



HAL
open science

Evolution of GOUNDRY, a cryptic subgroup of *Anopheles gambiae* s.l., and its impact on susceptibility to *Plasmodium* infection.

Jacob E. Crawford, Michelle M. Riehle, Kyriacos Markianos, Emmanuel Bischoff, Wamdaogo M. Guelbeogo, Awa Gneme, N’Fale Sagnon, Kenneth D Vernick, Rasmus Nielsen, Brian P. Lazzaro

► To cite this version:

Jacob E. Crawford, Michelle M. Riehle, Kyriacos Markianos, Emmanuel Bischoff, Wamdaogo M. Guelbeogo, et al.. Evolution of GOUNDRY, a cryptic subgroup of *Anopheles gambiae* s.l., and its impact on susceptibility to *Plasmodium* infection.. *Molecular Ecology*, 2016, 25 (7), pp.1494–1510. 10.1111/mec.13572 . pasteur-02008316

HAL Id: pasteur-02008316

<https://pasteur.hal.science/pasteur-02008316v1>

Submitted on 13 Mar 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1

2 Received Date : 06-May-2015

3 Revised Date : 02-Jan-2016

4 Accepted Date : 18-Jan-2016

5 Article type : Original Article

6

7

8 **Title:** Evolution of GOUNDRY, a cryptic subgroup of *Anopheles gambiae s.l.*, and its
9 impact on susceptibility to *Plasmodium* infection.

10

11 **Running Head:** Evolution of the *Anopheles* GOUNDRY subgroup.

12

13 **Authors:** Jacob E. Crawford^{1,2}, Michelle M. Riehle³, Kyriacos Markianos⁴, Emmanuel
14 Bischoff⁵, Wamdaogo M. Guelbeogo⁶, Awa Gneme⁶, N'Fale Sagnon⁶, Kenneth D.
15 Vernick⁵, Rasmus Nielsen^{2*}, Brian P. Lazzaro^{1*}.

16 * These authors contributed equally to this work.

17

18 **Affiliations:**

19 1. Department of Entomology, Cornell University, Ithaca, NY, USA

20 2. Department of Integrative Biology, University of California, Berkeley, Berkeley, CA,
21 USA

22 3. Department of Microbiology, University of Minnesota, St. Paul, MN, USA

23 4. Program in Genomics, Children's Hospital Boston, Harvard Medical School

24 5. Unit for Genetics and Genomics of Insect Vectors, Institut Pasteur, Paris, France

25 6. Centre National de Recherche et de Formation sur le Paludisme, 1487 Avenue de
26 l'Oubritenga, 01 BP 2208 Ouagadougou, Burkina Faso.

27

28 **Corresponding Author:** Jacob Crawford, Department of Integrative Biology,

29 University of California, Berkeley, Berkeley, CA, USA, j.crawford@berkeley.edu

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as [doi: 10.1111/mec.13572](https://doi.org/10.1111/mec.13572)

30

31 **Key Words:** *Anopheles gambiae*, malaria, population genetics, inbreeding, demography,
32 speciation

33

34 **Abstract:**

35 The recent discovery of a previously unknown genetic subgroup of *Anopheles gambiae*
36 *sensu lato* underscores our incomplete understanding of complexities of vector
37 population demographics in *Anopheles*. This subgroup, named GOUNDRY, does not rest
38 indoors as adults and is highly susceptible to *Plasmodium* infection in the laboratory.
39 Initial description of GOUNDRY suggested it differed from other known *Anopheles* taxa
40 in surprising and sometimes contradictory ways, raising a number of questions about its
41 age, population size, and relationship to known subgroups. To address these questions,
42 we sequenced the complete genomes of 12 wild-caught GOUNDRY specimens and
43 compared these genomes to a panel of *Anopheles* genomes. We show that GOUNDRY is
44 most closely related to *Anopheles coluzzii*, and the timing of cladogenesis is not recent,
45 substantially predating the advent of agriculture. We find a large region of the X
46 chromosome that has swept to fixation in GOUNDRY within the last 100 years, which
47 may be an inversion that serves as a partial barrier to contemporary gene flow.
48 Interestingly, we show that GOUNDRY has a history of inbreeding that is significantly
49 associated with susceptibility to *Plasmodium* infection in the laboratory. Our results
50 illuminate the genomic evolution of one of probably several cryptic, ecologically
51 specialized subgroups of *Anopheles* and provide a potent example of how vector
52 population dynamics may complicate efforts to control or eradicate malaria.

53

54

55

56 **Introduction:**

57 The continued devastating burden of malaria on human populations in sub-
58 Saharan Africa (Murray *et al.* 2012; WHO 2013) spurs ongoing searches for novel means
59 of controlling vector mosquitoes, including through genetic manipulation. However, it is
60 becoming increasingly appreciated that *Anopheles* species frequently form partially

61 reproductively isolated and ecologically differentiated subpopulations (Costantini *et al.*
62 2009; Gnémé *et al.* 2013; Lee *et al.* 2013; Fontaine *et al.* 2015), which could complicate
63 control efforts and extend disease transmission across seasons and micro-environmental
64 space. As an example, a recent study showed that subgroups of *Anopheles gambiae*
65 *sensu lato* have evolved distinct approaches for surviving the dry season resulting in the
66 presence of vector populations throughout an extended proportion of the year (Dao *et al.*
67 2014). Comprehensive genomic analysis of evolutionary origins, demography, and
68 adaptation will advance our understanding of such phenotypic divergence and its role in
69 the formation of new *Anopheles* subgroups. Furthermore, genomic analysis of population
70 diversity and genetic affinity among taxa can elucidate epidemiologically relevant aspects
71 of population ecology like breeding structure and ecological distribution that are
72 important for malaria control efforts.

73 Population structure analysis of a comprehensive *Anopheles* mosquito sampling
74 effort along a 400-km transect in the Sudan-Savanna ecological zone of central Burkina
75 Faso surprisingly revealed a previously unknown genetic cluster of *Anopheles gambiae*
76 *sensu lato* (Riehle *et al.* 2011). The new subgroup, named GOUNDRY, was found in
77 collections from larval pools but never in collections taken from inside human dwellings,
78 implying an exophilic adult resting habit. GOUNDRY mosquitoes are highly susceptible
79 to *Plasmodium* infection in the laboratory, but the feeding behavior of GOUNDRY adults
80 is unknown. Thus it is unclear whether the subgroup is a major vector of human malaria.

81 Current knowledge of GOUNDRY is incomplete, with previous genetic
82 understanding based on sparse microsatellite and SNP data (Riehle *et al.* 2011), but it is
83 essential to global public health to understand the evolution of new subgroups such as
84 GOUNDRY and how they may impact malaria control. GOUNDRY bears an atypical
85 genetic profile for *Anopheles* in the Sudan-Savanna zone of West Africa that raises
86 questions about its origins, such as whether it is a hybrid between *A. coluzzii* and *A.*
87 *gambiae*, as well as how old it is, and how reproductively isolated it is from other
88 *Anopheles* species. For example, the diagnostic SNPs that underlie one standard
89 approach for distinguishing between *A. coluzzii* (previously *A. gambiae* M form) and *A.*
90 *gambiae* (previously *A. gambiae* S form) were found to be segregating freely at Hardy-
91 Weinberg Equilibrium (HWE) in GOUNDRY mosquitoes (Riehle *et al.* 2011), implying

92 that the population is either hybrid or that it predates the *gambiae-coluzzii* species split.
93 Although high frequencies of hybrids diagnosed with these markers have been identified
94 in coastal regions of West Africa (Ndiath *et al.* 2008; Oliveira *et al.* 2008; Caputo *et al.*
95 2011), hybrid genotypes are quite rare (<1%) in the region where GOUNDRY was
96 collected (della Torre *et al.* 2001). An independent study used a slightly larger panel of
97 SNPs that differentiate *A. coluzzii* and *A. gambiae* in the pericentromeric regions of the X
98 chromosome and autosomes and found that typically diagnostic haplotypes were
99 segregating at HWE in GOUNDRY with evidence of recombination among them (Lee *et*
100 *al.* 2013). GOUNDRY also differed from typical *Anopheles s.l.* populations in the region
101 in karyotype frequencies of the large 2La chromosomal inversion. In the Sudan-Savanna
102 zone, the inverted allele of the 2La chromosomal inversion segregates near fixation in *A.*
103 *coluzzii* and *A. gambiae* (Coluzzi *et al.* 1979), but both forms of the inversion are
104 segregating at HWE frequencies in GOUNDRY (Riehle *et al.* 2011). Moreover, analysis
105 of microsatellites and SNP markers revealed considerable distinction between
106 GOUNDRY and other described *Anopheles* in the region and concluded that GOUNDRY
107 is a genetic outgroup to *A. gambiae* and *A. coluzzii* (Riehle *et al.* 2011). However,
108 GOUNDRY was less genetically variable than these other species, raising the possibility
109 that, among other potential explanations, its origin may be more recent.

110 To identify the evolutionary origins, age, and degree of genetic isolation from
111 other genetic subgroups of GOUNDRY, we analyzed full genome data from GOUNDRY
112 and multiple closely related *Anopheles* species as well as SNP chip and phenotype data
113 from an independent study (Mitri *et al.* 2015). We estimate the demographic history of
114 GOUNDRY and its potential importance for *Plasmodium* infections, and identify a
115 putative, novel X-linked chromosomal inversion in GOUNDRY that may be a barrier to
116 gene flow with closely related subgroups. We discuss these results in the context of
117 malaria control efforts.

118

119 **Materials and Methods**

120 ***Mosquito samples***

121 Mosquito sample collection and species/subgroup identification was previously
122 described for *A. coluzzii*, GOUNDRY, and *A. arabiensis* samples (Riehle *et al.* 2011).

123 Briefly, larvae and adults were collected from three villages in Burkina Faso in 2007 and
124 2008 (Table S1). Larvae were reared to adults in an insectary, and both field caught
125 adults and reared adults were harvested and stored for DNA collection. In addition to
126 standard species diagnostic assays, individuals were assigned to genetic subgroups using
127 genetic clustering analysis based on 3rd chromosome SNPs and microsatellites (Riehle *et*
128 *al.* 2011). One *A. gambiae* individual was also included in this study. This sample was
129 collected indoors as an adult in the village of Korabo in the Kissidougou prefecture in
130 Guinea in October 2012. Individuals were typed for species, molecular form and 2La
131 karyotype using a series of standard molecular diagnostics (Fanello *et al.* 2002; White *et*
132 *al.* 2007; Santolamazza *et al.* 2008). All *A. coluzzii* and *A. arabiensis* samples are 2La^{a/a}
133 homokaryotypes and the *A. gambiae* sample typed as a heterokaryotype (2La^{a/+}). As
134 discussed above, both forms of the 2La inversion are segregating in GOUNDRY, and we
135 chose to sequence eleven 2La^{+/+} GOUNDRY samples and one 2La^{a/a} sample
136 (GOUND_0446).

