptors were trimmed from amplicon reads using *Trim Galore!* version 0.6.5 and *cutadapt* version 2.7. The wrapper pipeline is available at github.com/YourePrettyGood/ParsingPipeline. We mapped reads to a 34 kbp region of chromosome 1 of the *H. cornu* genome that includes the amplicon SNPs (40,477,000 – 40, 511,000 bp) with *bwa mem* version 0.7.17, sorted BAM files with *SAMtools sort* version 1.9, and re-aligned indels using *GATK IndelRealigner* version 3.4. No marking of duplicates was done given the nature of amplicon sequencing data. To maximize genotyping efficiency and improve accuracy, we performed variant calling with two distinct pipelines: *SAMtools mpileup* version 1.7 (Li et al., 2009) plus *BCFtools call* version 1.7, and *GATK HaplotypeCaller* version 3.4. The mapping and indel re-alignment pipelines are available as part of the *MAP* (with flag *only_bwa*) and *IR* functions of the *PseudoreferencePipeline*. Using the same *PseudoreferencePipeline*, variant calling was performed using the *MPILEUP* function of *BCFtools* and *HC* function of *GATK*. FASTA sequences for each individual, where the genotyped SNPs were updated in the reference space, were then generated for both *BCFtools* and *GATK* variant calls using the above *PseudoreferencePipeline*'s *PSEUDOFASTA* function, with flags "*MPILEUP, no_markdup*" and "*HC, no_markdup*" respectively. The *BCFtools* SNP updating pipeline used *bcftools filter*, *query*, and *consensus* version 1.9, and we masked all sites where MQ <= 20 or QUAL <= 26 or DP <= 5. The HC SNP updating pipeline used *GATK SelectVariants* and *FastaAlternateReferenceMaker* version 3.4, and we masked all sites where MQ < 50, DP <= 5, GQ < 90 or RGQ < 90.

940 We then merged the variant calls from *BCFtools* and *GATK*, as well as the qPCR
941 genotyping results, and manually identified all missing or discrepant genotypes. We manually
942 curated these missing or discrepant genotype calls from the indel realigned BAM files using the
943 following criteria: for heterozygous calls, the site has to have at least two reads supporting each
944 allele, and for homozygous calls, the site has to have at least ten reads supporting the allele and
945 no reads supporting alternative alleles.
946
947 RNA-seq of salivary glands from aphids inducing red and green galls
948
949 We dissected salivary glands from fundatrices isolated from green and red galls in PBS,
950 gently pipetted the salivary glands from the dissection tray in < 0.5uL volume of PBS, and
951 deposited glands into 3 uL of Smart-seq2 lysis buffer (0.2%Triton-X 100, 0.1 U/uL RNasin®
952 Ribonuclease Inhibitor). RNA-seq libraries were prepared with a single-cell RNA-seq method
953 developed by the Janelia Quantitative Genomics core facility and described previously
954 (Cembrowski et al. 2018).
955 RNAseq libraries were prepared as described above for red and green gall samples
956 except that the entire 3uL sample of salivary glands in lysis buffer was provided as input.
957 Barcoded samples were pooled and sequenced on an Illumina NextSeq 550. We detected 9.0
958 million reads per sample on average. We replaced the original oligonucleotides with the
959 following oligonucleotides to generate unstranded reads from the entire transcript:

Name	Sequence	Scale	Purification
WGB_RTr1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN	250nm	PAGE
WGB_PCRr1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG	250nm	HPLC
WGB_PCRr2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG	250nm	HPLC
WGB_TSOr2	/5Biosg/GTCTCGTGGGCTCGGAGATGTGTATAAGAGACArGrGrG	250nmR	RNASE

960

961 Samples were PCR amplified for 18 cycles and the library was prepared using ¼ of the standard
962 Nextera XT sample size and 150 pg of cDNA.
963
964 Differential expression analyses of fundatrix salivary glands from red and green galls
965
966 All differential expression analyses for plant and aphid samples were performed in *R*
967 version 3.6.1 (R Core Team 2018) and all *R* Notebooks are provided on Github (will be posted
968 prior to publication). Adapters were trimmed using *cutadapt* version 2.7 and read counts per
969 transcript were calculated by mapping reads to the genome with *hisat2* version 2.1.0 (Kim,
970 Langmead, and Salzberg 2015) and counting reads per gene with *htseq-count* version 0.12.4
971 (Anders, Pyl, and Huber 2015). In R, technical replicates were examined and pooled, since all
972 replicates were very similar to each other. We performed exploratory data analysis using
973 interactive multidimensional scaling plots, using the command *glMDSPlot* from the package
974 *Glimma* (Su et al. 2017), and outlier samples were excluded from subsequent analyses.
975 Differentially expressed genes were identified using the *glmQLFTest* and associated functions
976 of the package *edgeR* (Robinson, McCarthy, and Smyth 2009). Volcano plots were generated
977 using the *EnhancedVolcano* command from the package *EnhancedVolcano* version 1.4.0
978 (Blighe, Rana, and Lewis 2020). Mean-Difference (MD) plots and multidimensional scaling
979 plots were generated using the *plotMD* and *plotMDS* functions, respectively, from the package
980 *limma* version 3.42.0 (Ritchie et al. 2015).
981
982 *Hamamelis virginiana* genome sequencing, assembly, and annotation
983

984 Leaves from a single tree of *Hamamelis virginiana* were sampled from the Janelia
985 Research Campus forest as follows. Branches containing leaves that were less than 50%
986 expanded were wrapped with aluminum foil and harvested after 40 hours. Leaves were cleared
987 of obvious contamination, including aphids and other insects, and then plunged into liquid N_2.
988 Samples were stored at -80°C and sent to the Arizona Genomics Institute, University of
989 Arizona on dry ice, which prepared HMW DNA from nuclei isolated from the frozen leaves.
990 The Janelia Quantitative Genomics core facility generated a 10X chromium linked-read library
991 from this DNA and sequenced the library on an Illumina NextSeq 550 to generate 608M linked
992 reads.

993 The *H. virginiana* genome was assembled with the *supernova* commands *run* and
994 *mkoutput* version 2.1.1, with options minsize=1000 and style=pseudohap (Weisenfeld et al.
995 2017). We used 332M reads in the assembly to achieve raw coverage of 56X as recommended
996 by the supernova instruction manual. BUSCO completeness analysis (Simao et al. 2015) was
997 performed using the *gVolante* Web interface (Nishimura, Hara, and Kuraku 2017) using the
998 plants database. Genome assembly and BUSCO statistics are reported below.

