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, Kyriacoss Markianos, Emmanuel Bischoff, Wamdaogo M. Guelbeogo, Awa Gneme, N’Fale Sagnon, Kenneth D Vernick, Rasmus Nielsen, Brian P. Lazzaro

To cite this version:

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

Running Head: Evolution of the Anopheles GONDROY subgroup.

Authors: Jacob E. Crawford¹,², Michelle M. Riehle³, Kyriacos Markianos⁴, Emmanuel Bischoff⁵, Wamdaogo M. Guelbeogo⁶, Awa Gneme⁶, N’Fale Sagnon⁶, Kenneth D. Vernick⁵, Rasmus Nielsen²*, Brian P. Lazzaro¹*.

* These authors contributed equally to this work.

Affiliations:
1. Department of Entomology, Cornell University, Ithaca, NY, USA
2. Department of Integrative Biology, University of California, Berkeley, Berkeley, CA, USA
3. Department of Microbiology, University of Minnesota, St. Paul, MN, USA
4. Program in Genomics, Children’s Hospital Boston, Harvard Medical School
5. Unit for Genetics and Genomics of Insect Vectors, Institut Pasteur, Paris, France
6. Centre National de Recherche et de Formation sur le Paludisme, 1487 Avenue de l’Oubritenga, 01 BP 2208 Ouagadougou, Burkina Faso.

Corresponding Author: Jacob Crawford, Department of Integrative Biology, 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
Key Words: *Anopheles gambiae*, malaria, population genetics, inbreeding, demography, speciation

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

Introduction:
The continued devastating burden of malaria on human populations in sub-Saharan Africa (Murray *et al.* 2012; WHO 2013) spurs ongoing searches for novel means of controlling vector mosquitoes, including through genetic manipulation. However, it is becoming increasingly appreciated that *Anopheles* species frequently form partially
reproductively isolated and ecologically differentiated subpopulations (Costantini et al. 2009; Gnémé et al. 2013; Lee et al. 2013; Fontaine et al. 2015), which could complicate control efforts and extend disease transmission across seasons and micro-environmental space. As an example, a recent study showed that subgroups of Anopheles gambiae sensu lato have evolved distinct approaches for surviving the dry season resulting in the presence of vector populations throughout an extended proportion of the year (Dao et al. 2014). Comprehensive genomic analysis of evolutionary origins, demography, and adaptation will advance our understanding of such phenotypic divergence and its role in the formation of new Anopheles subgroups. Furthermore, genomic analysis of population diversity and genetic affinity among taxa can elucidate epidemiologically relevant aspects of population ecology like breeding structure and ecological distribution that are important for malaria control efforts.

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

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

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

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

**Materials and Methods**

**Mosquito samples**

Mosquito sample collection and species/subgroup identification was previously described for *A. coluzzii*, GOUNDARY, and *A. arabiensis* samples (Riehle *et al.* 2011).
Briefly, larvae and adults were collected from three villages in Burkina Faso in 2007 and 2008 (Table S1). Larvae were reared to adults in an insectary, and both field caught adults and reared adults were harvested and stored for DNA collection. In addition to standard species diagnostic assays, individuals were assigned to genetic subgroups using genetic clustering analysis based on 3rd chromosome SNPs and microsatellites (Riehle et al. 2011). One A. gambiae individual was also included in this study. This sample was collected indoors as an adult in the village of Korabo in the Kissidougou prefecture in Guinea in October 2012. Individuals were typed for species, molecular form and 2La karyotype using a series of standard molecular diagnostics (Fanello et al. 2002; White et al. 2007; Santolamazza et al. 2008). All A. coluzzii and A. arabiensis samples are 2La<sup>aa</sup> homokaryotypes and the A. gambiae sample typed as a heterokaryotype (2La<sup>aa/+</sup>). As discussed above, both forms of the 2La inversion are segregating in GOUNDRY, and we chose to sequence eleven 2La<sup>aa/+</sup> GOUNDRY samples and one 2La<sup>aa</sup> sample (GROUND_0446).

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

A detailed description of the DNA extractions, sequencing, and processing has been included in a separate publication (Crawford et al. 2015), but briefly, genomic DNA was extracted using standard protocols and was sequenced using the Illumina HiSeq2000 platform by BGI (Shenzhen, China). Paired-end 100-bp reads were obtained for all samples. The *Anopheles gambiae* sample was sequenced on the same platform at the University of Minnesota Genomics Center core facility. Raw Illumina reads were deposited at NCBI SRA under BioProject ID PRJNA273873. Short-reads were aligned in two steps using BWA-mem (v0.7.4) alignment algorithm [(Li 2013); bwa.sourceforge.net]. First, reads were mapped to the *A. gambiae* PEST AgamP3 reference assembly [(Holt et al. 2002); vectorbase.org]. Second, reads were mapped to a new updated sequence where the major allele (frequency in sample ≥ 0.5) from each population were substituted into the PEST reference to make population specific references. Local realignment around indels was conducted with GATK v.2.5-2 (DePristo et al. 2011). Duplicates were removed using the SAMtools v.0.1.18 (Li et al. 2009) rmdup function. We applied a series of quality filters and identified a set of robust
genomic positions that were included in all downstream analysis. As a rule, heterochromatic regions as defined for *A. gambiae* (Sharakhova et al. 2010) were excluded from all analyses since short read mapping is known to be problematic in such regions.