137

138 ***DNA extractions, genome sequencing, short-read processing***

139 A detailed description of the DNA extractions, sequencing, and processing has
140 been included in a separate publication (Crawford *et al.* 2015), but briefly, genomic
141 DNA was extracted using standard protocols and was sequenced using the Illumina
142 HiSeq2000 platform by BGI (Shenzhen, China). Paired-end 100-bp reads were obtained
143 for all samples. The *Anopheles gambiae* sample was sequenced on the same platform at
144 the University of Minnesota Genomics Center core facility. Raw Illumina reads were
145 deposited at NCBI SRA under BioProject ID PRJNA273873. Short-reads were aligned
146 in two steps using BWA-mem (v0.7.4) alignment algorithm [(Li 2013); bio-
147 bwa.sourceforge.net]. First, reads were mapped to the *A. gambiae* PEST AgamP3
148 reference assembly [(Holt *et al.* 2002); vectorbase.org]. Second, reads were mapped to a
149 new updated sequence where the major allele (frequency in sample ≥ 0.5) from each
150 population were substituted into the PEST reference to make population specific
151 references. Local realignment around indels was conducted with GATK v.2.5-2
152 (DePristo *et al.* 2011). Duplicates were removed using the SAMtools v.0.1.18 (Li *et al.*
153 2009) *rmdup* function. We applied a series of quality filters and identified a set of robust

154 genomic positions that were included in all downstream analysis. As a rule,
155 heterochromatic regions as defined for *A. gambiae* (Sharakhova *et al.* 2010) were
156 excluded from all analyses since short read mapping is known to be problematic in such
157 regions.

158

159 ***Bioinformatics and population genetic analyses***

160 Detailed descriptions of additional methods, mostly involving standard
161 approaches and previously existing software, can be found in Appendix S1. Included are
162 descriptions of genotype calling, estimation of nucleotide diversity, fixed difference
163 calling, calculation of genetic divergence (D_{xy}) and the neighbor-joining tree, ancestral
164 sequence synthesis, demographic model inference, selective sweep dating, and putative
165 inversion breakpoint mapping.

166

167 ***Inbreeding analysis***

168 ***Estimating inbreeding coefficients***

169 Initial estimates of the global site frequency spectrum (SFS) in GOUNDRY
170 produced distributions of allele frequencies that deviated substantially from standard
171 equilibrium expectations, as well as from those observed in the *A. coluzzii* and *A.*
172 *arabiensis* groups. Most notably, the proportion of doubletons was nearly equal to that of
173 singletons in *A. gambiae* GOUNDRY (see Results). This observation is consistent with
174 widespread inbreeding in the GOUNDRY subgroup. We tested the hypothesis of
175 extensive inbreeding in two ways, with the goals of both characterizing the pattern of
176 inbreeding in this subgroup as well as obtaining inbreeding coefficients for each
177 individual that could then be used as priors for an inbreeding-aware genotype-calling
178 algorithm. We used the method of Vieira *et al.* (Vieira *et al.* 2013), which estimates
179 inbreeding coefficients in a probabilistic framework taking uncertainty of genotype
180 calling into account. This approach is implemented in a program called ngsF
181 (github.com/fgvieira/ngsF). ngsF estimates inbreeding coefficients for all individuals in
182 the sample jointly with the allele frequencies in each site using an Expectation-
183 Maximization (EM) algorithm (Vieira *et al.* 2013). We estimated minor allele
184 frequencies at each site (-doMaf 1) and defined sites as variable if their minor allele

185 frequency was estimated to be significantly different from zero using a minimum log
186 likelihood ratio statistic of 24, which corresponds approximately to a P value of 10^{-6} .
187 Genotype likelihoods were calculated at variable sites and used as input into ngsF using
188 default settings. For comparison, we estimated inbreeding coefficients for *A. coluzzii*,
189 GOUNDRY, and *A. arabiensis* using data from each chromosomal arm separately.

190

191 *Recalibrating the site-frequency spectrum and genotype calls*

192 We used the inbreeding coefficients obtained above for the GOUNDRY sample
193 as priors to obtain a second set of inbreeding-aware genotype calls and an updated global
194 SFS. We used ANGSD v.0.534 to make genotype calls as described above. However, in
195 this case, we used the `-indF` flag within ANGSD, which takes individual inbreeding
196 coefficients as priors instead of the global SFS (Vieira *et al.* 2013). Similarly, we used
197 the inferred inbreeding coefficients to obtain an inbreeding-aware global SFS. We
198 estimated the global SFS from genotype probabilities using `-realSFS 2` in ANGSD,
199 which is identical to `-realSFS 1` (Nielsen *et al.* 2012) except that it uses inbreeding
200 coefficients as priors for calculations of posterior probabilities (Vieira *et al.* 2013).

201

202 *IBD Tracts*

203 We examined the effects of inbreeding within diploid individuals using FEstim
204 (Leutenegger *et al.* 2006, 2011), which implements a maximum likelihood method within
205 a Hidden-Markov-Model that models dependencies along the genome. We used the
206 FSuite v.1.0.3 (Gazal *et al.* 2014) pipeline to generate submaps, estimate inbreeding
207 parameters using FEstim, identify IBD tracts, and plot IBD tracts using Circos v.0.67-6
208 (Krzywinski *et al.* 2009). To minimize linkage disequilibrium that creates non-
209 independence among SNPs while maximizing information content, we generated 20
210 independent random subsets of between 187 and 193 SNPs (or submaps) spaced at least 1
211 kb apart, and inbreeding parameters were inferred using all 20 submaps. We used allele
212 frequencies estimated using ANGSD above (`-doMaf` and `-indF`) for calculation of
213 emission probabilities in FEstim. We also used genetic maps for *Anopheles gambiae*
214 from Zheng *et al.* (Zheng *et al.* 1996). To convert data from Zheng *et al.* to dense
215 genetic maps, we mapped primers from that study onto the *Anopheles gambiae* PEST

216 reference using standard e-PCR approaches that map PCR primers onto a reference
217 sequence using computational sequence matching. Autosomal maps and code for
218 polynomial analysis were kindly provided by Russ Corbett-Detig
219 (github.com/tsackton/linked-selection/), and we performed e-PCR mapping for the X
220 chromosome. We fit a polynomial function to the genetic map for each chromosome and
221 used this function to convert the physical position of SNP marker to genetic distance. For
222 this analysis, we joined the left and right arms of chromosomes 2 and 3 by adjusting the
223 physical position of SNPs on the left arms by the full length of the right arm.

224 FSuite is designed for genotyping array data and does not allow any genotyping
225 errors. Therefore, we took additional steps to minimize the effects of genotyping errors.
226 First, we set a minimum minor allele frequency of 10% and included only genotypes with
227 95% posterior probability. Second, we set a liberal threshold of $1e^{-6}$ for the minimum
228 posterior probability required for considered IBD. Since this threshold allows many
229 small IBD tracts that are likely to be erroneous, we set a minimum size threshold of 0.1
230 cM for inclusion in the final set of IBD tracts.

231

232 *Ruling out bioinformatic and sequencing artifacts*

233 Since the observation of high rates of inbreeding stem directly from intermediate
234 coverage (~10X) next-generation sequencing data that can be prone to bioinformatic
235 errors and biases, we conducted several tests to determine whether such artifacts could
236 explain the observed inbreeding signal. One possible artifact could stem from mapping
237 or alignment biases against divergent next-generation reads that could lead to excess
238 homozygosity. If mapping is unbiased, the proportion of reference bases at heterozygous
239 sites should be distributed with a mean of 0.5. We find that the mean proportion of
240 reference bases at heterozygous sites is 0.4893 ($\sigma = 0.1646$) in *A. coluzzii* and 0.4757 ($\sigma =$
241 0.1581) in GOUNDRY indicating very similar read distributions in these populations
242 (Figure S1). Although both populations show a small deviation from 0.5 at biallelic sites,
243 this deviation cannot explain large regions of homozygosity in GOUNDRY.

244 We also asked whether excess homozygosity could stem from erroneous
245 assignment of homozygous genotypes at true heterozygous sites. Such errors could result
246 if short read depths were exceptionally low in some genomic regions. We calculated read

247 depths at sites in different genotype classes in GOUNDRY and find that the mean read
248 depth is 12.3569 ($\sigma = 5.3917$) at homozygous reference sites, 12.2156 ($\sigma = 5.1235$) at
249 homozygous alternative sites, and 12.6871 ($\sigma = 5.5163$) at heterozygous sites, indicating
250 that the distribution of read depth is very similar between all three classes (Figure S2).
251 We find a similar pattern in *A. coluzzii*, which shows no evidence of inbreeding. In this
252 population, the mean read depth is 10.4773 ($\sigma = 4.7555$) at homozygous reference sites,
253 11.0082 ($\sigma = 4.1849$) at homozygous alternative sites, and 10.6660 ($\sigma = 4.7755$) (Figure
254 S2). Moreover, the distributions of read depths at heterozygous sites and homozygous
255 sites are very similar in both the *A. coluzzii* and GOUNDRY populations (Figure S2).
256 These results strongly suggest that bioinformatic artifacts cannot explain the excess
257 homozygosity and IBD tracts observed in GOUNDRY.

258 Large variations in observed sequence diversity could also stem from issues
259 related to DNA sequencing. Importantly, the same DNA preparation and library
260 preparation protocols were used for GOUNDRY as well as *A. coluzzii* and *A. arabiensis*,
261 so the increased IBD observed in GOUNDRY is not likely attributable to a difference in
262 sample preparation. Low DNA input could also lead to artifacts in sequencing, but the
263 total mass of DNA used for library preparation did not differ between GOUNDRY and
264 the other samples that were all sequenced together (Table S1). In general, DNA mass
265 and handling was similar between GOUNDRY and the other populations examined here
266 that do not harbor long IBD tracts, suggesting that such differences cannot explain the
267 signals of increased inbreeding in GOUNDRY.

268

269

270

271 ***Inbreeding-phenotype association test***

272 Since 12 GOUNDRY genomes is not a large enough sample size for association
273 testing, we obtained SNP genotype data for 274 GOUNDRY individuals who had also
274 been phenotyped in a *Plasmodium* infection experiment as part of an independent study
275 (Mitri *et al.* 2015). Full details of Illumina SNP chip assay design and data collection and
276 infection experiments are available in that publication but will be summarized here.
277 Briefly, larvae were collected in three villages in central Burkina Faso, and raised to

278 adulthood in the laboratory where females were given *Plasmodium falciparum*-infectious
279 bloodmeals from local volunteers. Fully fed females were dissected 7-8 days later for
280 oocyst quantification and DNA extraction. For the Illumina SNP chip, SNPs were
281 identified from raw sequence reads generated for genome sequencing projects for *A.*
282 *coluzzii* and *A. gambiae* as well as from independent deep sequencing efforts. DNA
283 hybridization and genotype calling were conducted using standard procedures followed
284 by stringent quality filtering of genotype calls and independent confirmation using
285 duplicate hybridizations and independent Sequenom assays using a subset of SNPs. We
286 used a set of SNPs distributed approximately uniformly across the autosomes. Among
287 these sites, we included only sites (n = 678) that were variable in the GOUNDRY
288 subgroup.

289 We used ngsF (Vieira *et al.* 2013) to estimate inbreeding coefficients for each of
290 the 274 females using the SNP genotype data as input. Since genomic estimates of
291 inbreeding coefficients are statistically noisy with less than 1000 SNPs, we used a
292 bootstrap approach by sampling the SNPs with replacement to make 1000 new
293 bootstrapped datasets of the same size, estimating inbreeding coefficients using ngsF.
294 Point estimates of the inbreeding coefficients were obtained by taking the mean of log 10
295 transformed bootstrap values and re-transforming the mean value.