Assembly statistics	**Hvir_nuclei_sn_run2_2_pseudohap_mskd.smplchr**
# of scaffolds:	84,975
Total length:	907,642,797
Longest sequence:	7,097,227
N50 sequence length:	167,515
Sum length of sequences > 1M	142
Sum length of sequences > 1M (nt)	303391067 (33.4% of total length)

999

BUSCO Analysis	**Hvir_nuclei_sn_run2_2_pseudohap_mskd.smplchr**

Total # of core genes queried	1440
# of complete core genes detected	1309 (90.90%)
# of complete and partial core genes detected	1365 (94.79%)
# of missing core genes:	75 (5.21%)
Average # of orthologs per core genes:	1.05
% of detected core genes that have more than 1 ortholog:	3.90

1000

1001 The assembled genome reference was repeat masked with soft masking using
1002 *RepeatMasker* version 4.0.9 (Smit n.d.). Twenty-five RNA-seq libraries from galls and leaves
1003 were used for genome annotation. RNA-seq reads were adapter trimmed using *cutadapt* version
1004 2.7 and mapped to the genome using *HISAT2* version 2.1.0. Genome annotation was performed
1005 with *BRAKER* version 2.1.4 using the RNAseq data to provide intron hints (Altschul et al.
1006 1990; Barnett et al. 2011; Camacho et al. 2009; Hoff et al. 2016, 2019; Li et al. 2009;
1007 Lomsadze et al. 2014; Ritchie et al. 2015; Stanke et al. 2006, 2008) and 3' UTRs were
1008 predicted using *UTRme* (Radío et al. 2018).

1009

1010 *H. virginiana* RNA extraction and RNA-seq library preparation

1011

1012 RNA was extracted from frozen *H. virginiana* leaf or gall tissue as follows. Plant tissue
1013 frozen at -80°C was placed into ZR BashingBead Lysis Tubes (pre-chilled in liquid N_2) and
1014 pulverized to a fine powder in a Talboys High Throughput homogenizer (Troemer) with
1015 minimal thawing. Powdered plant tissue was suspended in extraction buffer (100 mM Tris-HCl,
1016 pH 7.5, 25 mM EDTA, 1.5 M NaCl, 2% (w/v) Hexadecyltrimethylammonium bromide, 10%
1017 Polyvinylpyrrolidone (w/v) and 0.3% (v/v) β-mercaptoethanol) and heated to 55°C for 8 min

1018 followed by centrifugation at 13,000 x g for 5 minutes at room temperature to remove insoluble
1019 debris (Jordon-Thaden et al. 2015). Total RNA was extracted from the supernatant using the
1020 Quick-RNA Plant Miniprep Kit (Zymo Research) with the inclusion of in-column DNAse I
1021 treatment. RNA-seq libraries were prepared with the Universal Plus mRNA-Seq kit (Nugen).
1022
1023 Differential expression analysis of red versus green galls
1024
1025 We performed differential expression analysis on red and green galls by collecting
1026 paired red and green gall samples from the same leaves. In total, we collected 17 red galls and
1027 23 green galls from 17 leaves. RNA was prepared as described above for plant material and
1028 RNA-seq libraries were prepared with a single-cell RNA-seq method modified for plant
1029 material described above.
1030 These RNAseq libraries contained on average 4.4 million mapped reads per sample.
1031 Reads were quality trimmed and mapped to the transcriptome as described above. Only genes
1032 with greater than 1 count per million in at least 15 samples were included in subsequent
1033 analyses. Red and green samples clustered together in a principal components analysis (Fig.
1034 S9A, B) and no samples were identified as outliers. The expression analysis model included the
1035 effect of leaf blocking.
1036
1037 Gall pigment extraction and analysis
1038
1039 Frozen gall tissue was ground to a powder under liquid nitrogen. We first tested for the
1040 presence of carotenoids by dehydrating gall tissue with methanol and extracting with

1041 hexane/acetone (Zhong et al. 2019). However, the lipophilic extract was colorless and all color
1042 remained in the polar phase and pellet, indicating that carotenoids do not contribute to red gall
1043 color.

1044 We therefore next tested for presence of anthocyanins. Approximately 20 mg of ground
1045 gall tissue was suspended in 100 µL methanol (Optima grade, Fisher Scientific) and 400 µl of
1046 5% aqueous formic acid (Optima grade, Fisher Scientific), vortexed for 30 seconds and
1047 centrifuged at 8000 x *g* for 2 min at 10° C. The supernatant containing pigment was filtered
1048 using a 0.2 µm, 13 mm diameter PTFE syringe filter to remove debris. Colorless pellet was
1049 discarded. Authentic anthocyanin pigment standards for malvidin 3,5-diglucoside chloride
1050 (Sigma Aldrich, St. Louis, MO, USA) and peonidin-3,5-diglucoside chloride (Carbosynth
1051 LLC,San Diego, CA, USA) were prepared at 1mg/mL in 5% aqueous formic acid.

1052 Pigment separation and identification alongside standards was performed on a reverse
1053 phase C18 column (Acquity Plus BEH, 50 mm × 2.1 mm, 1.7 µm particle size, Waters, Milford,
1054 MA) using an Agilent 1290 UHPLC coupled to an Agilent 6545 quadrupole time-of-flight mass
1055 spectrometer (Agilent Technologies, Santa Clara, CA, USA) using an ESI probe in positive ion
1056 mode. Five µl of filtered pigment extract or a 1:100 dilution of anthocyanin standard was
1057 injected Solvent (A) consisted of 5% aqueous formic acid and (B) 1:99 water/acetonitrile
1058 acidified with 5% aqueous formic acid (*v/v*) (B). The gradient conditions were as follows: 1 min
1059 hold at 0% B, 4 min linear increase to 20% B, 5 min linear increase to 40% B, ramp up to 95%
1060 B in 0.1 min, hold at 95% B for 5 min, return to 0% B and hold for 0.9 min. The column flow
1061 rate was 0.3 mL/min, and the column temperature 30° C. The MS source parameters for initial
1062 anthocyanin detection were as follows: capillary = 4000 V, nozzle = 2000 V, gas temperature =
1063 350° C, gas flow = 13 L/min, nebulizer = 30 psi, sheath gas temperature = 400° C, sheath gas

1064 flow = 12 L/min. DAD detection at 300 nm and 520 nm, and MS scanning from 50-1700 m/z at
1065 a rate of 2 spectra per second. Iterative fragmentation, followed by targeted MS/MS
1066 experiments, were performed using a collision energy = 35. Authentic standards confirmed the
1067 presence of peonidin-3,5-diglucoside and malvidin 3,5-diglucoside. The remaining anthocyanin
1068 species were identified using UV-Vis spectra, retention time relative to the other species in the
1069 sample, $[M]^+$ precursor ions and aglycone fragment ions matching the respective entries in the
1070 RIKEN database (Sawada et al. 2012).
1071
1072 Differential expression analysis of galls versus leaves
1073
1074 We performed RNA seq on 36 gall samples and 17 adjacent leaf samples. These gall
1075 samples did not overlap with the gall samples used in red versus green gall comparison
1076 described earlier. For larger galls, RNA was isolated separately from basal, medial, and apical
1077 gall regions (Table TAR Methods S8). Libraries were sequenced on an Illumina NextSeq 550 to
1078 generate 150 bp paired-end reads with an average of 8.1 million mapped reads per sample. Only
1079 genes expressed at greater than 1 count per million in at least 18 samples were included in
1080 subsequent analysis. Gall and leaf samples clustered separately in a principal components
1081 analysis (Fig. S9C, D) and no samples were excluded as outliers. We included only samples
1082 where there are paired gall and leaf samples from the same leaf and potential leaf effects were
1083 modeled in the different expression analysis.
1084 To facilitate Gene Ontology (GO) analysis, the UniProt IDs of the differentially
1085 expressed genes were obtained by mapping the coding sequence of the *H. virginana* genome to
1086 the UniProt / Swiss-Prot database (Bateman 2019) using Protein-Protein BLAST 2.7.1