**Bioinformatics and population genetic analyses**

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

**Inbreeding analysis**

*Estimating inbreeding coefficients*

Initial estimates of the global site frequency spectrum (SFS) in GOUNDRY produced distributions of allele frequencies that deviated substantially from standard equilibrium expectations, as well as from those observed in the *A. coluzzii* and *A. arabiensis* groups. Most notably, the proportion of doubletons was nearly equal to that of singletons in *A. gambiae* GOUNDRY (see Results). This observation is consistent with widespread inbreeding in the GOUNDRY subgroup. We tested the hypothesis of extensive inbreeding in two ways, with the goals of both characterizing the pattern of inbreeding in this subgroup as well as obtaining inbreeding coefficients for each individual that could then be used as priors for an inbreeding-aware genotype-calling algorithm. We used the method of Vieira et al. (Vieira et al. 2013), which estimates inbreeding coefficients in a probabilistic framework taking uncertainty of genotype calling into account. This approach is implemented in a program called ngsF (github.com/fgvieira/ngsF). ngsF estimates inbreeding coefficients for all individuals in the sample jointly with the allele frequencies in each site using an Expectation-Maximization (EM) algorithm (Vieira et al. 2013). We estimated minor allele frequencies at each site (-doMaf 1) and defined sites as variable if their minor allele
frequency was estimated to be significantly different from zero using a minimum log
likelihood ratio statistic of 24, which corresponds approximately to a $P$ value of $10^{-6}$.
Genotype likelihoods were calculated at variable sites and used as input into ngsF using
default settings. For comparison, we estimated inbreeding coefficients for *A. coluzzii*,
GROUNDY, and *A. arabiensis* using data from each chromosomal arm separately.

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

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

**IBD Tracts**

We examined the effects of inbreeding within diploid individuals using FEstim
(Leutenegger et al. 2006, 2011), which implements a maximum likelihood method within
a Hidden-Markov-Model that models dependencies along the genome. We used the
FSuite v.1.0.3 (Gazal et al. 2014) pipeline to generate submaps, estimate inbreeding
parameters using FEstim, identify IBD tracts, and plot IBD tracts using Circos v.0.67-6
(Krzywinski et al. 2009). To minimize linkage disequilibrium that creates non-
independence among SNPs while maximizing information content, we generated 20
independent random subsets of between 187 and 193 SNPs (or submaps) spaced at least 1
kb apart, and inbreeding parameters were inferred using all 20 submaps. We used allele
frequencies estimated using ANGSD above (–doMaf and -indF) for calculation of
emission probabilities in FEstim. We also used genetic maps for *Anopheles gambiae*
from Zheng et al. (Zheng et al. 1996). To convert data from Zheng et al. to dense
genetic maps, we mapped primers from that study onto the *Anopheles gambiae* PEST
reference using standard e-PCR approaches that map PCR primers onto a reference sequence using computational sequence matching. Autosomal maps and code for polynomial analysis were kindly provided by Russ Corbett-Detig (github.com/tsackton/linked-selection/), and we performed e-PCR mapping for the X chromosome. We fit a polynomial function to the genetic map for each chromosome and used this function to convert the physical position of SNP marker to genetic distance. For this analysis, we joined the left and right arms of chromosomes 2 and 3 by adjusting the physical position of SNPs on the left arms by the full length of the right arm.

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

**Ruling out bioinformatic and sequencing artifacts**

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

We also asked whether excess homozygosity could stem from erroneous assignment of homozygous genotypes at true heterozygous sites. Such errors could result if short read depths were exceptionally low in some genomic regions. We calculated read
depths at sites in different genotype classes in GOUNDRY and find that the mean read depth is 12.3569 (σ = 5.3917) at homozygous reference sites, 12.2156 (σ = 5.1235) at homozygous alternative sites, and 12.6871 (σ = 5.5163) at heterozygous sites, indicating that the distribution of read depth is very similar between all three classes (Figure S2).

We find a similar pattern in A. coluzzii, which shows no evidence of inbreeding. In this population, the mean read depth is 10.4773 (σ = 4.7555) at homozygous reference sites, 11.0082 (σ = 4.1849) at homozygous alternative sites, and 10.6660 (σ = 4.7755) (Figure S2). Moreover, the distributions of read depths at heterozygous sites and homozygous sites are very similar in both the A. coluzzii and GOUNDRY populations (Figure S2). These results strongly suggest that bioinformatic artifacts cannot explain the excess homozygosity and IBD tracts observed in GOUNDRY.

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

**Inbreeding-phenotype association test**

Since 12 GOUNDRY genomes is not a large enough sample size for association testing, we obtained SNP genotype data for 274 GOUNDRY individuals who had also been phenotyped in a *Plasmodium* infection experiment as part of an independent study (Mitri et al. 2015). Full details of Illumina SNP chip assay design and data collection and infection experiments are available in that publication but will be summarized here. Briefly, larvae were collected in three villages in central Burkina Faso, and raised to
adulthood in the laboratory where females were given *Plasmodium falciparum*-infectious bloodmeals from local volunteers. Fully fed females were dissected 7-8 days later for oocyst quantification and DNA extraction. For the Illumina SNP chip, SNPs were identified from raw sequence reads generated for genome sequencing projects for A. *coluzzii* and A. *gambiae* as well as from independent deep sequencing efforts. DNA hybridization and genotype calling were conducted using standard procedures followed by stringent quality filtering of genotype calls and independent confirmation using duplicate hybridizations and independent Sequenom assays using a subset of SNPs. We used a set of SNPs distributed approximately uniformly across the autosomes. Among these sites, we included only sites (n = 678) that were variable in the GOUNDRY subgroup.