296 To test whether infection prevalence was higher in inbred individuals, we used a
297 two-by-two χ^2 test. The table cells corresponded to ‘infected’ and ‘not infected’
298 phenotypes as well as high and low inbreeding coefficients. Since the distribution of
299 inbreeding coefficients was not bimodal, we categorized individuals as either ‘high’ or
300 ‘low’ inbreeding levels based on whether their inbreeding coefficient was above or below
301 a cutoff value, respectively. We included two cutoff values: 1) F = the median
302 coefficient value of 0.026, and 2) The maximum F estimated from genome sequencing in
303 *A. coluzzii*, which does not show signs of inbreeding. To establish statistical significance
304 while preserving correlations among mosquitoes within each blood donor cohort, we
305 randomly permuted infection phenotype among mosquitoes within donor cohort, and
306 recalculated the χ^2 value. We compared the empirical χ^2 value to 10^4 values from
307 permuted datasets in a one-tailed statistical test. We further tested the association and the
308 effect of blood-donor using the Cochran-Mantel-Haenszel procedure (cmh.test in R)

309 that directly accounts for additional factors within the contingency test. To test for
310 correlations between inbreeding coefficients and the number of oocysts (infection
311 intensity), we fit a linear model to relate inbreeding coefficients (log transformed) to the
312 number of oocysts (log transformed) with blood-donor as a factor in the model. Only
313 mosquitoes with at least one oocyst were included in this part of the analysis.

314

315 **Results**

316 *Genome Sequencing and Population Genetic Analysis*

317 We have completely sequenced the genomes of 12 field-captured female
318 *Anopheles* GOUNDRY mosquitoes from Burkina Faso and Guinea using the Illumina
319 HiSeq2000 platform. We compared these genomes to full genomes from *A. coluzzii*
320 (n=10), *A. gambiae* (n=1) and *Anopheles arabiensis* (n=9). Most individuals were
321 sequenced to an average read depth of 9.79x, while one individual each from
322 GOUNDRY, *A. coluzzii*, and *A. gambiae* was sequenced to at least 16.44x (Table S1).
323 We also used publicly available genome sequences from *Anopheles merus* (*Anopheles*
324 *gambiae* 1000 Genomes Project) as an outgroup. We conducted population genetic
325 analysis of aligned short-read data using genotype likelihoods and genotype calls
326 calculated using the probabilistic inference framework ANGSD (Korneliussen *et al.*
327 2014).

328

329 *Genetic Relatedness Among Species and Subgroups*

330 To determine the genetic relationship of the GOUNDRY subgroup to other
331 known species and subgroups of *Anopheles*, we calculated an unrooted neighbor-joining
332 tree based on genome-wide genetic distance (D_{xy}) at intergenic sites (Figure 1). Previous
333 findings indicated that the recently discovered GOUNDRY subgroup of *A. gambiae* is a
334 genetic outgroup to *A. coluzzii* (formerly known as M molecular form) and *A. gambiae*
335 (formerly S form) (Riehle *et al.* 2011). However, our data indicate that GOUNDRY is
336 actually genetically closer to *A. coluzzii* ($D_{GAc} = 0.0109$; 100% bootstrap support) than
337 either group is to *A. gambiae* ($D_{GAg} = 0.0149$; $D_{AcAg} = 0.0143$).

338

339 It has been speculated that GOUNDRY may be a recently formed backcrossed
339 hybrid of *A. coluzzii* and *A. gambiae* (Lee *et al.* 2013). This hypothesis also predicts that

340 GOUNDRY will be segregating chromosomes that are mosaics of haplotypes derived
341 from *A. coluzzii* and *A. gambiae*, and therefore most, if not all, polymorphisms found in
342 GOUNDRY should also be found in one of these putative parental taxa. In contrast, we
343 find 7,383 fixed differences between *A. coluzzii* and GOUNDRY [excluding 2L since it is
344 dominated by the large 2La inversion known to have crossed species boundaries
345 (Fontaine *et al.* 2015)], of which 27% are putatively GOUNDRY-specific alleles not
346 shared with *A. gambiae*, *A. arabiensis*, or *A. merus*. GOUNDRY shares an allele with the
347 *A. gambiae* individual sampled here at an additional 31.5% of the fixed sites, although
348 this number may increase if more *A. gambiae* samples are included. These results do not
349 exclude the possibility of gene flow between GOUNDRY and *A. gambiae*, but they fail to
350 support the hypothesis that GOUNDRY is simply a very recent hybrid of *A. coluzzii* and
351 *A. gambiae*. Instead, the substantial number of putatively GOUNDRY-specific fixed
352 alleles support GOUNDRY as a unique subgroup that may have originated as an offshoot
353 of *A. coluzzii* and experienced subsequent gene flow from *A. gambiae*.

354

355 ***Origins of GOUNDRY***

356 It has been hypothesized that the advent of agriculture in sub-Saharan Africa ~5-
357 10 kya played a role in driving diversification and expansion of *Anopheles* mosquitoes
358 (Coluzzi *et al.* 2002). The two-dimensional site-frequency spectrum reveals substantial
359 differentiation in allele frequencies between GOUNDRY and *A. coluzzii* with many fixed
360 differences differentiating these groups and is not compatible with a very recent origin of
361 GOUNDRY (Figure 2). To test whether the origin of GOUNDRY could have been
362 associated with habitat modification driven by agriculture, we fit four population
363 historical models with increasing complexity (Figure 2; Table 1; Methods) to the two-
364 dimensional site frequency spectrum for GOUNDRY and *A. coluzzii* using *dadi*
365 (Gutenkunst *et al.* 2009). The 2D spectra from the empirical data and the best-fit model
366 for each demographic model are presented in Figure 2. We first fit a simple one-epoch,
367 split model with no migration. The maximum-likelihood model under this scenario gave
368 a poor fit to the empirical data with a likelihood value (L) of -176,635.8. We then added
369 asymmetrical migration to the model (one-epoch, split with migration), which resulted in
370 a nearly three-fold improvement of the likelihood value improvement of the fit of the

371 model to the data with $L_{1\text{-ep-splt-mig}} = -59.896.75$ -, providing strong evidence that
372 migration has played a key role in the history of these taxa. Residual differences between
373 the 2D spectra from the model and the data (Figure 2), however, were unevenly
374 distributed across the spectra, suggesting that one-epoch models are missing potentially
375 important features of the demographic history. To improve flexibility in the model fitting,
376 we fit both two-epoch and three-epoch population split-with-migration models (Table 1).
377 Interestingly, the adding a second epoch did not result in a substantial improvement of
378 the fit to the data as indicated by the remaining large residuals and decreased likelihood
379 value ($L_{2\text{-ep-splt-mig}} = -59,949.49$) relative to the one-epoch model. Adding a third epoch,
380 however, achieved a considerable improvement of the fit to the data ($L_{3\text{-ep-splt-mig}} = -$
381 $49,023.85$). Residuals indicating differences between the model and data are also
382 presented and suggest that deviations between spectra associated with the model and data
383 are well correlated.

384 The best-fitting three-epoch-split-with-migration model (Table 1) predicts that
385 these subgroups diverged $\sim 111,200$ ya (95% CI 96,718 – 125,010), followed by a 100-
386 fold reduction in the size of both subgroups after isolation (Methods). The timing of this
387 model rejects any role of modern agriculture in subgroup division, although it should be
388 noted that estimates of such old split times inherently carry considerable uncertainty. Our
389 inferred model is inconsistent by an order of magnitude with agriculture as a driving
390 force in cladogenesis and is more consistent with habitat fragmentation and loss due to
391 natural causes, potentially including climatic shifts such as changes in pluviometry that
392 would lead to increased population size. The model supports a >500 -fold population
393 growth in *A. coluzzii* and 19-fold growth in GOUNDRY with extensive gene flow
394 between them $\sim 85,300$ ya, consistent with a re-establishment of contiguous habitat and
395 abundant availability of bloodmeal hosts. Interestingly, the model supports additional
396 population growth in both subgroups in the most recent epoch, which spans the last
397 10,000 years and coincides with the advent of agriculture. Any hybridization related to
398 secondary contact during this period has not led to complete homogenization, as we
399 conservatively identified nearly 8,000 fixed nucleotide differences distributed across the
400 genomes of the two subgroups.

401 The dates reported here depend on assumptions about both the physiological
402 mutation rate as well as the number of generations per year, neither of which are well
403 known in *Anopheles*. As such, the details of these results would differ somewhat if
404 different estimates were used. However, we would have to invoke extreme values of
405 these parameters that are outside reasonable expectation in order to obtain estimates for
406 the time of the GOUNDRY- *A. coluzzii* split that coincides with the advent of agriculture.
407 Overall, the model suggests that the origin of GOUNDRY is not recent and both
408 GOUNDRY as well as *A. coluzzii* have both undergone bouts of population growth and
409 increased rates of hybridization in more recent evolutionary time.

410 The initial description of GOUNDRY (Riehle *et al.* 2011) suggested that it
411 harbored lower allelic diversity than other sampled subgroups potentially suggesting a
412 small effective population size while being proportionally more numerous than other
413 subgroups at the time and place of collection. Our model suggests that the recent
414 effective population size is approximately 98,400 (95% CI 55,100 – 158,500) compared
415 to a recent *A. coluzzii* effective size of approximately 1,558,000 (95% CI 848,000 –
416 2,508,000). The disparity between recent effective sizes of these two subgroups suggests
417 that, while GOUNDRY may have been locally abundant at the time and place of the
418 initial study, it is not likely to be geographically widespread on a scale similar to *A.*
419 *coluzzii*.

420

421 ***Novel X-linked chromosomal inversion in GOUNDRY***

422 A large cluster of fixed differences (~ 530; Figure S3) identified between
423 GOUNDRY and *A. coluzzii* falls within a 1.67 Mb region on the X chromosome that is
424 nearly absent of polymorphism (Figure 3), despite sequence read coverage comparable to
425 neighboring genomic regions (Figure S3). The remarkably large size of the region
426 devoid of diversity would imply exceptionally strong positive selection under standard
427 rates of meiotic recombination. For comparison, previously identified strong sweeps
428 associated with insecticide resistance span approximately 40 Kb and 100 kb in freely
429 recombining genomic regions of *Drosophila melanogaster* and *D. simulans*, respectively
430 (Schlenke & Begun 2004; Aminetzach *et al.* 2005). The swept region in GOUNDRY is
431 marked by especially sharp edges (Figure 3), implying that recombination has been

432 suppressed at the boundaries this region. Collectively, these observations suggest that the
433 swept region may be a small chromosomal inversion, which we have named *Xh* in
434 keeping with inversion naming conventions in the *Anopheles* system. Notably, this
435 pattern is virtually identical to the pattern of diversity in a confirmed X-linked inversion
436 discovered in African populations of *D. melanogaster* (Corbett-Detig & Hartl 2012). The
437 *Xh* region in GOUNDRY includes 92 predicted protein coding sequences (Table S2),
438 including the *white* gene, two members of the gene family encoding the TWDL cuticular
439 protein family (*TWDL8* and *TWDL9*), and five genes annotated with immune function
440 (*CLIPC4*, *CLIPC5*, *CLIPC6*, *CLIPC10*, *PGRPS1*). The lack of diversity in the region
441 implies that the presumed *Xh* inversion has a single recent origin and was quickly swept
442 to fixation in GOUNDRY. We estimated the age of the haplotype inside the sweep
443 region to be 78 years with a standard deviation of 9.15 by assuming that all segregating
444 polymorphisms in the region postdate fixation of the haplotype (see Methods). Such
445 extraordinarily recent adaptation is consistent with the selection pressures related to 19th
446 and 20th century human activity such as insecticide pressure or widespread habitat
447 modification.

448

449 ***Xh is a barrier to introgression***

450 Chromosomal inversions are thought to play important roles as barriers to gene
451 flow between taxa diverging with ongoing gene flow (Rieseberg 2001; Noor *et al.* 2001;
452 Navarro & Barton 2003), so we hypothesized that this putative X-linked chromosomal
453 inversion in GOUNDRY may serve as a barrier to gene flow with *A. coluzzii*. If this
454 inversion has acted as a barrier to gene flow with *A. coluzzii*, or taxa undergoing
455 secondary contact after divergence, we would expect the X chromosome to be more
456 diverged than the autosome and the inversion would be more diverged than other regions
457 of the X chromosome.