1087 (Altschul et al. 1990; Camacho et al. 2009) and by extracting the differentially expressed genes
1088 (Fig. S9E). We used the WebGestalt 2019 webtool (Liao et al. 2019) to perform GO analysis on
1089 the differentially expressed genes.
1090
1091 Differential expression analysis of *H. cornu* organs and life stages
1092
1093 RNAseq libraries were generated for fundatrix salivary glands (N = 20) and whole
1094 bodies (N = 8), G2 salivary glands (N = 6) and carcasses (N = 3), G5 salivary glands (N = 6)
1095 and carcasses (N = 2), and G7 salivary glands (N = 5). Libraries were generated as described
1096 above for salivary glands except that RNA samples of carcasses and whole bodies were
1097 prepared using the Arcturus PicoPure RNA Isolation Kit (Applied Biosystems). Only genes
1098 expressed with at least 1 count per million in at least 29 samples were included in subsequent
1099 analyses.
1100
1101 Bioinformatic identification of *bicycle* genes in *H. cornu*
1102
1103 Genes that were upregulated specifically in the salivary glands of the fundatrix
1104 generation were prime candidates for inducing galls. We therefore identified genes that were
1105 upregulated both in salivary glands of fundatrices versus sexuals and in fundatrix salivary
1106 glands. These differentially expressed genes were then separated into genes with and without
1107 homologs containing some functional annotation. Homologs with previous functional
1108 annotations were identified using three methods: we performed (1) translated query-protein
1109 (*blastx*) and (2) protein-protein (*blastp*) based homology searches using BLAST 2.7.1 (Altschul

1110 et al. 1990; Camacho et al. 2009) against the UniProt / Swiss-Prot database (Bateman 2019)
1111 (Bateman 2019), and (3) Hidden-Markov based searches with the predicted proteins using
1112 *hmmscan* in *HMMER* version 3.1b2 (Eddy 2011) against the *pfam* database (Finn et al. 2014).
1113 For all predicted proteins, we also searched for secretion signal peptides using *SignalP-5.0*
1114 (Almagro Armenteros et al. 2019) and for transmembrane domains using *tmhmm* version 2.0
1115 (Krogh et al. 2001). Gene Ontology analysis of genes with annotations that were enriched in
1116 fundatrix salivary glands was performed by searching for *Drosophila melanogaster* homologs
1117 of differentially expressed genes and using these *D. melanogaster* homologs as input into gene
1118 ontology analysis.

1119 To determine whether any of these genes without detectable homologs in existing
1120 protein databases were homologous to each another, we performed sensitive homology searches
1121 of all-against-all of these genes using *jackhmmer* in *HMMER* version 3.1b2. We performed
1122 hierarchical clustering on the quantitative results of the *jackhmmer* analysis by first calculating
1123 distances amongst genes with the *dist* function using method *canberra* and clustering using the
1124 *hclust* function with method *ward.D2*, both from the library stats in *R* (R Core Team 2018). We
1125 aligned sequences of the clustered homologs using *MAFFT* version 7.419 with default
1126 parameters (Katoh 2002; Katoh and Standley 2013), trimmed aligned sequences using *trimAI*
1127 (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009) with parameters -*gt* 0.50, and
1128 generated sequence logos by importing alignments using the functions *read.alignment* and
1129 *ggseqlogo* in the *R* packages *seqinr* (Charif and Lobry 2007) and *ggseqlogo* (Wagih 2017).
1130 After identification of the *bicycle* genes, we searched for additional *bicycle* genes in the entire
1131 *H. cornu* genome, which might not have been enriched in fundatrix salivary glands, using

1132 *jackhmmer* followed by hierarchical clustering to identify additional putative homologs. As
1133 described above, we manually annotated all these putative *bicycle* genes.
1134
1135 Differential expression analysis of a single *H. cornu* fundatrix with a *dgc*^{Green} genotype inducing
1136 a red gall
1137
1138 Approximately 2.1% of fundatrices inducing red galls were homozygous for ancestral
1139 "green" alleles at all of the 11 *dgc* SNPs (Fig. 2E). Five such individuals were found in our
1140 original GWAS study and we did not observe any variants in the *dgc* gene region that were
1141 associated specifically with these individuals, suggesting that they carried variants elsewhere in
1142 the genome that caused them to generate red galls. Since isolating salivary glands from
1143 fundatrices is challenging and time consuming, we were unable to systematically examine
1144 transcriptome changes in the salivary glands of the rare individuals homozygous for *dgc*^{Green}
1145 that induced red galls. However, we fortuitously isolated salivary glands from one fundatrix
1146 from a red gall that was homozygous for *dgc*^{Green}. We performed whole-transcriptome
1147 sequencing of the salivary glands from this one individual and compared expression levels of
1148 all genes in the *bicycle* gene paralog group to which *dgc* belongs. We also examined the *dgc*
1149 transcripts produced by this individual and found no exonic SNPs in the *dgc* transcripts
1150 produced by this individual, indicating that this individual probably expressed a functional *dgc*
1151 transcript.
1152
1153 *H. cornu* and *H. hamamelidis* polymorphism and divergence measurements in GWAS region
1154

1155 To summarize the polymorphism and divergence patterns in *H. cornu* and *H.*
1156 *hamamelidis* in the target enrichment region, we generated SNP updated FASTA sequences for
1157 each individual by mapping the *H. cornu* reads to the *H. cornu* genome reference and the *H.*
1158 *hamamelidis* reads to a *H. hamamelidis* SNP-updated reference genome in *H. cornu* genome
1159 coordinate space. We used the same genotyping pipeline as described in the enrichment region
1160 GWAS, using the *PseudoreferencePipeline*, and masked sites were DP <= 10, MQ <= 20.0, or
1161 QUAL <= 29.5. High coverage individuals of *H. cornu* (n=90) and *H. hamamelidis* (N=92)
1162 were included in this analysis, including both of the color phenotypes. Polymorphism within
1163 each species and divergence (Dxy) between species were calculated using a custom script
1164 *calculateDiversity* (available at
1165 https://github.com/YourePrettyGood/DyakInversions/tree/master/tools). Windowed
1166 measurements of each of these statistics are generated with the custom script
1167 *nonOverlappingWindows.cpp*, using window size 3000bp (script available at
1168 github.com/YourePrettyGood/RandomScripts). Sites that are missing genotypes due to masking
1169 in 50% or more of the samples are not included in the windowed average and windows where
1170 50% or more of the sites are missing are excluded from the plot in Fig. S13E.
1171
1172 <u>Whole-genome polymorphism and divergence measurements in *H. cornu* and *H. hamamelidis*</u>
1173
1174 To examine genome-wide pattern of polymorphism and divergence, we used the same
1175 whole-genome sequencing data as in the genome-wide association study for gall color. This
1176 data set includes samples from fundatrices isolated from 43 green galls and 47 red galls for *H.*
1177 *cornu* and 48 galls for *H. hamamelidis*. Reads were mapped using *bwa mem* version 0.7.17-