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

To test whether infection prevalence was higher in inbred individuals, we used a two-by-two $\chi^2$ test. The table cells corresponded to ‘infected’ and ‘not infected’ phenotypes as well as high and low inbreeding coefficients. Since the distribution of inbreeding coefficients was not bimodal, we categorized individuals as either ‘high’ or ‘low’ inbreeding levels based on whether their inbreeding coefficient was above or below a cutoff value, respectively. We included two cutoff values: 1) $F = \text{the median coefficient value of 0.026}$, and 2) The maximum $F$ estimated from genome sequencing in A. *coluzzii*, which does not show signs of inbreeding. To establish statistical significance while preserving correlations among mosquitoes within each blood donor cohort, we randomly permuted infection phenotype among mosquitoes within donor cohort, and recalculated the $\chi^2$ value. We compared the empirical $\chi^2$ value to $10^4$ values from permuted datasets in a one-tailed statistical test. We further tested the association and the effect of blood-donor using the Cochran-Mantel-Haenszel procedure (cmh.test in R)
that directly accounts for additional factors within the contingency test. To test for
correlations between inbreeding coefficients and the number of oocysts (infection
intensity), we fit a linear model to relate inbreeding coefficients (log transformed) to the
number of oocysts (log transformed) with blood-donor as a factor in the model. Only
mosquitoes with at least one oocyst were included in this part of the analysis.

Results

Genome Sequencing and Population Genetic Analysis

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

Genetic Relatedness Among Species and Subgroups

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

It has been speculated that GOUNDRY may be a recently formed backcrossed
hybrid of A. coluzzii and A. gambiae (Lee et al. 2013). This hypothesis also predicts that
GROUNDRY will be segregating chromosomes that are mosaics of haplotypes derived from *A. coluzzii* and *A. gambiae*, and therefore most, if not all, polymorphisms found in GOUNDRY should also be found in one of these putative parental taxa. In contrast, we find 7,383 fixed differences between *A. coluzzii* and GOUNDRY [excluding 2L since it is dominated by the large 2La inversion known to have crossed species boundaries (Fontaine et al. 2015)], of which 27% are putatively GOUNDRY-specific alleles not shared with *A. gambiae*, *A. arabiensis*, or *A. merus*. GOUNDRY shares an allele with the *A. gambiae* individual sampled here at an additional 31.5% of the fixed sites, although this number may increase if more *A. gambiae* samples are included. These results do not exclude the possibility of gene flow between GOUNDRY and *A. gambiae*, but they fail to support the hypothesis that GOUNDRY is simply a very recent hybrid of *A. coluzzii* and *A. gambiae*. Instead, the substantial number of putatively GOUNDRY-specific fixed alleles support GOUNDRY as a unique subgroup that may have originated as an offshoot of *A. coluzzii* and experienced subsequent gene flow from *A. gambiae*.

**Origins of GOUNDRY**

It has been hypothesized that the advent of agriculture in sub-Saharan Africa ~5-10 kya played a role in driving diversification and expansion of *Anopheles* mosquitoes (Coluzzi et al. 2002). The two-dimensional site-frequency spectrum reveals substantial differentiation in allele frequencies between GOUNDRY and *A. coluzzii* with many fixed differences differentiating these groups and is not compatible with a very recent origin of GOUNDRY (Figure 2). To test whether the origin of GOUNDRY could have been associated with habitat modification driven by agriculture, we fit four population historical models with increasing complexity (Figure 2; Table 1; Methods) to the two-dimensional site frequency spectrum for GOUNDRY and *A. coluzzii* using dadi (Gutenkunst et al. 2009). The 2D spectra from the empirical data and the best-fit model for each demographic model are presented in Figure 2. We first fit a simple one-epoch, split model with no migration. The maximum-likelihood model under this scenario gave a poor fit to the empirical data with a likelihood value ($L$) of -176,635.8. We then added asymmetrical migration to the model (one-epoch, split with migration), which resulted in a nearly three-fold improvement of the likelihood value improvement of the fit of the
model to the data with $L_{1\text{-ep-split-mig}} = -59.896.75$, providing strong evidence that migration has played a key role in the history of these taxa. Residual differences between the 2D spectra from the model and the data (Figure 2), however, were unevenly distributed across the spectra, suggesting that one-epoch models are missing potentially important features of the demographic history. To improve flexibility in the model fitting, we fit both two-epoch and three-epoch population split-with-migration models (Table 1). Interestingly, the adding a second epoch did not result in a substantial improvement of the fit to the data as indicated by the remaining large residuals and decreased likelihood value ($L_{2\text{-ep-split-mig}} = -59,949.49$) relative to the one-epoch model. Adding a third epoch, however, achieved a considerable improvement of the fit to the data ($L_{3\text{-ep-split-mig}} = -49,023.85$). Residuals indicating differences between the model and data are also presented and suggest that deviations between spectra associated with the model and data are well correlated.