458 One approach to estimate differences in divergence among genomic regions is to
459 compare divergence between a focal pair of subgroups (GOUNDRY and *A. coluzzii*) to
460 divergence between one of the focal groups and an outgroup (GOUNDRY and *A.*
461 *gambiae*) in order to scale divergence levels by differences among regions in mutation
462 rate and the effects of selection on linked sites. This approach estimates what is known

463 as Relative Node Depth ($RND = D_{GAc}/D_{GAg}$, where subscripts G, Ac, and Ag indicate
464 GOUNDRY, *A. coluzzii*, and *A. gambiae* respectively), and a higher RND indicates
465 greater divergence between the focal groups (Feder *et al.* 2005). We find that RND is
466 0.7797 on the autosomes and 0.8058 on the X, indicating higher genetic divergence
467 between GOUNDRY and *A. coluzzii* on X relative to the autosomes. To explicitly test
468 whether such a pattern could be obtained under a pure split model with no gene flow, we
469 obtained expected values of Relative Node Depth (RND) assuming a phylogeny where *A.*
470 *coluzzii* and GOUNDRY form a clade with *A. gambiae* as the outgroup (Methods).

471 Our analytical results support the hypothesis that D_{GAc} is downwardly biased on
472 the autosomes relative to D_{GAc} on the X as a result of higher rates of gene flow on the
473 autosomes relative to the X. We find that under some parameter combinations (Figure 4),
474 RND decreases with increasing effective *A.coluzzii*-GOUNDRY effective population
475 size, which could result in a smaller RND value on the autosomes since the autosomes
476 should have an effective size at least as big as the X. However, most parameter
477 combinations suggest that this pattern is unexpected (i.e. most regions of the curves
478 predict that RND should increase with increasing effective population size), and the
479 estimate for the ancestral effective size of *A.coluzzii*-GOUNDRY we obtained in a
480 separate demographic analysis above suggests that these subgroups exist in a parameter
481 space where the RND function is consistently increasing with increasing effective sizes.

482 To test the second expectation that the inversion is more diverged than other
483 regions on the X chromosome, we compared divergence with *A. coluzzii* in windows
484 inside and outside of the inverted region. Absolute sequence divergence (D_{xy}) is not
485 sensitive to detect differential gene flow for relatively recent changes in gene flow
486 (Cruickshank & Hahn 2014), and we expect that the putative *Xh* inversion is likely too
487 young for measurable differences to have accumulated, so we tested for excess
488 divergence in the *Xh* inversion using a more sensitive approach. For comparison, we find
489 that the inverted region is significantly more diverged between *A. coluzzii* and
490 GOUNDRY relative to the remaining X chromosome ($\bar{D}_{G:Ac}(Xh) = 0.0103$, $\bar{D}_{G:Ac}$ (non-
491 *Xh*) = 0.0071; M-W $P < 2.2 \times 10^{-16}$), but nucleotide diversity in *A. coluzzii* is also
492 significantly higher in this region ($\pi_{Ac}(Xh) = 0.0080$, π_{Ac} (non-*Xh*) = 0.0061; M-W $P <$
493 5.49×10^{-14}), implying that the increased divergence could be partially explained by

494 increased mutation rate in this region. However, when absolute divergence along the X
495 chromosome is explicitly scaled by the mutation rate inferred from levels of
496 polymorphism in the *A. coluzzii* sample (D_a), the putatively adaptive *Xh* inversion
497 between GOUNDRY and *A. coluzzii* is proportionally much more divergent than is the
498 remainder of the X chromosome (\bar{D}_x (*Xh*) = 0.0022, \bar{D}_x (non-*Xh*) = 0.0013; M-W $P <$
499 4.89×10^{-08} ; Figure 5). Although relative measures of divergence, such as D_a , are known,
500 for example, to be confounded by reductions in nucleotide diversity related to natural
501 selection on linked sites (Charlesworth 1998; Noor & Bennett 2009), we believe that this
502 analysis is robust to these concerns because the comparison is among only X-linked
503 windows and the region of interest is in a region of the chromosome that is highly diverse
504 in subgroups where there is no evidence of selective sweeps (Figure 3).

505 Both of these tests indicate that sequence divergence between *A. coluzzii* and
506 GOUNDRY is greater inside the putative inversion relative to the X as a whole, which
507 likely reflects both the accumulation of a small number of new private mutations inside
508 the inversion as well as a greater proportion of shared polymorphisms outside the
509 inversion, consistent with higher rates of introgression outside the inversion. Taken
510 together with the demographic inference, the above results suggest that, after initial
511 ecological divergence between these taxa approximately 100,000 years ago, this genomic
512 barrier to introgression has established in the face of ongoing hybridization only within
513 the last 100 years, presumably owing to the accumulation and extended effects of locally
514 adapted loci or genetic incompatibility factors within the large swept/inverted *Xh* region
515 on the GOUNDRY X chromosome, meiotic drive, or aneuploidy resulting from
516 nondisjunction in heterokaryotypes.

517

518 ***GOUNDRY is inbred***

519 Unexpectedly, we found that GOUNDRY exhibits a deficiency of heterozygotes
520 relative to Hardy-Weinberg expectations and extensive regions of Identity-By-Descent
521 (IBD), a pattern that is not observed in any of our other *Anopheles* collections. Individual
522 diploid GOUNDRY genomes are checkered with footprints of IBD, even though the
523 genome as a whole harbors substantial genetic variation indicating a relatively large
524 genetic (effective) population size (Figure 6a). The observation of stochastic tracts of

525 IBD is most consistent with an unusually high rate of close inbreeding. To explicitly test
526 for elevated inbreeding coefficients (F), we used a maximum likelihood framework to
527 infer F for each individual without calling genotypes. We found that values of F range
528 from 0.0087 to 0.2106 genome wide (Figure S4). In contrast, estimates of inbreeding
529 coefficients for 10 *A. coluzzii* genomes and 9 *A. arabiensis* genomes were consistently
530 low ($F_{Ac} < 0.03$; $F_{Aa} < 0.04$). The relatively high inbreeding coefficients in GOUNDRY
531 suggest that this population has a history of mating among relatively closely related
532 individuals.

533 The lengths of these tracts provide information about the timing and nature of
534 inbreeding in the population since recombination is expected to break up large tracts
535 generated by recent inbreeding. All 12 GOUNDRY genomes analyzed here are marked
536 by IBD tracts of various lengths, and the specific chromosomal locations of the IBD
537 regions are random and vary among the sequenced GOUNDRY individuals (Figure 6b).
538 While many IBD tracts are relatively short, several individuals harbor tracts that span 30-
539 40 cM (Figure 6c). This mixture of tract lengths is most consistent with both a
540 generations-old history of inbreeding (short tracts) as well as the possibility of mating
541 among half-siblings or first-cousins (long tracts).

542

543 ***Effect of inbreeding on Plasmodium-resistance***

544 Inbreeding is known to have detrimental effects on various phenotypes, including
545 resistance to parasite infection (Hamilton *et al.* 1990; Luong *et al.* 2007). To test whether
546 inbreeding in GOUNDRY increases intrinsic susceptibility to *Plasmodium falciparum*
547 infection in this group, we studied a larger panel of 274 GOUNDRY females that were
548 experimentally infected with local wild isolates of *P. falciparum* and genotyped at 1,436
549 SNPs across the genome (Mitri *et al.* 2015). After filtering, we estimated inbreeding
550 coefficients with the program *ngsF* (Vieira *et al.* 2013) using 678 autosomal variable sites
551 (Methods) and found that F ranges from 0 to 0.3797 in this sample of GOUNDRY
552 females (Figure S5). Although it is possible that some GOUNDRY individuals are truly
553 not inbred, all 12 GOUNDRY individuals subjected to whole genome sequencing showed
554 significant evidence of inbreeding, so we suspect that the relatively sparse genotyping (1

555 per ~400 kb) assay used on this panel of mosquitoes failed to capture IBD tracts in some
556 individuals.

557 Blood feeding experiments were conducted using five human *Plasmodium*
558 gametocyte donors, and blood donor had a significant effect on both infection prevalence
559 (ANOVA; $P = 1.593 \times 10^{-9}$) and intensity (ANOVA; $P = 1.194 \times 10^{-13}$). Importantly, the
560 distributions of mosquito inbreeding coefficients did not differ significantly between
561 blood donor cohorts (ANOVA, $P = 0.0934$).

562 Of the females that fed on infectious blood-meals, 104 (37.9%) had no parasites at
563 the time of dissection, and we asked whether this infection prevalence is statistically
564 associated with inbreeding in the mosquito host. Inspection of the distribution of F in
565 this sample indicates that categorization of individuals as inbred or outbred is difficult
566 since a substantial proportion of individuals were assigned values of F close to 0 (Figures
567 S5 and S6) and even individuals from outbred populations such as *A. coluzzii* and *A.*
568 *arabiensis* can have estimates of F as high as 0.03 or 0.04 (Figure S4). Therefore, we
569 used the median value of F estimated from genome-wide SNPs in GOUNDRY (0.026)
570 and categorized mosquitoes as more inbred ($F > 0.026$) or less inbred ($F \leq 0.026$). We
571 used a χ^2 test with this categorization approach to test whether higher inbreeding
572 significantly associated with higher infection prevalence and find that females with
573 higher inbreeding coefficients are overrepresented in the 'infected' class ($P = 0.0205$;
574 Table 2). We also used the Cochran-Mantel-Haenszel procedure to directly account for
575 blood-donor in the test for association and found very similar results ($P = 0.025$ for
576 median cutoff). As an alternative assignment approach, we defined the inbreeding
577 categories using the highest inbreeding coefficient value obtained from full genome
578 sequencing of an outbred population, *A. coluzzii* ($F = 0.0292$), which should be more
579 robust to statistical uncertainty than estimates from the SNP chip data, and find that the
580 association is on the borderline of significance ($P = 0.0546$; Table 2). These analyses
581 indicate that an increase in the proportion of genes with alleles that are Identical by
582 Descent may decrease the ability of adult female mosquitoes to resist parasite infection,
583 although the effect is small enough that detection of the association is sensitive to how
584 the distribution of F is categorized.

585 We also asked whether the degree of inbreeding has an effect on the intensity of
586 infection (number of oocysts per midgut). Of the 274 females that fed on natural
587 gametocytemic blood samples and were assayed for infection status, 170 harbored at least
588 one oocyst, while the remaining 104 females were uninfected, corresponding to an
589 infection rate of 0.62. Among the infected females, infection intensity varied from 1 to
590 38 with a mean of 5.73 oocysts per individual. We fit linear models for mosquitoes fed
591 on each blood donor separately and find no significant correlation ($P > 0.05$) between
592 inbreeding coefficients and infection intensity.

593

594 **Discussion**

595 It is not known how many such cryptic subpopulations of *Anopheles* exist or how
596 much gene flow they share with described subgroups, although there is evidence gene
597 flow may be common (Lee *et al.* 2013). Epidemiological modeling and vector-based
598 malaria control strategies must account for populations like GOUNDRY if they are to
599 effectively predict disease dynamics and responses to intervention (Griffin *et al.* 2010).
600 Failure to account for such subpopulations will undermine malaria control efforts, as in
601 the case of the Garki malaria control project in Nigeria in the 1970s that did not account
602 for genetic variation in adult resting behavior and missed outdoor resting adults
603 (Molineaux *et al.* 1980) .