r1188. The *H. cornu* reads were mapped to the *H. cornu* reference genome and the *H. hamamelidis* reads were mapped to the *H. hamamelidis* SNPs updated reference genome, which are in the same coordinate space. We joint genotyped each species separately using *bcftools* version 1.9 *mpileup* and *call -m* to generate raw, multi-sample VCF for each species. Sites filtering for QUAL >= 20 was then done using *bcftools filter*, and insertions and deletions sites were removed using *VCFtools* version 0.1.16 *–remove-indels*. To generate a BED file of sites lacking genotypes for each sample, *bcftools view* with flag *-g ^miss* was used to extract only the genotyped sites for each sample, *bcftools query* was used to generate the location of all SNPs and reference calls, and *bedtools* version 2.29.2 *complement* was used to generate the complementary positions to these genotyped calls as the positions to be masked. Consensus sequences were generated for each sample using *bcftools consensus*, with the flag *–iupac-codes*.

Polymorphism within each species and divergence (Dxy) between species were calculated using a custom script *calculateDiversity* (available at https://github.com/YourePrettyGood/DyakInversions/tree/master/tools). Windowed measurements of each of these statistics were generated with the custom script *nonOverlappingWindows.cpp*, using window size 1000bp (script available at github.com/YourePrettyGood/RandomScripts). Windows were designated as "*bicycle*" if they overlapped in coordinates with any *bicycle* genes; otherwise they were designated as "non-*bicycle*". Sites masked in 80% or more of the samples were excluded in the windowed average and windows with 80% or more of the sites missing were excluded from genome-wide summary statistic calculations. This genotyping rate filter is more lenient to accommodate the higher amount of missing data due to the shallower sequencing depth.

1201 Bioinformatic analysis of *bicycle* gene structure and evolution

1202

1203 The *H. cornu* median exon size and number of exons per *bicycle* gene were calculated
1204 from the GFF annotation file Augustus.updated_w_annots.21Aug20.gff3. The *bicycle* genes
1205 alignment for the divergence tree was generated using *FastTree* version 2.1.11 (Price, Dehal,
1206 and Arkin 2010) and plotted using *ggtree* version 2.2.2 (Yu 2020) in R.

1207 To calculate DnDs between *H. cornu* and *H. hamamelidis*, we first generated a masked
1208 *H. hamamelidis* reference genome by updating the *H. cornu* reference genome with *H.*
1209 *hamamelidis* SNPs. We randomly subset *H. hamamelidis* reads from 150bp PE 10X linked
1210 reads sequencing data to roughly 65X coverage using *seqtk* version 1.3 *sample -s100*
1211 (https://github.com/lh3/seqtk) and trimmed adapters using *Long Ranger* version 2.2.2 *basic*
1212 (10X Genomics). We then mapped the *H. hamamelidis* reads to the *H. cornu* reference and
1213 updated the SNPs using the *PseudoreferencePipeline*. Sites where MQ <= 20, QUAL <= 26 or
1214 DP <= 5 were masked, resulting in 21.0% of the genome being masked. This *H. hamamelidis*
1215 reference genome was used in the DnDs calculation.

1216 DnDs between all orthologous genes in *H. cornu* and *H. hamamelidis* was calculated
1217 using the *codeml* function from the *PAML* package, version 4.9j (Yang 2007). The CDS
1218 sequences were extracted from both reference genomes for each gene using
1219 constructCDSesFromGFF3.pl and we split all degenerate bases into the two alleles to generate
1220 two pseudo-haplotypes using *fakeHaplotype.pl -s 42* (both scripts available at
1221 github.com/YourePrettyGood/RandomScripts). Haplotype 1 was used as input into *codeml* for
1222 both species. The settings used in *codeml* control file are: runmode=-2, seqtype=1,
1223 CodonFreq=0, clock=0, model=0, NSsites=0, fix_kappa=0, kappa=2, fix_omega=0, omega=1,

1224 Small_Diff=0.5e-6, method=0, fix_blength=0. Then, to evaluate whether DnDs is significantly
1225 different from 1 for each gene, we additionally ran a null model with the same control file,
1226 except with fix_omega=1. We then calculated the likelihood scores between the two models as
1227 2×(lnL1-lnL2) and compared it to the 95th percentile of the chi-square distribution with 1
1228 degree of freedom. Mann-Whitney U test comparing DnDs of *bicycle* and non-*bicycle* genes
1229 was performed using the *wilcox.test* function in *R* version 3.6.1.

1230 To perform McDonald-Kreitman tests for *bicycle* genes, we generated population
1231 sample of the *bicycle* coding regions for *H. cornu* by mapping RNA-seq data from 21 fundatrix
1232 salivary gland samples, which provided high coverage. We performed genotyping using the
1233 *STAR*, *IRRNA*, and *HC* functions in the *PseudoreferencePipeline*. This pipeline uses *STAR*
1234 version 2.7.3a for mapping and *IndelRealigner* and *HaplotypeCaller* in *GATK* version 3.4 for
1235 indel realignment and variant calling. SNP-updated FASTA sequences were generated for each
1236 individual from the genotype VCF using the *PSEUDOFASTA* function of the
1237 *PseudoreferencePipeline* without any additional masking. Then, for each individual, we split all
1238 degenerate bases into the two alleles to generate two pseudo-haplotypes using *fakeHaplotype.pl*
1239 *-s 42*. We calculated the number of synonymous and nonsynonymous polymorphisms (P_S and
1240 P_N, respectively) and divergent substitutions (D_S and D_N, respectively) using *Polymorphorama*
1241 version 6
1242 (https://ib.berkeley.edu/labs/bachtrog/data/polyMORPHOrama/polyMORPHOrama.html)
1243 (Andolfatto 2007) for multiple categories of genes significantly over-expressed in the fundatrix
1244 salivary gland. The fraction of non-synonymous divergent substitutions that were fixed by
1245 positive selection (α; Fay, Wyckoff, and Wu 2002; Smith and Eyre-Walker 2002) was
1246 estimated using the Cochran-Mantel-Haenszel test framework (Shapiro et al. 2007). Using the