The best-fitting three-epoch-split-with-migration model (Table 1) predicts that these subgroups diverged ~111,200 ya (95% CI 96,718 – 125,010), followed by a 100-fold reduction in the size of both subgroups after isolation (Methods). The timing of this model rejects any role of modern agriculture in subgroup division, although it should be noted that estimates of such old split times inherently carry considerable uncertainty. Our inferred model is inconsistent by an order of magnitude with agriculture as a driving force in cladogenesis and is more consistent with habitat fragmentation and loss due to natural causes, potentially including climatic shifts such as changes in pluviometry that would lead to increased population size. The model supports a >500-fold population growth in A. coluzzii and 19-fold growth in GOUNDRY with extensive gene flow between them ~85,300 ya, consistent with a re-establishment of contiguous habitat and abundant availability of bloodmeal hosts. Interestingly, the model supports additional population growth in both subgroups in the most recent epoch, which spans the last 10,000 years and coincides with the advent of agriculture. Any hybridization related to secondary contact during this period has not led to complete homogenization, as we conservatively identified nearly 8,000 fixed nucleotide differences distributed across the genomes of the two subgroups.
The dates reported here depend on assumptions about both the physiological mutation rate as well as the number of generations per year, neither of which are well known in Anopheles. As such, the details of these results would differ somewhat if different estimates were used. However, we would have to invoke extreme values of these parameters that are outside reasonable expectation in order to obtain estimates for the time of the GOUNDRY- A. coluzzii split that coincides with the advent of agriculture. Overall, the model suggests that the origin of GOUNDRY is not recent and both GOUNDRY as well as A. coluzzii have both undergone bouts of population growth and increased rates of hybridization in more recent evolutionary time.

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

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

A large cluster of fixed differences (~ 530; Figure S3) identified between GOUNDRY and A. coluzzii falls within a 1.67 Mb region on the X chromosome that is nearly absent of polymorphism (Figure 3), despite sequence read coverage comparable to neighboring genomic regions (Figure S3). The remarkably large size of the region devoid of diversity would imply exceptionally strong positive selection under standard rates of meiotic recombination. For comparison, previously identified strong sweeps associated with insecticide resistance span approximately 40 Kb and 100 kb in freely recombinating genomic regions of Drosophila melanogaster and D. simulans, respectively (Schlenke & Begun 2004; Aminetzach et al. 2005). The swept region in GOUNDRY is marked by especially sharp edges (Figure 3), implying that recombination has been
suppressed at the boundaries this region. Collectively, these observations suggest that the swept region may be a small chromosomal inversion, which we have named Xh in keeping with inversion naming conventions in the Anopheles system. Notably, this pattern is virtually identical to the pattern of diversity in a confirmed X-linked inversion discovered in African populations of D. melanogaster (Corbett-Detig & Hartl 2012). The Xh region in GOUNDRY includes 92 predicted protein coding sequences (Table S2), including the white gene, two members of the gene family encoding the TWDL cuticular protein family (TWDL8 and TWDL9), and five genes annotated with immune function (CLIPC4, CLIPC5, CLIPC6, CLIPC10, PGRPS1). The lack of diversity in the region implies that the presumed Xh inversion has a single recent origin and was quickly swept to fixation in GOUNDRY. We estimated the age of the haplotype inside the sweep region to be 78 years with a standard deviation of 9.15 by assuming that all segregating polymorphisms in the region postdate fixation of the haplotype (see Methods). Such extraordinarily recent adaptation is consistent with the selection pressures related to 19th and 20th century human activity such as insecticide pressure or widespread habitat modification.

**Xh is a barrier to introgression**

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

One approach to estimate differences in divergence among genomic regions is to compare divergence between a focal pair of subgroups (GOUNDRY and A. coluzzii) to divergence between one of the focal groups and an outgroup (GOUNDRY and A. gambiae) in order to scale divergence levels by differences among regions in mutation rate and the effects of selection on linked sites. This approach estimates what is known
as Relative Node Depth (RND = $D_{DAc}/D_{DAg}$, where subscripts G, Ac, and Ag indicate GOUNDRY, A. coluzzii, and A. gambiae respectively), and a higher RND indicates greater divergence between the focal groups (Feder et al. 2005). We find that RND is 0.7797 on the autosomes and 0.8058 on the X, indicating higher genetic divergence between GOUNDRY and A. coluzzii on X relative to the autosomes. To explicitly test whether such a pattern could be obtained under a pure split model with no gene flow, we obtained expected values of Relative Node Depth (RND) assuming a phylogeny where A. coluzzii and GOUNDRY form a clade with A. gambiae as the outgroup (Methods).

