Published in final edited form as:
Science. 2010 October 22; 330(6003): 512–514. doi:10.1126/science.1195755.

Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences

M.K.N. Lawniczak[1,*], **S. Emrich**[2,*], **A. K. Holloway**[3], **A.P. Regier**[2], **M. Olson**[2], **B. White**[4], **S. Redmond**[1], **L. Fulton**[5], **E. Appelbaum**[5], **J. Godfrey**[5], **C. Farmer**[5], **A. Chinwalla**[5], **S-P. Yang**[5], **P. Minx**[5], **J. Nelson**[5], **K. Kyung**[5], **B.P. Walenz**[6], **E. Garcia-Hernandez**[6], **M. Aguiar**[6], **L.D. Viswanathan**[6], **Y-H. Rogers**[6], **R.L. Strausberg**[6], **C.A. Saski**[7], **D. Lawson**[8], **F.H. Collins**[4], **F.C. Kafatos**[1], **G.K. Christophides**[1], **S.W. Clifton**[5], **E.F. Kirkness**[6], and **N.J. Besansky**[4]

[1]Division of Cell and Molecular Biology, Imperial College London, South Kensington Campus, London SW4 2AZ UK

[2]Department of Computer Science and Engineering and Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN 46556 USA

[3]J. David Gladstone Institutes, San Francisco, CA 94158 USA

[4]Eck Institute for Global Health, Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556 USA

[5]The Genome Center at Washington University, St Louis, MO 63108 USA

[6]The J. Craig Venter Institute, Rockville, MD 20850 USA

[7]Clemson University Genomics Institute, Clemson University, Clemson, SC 29634 USA

[8]The European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD UK

Abstract

The Afrotropical mosquito *Anopheles gambiae* sensu stricto (*A. gambiae*), a major vector of malaria, is currently undergoing speciation into the M and S molecular forms. These forms have diverged in larval ecology and reproductive behavior through unknown genetic mechanisms, despite considerable levels of hybridization. Previous genome-wide scans using gene-based microarrays uncovered divergence between M and S that was largely confined to gene-poor pericentromeric regions, prompting a speciation-with-ongoing-gene-flow model that implicated only ~3% of the genome near centromeres in the speciation process. Here, based on the complete M and S genome sequences, we report widespread and heterogeneous genomic divergence inconsistent with appreciable levels of inter-form gene flow, suggesting a more advanced speciation process and greater challenges to identify genes critical to initiating that process.

Population-based genome sequences provide a rich foundation for “reverse ecology” (1). By analogy to reverse genetics, reverse ecology uses population genomic data to infer the

[*]These authors contributed equally.

Supporting Online Material
www.sciencemag.org
Materials and Methods
Figs. S1 to S8
Tables S1 to S5
References
Database S1

genetic basis of adaptive phenotypes, even if the relevant phenotypes are not yet known. This approach can be especially powerful for gaining insight into the genetic basis of ecological speciation, a process whereby barriers to gene flow evolve between populations as by-products of strong, ecologically-based, divergent selection (2). Here, we apply reverse ecology to study incipient speciation within *Anopheles gambiae*, one of the most efficient vectors of human malaria. The complex population structure of *A. gambiae*, exemplified by the emergence of the M and S molecular forms (3), poses significant challenges for malaria epidemiology and control, as underlying differences in behavior and physiology may affect disease transmission and compromise antivector measures. Genome-wide analysis of M and S can provide insight into the mechanisms promoting their divergence, and open new avenues for malaria vector control.

Morphologically, M and S are indistinguishable at all life-stages, and can only be recognized by fixed differences in the ribosomal DNA genes (3). Geographically and microspatially, both forms co-occur across much of West and Central Africa (4), and in areas where they are sympatric, adults may be found resting in the same houses and even flying in the same mating swarms (5, 6). Assortative mating limits gene flow between forms (5, 6), but appreciable hybridization still occurs (4, 7–10) without intrinsic hybrid inviability or sterility (11). Although the aquatic larvae of both forms also may be collected from the same breeding site, S form larvae are associated with ephemeral and largely predator-free pools of rain water, while M form larvae exploit longer-lived but predator-rich anthropogenic habitats (12). Thus, persistence of M and S despite hybridization may be driven by ecologically-dependent fitness trade-offs in the alternative larval habitats to which they are adapting (12).