1247 *mantelhaen.test* function in *R* version 3.6.1, we estimated a common odds ratio ($\frac{D_s P_n}{D_n P_s}$) across a
1248 series of McDonald-Kreitman tables (McDonald and Kreitman 1991), and estimated $\alpha = 1 -$
1249 $\frac{D_s P_n}{D_n P_s}$. A MAF filter of 0.03 was applied to the polymorphism counts to reduce the downward
1250 bias in the estimator due to segregating weakly deleterious amino acid polymorphisms(Fay et
1251 al. 2002).
1252
1253 Genome-wide tests of selection
1254
1255 To scan for genome-wide signatures of positive selection in *H. cornu* and *H.*
1256 *hamamelidis*, we used the joint-genotyped VCF for each species as described above. We
1257 filtered for QUAL >= 20 using *bcftools filter,* and 80% maximum missing genotype and
1258 biallelic sites using *VCFtools* version 0.1.16 *-max-missing 0.2*, *-max-alleles 2*, and *-min-alleles*
1259 *2*. We then performed the selection scan using *SweeD-P* version 3.1 (Pavlidis et al. 2013), with
1260 one grid point per kilobase and *-folded* flag.
1261 To determine the significance cutoff for the composite likelihood ratio (CLR) output by
1262 *SweeD*, we simulated 10 Mbp regions under neutrality for 100 haplotypes for each species
1263 using *MaCS* version 0.4f (Chen, Marjoram, and Wall 2008). The effective population size was
1264 set to 1,174,297 and 1,115,822 for *H. cornu* and *H. hamamelidis*, respectively. The mutation
1265 and recombination rates were set to 2.8e-9 and 1.1e-8, respectively, per bp per generation. The
1266 simulation output was converted to multi-sample VCF format using a custom script and run
1267 through *SweeD-P* using the same parameters as for the observed data. We ran ten iterations of
1268 the simulation for each species and combined the results to account for stochasticity of the

simulations. The CLR cutoffs for the observed data were chosen as the 99th quantiles of the CLR values from the neutral simulations of their respective species.

Supplementary Text

Chromosomal aberrations cannot explain association of 11 SNPs with gall color

We considered the possibility that DNA aberrations linked to the 11 associated SNPs might explain the red-green gall polymorphism. We found no evidence for large insertions, deletions, or other chromosomal aberrations in this region (Fig. S1E-I). In addition, we performed a genome-free association study (Rahman et al. 2018) and found strong linkage only for a subset of the 11 SNPs identified in the original GWAS (Fig. S3), further supporting the inference that these SNPs, and not other genetic changes, are associated with gall color.

A small fraction of fundatrices likely carry variants not linked to *dgc* that induce red galls

Two percent of fundatrices from red galls were homozygous for the *dgc*^{Green} alleles at all 11 SNPs (Fig. 2E). These individuals do not carry exonic SNPs in *dgc* that would alter *dgc* function. We fortuitously performed RNA-seq on the salivary glands from one of these rare fundatrices and found that *dgc* was expressed at high levels (Fig. S7C), which suggests that the rare individuals that make red galls and are homozygous for *dgc*^{Green} alleles harbor variants at loci other than *dgc* that generate red galls.

1292 Gene ontology analysis of plant genes differentially expression in galls is consistent with the
1293 observed cell biology of gall development
1294
1295 Genes associated with cell division were strongly upregulated in galls (Fig. S9F),
1296 consistent with the extensive growth observed in gall tissue (Figure 1G-I). Genes associated
1297 with chloroplasts and photosynthesis were strongly downregulated in gall tissue (Fig. S9F),
1298 probably reflecting the reprogramming of plant leaf tissue from nutrient exporters to nutrient
1299 sinks (Inbar, Eshel, and Wool 1995). We also observed downregulation of genes related to
1300 immune response and multiple defense mechanisms in gall tissue (Fig. S9F), which may reflect
1301 suppression of the plant's defense responses to infection. Some of these patterns are similar to
1302 patterns of differential expression observed in other gall systems (Hearn et al. 2019; Martinson,
1303 Werren, and Egan 2020). The pattern of differential expression does not appear to reflect a plant
1304 wound response, because wound response genes are not significantly over-expressed in gall
1305 tissue (Chi-square P-value = 0.367) and are biased toward under-represented.
1306
1307 Evidence for roles of homologs of genes previously proposed as gall effector genes in *H. cornu*
1308
1309 Previously, the SSGP-71s effector gene family was described from the Hessian fly
1310 genome (Zhao et al. 2015) and it was suggested that these encode E3-ligase-mimicking effector
1311 proteins. We identified four homologs of the SSGP-71s effector gene family in *H. cornu* by
1312 searching *H. cornu* protein sequences using a hidden marker model search with a protein profile
1313 generated using the 426 SSGP-71 genes from Document S2, SSGP-71s tab, from Zhao et al.
1314 2015 (hmmsearch with HMMER v3.2.1). However, none of these genes were differentially

1315 expressed in the fundatrix salivary gland. We also identified seven ubiquitin E3-ligase genes
1316 (Zhao et al. 2015) in *H. cornu*, two of which are upregulated in fundatrix salivary gland.
1317 However, none of these seven genes contain N-terminal secretion signals, making it unlikely
1318 that they could be injected into plants. Thus, these genes are unlikely to contribute to *H. cornu*
1319 gall development.

1320 Ring domain proteins usually mediate E2 ubiquitin ligase activity, which can modify
1321 protein function or target proteins for proteosomal degradation (Metzger et al. 2014). The
1322 genome of a galling phylloxeran, a species from the sister family to aphids, has been shown to
1323 encode secreted RING domain proteins (Zhao, Rispe, and Nabity 2019), although it is not yet
1324 clear how the phylloxeran secreted RING domain proteins act and whether they are involved in
1325 gall-specific processes. We identified two *H. cornu* secreted RING domain proteins based on
1326 their similarity to proteins in the PFAM database (Finn et al. 2014), but only one was
1327 upregulated weakly in fundatrix salivary glands. Thus, secreted RING domain proteins are
1328 unlikely to contribute to *H. cornu* gall development.

1329

1330 Characterization of *H. cornu* and *H. hamamelidis*

1331

1332 The sister species *H. cornu* and *H. hamamelidis* produce similar looking galls in
1333 adjacent geographic areas and were long confused as populations of a single species (von
1334 Dohlen and Stoetzel 1991). However, *H. cornu* exhibits a life cycle where aphids alternate
1335 between *Hamamelis virginiana* and *Betula nigra* (Fig. 4A), whereas *H. hamamelidis* does not
1336 host alternate to *B. nigra* and displays a truncated life cycle, where the offspring of the
1337 fundatrix develop as sexuparae (the equivalent of G6 in Fig. 4A) and deposit sexuals on *H.*

virginiana in the autumn. Populations of *H. cornu* tend to live in the lowlands with prolonged summers and *H. hamamelidis* tend to live in the highlands and more northern latitudes, which experience shorted summers. In some locations, both species can be found making galls on the same trees.

Previous mtDNA sequencing supported the hypothesis that these are distinct species (von Dohlen, Kurosu, and Aoki 2002). Genome-wide divergence (Dxy) between these species was estimated to be 0.0189 and polymorphism was 0.0132 and 0.0125 for *H. cornu* and *H. hamamelidis*, respectively. Assuming a mutation rate μ=2.8e-9 (Keightley et al. 2014), we estimated effective population sizes using $\theta = 4N_e\mu$ for *H. cornu* and *H. hamamelidis* as 1,174,297 and 1,115,822, respectively.