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

To test the second expectation that the inversion is more diverged than other regions on the X chromosome, we compared divergence with A. coluzzii in windows inside and outside of the inverted region. Absolute sequence divergence ($D_{xy}$) is not sensitive to detect differential gene flow for relatively recent changes in gene flow (Cruickshank & Hahn 2014), and we expect that the putative Xh inversion is likely too young for measurable differences to have accumulated, so we tested for excess divergence in the Xh inversion using a more sensitive approach. For comparison, we find that the inverted region is significantly more diverged between A. coluzzii and GOUNDRY relative to the remaining X chromosome ($\bar{D}_{G ALk} (Xh) = 0.0103, \bar{D}_{G ALk} (non-Xh) = 0.0071; M-W P < 2.2\times10^{-16}$), but nucleotide diversity in A. coluzzii is also significantly higher in this region ($\pi_{Ac} (Xh) = 0.0080, \pi_{Ac} (non-Xh) = 0.0061; M-W P < 5.49\times10^{-14}$), implying that the increased divergence could be partially explained by
increased mutation rate in this region. However, when absolute divergence along the X chromosome is explicitly scaled by the mutation rate inferred from levels of polymorphism in the A. coluzzii sample ($D_a$), the putatively adaptive Xh inversion between GOUNDRY and A. coluzzii is proportionally much more divergent than is the remainder of the X chromosome ($\bar{D}_e (Xh) = 0.0022$, $\bar{D}_e$ (non-Xh) = 0.0013; M-W $P < 4.89 \times 10^{-8}$; Figure 5). Although relative measures of divergence, such as $D_a$, are known, for example, to be confounded by reductions in nucleotide diversity related to natural selection on linked sites (Charlesworth 1998; Noor & Bennett 2009), we believe that this analysis is robust to these concerns because the comparison is among only X-linked windows and the region of interest is in a region of the chromosome that is highly diverse in subgroups where there is no evidence of selective sweeps (Figure 3).

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

**GOUNDRY is inbred**

Unexpectedly, we found that GOUNDRY exhibits a deficiency of heterozygotes relative to Hardy-Weinberg expectations and extensive regions of Identity-By-Descent (IBD), a pattern that is not observed in any of our other Anopheles collections. Individual diploid GOUNDRY genomes are checkered with footprints of IBD, even though the genome as a whole harbors substantial genetic variation indicating a relatively large genetic (effective) population size (Figure 6a). The observation of stochastic tracts of
IBD is most consistent with an unusually high rate of close inbreeding. To explicitly test for elevated inbreeding coefficients ($F$), we used a maximum likelihood framework to infer $F$ for each individual without calling genotypes. We found that values of $F$ range from 0.0087 to 0.2106 genome wide (Figure S4). In contrast, estimates of inbreeding coefficients for 10 *A. coluzzii* genomes and 9 *A. arabiensis* genomes were consistently low ($F_{Ac} < 0.03; F_{Aa} < 0.04$). The relatively high inbreeding coefficients in GOUNDRY suggest that this population has a history of mating among relatively closely related individuals.

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

**Effect of inbreeding on Plasmodium-resistance**

Inbreeding is known to have detrimental effects on various phenotypes, including resistance to parasite infection (Hamilton et al. 1990; Luong et al. 2007). To test whether inbreeding in GOUNDRY increases intrinsic susceptibility to *Plasmodium falciparum* infection in this group, we studied a larger panel of 274 GOUNDRY females that were experimentally infected with local wild isolates of *P. falciparum* and genotyped at 1,436 SNPs across the genome (Mitri et al. 2015). After filtering, we estimated inbreeding coefficients with the program *ngsF* (Vieira et al. 2013) using 678 autosomal variable sites (Methods) and found that $F$ ranges from 0 to 0.3797 in this sample of GOUNDRY females (Figure S5). Although it is possible that some GOUNDRY individuals are truly not inbred, all 12 GOUNDRY individuals subjected to whole genome sequencing showed significant evidence of inbreeding, so we suspect that the relatively sparse genotyping (1
per ~400 kb) assay used on this panel of mosquitoes failed to capture IBD tracts in some 
individuals.

Blood feeding experiments were conducted using five human *Plasmodium*

gametocyte donors, and blood donor had a significant effect on both infection prevalence
(ANOVA; $P = 1.593 \times 10^{-9}$) and intensity (ANOVA; $P = 1.194 \times 10^{-13}$). Importantly, the 
distributions of mosquito inbreeding coefficients did not differ significantly between
blood donor cohorts (ANOVA, $P = 0.0934$).

Of the females that fed on infectious blood-meals, 104 (37.9%) had no parasites at
the time of dissection, and we asked whether this infection prevalence is statistically
associated with inbreeding in the mosquito host. Inspection of the distribution of $F$ in this sample indicates that categorization of individuals as inbred or outbred is difficult
since a substantial proportion of individuals were assigned values of $F$ close to 0 (Figures S5 and S6) and even individuals from outbred populations such as *A. coluzzii* and *A.
arabiensis* can have estimates of $F$ as high as 0.03 or 0.04 (Figure S4). Therefore, we
used the median value of $F$ estimated from genome-wide SNPs in GROUNDY (0.026)
and categorized mosquitoes as more inbred ($F > 0.026$) or less inbred ($F \leq 0.026$). We
used a $\chi^2$ test with this categorization approach to test whether higher inbreeding
significantly associated with higher infection prevalence and find that females with
higher inbreeding coefficients are overrepresented in the ‘infected’ class ($P = 0.0205$; Table 2). We also used the Cochran-Mantel-Haenszel procedure to directly account for
blood-donor in the test for association and found very similar results ($P = 0.025$ for
median cutoff). As an alternative assignment approach, we defined the inbreeding
categories using the highest inbreeding coefficient value obtained from full genome
sequencing of an outbred population, *A. coluzzii* ($F = 0.0292$), which should be more
robust to statistical uncertainty than estimates from the SNP chip data, and find that the
association is on the borderline of significance ($P = 0.0546$; Table 2). These analyses
indicate that an increase in the proportion of genes with alleles that are Identical by
Descent may decrease the ability of adult female mosquitoes to resist parasite infection,
although the effect is small enough that detection of the association is sensitive to how
the distribution of $F$ is categorized.
We also asked whether the degree of inbreeding has an effect on the intensity of infection (number of oocysts per midgut). Of the 274 females that fed on natural gametocytemic blood samples and were assayed for infection status, 170 harbored at least one oocyst, while the remaining 104 females were uninfected, corresponding to an infection rate of 0.62. Among the infected females, infection intensity varied from 1 to 38 with a mean of 5.73 oocysts per individual. We fit linear models for mosquitoes fed on each blood donor separately and find no significant correlation ($P > 0.05$) between inbreeding coefficients and infection intensity.