Under a model of speciation in the presence of gene flow, genomic divergence between incipient species should be limited to regions containing the genes which confer differential adaptations or are involved in reproductive isolation (13). Consistent with this expectation, scans of genomic divergence between M and S at the resolution of gene-based microarrays revealed elevated divergence near the centromeres of all three independently assorting chromosomes, and almost nowhere else (14, 15). Given the assumption of appreciable genetic exchange through hybridization, this pattern suggested that the genes causing ecological and behavioral isolation were located in the centromeric “speciation islands” (14). The small number, size and gene content of these islands implied that speciation of M and S was very recent and involved only a few genes in a few isolated chromosomal regions-- an influential model for speciation with gene flow (13, 16, 17). The complete genome sequences of *A. gambiae* M and S forms reported here provide much higher resolution than previous studies to address how genomes diverge during speciation.

Sequences were determined from colonies established in 2005 from Mali, where the rate of natural M-S hybridization (~1%) is theoretically high enough for introgression to homogenize neutral variation between genomes (18) in the absence of countervailing selection. Both colonies were homosequential and homozygous with respect to all known chromosomal inversions with the exception of 2L*a* and 2R*c* (19). Independent draft genome assemblies were generated based on ~2.7 million Sanger traces (19). Both assemblies were performed independently of the reference *A. gambiae* PEST genome (20), which is a chimera of the M and S forms. Genome assembly metrics were similar between M and S (table S1) (19). Lower coverage (~6× in M/S vs. ~10× in PEST) contributed to assembly gaps, motivating alignment of the M and S scaffolds to the PEST assembly for transfer of genomic coordinates and gene annotations (www.vectorbase.org; table S2) (19). Importantly, we confirmed the major trends of M-S divergence by direct alignment of M and S scaffolds to each other (fig. S1) (19).

More than two million single nucleotide polymorphisms (SNPs) per form, and more than 150,000 fixed differences between forms were identified in the sequence data using strict coverage and quality restrictions (table S3) (19). The chromosomes show significantly different patterns of divergence, with chromosome 2 showing proportionally more fixed differences than chromosome 3, and chromosome X showing the highest proportion of fixed differences [further explored in (19)] (table S3). The spatial distribution of polymorphism and divergence along chromosome arms also was investigated, using sliding window analyses to minimize noise from individual site-based divergence estimates (Fig. 1, figs. S2 to S6) (19, 21). Significant outlier divergence values falling in the top 1% of the empirical distribution (fig. S7) (19, 22) are spread heterogeneously across the entire genome, not confined mostly to pericentromeric regions as observed in gene-based microarray studies (14, 15, 19).

The 436 genes overlapping with the top percentile of diverged 1-kb windows were tested for functional enrichment based on their gene ontology terms (database S1) (19). The 1-kb window size, smaller than the average gene size (~5.7 kb including introns), mitigates the potentially confounding effect of physical clustering of functionally related members of gene families in *A. gambiae*. Functions related to G-protein-coupled receptor (GPCR) signaling, particularly neurohormone signaling, are significantly over-represented in genomic regions of highest divergence (table S4). The neurohormone subfamily of GPCRs bind biogenic amines, neuropeptides, and protein hormone ligands, which in insects control development, feeding, reproduction and complex behaviors (*e.g.*, locomotion) that potentially bear on niche adaptation and mate recognition.

We also examined genes for evidence of divergence. Genes showing evidence of directional selection within forms, or amino acid fixations between forms (database S1, figs. S2 to S6) (19), occur throughout the genome, suggesting that differential adaptation of M and S to their specific ecologies could involve an appreciable number of genes outside of pericentromeric regions. Some genomic regions appear to have experienced strong and recent selective sweeps, as illustrated by elevated divergence coupled with reductions in shared and private polymorphism (Fig. 1, figs. S2 to 6). The most notable such region is on 2L (near Mb 25) centered on the *resistance to dieldrin* (*Rdl*) gene, which has been previously associated with insecticide resistance in *A. gambiae* and other insects (Fig. 1, fig. S4) (23). In fact, M and S appear to carry different “resistant” substitutions (Ala296Ser in M, Ala296Gly in S) at *Rdl* (23) suggesting independent selective sweeps. Another notable region occurs on 3R near position ~40 Mb (Fig. 1, fig. S4) and contains seven odorant receptors (ORs) whose closest match to the proteome of the fruit fly *Drosophila melanogaster* is OR67d. The single copy of this gene in *Drosophila* serves as the pheromone receptor for cis-vaccenyl acetate which mediates both social aggregation and female sexual receptivity (24), tempting speculation that these genes might play similar yet species-specific roles in M and S.