604 Here, we present an analysis of complete genome sequences from the newly
605 discovered cryptic GOUNDRY subgroup of *A. gambiae*. Our results help clarify some
606 outstanding questions raised by the initial description of this subgroup. We show that, in
607 contrast to initial suggestions (Riehle *et al.* 2011), GOUNDRY subgroup of *A. gambiae*
608 falls genetically within *Anopheles gambiae sensu lato* and is not an outgroup.
609 GOUNDRY shows strongest genetic affinity with *A. coluzzii* and therefore may be an
610 ecologically specialized subgroup of *A. coluzzii*. The discrepancy between our findings
611 and previously published results is likely due to the fact that the first description was
612 based on a small number of microsatellite markers and SNPs and was based on
613 differences in allele frequency, while the current study is based on absolute sequence
614 divergence calculated from whole genome sequencing data, and therefore included both
615 shared and private mutations. Our demographic analysis suggests that GOUNDRY has

616 existed for approximately 100,000 years and represents a recent example of the frequent
617 speciation dynamics in *Anopheles* that appears to be common (Crawford *et al.* 2015;
618 Fontaine *et al.* 2015). Since GOUNDRY was identified using an outdoor sampling
619 approach not common in previous studies, it was unclear whether or not this subgroup
620 may be more broadly distributed and just un-sampled. We estimate that the recent
621 (effective) population size of GOUNDRY is approximately 5% that of *A. coluzzii*,
622 suggesting that GOUNDRY is likely restricted to a relatively small region of the Sudan-
623 Savanna zone in West Africa.

624 In addition to thousands of mutations found to be putatively unique to
625 GOUNDRY, we identified a large GOUNDRY-specific genetic marker in the form of a
626 new putative X-linked chromosomal inversion that originated and fixed within
627 GOUNDRY within the last 100 years. It remains unknown whether positive selection or
628 meiotic drive has driven this inverted haplotype to high frequency and ultimately fixation
629 in GOUNDRY, but our results suggest that it may serve as a recent barrier to gene flow
630 with *A. coluzzii*, and potentially other taxa as well. Collectively, the data show that
631 nucleotide-diversity corrected divergence is higher inside the putative inverted region, the
632 inverted region as a chromosomal segment is the most diverged of all segments of the
633 same size on the X chromosome, and the X chromosome as a whole is more diverged
634 among GOUNDRY and *A. coluzzii* relative to the autosomes. The most parsimonious
635 explanation for these patterns is that, although very few new mutations have accumulated
636 inside of *Xh* since its origin less than 100 years ago, ongoing gene flow between *A.*
637 *coluzzii* and GOUNDRY has led to a greater density of shared polymorphism and
638 therefore lower sequence divergence in non-inverted regions of the X chromosome
639 relative to the inversion, especially distal to the inversion breakpoints. These results lead
640 us to conclude that while cladogenesis of GOUNDRY and *A. coluzzii* ~100 kya by other
641 means established some degree of temporally fluctuating reproductive isolation, the
642 recently derived *Xh* putative inversion now serves as a genomic barrier to gene flow, and
643 the effects of selection against migrant haplotypes or lack of recombination with non-
644 inverted chromosomes have begun to extend to linked sites outside the inversion
645 breakpoints.

646 The observation that GOUNDRY is more closely related at the genome level to *A.*
647 *coluzzii* than to *A. gambiae* could be biased by higher rates of gene flow between
648 GOUNDRY and *A. coluzzii* as well as sampling bias caused by the fact that *A. gambiae* is
649 represented by only a single individual that was sampled from a different country.
650 Although we cannot formally rule out the possibility that GOUNDRY originated as
651 something other than a subpopulation of *A. coluzzii* and later experienced substantial
652 gene flow from *A. coluzzii* that led to genetic affinity in our analysis, the most
653 parsimonious explanation is that it is a subgroup that originated from *A. coluzzii* that has
654 experienced gene flow from multiple sympatric taxa over its history. The most
655 compelling piece of evidence that GOUNDRY is not a recent *A. coluzzii*-*A. gambiae*
656 hybrid-backcross is the presence of the large fixed haplotype on the X chromosome in
657 GOUNDRY that is not expected under the recent backcross model. In support of this
658 notion, a recently published study (Fontaine *et al.* 2015) constructed similar distance
659 based trees using samples of *A. gambiae* from across the continent and found that
660 geographically disparate individuals were consistently interdigitated while excluding *A.*
661 *coluzzii*, suggesting that species assignment was more important than geography.

662 An additional potential concern regarding our estimate of the demographic
663 modeling and our conclusion that the X chromosome is more diverged than the autosome
664 between GOUNDRY and *A. coluzzii* stems from introgression between *A. gambiae* and
665 both GOUNDRY and *A. coluzzii*. We showed in a companion manuscript that
666 GOUNDRY has introgressed with *A. gambiae* in the evolutionarily recent past (Crawford
667 *et al.* 2015), and the presence of *A. gambiae* haplotypes in GOUNDRY could bias our
668 demographic estimate of the split time since this introgression was not explicitly modeled.
669 A four-taxon model including *A. gambiae* and *A. arabiensis* would probably improve our
670 estimates, but the dimensionality of such a model would increase dramatically and would
671 require much more sequence data than is available in the current study. Introgression
672 with *A. gambiae* could also compromise our RND analysis in which this group was used
673 as an outgroup. For example, higher introgression between *A. gambiae* and the ingroups
674 on the X relative to the autosome could result in an underestimate of the mutation rate on
675 the X chromosome and thus an inflation of the ingroup divergence. However, we showed
676 in a companion manuscript (Crawford *et al.* 2015) that signals of introgressed haplotypes

677 are concentrated on the autosome and absent from the X, suggesting that RND scaling
678 may be downwardly biased on the autosomes rather than the X. For these reasons, *A.*
679 *gambiae* is not an ideal outgroup for an RND analysis, but it is suitable for our purposes
680 and a low rate of introgression from this taxon is not likely to bias our results.

681 Perhaps the most unexpected feature of GOUNDRY is the high degree of
682 inbreeding in this population. We emphasize that the deficit of heterozygosity and
683 presence of unusually long IBD tracts that we observe in GOUNDRY are not a typical
684 function of persistently small population size. The inbreeding that we see here is
685 different from the strong drift that would be associated with small effective population
686 sizes over many generations, and which would manifest as generally low levels of
687 nucleotide diversity across the genome. Instead, the observed pattern indicates that some
688 proportion of individuals in an otherwise relatively large population tend to mate with
689 closely related individuals. Although IBD patterns in GOUNDRY are not consistent with
690 a long term small population size, in principle it could reflect a very recent and severe
691 reduction in population size, perhaps related to a strong insecticide pressure. The full
692 insecticide resistance profile of GOUNDRY is unknown. It was shown previously that a
693 resistance allele at *kdr* is segregating in this population (Riehle *et al.* 2011), although the
694 resistant and susceptible alleles are segregating at HWE, and the *kdr* allele is segregating
695 at a similar frequency in our sample (Table S1). This suggests that this locus has not
696 been subject to recent severe selection pressure in GOUNDRY.

697 We propose four hypotheses to explain the inbreeding signal in GOUNDRY.
698 Two hypotheses involve the evolution of modified mating biology where GOUNDRY
699 individuals 1) have preference for mating with related individuals, or 2) mate
700 immediately after eclosion. Two additional hypotheses involve the spatial distribution of
701 mating where GOUNDRY individuals either return to their larval habitat to mate or
702 suitable habitats are rare so they return to the same habitat by necessity. In both
703 scenarios, GOUNDRY would exist as a series of micro-populations, perhaps related to
704 habitat fragmentation, where the likelihood of mating with a related individual is higher
705 than that of larger populations such as *A. coluzzii* or *A. gambiae*. The first two
706 hypotheses are biologically less plausible and are not supported by the patterns of IBD
707 tracts since we do not observe a ‘mate preference’ locus that is inbred in all individuals or

708 uniformly long IBD tracts as predicted by these scenarios. The spatial distribution
709 hypotheses predict a distribution of mixed sized IBD tract lengths reflecting mating
710 between both close and more distant relatives by chance. Our data are consistent with the
711 spatial hypotheses, although additional field studies are needed to identify suitable
712 GOUNDRY habitat and test these hypotheses directly. Such dynamics have not been
713 previously observed in mosquito populations, which are thought to typically be large and
714 outbred.

715 Inbreeding is known to have negative fitness consequences in some cases
716 (Hamilton *et al.* 1990; Luong *et al.* 2007). Detrimental effects of inbreeding can be
717 caused either directly when individuals become homozygous for less fit alleles at a given
718 gene or indirectly when overall vigor of an individual is reduced due to exposure of
719 multiple small effect recessive mutations (Charlesworth & Charlesworth 1987). Reduced
720 immune performance is one possible effect of inbreeding, which could have implications
721 for public health if *Anopheles* mosquitoes become more effective vectors of *Plasmodium*
722 parasites. We show here that the degree of inbreeding is positively, albeit weakly,
723 associated with infection prevalence. Our results show that the odds of an individual
724 with even moderate inbreeding coefficient getting infected are 65% greater than for
725 individuals with very low inbreeding coefficients. That we observed a significant
726 association at all is surprising given the coarse and noisy estimates of both relevant
727 parameters. Experimental *Plasmodium* infections are notoriously difficult to control and
728 highly variable even among sibling females (Medley *et al.* 1993; Niare *et al.* 2002).
729 Moreover, our estimates of inbreeding coefficients are based on a relatively small number
730 of variable sites (~650), which corresponds to an average SNP density of 1 per ~400 kb.
731 Given the large number of IBD tracts that are smaller than 400 kb (Figure 6), our
732 estimates are likely to miss many smaller IBD tracts and thus be underestimates of true
733 levels of IBD within these genomes. As such, improved estimates of inbreeding may or
734 may not bolster the significant trend indicating an effect of inbreeding on infection status.
735 Inbreeding coefficients did not, however, explain variation among individuals in the
736 intensity of infection, although increasing the sample size and accuracy of the inbreeding
737 coefficients may change this conclusion. While it remains possible that our rough
738 parameter estimates inhibit this level precise correlation, a single *Plasmodium* oocyst can

739 be sufficient for successful transmission of the parasite. Thus, an increased odds of
740 getting infected, regardless of how intense the infection becomes, could still have serious
741 epidemiological consequences. More work is needed to determine the ecological and
742 population dynamics leading to inbreeding in GOUNDRY, but it is possible that
743 anthropogenic interventions such as intense insecticide and bed-net eradication
744 campaigns, could in principle lead to increased inbreeding in other populations as well.
745 Such inbreeding could be especially problematic if it causes, as our results suggest,
746 increased efficiency in parasite transmission among the remaining small pockets of
747 mosquitoes that escape eradication. If this is the case, the combination between the
748 potential side effects of intense eradication efforts and ecological specialization of
749 subgroups across time and environmental space may make complete interruption of local
750 parasite transmission difficult.

751 In many ways, GOUNDRY has proven to be an atypical subgroup within the well
752 studied *Anopheles gambiae* species complex underscoring our incomplete understanding
753 of vector population dynamics in this system. This study has provided answers to some
754 of the outstanding questions raised around this subgroup while generating still new
755 questions that are difficult to reconcile. Our data suggest that GOUNDRY has existed as
756 an offshoot population from *A. coluzzii* for many generations, hybridizing with its
757 parental population for a substantial portion of its history, yet the most prominent
758 genomic barrier to introgression established only very recently. The process and
759 mechanisms that have kept these two taxa from collapsing back to a single gene pool
760 over their history remains unclear and warrants further study. Moreover, we find
761 evidence for a history of extensive inbreeding within GOUNDRY that we hypothesize
762 could be explained by microstructure creating local breeding demes, yet this population is
763 thought to be exophilic and thus likely less clustered. Whether GOUNDRY has
764 specialized within a rare and patchy ecological niche, has become less likely to fly long
765 distances, or has evolved in some other way that can explain this pattern remains an open
766 question for future study. Additional field studies and genetic analysis of this subgroup
767 are sure to help clarify many of these questions and help to understand the ecological and
768 evolutionary dynamics of populations with relevance to human health and otherwise.