**Discussion**

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

Here, we present an analysis of complete genome sequences from the newly discovered cryptic GOUNDRY subgroup of *A. gambiae*. Our results help clarify some outstanding questions raised by the initial description of this subgroup. We show that, in contrast to initial suggestions (Riehle *et al.* 2011), GOUNDRY subgroup of *A. gambiae* falls genetically within *Anopheles gambiae sensu lato* and is not an outgroup. GOUNDRY shows strongest genetic affinity with *A. coluzzii* and therefore may be an ecologically specialized subgroup of *A. coluzzii*. The discrepancy between our findings and previously published results is likely due to the fact that the first description was based on a small number of microsatellite markers and SNPs and was based on differences in allele frequency, while the current study is based on absolute sequence divergence calculated from whole genome sequencing data, and therefore included both shared and private mutations. Our demographic analysis suggests that GOUNDRY has
existed for approximately 100,000 years and represents a recent example of the frequent
speciation dynamics in *Anopheles* that appears to be common (Crawford *et al.* 2015;
Fontaine *et al.* 2015). Since GOUNDRY was identified using an outdoor sampling
approach not common in previous studies, it was unclear whether or not this subgroup
may be more broadly distributed and just un-sampled. We estimate that the recent
(effective) population size of GOUNDRY is approximately 5% that of *A. coluzzii*,
suggesting that GOUNDRY is likely restricted to a relatively small region of the Sudan-
Savanna zone in West Africa.

In addition to thousands of mutations found to be putatively unique to
GROUNDY, we identified a large GOUNDRY-specific genetic marker in the form of a
new putative X-linked chromosomal inversion that originated and fixed within
GROUNDY within the last 100 years. It remains unknown whether positive selection or
meiotic drive has driven this inverted haplotype to high frequency and ultimately fixation
in GOUNDY, but our results suggest that it may serve as a recent barrier to gene flow
with *A. coluzzii*, and potentially other taxa as well. Collectively, the data show that
nucleotide-diversity corrected divergence is higher inside the putative inverted region, the
inverted region as a chromosomal segment is the most diverged of all segments of the
same size on the X chromosome, and the X chromosome as a whole is more diverged
among GOUNDRY and *A. coluzzii* relative to the autosomes. The most parsimonious
explanation for these patterns is that, although very few new mutations have accumulated
inside of *Xh* since its origin less than 100 years ago, ongoing gene flow between *A.
* coluzzii and GOUNDRY has led to a greater density of shared polymorphism and
therefore lower sequence divergence in non-inverted regions of the X chromosome
relative to the inversion, especially distal to the inversion breakpoints. These results lead
us to conclude that while cladogenesis of GOUNDRY and *A. coluzzii* ~100 kya by other
means established some degree of temporally fluctuating reproductive isolation, the
recently derived *Xh* putative inversion now serves as a genomic barrier to gene flow, and
the effects of selection against migrant haplotypes or lack of recombination with non-
inverted chromosomes have begun to extend to linked sites outside the inversion
breakpoints.
The observation that GOUNDRY is more closely related at the genome level to *A. coluzzii* than to *A. gambiae* could be biased by higher rates of gene flow between GOUNDRY and *A. coluzzii* as well as sampling bias caused by the fact that *A. gambiae* is represented by only a single individual that was sampled from a different country. Although we cannot formally rule out the possibility that GOUNDRY originated as something other than a subpopulation of *A. coluzzii* and later experienced substantial gene flow from *A. coluzzii* that led to genetic affinity in our analysis, the most parsimonious explanation is that it is a subgroup that originated from *A. coluzzii* that has experienced gene flow from multiple sympatric taxa over its history. The most compelling piece of evidence that GOUNDRY is not a recent *A. coluzzii*-*A. gambiae* hybrid-backcross is the presence of the large fixed haplotype on the X chromosome in GOUNDRY that is not expected under the recent backcross model. In support of this notion, a recently published study (Fontaine *et al.* 2015) constructed similar distance based trees using samples of *A. gambiae* from across the continent and found that geographically disparate individuals were consistently interdigitated while excluding *A. coluzzii*, suggesting that species assignment was more important than geography.