References

Almagro Armenteros, José Juan, Konstantinos D. Tsirigos, Casper Kaae Sønderby, Thomas Nordahl Petersen, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2019a. "SignalP 5.0 Improves Signal Peptide Predictions Using Deep Neural Networks." *Nature Biotechnology* 37(4):420–23. doi: 10.1038/s41587-019-0036-z.

Almagro Armenteros, José Juan, Konstantinos D. Tsirigos, Casper Kaae Sønderby, Thomas Nordahl Petersen, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2019b. "SignalP 5.0 Improves Signal Peptide Predictions Using Deep Neural Networks." *Nature Biotechnology* 37(4):420–23. doi: 10.1038/s41587-019-0036-z.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215(3):403–10. doi: 10.1016/S0022-2836(05)80360-2.

Anders, S., P. T. Pyl, and W. Huber. 2015. "HTSeq--a Python Framework to Work with High-Throughput Sequencing Data." *Bioinformatics* 31(2):166–69. doi: 10.1093/bioinformatics/btu638.

Andolfatto, Peter. 2007. "Hitchhiking Effects of Recurrent Beneficial Amino Acid Substitutions in the Drosophila Melanogaster Genome." *Genome Research* 17(12):1755–62. doi: 10.1101/gr.6691007.

Barnett, Derek W., Erik K. Garrison, Aaron R. Quinlan, Michael P. Stŕmberg, and Gabor T. Marth. 2011. “Bamtools: A C++ API and Toolkit for Analyzing and Managing BAM Files.” *Bioinformatics* 27(12):1691–92. doi: 10.1093/bioinformatics/btr174.

Bateman, Alex. 2019. “UniProt: A Worldwide Hub of Protein Knowledge.” *Nucleic Acids Research* 47(D1):D506–15. doi: 10.1093/nar/gky1049.

Blackman, R. L., and V. F. Eastop. 1994. *Aphids on the World’s Trees: An Identification and Information Guide*. Wallingford: CAB International.

Blighe, Kevin, Sharmila Rana, and Myles Lewis. 2020. “EnhancedVolcano: Publication-Ready Volcano Plots with Enhanced Colouring and Labeling.”

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. “BLAST+: Architecture and Applications.” *BMC Bioinformatics* 10:1–9. doi: 10.1186/1471-2105-10-421.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. “TrimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses.” *Bioinformatics* 25(15):1972–73. doi: 10.1093/bioinformatics/btp348.

Cembrowski, Mark S., Matthew G. Phillips, Salvatore F. DiLisio, Brenda C. Shields, Johan Winnubst, Jayaram Chandrashekar, Erhan Bas, and Nelson Spruston. 2018. “Dissociable Structural and Functional Hippocampal Outputs via Distinct Subiculum Cell Classes.” *Cell* 173(5):1280-1292.e18. doi: 10.1016/j.cell.2018.03.031.

Charif, D., and J. R. Lobry. 2007. “Seqin{R} 1.0-2: A Contributed Package to the {R} Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis.” Pp. 207–32 in *Structural approaches to sequence evolution: Molecules, networks, populations*, *Biological and Medical Physics, Biomedical Engineering*, edited by U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo. New York: Springer Verlag.

Chen, G. K., P. Marjoram, and J. D. Wall. 2008. “Fast and Flexible Simulation of DNA Sequence Data.” *Genome Research* 19(1):136–42. doi: 10.1101/gr.083634.108.

Clayton, David. 2019. “SnpStats: SnpMatrix and XSnpMatrix Classes and Methods.”

von Dohlen, C. D., and M. B. Stoetzel. 1991. “Separation and Redescription of Hormaphis Hamamelidis (Fitch 1851) and Hormaphis Cornu (Shimer 1867) (Homoptera: Aphididae) on Witch-Hazel in the Eastern United States.” *Proc. Entomol. Soc. Wash.* 93:533–48.

von Dohlen, Carol D., Utako Kurosu, and Shigeyuki Aoki. 2002. “Phylogenetics and Evolution of the Eastern Asian-Eastern North American Disjunct Aphid Tribe, Hormaphidini (Hemiptera: Aphididae).” *Molecular Phylogenetics and Evolution* 23(2):257–67. doi: 10.1016/S1055-7903(02)00025-8.

Drozdetskiy, Alexey, Christian Cole, James Procter, and Geoffrey J. Barton. 2015. “JPred4: A Protein Secondary Structure Prediction Server.” *Nucleic Acids Research* 43(W1):W389–94. doi: 10.1093/nar/gkv332.

Eddy, Sean R. 2011. “Accelerated Profile HMM Searches.” *PLoS Computational Biology* 7(10). doi: 10.1371/journal.pcbi.1002195.

Fay, Justin C., Gerald J. Wyckoff, and Chung-I. Wu. 2002. “Testing the Neutral Theory of Molecular Evolution with Genomic Data from Drosophila.” *Nature* 415:1024–26.

Finn, Robert D., Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, Erik L. L. Sonnhammer, John Tate, and Marco Punta. 2014. “Pfam: The Protein Families Database.” *Nucleic Acids Research* 42(D1):222–30. doi: 10.1093/nar/gkt1223.

Hearn, Jack, Mark Blaxter, Karsten Schönrogge, José-Luis Nieves-Aldrey, Juli Pujade-Villar, Elisabeth Huguet, Jean-Michel Drezen, Joseph D. Shorthouse, and Graham N. Stone. 2019. “Genomic Dissection of an Extended Phenotype: Oak Galling by a Cynipid Gall Wasp.” *PLOS Genetics* 15(11):e1008398. doi: 10.1371/journal.pgen.1008398.

Hennig, Bianca P., Lars Velten, Ines Racke, Chelsea Szu Tu, Matthias Thoms, Vladimir Rybin, Hüseyin Besir, Kim Remans, and Lars M. Steinmetz. 2018. “Large-Scale Low-Cost NGS Library Preparation Using a Robust Tn5 Purification and Tagmentation Protocol.” *G3: Genes, Genomes, Genetics* 8(1):79–89. doi: 10.1534/g3.117.300257.

Hoff, Katharina J., Simone Lange, Alexandre Lomsadze, Mark Borodovsky, and Mario Stanke. 2016. “BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS.” *Bioinformatics* 32(5):767–69. doi: 10.1093/bioinformatics/btv661.

Hoff, Katharina J., Alexandre Lomsadze, Mark Borodovsky, and Mario Stanke. 2019. *Whole-Genome Annotation with BRAKER*. Vol. 1962.

Inbar, M., A. Eshel, and D. Wool. 1995. “Interspecific Competition among Phloem-Feeding Insects Mediated by Induced Host-Plant Sinks.” *Ecology* 76(5):1506–15. doi: 10.2307/1938152.

Jurka, Jerzy. 1998. “Repeats in Genomic DNA: Mining and Meaning.” *Current Opinion in Structural Biology* 8(3):333–37. doi: 10.1016/S0959-440X(98)80067-5.

Katoh, K. 2002. “MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform.” *Nucleic Acids Research* 30(14):3059–66. doi: 10.1093/nar/gkf436.