769 **References:**

770 Aminetzach YT, Macpherson JM, Petrov DA (2005) Pesticide resistance via
771 transposition-mediated adaptive gene truncation in *Drosophila*. *Science*
772 (*New York, N.Y.*), **309**, 764–767.

773 Caputo B, Santolamazza F, Vicente JL *et al.* (2011) The “far-west” of *Anopheles*
774 *gambiae* molecular forms. *PloS one*, **6**, e16415.

775 Charlesworth B (1998) Measures of divergence between populations and the effect
776 of forces that reduce variability. *Molecular Biology and Evolution*, **15**, 538–
777 543.

778 Charlesworth D, Charlesworth B (1987) Inbreeding Depression and its Evolutionary
779 Consequences. *Annual Review of Ecology and Systematics*, **18**, 237–268.

780 Coluzzi M, Sabatini A, Petrarca V, Di Deco MA (1979) Chromosomal differentiation
781 and adaptation to human environments in the *Anopheles gambiae* complex.
782 *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **73**, 483–
783 497.

784 Coluzzi M, Sabatini A, della Torre A, Di Deco MA, Petrarca V (2002) A polytene
785 chromosome analysis of the *Anopheles gambiae* species complex. *Science*
786 (*New York, N.Y.*), **298**, 1415–1418.

787 Corbett-Detig RB, Hartl DL (2012) Population genomics of inversion polymorphisms
788 in *Drosophila melanogaster*. *PLoS genetics*, **8**, e1003056.

789 Costantini C, Ayala D, Guelbeogo WM *et al.* (2009) Living at the edge: biogeographic
790 patterns of habitat segregation conform to speciation by niche expansion in
791 *Anopheles gambiae*. *BMC Ecology*, **9**, 16.

792 Crawford JE, Riehle MM, Guelbeogo WM *et al.* (2015) Reticulate Speciation and
793 Barriers to Introgression in the *Anopheles gambiae* Species Complex.
794 *Genome Biology and Evolution*, **7**, 3116–3131.

795 Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of
796 speciation are due to reduced diversity, not reduced gene flow. *Molecular*
797 *Ecology*, **23**, 3133–3157.

798 Dao A, Yaro AS, Diallo M *et al.* (2014) Signatures of aestivation and migration in
799 Sahelian malaria mosquito populations. *Nature*, **516**, 387–390.

800 DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery
801 and genotyping using next-generation DNA sequencing data. *Nature genetics*,
802 **43**, 491–498.

803 Fanello C, Santolamazza F, della Torre A (2002) Simultaneous identification of
804 species and molecular forms of the *Anopheles gambiae* complex by PCR-
805 RFLP. *Medical and veterinary entomology*, **16**, 461–464.

806 Feder JL, Xie X, Rull J *et al.* (2005) Mayr, Dobzhansky, and Bush and the complexities
807 of sympatric speciation in *Rhagoletis*. *Proceedings of the National Academy*
808 *of Sciences of the United States of America*, **102 Suppl 1**, 6573–6580.

809 Fontaine MC, Pease JB, Steele A *et al.* (2015) Mosquito genomics. Extensive
810 introgression in a malaria vector species complex revealed by phylogenomics.
811 *Science (New York, N.Y.)*, **347**, 1258524.

812 Gazal S, Sahbatou M, Babron M-C, Génin E, Leutenegger A-L (2014) FSuite:
813 exploiting inbreeding in dense SNP chip and exome data. *Bioinformatics*
814 *(Oxford, England)*, **30**, 1940–1941.

815 Gnémé A, Guelbéogo WM, Riehle MM *et al.* (2013) Equivalent susceptibility of
816 *Anopheles gambiae* M and S molecular forms and *Anopheles arabiensis* to
817 *Plasmodium falciparum* infection in Burkina Faso. *Malaria Journal*, **12**, 204.

818 Griffin JT, Hollingsworth TD, Okell LC *et al.* (2010) Reducing *Plasmodium falciparum*
819 Malaria Transmission in Africa: A Model-Based Evaluation of Intervention
820 Strategies. *PLoS Med*, **7**, e1000324.

821 Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the
822 Joint Demographic History of Multiple Populations from Multidimensional
823 SNP Frequency Data. *PLoS Genet*, **5**, e1000695.

824 Hamilton WD, Axelrod R, Tanese R (1990) Sexual reproduction as an adaptation to
825 resist parasites (a review). *Proceedings of the National Academy of Sciences*
826 *of the United States of America*, **87**, 3566–3573.

827 Holt RA, Subramanian GM, Halpern A *et al.* (2002) The genome sequence of the
828 malaria mosquito *Anopheles gambiae*. *Science (New York, N.Y.)*, **298**, 129–
829 149.

830 Korneliussen T, Albrechtsen A, Nielsen R (2014) ANGSD: Analysis of Next
831 Generation Sequencing Data. *BMC bioinformatics*, **15**, 356.

832 Krzywinski MI, Schein JE, Birol I *et al.* (2009) Circos: An information aesthetic for
833 comparative genomics. *Genome Research*.

834 Lee Y, Marsden CD, Norris LC *et al.* (2013) Spatiotemporal dynamics of gene flow
835 and hybrid fitness between the M and S forms of the malaria mosquito,
836 *Anopheles gambiae*. *Proceedings of the National Academy of Sciences*,
837 201316851.

838 Leutenegger A-L, Labalme A, Genin E *et al.* (2006) Using genomic inbreeding
839 coefficient estimates for homozygosity mapping of rare recessive traits:
840 application to Taybi-Linder syndrome. *American Journal of Human Genetics*,
841 **79**, 62–66.

842 Leutenegger A-L, Sahbatou M, Gazal S, Cann H, Génin E (2011) Consanguinity
843 around the world: what do the genomic data of the HGDP-CEPH diversity
844 panel tell us? *European journal of human genetics: EJHG*, **19**, 583–587.

845 Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with
846 BWA-MEM. *arXiv:1303.3997 [q-bio]*.

847 Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format
848 and SAMtools. *Bioinformatics (Oxford, England)*, **25**, 2078–2079.

849 Luong LT, Heath BD, Polak M (2007) Host inbreeding increases susceptibility to
850 ectoparasitism. *Journal of Evolutionary Biology*, **20**, 79–86.

851 Medley GF, Sinden RE, Fleck S *et al.* (1993) Heterogeneity in patterns of malarial
852 oocyst infections in the mosquito vector. *Parasitology*, **106 (Pt 5)**, 441–449.

853 Mitri C, Markianos K, Guelbeogo WM *et al.* (2015) The kdr-bearing haplotype and
854 susceptibility to *Plasmodium falciparum* in *Anopheles gambiae*: genetic
855 correlation and functional testing. *Malaria Journal*, **14**, 391.

856 Molineaux L, Gramiccia G, Organization WH (1980) The Garki project : research on
857 the epidemiology and control of malaria in the Sudan savanna of West Africa.

858 Murray CJ, Rosenfeld LC, Lim SS *et al.* (2012) Global malaria mortality between 1980
859 and 2010: a systematic analysis. *The Lancet*, **379**, 413–431.

860 Navarro A, Barton NH (2003) Chromosomal Speciation and Molecular Divergence--
861 Accelerated Evolution in Rearranged Chromosomes. *Science*, **300**, 321–324.

862 Ndiath MO, Brengues C, Konate L *et al.* (2008) Dynamics of transmission of
863 Plasmodium falciparum by Anopheles arabiensis and the molecular forms M
864 and S of Anopheles gambiae in Dielmo, Senegal. *Malaria journal*, **7**, 136.

865 Niare O, Markianos K, Volz J *et al.* (2002) *Genetic loci affecting resistance to human*
866 *malaria parasites in a West African mosquito vector population.*

867 Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J (2012) SNP calling, genotype
868 calling, and sample allele frequency estimation from New-Generation
869 Sequencing data. *PloS one*, **7**, e37558.

870 Noor MAF, Bennett SM (2009) Islands of speciation or mirages in the desert?
871 Examining the role of restricted recombination in maintaining species.
872 *Heredity*, **103**, 439–444.

873 Noor MA, Grams KL, Bertucci LA, Reiland J (2001) Chromosomal inversions and the
874 reproductive isolation of species. *Proceedings of the National Academy of*
875 *Sciences of the United States of America*, **98**, 12084–12088.

876 Oliveira E, Salgueiro P, Palsson K *et al.* (2008) High levels of hybridization between
877 molecular forms of Anopheles gambiae from Guinea Bissau. *Journal of*
878 *medical entomology*, **45**, 1057–1063.

879 Riehle MM, Guelbeogo WM, Gneme A *et al.* (2011) A cryptic subgroup of Anopheles
880 gambiae is highly susceptible to human malaria parasites. *Science (New York,*
881 *N.Y.)*, **331**, 596–598.

882 Rieseberg LH (2001) Chromosomal rearrangements and speciation. *Trends in*
883 *ecology & evolution*, **16**, 351–358.

884 Santolamazza F, Mancini E, Simard F *et al.* (2008) Insertion polymorphisms of
885 SINE200 retrotransposons within speciation islands of Anopheles gambiae
886 molecular forms. *Malaria journal*, **7**, 163.

887 Schlenke TA, Begun DJ (2004) Strong selective sweep associated with a transposon
888 insertion in *Drosophila simulans*. *Proceedings of the National Academy of*
889 *Sciences of the United States of America*, **101**, 1626–1631.

890 Sharakhova MV, George P, Brusentsova IV *et al.* (2010) Genome mapping and
891 characterization of the *Anopheles gambiae* heterochromatin. *BMC genomics*,
892 **11**, 459.

893 della Torre A, Fanello C, Akogbeto M *et al.* (2001) Molecular evidence of incipient
894 speciation within *Anopheles gambiae* s.s. in West Africa. *Insect Molecular*
895 *Biology*, **10**, 9–18.

896 Vieira FG, Fumagalli M, Albrechtsen A, Nielsen R (2013) Estimating inbreeding
897 coefficients from NGS data: Impact on genotype calling and allele frequency
898 estimation. *Genome research*, **23**, 1852–1861.

899 White BJ, Santolamazza F, Kamau L *et al.* (2007) Molecular karyotyping of the 2La
900 inversion in *Anopheles gambiae*. *The American journal of tropical medicine*
901 *and hygiene*, **76**, 334–339.

902 WHO (2013) World Malaria Report.

903 Zheng L, Benedict MQ, Cornel AJ, Collins FH, Kafatos FC (1996) An integrated genetic
904 map of the African human malaria vector mosquito, *Anopheles gambiae*.
905 *Genetics*, **143**, 941–952.

906

907 **Acknowledgements**

908 We thank Matteo Fumagalli, Filipe Vieira, and Tyler Linderoth for assistance with next
909 generation sequence data analyses and ANGSD. We thank members of the Nielsen
910 group for helpful discussions on various aspects of this work and comments on an earlier
911 version of this manuscript. We also thank multiple anonymous reviewers. We are
912 thankful for the use of the Extreme Science and Engineering Discovery Environment
913 (XSEDE), which is supported by National Science Foundation grant number OCI-
914 1053575. This work was also supported by a National Institutes of Health Ruth L.
915 Kirschstein National Research Service Award and a Cornell Center for Comparative and
916 Population Genomics Graduate Fellowship to JEC.

917

918 **Data Accessibility**

919 Sequence data generated for this study can be accessed through the Short Read Archive at
920 NCBI under BioProject ID PRJNA273873.