An additional potential concern regarding our estimate of the demographic modeling and our conclusion that the X chromosome is more diverged than the autosome between GOUNDRY and *A. coluzzii* stems from introgression between *A. gambiae* and both GOUNDRY and *A. coluzzii*. We showed in a companion manuscript that GOUNDRY has introgressed with *A. gambiae* in the evolutionarily recent past (Crawford *et al.* 2015), and the presence of *A. gambiae* haplotypes in GOUNDRY could bias our demographic estimate of the split time since this introgression was not explicitly modeled. A four-taxon model including *A. gambiae* and *A. arabiensis* would probably improve our estimates, but the dimensionality of such a model would increase dramatically and would require much more sequence data than is available in the current study. Introgression with *A. gambiae* could also compromise our RND analysis in which this group was used as an outgroup. For example, higher introgression between *A. gambiae* and the ingroups on the X relative to the autosome could result in an underestimate of the mutation rate on the X chromosome and thus an inflation of the ingroup divergence. However, we showed in a companion manuscript (Crawford *et al.* 2015) that signals of introgressed haplotypes
are concentrated on the autosome and absent from the X, suggesting that RND scaling
may be downwardly biased on the autosomes rather than the X. For these reasons, *A.
gambiae* is not an ideal outgroup for an RND analysis, but it is suitable for our purposes
and a low rate of introgression from this taxon is not likely to bias our results.

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

We propose four hypotheses to explain the inbreeding signal in GOUNDRY.

Two hypotheses involve the evolution of modified mating biology where GOUNDRY
individuals 1) have preference for mating with related individuals, or 2) mate
immediately after eclosion. Two additional hypotheses involve the spatial distribution of
mating where GOUNDRY individuals either return to their larval habitat to mate or
suitable habitats are rare so they return to the same habitat by necessity. In both
scenarios, GOUNDRY would exist as a series of micro-populations, perhaps related to
habitat fragmentation, where the likelihood of mating with a related individual is higher
than that of larger populations such as *A. coluzzii* or *A. gambiae*. The first two
hypotheses are biologically less plausible and are not supported by the patterns of IBD
tracts since we do not observe a ‘mate preference’ locus that is inbred in all individuals or
uniformly long IBD tracts as predicted by these scenarios. The spatial distribution hypotheses predict a distribution of mixed sized IBD tract lengths reflecting mating between both close and more distant relatives by chance. Our data are consistent with the spatial hypotheses, although additional field studies are needed to identify suitable GOUNDRY habitat and test these hypotheses directly. Such dynamics have not been previously observed in mosquito populations, which are thought to typically be large and outbred.

Inbreeding is known to have negative fitness consequences in some cases (Hamilton et al. 1990; Luong et al. 2007). Detrimental effects of inbreeding can be caused either directly when individuals become homozygous for less fit alleles at a given gene or indirectly when overall vigor of an individual is reduced due to exposure of multiple small effect recessive mutations (Charlesworth & Charlesworth 1987). Reduced immune performance is one possible effect of inbreeding, which could have implications for public health if Anopheles mosquitoes become more effective vectors of Plasmodium parasites. We show here that the degree of inbreeding is positively, albeit weakly, associated with infection prevalence. Our results show that the odds of an individual with even moderate inbreeding coefficient getting infected are 65% greater than for individuals with very low inbreeding coefficients. That we observed a significant association at all is surprising given the coarse and noisy estimates of both relevant parameters. Experimental Plasmodium infections are notoriously difficult to control and highly variable even among sibling females (Medley et al. 1993; Niare et al. 2002). Moreover, our estimates of inbreeding coefficients are based on a relatively small number of variable sites (~650), which corresponds to an average SNP density of 1 per ~400 kb. Given the large number of IBD tracts that are smaller than 400 kb (Figure 6), our estimates are likely to miss many smaller IBD tracts and thus be underestimates of true levels of IBD within these genomes. As such, improved estimates of inbreeding may or may not bolster the significant trend indicating an effect of inbreeding on infection status. Inbreeding coefficients did not, however, explain variation among individuals in the intensity of infection, although increasing the sample size and accuracy of the inbreeding coefficients may change this conclusion. While it remains possible that our rough parameter estimates inhibit this level precise correlation, a single Plasmodium oocyst can
be sufficient for successful transmission of the parasite. Thus, an increased odds of
getting infected, regardless of how intense the infection becomes, could still have serious
epidemiological consequences. More work is needed to determine the ecological and
population dynamics leading to inbreeding in GOUNDRY, but it is possible that
anthropogenic interventions such as intense insecticide and bed-net eradication
campaigns, could in principle lead to increased inbreeding in other populations as well.
Such inbreeding could be especially problematic if it causes, as our results suggest,
increased efficiency in parasite transmission among the remaining small pockets of
mosquitoes that escape eradication. If this is the case, the combination between the
potential side effects of intense eradication efforts and ecological specialization of
subgroups across time and environmental space may make complete interruption of local
parasite transmission difficult.

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

**References:**


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

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

This article is protected by copyright. All rights reserved
Author Contributions
JEC wrote software and analyzed data; KDV, MMR, WMG, AG, NS contributed new reagents and analytical tools; JEC, BPL, KDV, MMR, and RN designed research. JEC wrote the manuscript with contributions from the rest of the authors.