Katoh, Kazutaka, and Daron M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution* 30(4):772–80. doi: 10.1093/molbev/mst010.

Keightley, Peter D., Rob W. Ness, Daniel L. Halligan, and Penelope R. Haddrill. 2014. "Estimation of the Spontaneous Mutation Rate per Nucleotide Site in a Drosophila Melanogaster Full-Sib Family." *Genetics* 196(1):313–20. doi: 10.1534/genetics.113.158758.

Kim, Daehwan, Ben Langmead, and Steven L. Salzberg. 2015. "HISAT: A Fast Spliced Aligner with Low Memory Requirements." *Nature Methods* 12(4):357–60. doi: 10.1038/nmeth.3317.

Knaus, Brian J., and Niklaus J. Grünwald. 2017. "VCFR : A Package to Manipulate and Visualize Variant Call Format Data in R." *Molecular Ecology Resources* 17(1):44–53. doi: 10.1111/1755-0998.12549.

Krogh, Anders, Björn Larsson, Gunnar Von Heijne, and Erik L. L. Sonnhammer. 2001. "Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes." *Journal of Molecular Biology* 305(3):567–80. doi: 10.1006/jmbi.2000.4315.

Lewis, S. E., S. M. Searle, N. Harris, M. Gibson, V. Lyer, J. Richter, C. Wiel, L. Bayraktaroglir, E. Birney, M. A. Crosby, J. S. Kaminker, B. B. Matthews, S. E. Prochnik, C. D. Smithy, J. L. Tupy, G. M. Rubin, S. Misra, C. J. Mungall, and M. E. Clamp. 2002. "Apollo: A Sequence Annotation Editor." *Genome Biology* 3(12):1–14. doi: 10.1186/gb-2002-3-12-research0082.

Li, Heng. 2011. "A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data." *Bioinformatics* 27(21):2987–93. doi: 10.1093/bioinformatics/btr509.

Li, Heng. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM." *ArXiv* 1303.3997.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25(16):2078–79. doi: 10.1093/bioinformatics/btp352.

Liao, Yuxing, Jing Wang, Eric J. Jaehnig, Zhiao Shi, and Bing Zhang. 2019. "WebGestalt 2019: Gene Set Analysis Toolkit with Revamped UIs and APIs." *Nucleic Acids Research* 47(W1):W199–205. doi: 10.1093/nar/gkz401.

Lomsadze, Alexandre, Paul D. Burns, and Mark Borodovsky. 2014. "Integration of Mapped RNA-Seq Reads into Automatic Training of Eukaryotic Gene Finding Algorithm." *Nucleic Acids Research* 42(15):1–8. doi: 10.1093/nar/gku557.

Mapleson, Daniel, Gonzalo Garcia Accinelli, George Kettleborough, Jonathan Wright, and Bernardo J. Clavijo. 2017. "KAT: A K-Mer Analysis Toolkit to Quality Control NGS Datasets and Genome Assemblies." *Bioinformatics* 33(4):574–76. doi: 10.1093/bioinformatics/btw663.

Martin, Marcel. 2011. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." *EMBnet.Journal* 17(1):10–10. doi: 10.14806/ej.17.1.200.

Martinson, Ellen, John Werren, and Scott Egan. 2020. *Tissue-Specific Gene Expression Shows Cynipid Wasps Repurpose Host Gene Networks to Create Complex and Novel Parasite-Specific Organs on Oaks*. *preprint*. Preprints.

McDonald, J. H., and M. Kreitman. 1991. "Adaptive Protein Evolution at the Adh Locus in Drosophila." *Nature* 351(6328):652–54. doi: 10.1038/351652a0.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data." *Genome Research* 20(9):1297–1303. doi: 10.1101/gr.107524.110.

Metzger, Meredith B., Jonathan N. Pruneda, Rachel E. Klevit, and Allan M. Weissman. 2014. "RING-Type E3 Ligases: Master Manipulators of E2 Ubiquitin-Conjugating Enzymes and Ubiquitination." *Biochimica et Biophysica Acta - Molecular Cell Research* 1843(1):47–60. doi: 10.1016/j.bbamcr.2013.05.026.

Nishimura, Osamu, Yuichiro Hara, and Shigehiro Kuraku. 2017. "GVolante for Standardizing Completeness Assessment of Genome and Transcriptome Assemblies." *Bioinformatics* 33(22):3635–37. doi: 10.1093/bioinformatics/btx445.

Ott, Swidbert R. 2008. "Confocal Microscopy in Large Insect Brains: Zinc-Formaldehyde Fixation Improves Synapsin Immunostaining and Preservation of Morphology in Whole-Mounts." *Journal of Neuroscience Methods* 172(2):220–30. doi: 10.1016/j.jneumeth.2008.04.031.

Pavlidis, Pavlos, Daniel Živković, Alexandros Stamatakis, and Nikolaos Alachiotis. 2013. "SweeD: Likelihood-Based Detection of Selective Sweeps in Thousands of Genomes." *Molecular Biology and Evolution* 30(9):2224–34. doi: 10.1093/molbev/mst112.

Phanstiel, Douglas H., Alan P. Boyle, Carlos L. Araya, and Michael P. Snyder. 2014. "Sushi.R: Flexible, Quantitative and Integrative Genomic Visualizations for Publication-Quality Multi-Panel Figures." *Bioinformatics* 30(19):2808–10. doi: 10.1093/bioinformatics/btu379.

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments." *PLOS ONE* 5(3):e9490. doi: 10.1371/journal.pone.0009490.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. De Bakker, Mark J. Daly, and Pak C. Sham. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *American Journal of Human Genetics* 81(3):559–75. doi: 10.1086/519795.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26(6):841–42. doi: 10.1093/bioinformatics/btq033.

R Core Team. 2018. “R: A Language and Environment for Statistical Computing.”

Radío, Santiago, Rafael Sebastián Fort, Beatriz Garat, José Sotelo-Silveira, and Pablo Smircich. 2018. “UTRme: A Scoring-Based Tool to Annotate Untranslated Regions in Trypanosomatid Genomes.” *Frontiers in Genetics* 9(December):671–671. doi: 10.3389/fgene.2018.00671.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. “Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies.” *Nucleic Acids Research* 43(7):e47–e47. doi: 10.1093/nar/gkv007.

Robinson, James T., Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. 2011. “Integrative Genomics Viewer.” *Nature Biotechnology* 29(1):24–26. doi: 10.1038/nbt0111-24.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2009. “EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26(1):139–40. doi: 10.1093/bioinformatics/btp616.

Sawada, Yuji, Ryo Nakabayashi, Yutaka Yamada, Makoto Suzuki, Muneo Sato, Akane Sakata, Kenji Akiyama, Tetsuya Sakurai, Fumio Matsuda, Toshio Aoki, Masami Yokota Hirai, and Kazuki Saito. 2012. “RIKEN Tandem Mass Spectral Database (ReSpect) for Phytochemicals: A Plant-Specific MS/MS-Based Data Resource and Database.” *Phytochemistry* 82:38–45. doi: 10.1016/j.phytochem.2012.07.007.