921

922 **Author Contributions**

923 JEC wrote software and analyzed data; KDV, MMR, WMG, AG, NS contributed new
924 reagents and analytical tools; JEC, BPL, KDV, MMR, and RN designed research. JEC
925 wrote the manuscript with contributions from the rest of the authors.

926

927

928 **Tables**

929 **Table 1:** Optimized parameter values and confidence intervals from the maximum-
930 likelihood demographic model for GOUNDRY and *A. coluzzii*. See Figure 2 for
931 parameter descriptions.

Parameter	Optimized Value	95% Confidence Interval	
		Lower	Upper
$\theta (4N_A\mu L)^a$	180,914		
N_A^b	126,252	88,916	150,976
<i>Split Times</i>			
t_1	259,220	114,087	396,171
t_2	754,204	741,946	766,215
t_3	99,235	93,113	105,754
t_{TOT}	1,112,660	967,181	1,250,106
<i>Population sizes</i>			
N_{Ac1}/N_A	0.01	0.01	0.01
N_{Ac2}/N_A	5.74	4.58	7.65
N_{Ac3}/N_A	12.34	9.54	16.61
N_{G1}/N_A	0.01	0.00*	0.08
N_{G2}/N_A	0.19	0.11	0.32
N_{G3}/N_A	0.78	0.62	1.05
<i>Migration rates</i> $(4Nm)^c$			
G1 into Ac1	0.010	0.0047	0.02

G2 into Ac2	1.29	1.25	1.33
G3 into Ac3	0.37	0.08	0.74
Ac1 into G1	0.080	0.00*	0.64
Ac2 into G2	1.17×10^{-4}	0.00*	0.0015
Ac3 into G3	1.50	1.44	1.55

a – Instead of estimating a confidence interval for θ which itself is not model parameter, we solved for N_A and calculated a confidence for this parameter. In the implementation of *dadi*, N_A is used to scale all other parameters in the model.

b – Ancestral population size was calculated from the estimate of θ by dividing this value by 4 times the number of sites times the mutation rate (see Methods above).

c – Values of $4Nm$ were calculated by multiplying the migration rate reported by *dadi* ($2N_{Am}$) by 2 times the ratio of the effective size of the recipient population (e.g. N_{G1}) over N_A .

* indicates cases where the lower bound of the 95% CI was negative. This is not meaningful, so we set these values to 0.

932

933 **Table 2:** Association between inbreeding coefficients and *Plasmodium* infection

934 prevalence.

Cutoff ^b	Inbreeding Level ^a				X^2 value	P value ^c
	Low		High			
	Infected	Not infected	Infected	Not infected		
0.0260	77	60	93	44	3.4870	0.0205
0.0292	81	60	89	44	2.2200	0.0546

a – Inbreeding coefficient class

b – Cutoff used to assign individuals to low or high inbreeding class. Individuals were assigned to low class if their F value was less than or equal to this cutoff. See text for explanation choice of cutoff values.

c – P values were calculated by comparing the empirical X^2 value to X^2 values obtained from 10^4 permuted datasets in a one-tailed test (See Methods).

935

936 **Figure Legends**

937

938 **Figure 1: Average genetic relationships among species and subgroups in *Anopheles***

939 ***gambiae* species complex.** Unrooted neighbor-joining tree calculated with the *ape*
940 package in R and drawn with Geneious software. Branches indicate genetic distance
941 (D_{xy}) calculated using intergenic sites (see Methods) with scale bar for reference.

942 Bootstrap support percentages are indicated on all internal nodes. Branch lengths and
943 95% CIs indicated for branches leading to *A. merus* and *A. arabiensis*.

944

945 **Figure 2: 2D-Site frequency spectrum and demographic model fitting of**

946 **GOUNDRY and *A. coluzzii*.** **A)** Three-epoch demographic model. One and two-epoch
947 models have parameters from only first (Epoch 1) or first and second epochs, respectively.
948 N parameters indicate effective population sizes. The duration of each epoch is specified
949 by t parameters. Migration parameters ($2Nm$) are included as functions of the ratio of
950 epoch-specific effective sizes relative to the ancestral effective size. We included
951 separate migration parameters for *A. coluzzii* into GOUNDRY migration ($2N_{Am_{GC}}$) and
952 GOUNDRY into *A. coluzzii* ($2N_{Am_{CG}}$). **B)** Autosomal, unfolded two-dimensional site-
953 frequency spectrum (2D-sfs) for GOUNDRY and *A. coluzzii* for empirical data. **C)** 2D-
954 sfs (top row) for maximum-likelihood models under four demographic models. Residuals
955 are calculated for each model comparison (bottom row) as the normalized difference
956 between the model and the data (model – data), such that red colors indicate an excess
957 number of SNPs predicted by the model. See Table 1 for parameter values of the best-fit
958 models under each demographic scenarios.

959

960 **Figure 3: Chromosomal distributions of nucleotide diversity (π) at inter-genic sites**

961 (LOESS-smoothed with span of 1% using 10 kb non-overlapping windows). Low
962 complexity and heterochromatic regions were excluded. The strong reduction of
963 diversity on the X chromosome in GOUNDRY (Mb 8.47 – 10.1) corresponds to putative
964 chromosomal inversion *Xh*.

965

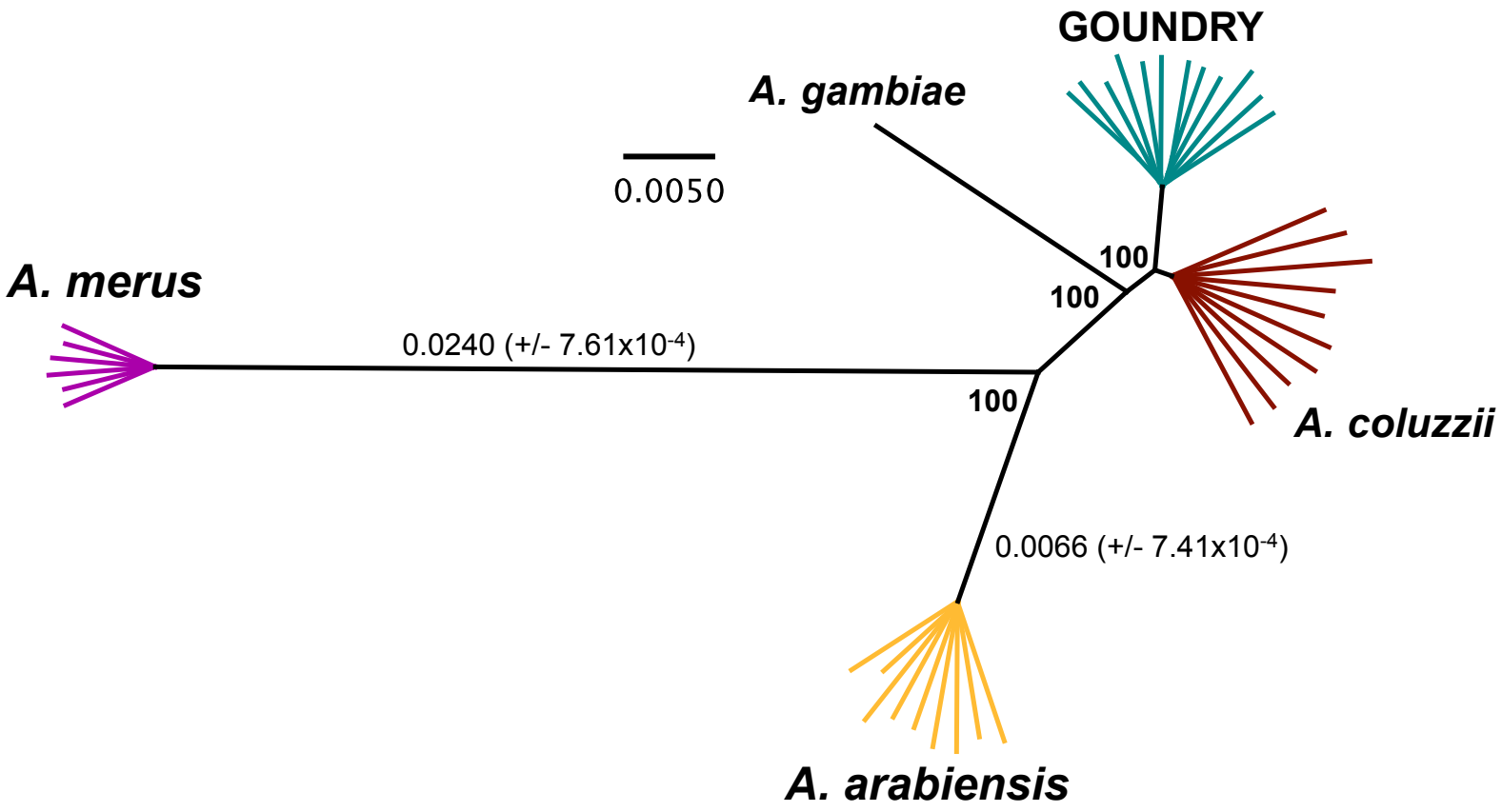
966 **Figure 4: Modeling expected values of Relative Node Depth (D_{GAc}/D_{GAg}).** **A)**
967 Expected values of RND when ancestral population sizes are assumed to be equal.
968 Colors indicate the expectations under different relative split times. **B)** Expected values
969 with t_{GAg} split time fixed to 1.1 (top) times the split time between GOUNDRY and *A.*
970 *coluzzii* (t_{GAc}) or 1.5 times (bottom). Colors indicated relative effective sizes of ancestral
971 populations. Values are plotted as a function of the GOUNDRY-*A.coluzzii* effective size
972 (x-axis). Grey bar indicates 95% confidence interval demographic estimate for
973 GOUNDRY-*A. coluzzii* ancestral size (see Methods).