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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Optimized Value</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Lower</td>
</tr>
<tr>
<td>$\theta (4N_A\mu L)^a$</td>
<td>180,914</td>
<td></td>
</tr>
<tr>
<td>$N_A^b$</td>
<td>126,252</td>
<td>88,916</td>
</tr>
<tr>
<td>Split Times</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t_1$</td>
<td>259,220</td>
<td>114,087</td>
</tr>
<tr>
<td>$t_2$</td>
<td>754,204</td>
<td>741,946</td>
</tr>
<tr>
<td>$t_3$</td>
<td>99,235</td>
<td>93,113</td>
</tr>
<tr>
<td>$t_{TOT}$</td>
<td>1,112,660</td>
<td>967,181</td>
</tr>
<tr>
<td>Population sizes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_{Ac1}/N_A$</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>$N_{Ac2}/N_A$</td>
<td>5.74</td>
<td>4.58</td>
</tr>
<tr>
<td>$N_{Ac3}/N_A$</td>
<td>12.34</td>
<td>9.54</td>
</tr>
<tr>
<td>$N_{G1}/N_A$</td>
<td>0.01</td>
<td>0.00*</td>
</tr>
<tr>
<td>$N_{G2}/N_A$</td>
<td>0.19</td>
<td>0.11</td>
</tr>
<tr>
<td>$N_{G3}/N_A$</td>
<td>0.78</td>
<td>0.62</td>
</tr>
<tr>
<td>Migration rates</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4Nm)$^c$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G1 into Ac1</td>
<td>0.010</td>
<td>0.0047</td>
</tr>
</tbody>
</table>
Table 2: Association between inbreeding coefficients and *Plasmodium* infection prevalence.

<table>
<thead>
<tr>
<th>Cutoff</th>
<th>Inbreeding Level</th>
<th>(X^2) value</th>
<th>(P) value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>0.0260</td>
<td>Infected 77</td>
<td>Infected 93</td>
<td>3.4870</td>
</tr>
<tr>
<td></td>
<td>Not infected 60</td>
<td>Not infected 44</td>
<td></td>
</tr>
<tr>
<td>0.0292</td>
<td>Infected 81</td>
<td>Infected 89</td>
<td>2.2200</td>
</tr>
<tr>
<td></td>
<td>Not infected 60</td>
<td>Not infected 44</td>
<td></td>
</tr>
</tbody>
</table>

- **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).
Figure Legends

Figure 1: Average genetic relationships among species and subgroups in *Anopheles gambiae* species complex. Unrooted neighbor-joining tree calculated with the *ape* package in R and drawn with Geneious software. Branches indicate genetic distance ($D_{xy}$) calculated using intergenic sites (see Methods) with scale bar for reference. Bootstrap support percentages are indicated on all internal nodes. Branch lengths and 95% CIs indicated for branches leading to *A. merus* and *A. arabiensis*.

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

Figure 3: Chromosomal distributions of nucleotide diversity ($\pi$) at inter-genic sites (LOESS-smoothed with span of 1% using 10 kb non-overlapping windows). Low complexity and heterochromatic regions were excluded. The strong reduction of diversity on the X chromosome in GOUNDRY (Mb 8.47 – 10.1) corresponds to putative chromosomal inversion $X_h$. 
Figure 4: Modeling expected values of Relative Node Depth ($D_{GA}/D_{GAg}$). A) Expected values of RND when ancestral population sizes are assumed to be equal. Colors indicate the expectations under different relative split times. B) Expected values with $t_{GAg}$ split time fixed to 1.1 (top) times the split time between GOUNDROY and $A. coluzzii$ ($t_{GA}$) or 1.5 times (bottom). Colors indicated relative effective sizes of ancestral populations. Values are plotted as a function of the GOUNDROY-$A. coluzzii$ effective size (x-axis). Grey bar indicates 95% confidence interval demographic estimate for GOUNDROY-$A. coluzzii$ ancestral size (see Methods).

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

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

A. gambiae

A. coluzzii

A. merus

GOUNDRY

0.0050

0.0240 (+/- 7.61x10^-4)

0.0066 (+/- 7.41x10^-4)
A

Size of GOUNDRY- \textit{A. coluzzii} ancestor

\begin{align*}
\frac{D_{GAc}}{D_{GAg}} & = 0.01 \\
\frac{D_{GAc}}{D_{GAg}} & = 0.05 \\
\frac{D_{GAc}}{D_{GAg}} & = 0.1 \\
\frac{D_{GAc}}{D_{GAg}} & = 0.2 \\
\frac{D_{GAc}}{D_{GAg}} & = 0.5 \\
\frac{D_{GAc}}{D_{GAg}} & = 1.0
\end{align*}

\begin{align*}
t_{GAc} & = 1.01 \times t_{GAc} \\
t_{GAc} & = 1.1 \times t_{GAc} \\
t_{GAc} & = 1.2 \times t_{GAc} \\
t_{GAc} & = 1.4 \times t_{GAc} \\
t_{GAc} & = 1.7 \times t_{GAc} \\
t_{GAc} & = 2.0 \times t_{GAc}
\end{align*}

Size of GOUNDRY- \textit{A. coluzzii} ancestor

\begin{align*}
t_{GS} & = 1.5 \times t_{GAc} \\
t_{GAg} & = 1.1 \times t_{GAc} \\
t_{GAg} & = 1.2 \times t_{GAc} \\
t_{GAg} & = 1.4 \times t_{GAc} \\
t_{GAg} & = 1.7 \times t_{GAc} \\
t_{GAg} & = 2.0 \times t_{GAc}
\end{align*}