Schindelin, Johannes, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, Jean-Yves Yves Tinevez, Daniel James White, Volker Hartenstein, Kevin Eliceiri, Pavel Tomancak, and Albert Cardona. 2012. “Fiji: An Open-Source Platform for Biological-Image Analysis.” *Nat Methods* 9(7):676–82. doi: 10.1038/nmeth.2019.

Shapiro, Joshua A., Wei Huang, Chenhui Zhang, Melissa J. Hubisz, Jian Lu, David A. Turissini, Shu Fang, Hurng-Yi Wang, Richard R. Hudson, Rasmus Nielsen, Zhu Chen, and Chung-I. Wu. 2007. “Adaptive Genic Evolution in the Drosophila Genomes.” *Proceedings of the National Academy of Sciences* 104(7):2271–76. doi: 10.1073/pnas.0610385104.

Shin, Ji-Hyung, Sigal Blay, Jinko Graham, and Brad McNeney. 2006. “**LDheatmap** : An *R* Function for Graphical Display of Pairwise Linkage Disequilibria Between Single Nucleotide Polymorphisms.” *Journal of Statistical Software* 16(Code Snippet 3). doi: 10.18637/jss.v016.c03.

Simao, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. “BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs.” *Bioinformatics* 31(19):3210–12. doi: 10.1093/bioinformatics/btv351.

Smit, A. F. A., Hubley, R. ,. Green, P. n.d. *RepeatMasker Open-4.0*.

Smith, N. G., and A. Eyre-Walker. 2002. “Adaptive Protein Evolution in Drosophila.” *Nature* 415(6875):1022–24. doi: 10.1038/4151022a.

Stanke, Mario, Mark Diekhans, Robert Baertsch, and David Haussler. 2008. “Using Native and Syntenically Mapped CDNA Alignments to Improve de Novo Gene Finding.” *Bioinformatics* 24(5):637–44. doi: 10.1093/bioinformatics/btn013.

Stanke, Mario, Oliver Schöffmann, Burkhard Morgenstern, and Stephan Waack. 2006. “Gene Prediction in Eukaryotes with a Generalized Hidden Markov Model That Uses Hints from External Sources.” *BMC Bioinformatics* 7:1–11. doi: 10.1186/1471-2105-7-62.

Su, Shian, Charity W. Law, Casey Ah-Cann, Marie-Liesse Asselin-Labat, Marnie E. Blewitt, and Matthew E. Ritchie. 2017. “Glimma: Interactive Graphics for Gene Expression Analysis.” *Bioinformatics* 33(13):2050–52.

Thorvaldsdóttir, Helga, James T. Robinson, and Jill P. Mesirov. 2013. “Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration.” *Briefings in Bioinformatics* 14(2):178–92. doi: 10.1093/bib/bbs017.

Wagih, Omar. 2017. “Ggseqlogo: A Versatile R Package for Drawing Sequence Logos.” *Bioinformatics* 33(22):3645–47. doi: 10.1093/bioinformatics/btx469.

Weisenfeld, Neil I., Vijay Kumar, Preyas Shah, Deanna M. Church, and David B. Jaffe. 2017. “Direct Determination of Diploid Genome Sequences.” *Genome Research* 27(5):757–67. doi: 10.1101/gr.235812.118.

Yang, Z. 2007. “PAML 4: Phylogenetic Analysis by Maximum Likelihood.” *Molecular Biology and Evolution* 24(8):1586–91. doi: 10.1093/molbev/msm088.

Yu, Guangchuang. 2020. “Using Ggtree to Visualize Data on Tree-Like Structures.” *Current Protocols in Bioinformatics* 69(1):e96. doi: 10.1002/cpbi.96.

Zhao, Chaoyang, Lucio Navarro Escalante, Hang Chen, Thiago R. Benatti, Jiaxin Qu, Sanjay Chellapilla, Robert M. Waterhouse, David Wheeler, Martin N. Andersson, Riyue Bao, Matthew Batterton, Susanta K. Behura, Kerstin P. Blankenburg, Doina Caragea, James C. Carolan, Marcus Coyle, Mustapha El-Bouhssini, Liezl Francisco, Markus Friedrich, Navdeep Gill, Tony Grace, Cornelis J. P. Grimmelikhuijzen, Yi Han, Frank Hauser, Nicolae Herndon, Michael Holder, Panagiotis Ioannidis, Laronda Jackson, Mehwish Javaid, Shalini N. Jhangiani, Alisha J. Johnson, Divya Kalra, Viktoriya Korchina, Christie L. Kovar, Fremiet Lara, Sandra L. Lee, Xuming Liu, Christer Löfstedt, Robert Mata, Tittu Mathew, Donna M. Muzny, Swapnil Nagar, Lynne V. Nazareth, Geoffrey Okwuonu, Fiona Ongeri, Lora Perales, Brittany F. Peterson, Ling Ling Pu, Hugh M. Robertson, Brandon J. Schemerhorn, Steven E. Scherer, Jacob T. Shreve, Denard Simmons, Subhashree Subramanyam, Rebecca L. Thornton, Kun Xue, George M. Weissenberger, Christie E. Williams, Kim C. Worley, Dianhui Zhu, Yiming Zhu, Marion O. Harris, Richard H. Shukle, John H. Werren, Evgeny M. Zdobnov, Ming Shun Chen, Susan J. Brown,

1576 Jeffery J. Stuart, and Stephen Richards. 2015. “A Massive Expansion of Effector Genes
1577 Underlies Gall-Formation in the Wheat Pest Mayetiola Destructor.” *Current Biology* 25(5):613–
1578 20. doi: 10.1016/j.cub.2014.12.057.

1579 Zhao, Chaoyang, Claude Rispe, and Paul D. Nabity. 2019. “Secretory RING Finger Proteins
1580 Function as Effectors in a Grapevine Galling Insect.” *BMC Genomics* 20(1):1–12. doi:
1581 10.1186/s12864-019-6313-x.

1582 Zheng, X., D. Levine, J. Shen, S. M. Gogarten, C. Laurie, and B. S. Weir. 2012. “A High-
1583 Performance Computing Toolset for Relatedness and Principal Component Analysis of SNP
1584 Data.” *Bioinformatics* 28(24):3326–28. doi: 10.1093/bioinformatics/bts606.

1585 Zhong, Siqiong, Mariona Vendrell-Pacheco, Brian Heskitt, Chureeporn Chitchumroonchokchai,
1586 Mark Failla, Sudhir K. Sastry, David M. Francis, Olga Martin-Belloso, Pedro Elez-Martínez, and
1587 Rachel E. Kopec. 2019. “Novel Processing Technologies as Compared to Thermal Treatment on
1588 the Bioaccessibility and Caco-2 Cell Uptake of Carotenoids from Tomato and Kale-Based
1589 Juices.” *Journal of Agricultural and Food Chemistry* 67(36):10185–94. doi:
1590 10.1021/acs.jafc.9b03666.

1591

1592 **QUANTIFICATION AND STATISTICAL ANALYSIS**

1593 All quantitative methods and statistical analyses were explained in the **Method Details** section
1594