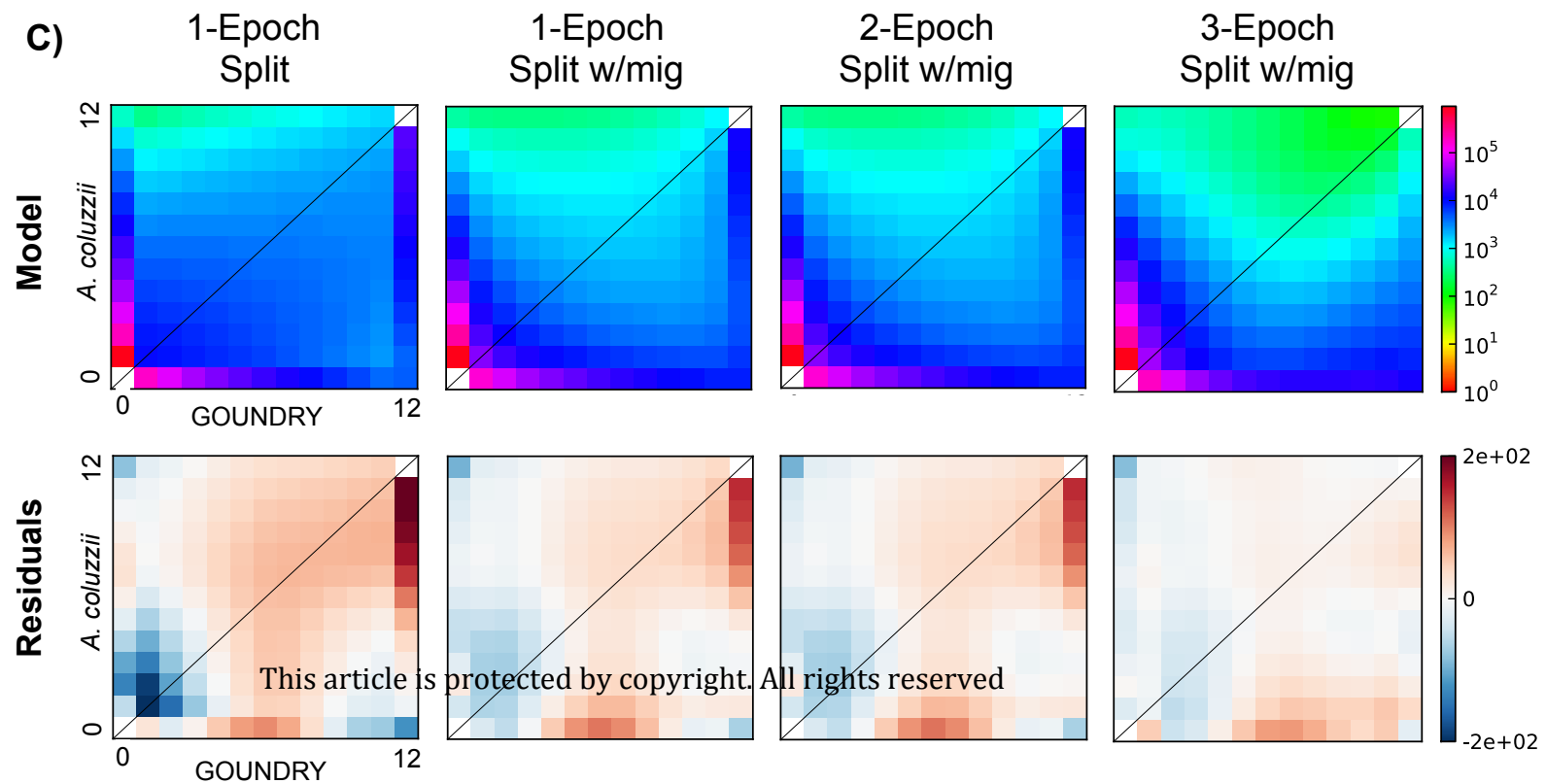
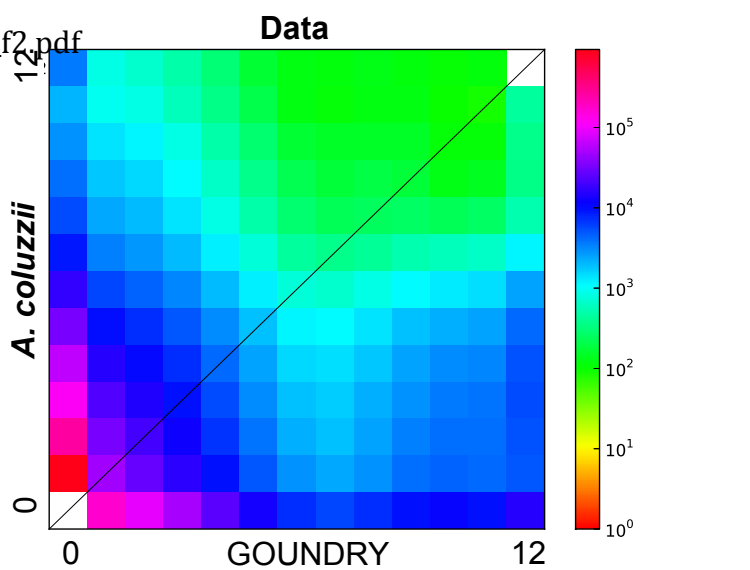
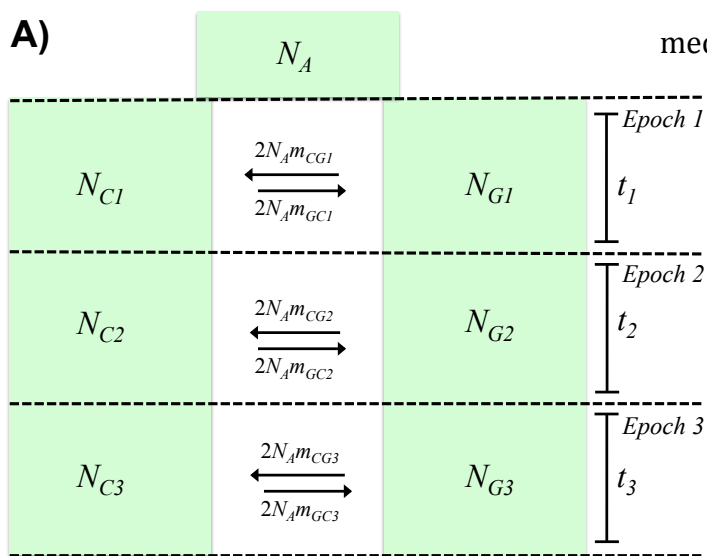
974

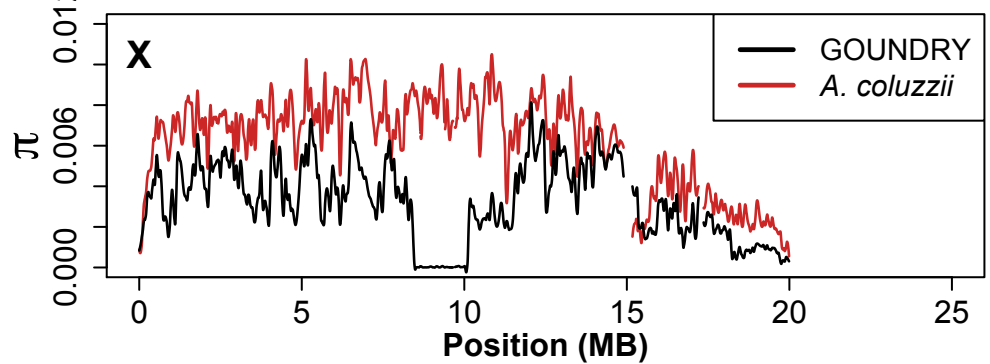
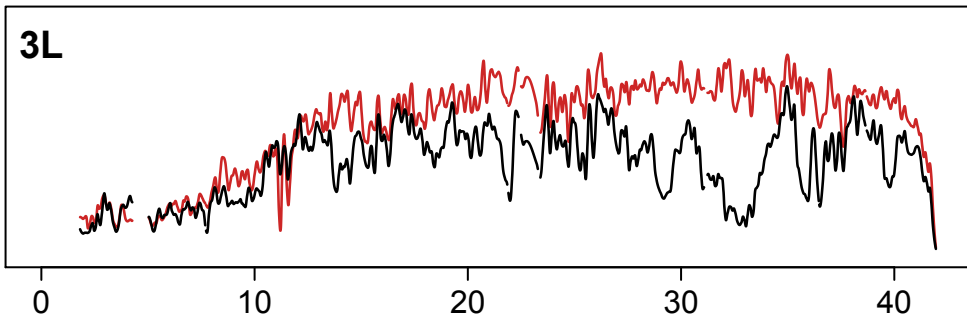
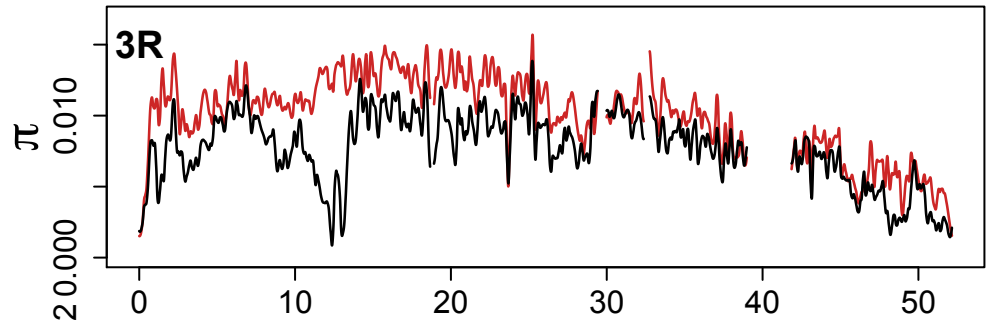
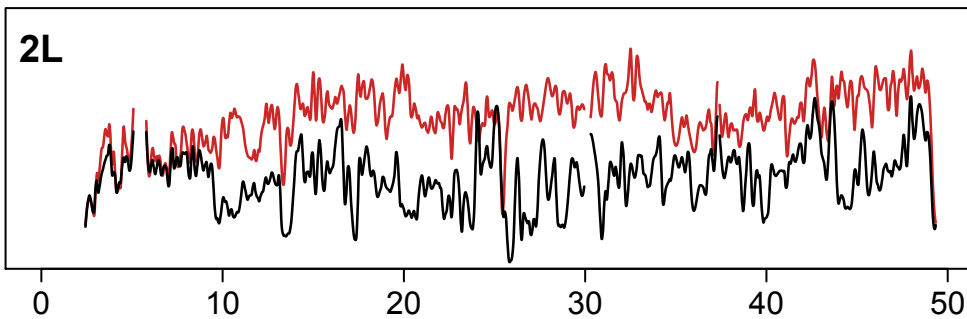
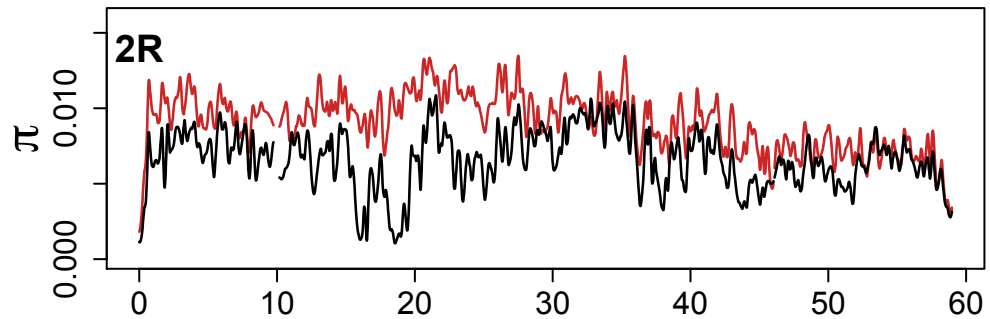
975 **Figure 5: Relative genetic divergence (D_a) between GOUNDRY and *A. coluzzii*.** D_a
976 plotted as a function of nucleotide diversity (*A. coluzzii*) using only intergenic sites in
977 non-overlapping 10 kb windows. Low complexity and heterochromatic regions were
978 excluded. X-Free: freely recombining regions on X chromosome. X-Inv: region inside
979 putative *Xh* chromosomal inversion. Non-parametric Mann-Whitney test indicates that
980 relative divergence (D_a) is significantly higher inside *Xh* ($P < 2.2 \times 10^{-16}$), consistent with
981 this region acting as barrier to gene flow.

982

983 **Figure 6: GOUNDRY genomes harbor long tracts of Identity-By-Descent.** **A)**
984 Comparison between rates of IBD in one representative *A. coluzzii* diploid (black; 'Ac' in
985 figure) and one representative GOUNDRY diploid on the 3L chromosomal arm (Orange;
986 'G' in figure) plotted in physical distance. Top panel shows Loess-smoothed estimate of
987 heterozygosity in 1 kb windows and bottom panel shows IBD tracts called with FSuite
988 (Methods). *A. coluzzi* individuals do not harbor long IBD tracts, and heterozygosity
989 within GOUNDRY individuals is comparable to heterozygosity in *A. coluzzii* except in
990 long regions of homozygosity. **B)** Genetic position and size of IBD regions (orange
991 bands) called with FSuite. **C)** Genetic position and size of IBD tracts called with FSuite
992 for six additional GOUNDRY individuals. Small breaks in long IBD tracts reflect rare
993 genotype errors causing erroneous break in IBD tract.







Position (MB)