Importantly, the pattern of genome-wide divergence inferred from colony-based genomic sequences is present in natural populations of M and S from the same region of Mali, based on a newly developed SNP genotyping array whose design included a subset of 400,000 SNPs derived from the M and S genome sequences (25). Indeed, visual and statistical concordance of patterns of divergence (fig. S8; table S5) between the two datasets indicates that, at least in Mali, the widespread genomic divergence observed between M and S is not an artifact of laboratory culture (19). Future genome-wide studies spanning different geographic locations will be necessary to provide insight into whether and how this pattern varies spatially. Further population genomic sequencing by current short-read technologies will benefit from read-mapping to the independent M and S genome assemblies reported here.

The widely adopted model of ongoing speciation-with-gene-flow for M and S (14) posits that frequent hybridization leads to M-S genome homogenization in all except a few small regions near centromeres (“speciation islands”), which are barred from introgression because they contribute to differential fitness (*i.e.*, ecological and reproductive isolation). Detection of much more widespread genomic divergence based on genotyping (25) and whole genome sequencing supports a very different model, in which realized gene flow between forms is currently much lower, and the process of speciation more advanced, than previously recognized, with the corollary that identification of genetic changes instrumental and not merely incidental to their ecological and behavioral divergence will be more difficult than initially hoped. However, powerful resources in the form of independently assembled M and S genomes and a SNP genotyping array (25) are now available for detecting morphologically cryptic vector subdivisions, probing their molecular basis, and ultimately developing innovative malaria interventions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank W. M. Gelbart, an early advocate of this project, and J.L. Feder, P. Nosil and M.W. Hahn for critical review. P. Howell of MR4 provided mosquitoes. Funding for genome sequencing of M (U54-HG00379) and S (U54-HG03068) was provided by NHGRI. N.J.B. was supported by NIH (RO1 AI63508 and AI076584). M.K.N.L. was supported by a BBSRC research grant BB/E002641/1 to G.K.C. Sequence data are deposited with GenBank (accessions ABKP00000000 and ABKQ00000000).

References and Notes

1. Li YF, Costello JC, Holloway AK, Hahn MW. Evolution. 2008; 62:2984. [PubMed: 18752601]
2. Rundle HD, Nosil P. Ecology Letters. 2005; 8:336.
3. della Torre A, et al. Insect Mol Biol. 2001; 10:9. [PubMed: 11240632]
4. Della Torre A, Tu Z, Petrarca V. Insect Biochem Mol Biol. 2005; 35:755. [PubMed: 15894192]
5. Diabate A, et al. Proc Biol Sci. 2009; 276:4215. [PubMed: 19734189]
6. Diabate A, et al. J Med Entomol. 2006; 43:480. [PubMed: 16739404]
7. Tripet F, et al. Mol Ecol. 2001; 10:1725. [PubMed: 11472539]
8. Caputo B, et al. Malar J. 2008; 7:182. [PubMed: 18803885]
9. Oliveira E, et al. J Med Entomol. 2008; 45:1057. [PubMed: 19058629]
10. Costantini C, et al. BMC Ecology. 2009; 9:16. [PubMed: 19460144]
11. Diabate A, Dabire RK, Millogo N, Lehmann T. J Med Entomol. 2007; 44:60. [PubMed: 17294921]
12. Lehmann T, Diabate A. Infect Genet Evol. 2008; 8:737. [PubMed: 18640289]
13. Nosil P, Funk DJ, Ortiz-Barrientos D. Mol Ecol. 2009; 18:375. [PubMed: 19143936]
14. Turner TL, Hahn MW, Nuzhdin SV. PLoS Biol. 2005; 3:e285. [PubMed: 16076241]
15. White BJ, Cheng C, Simard F, Costantini C, Besansky NJ. Mol Ecol. 2010; 19:925. [PubMed: 20149091]
16. Carneiro M, Ferrand N, Nachman MW. Genetics. 2009; 181:593. [PubMed: 19015539]
17. Feder JL, Nosil P. Evolution. 2009
18. Slatkin M. Science. 1987; 236:787. [PubMed: 3576198]
19. Materials and methods are available as supporting material on *Science* Online.
20. Holt RA, et al. Science. 2002; 298:129. [PubMed: 12364791]
21. Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG. Genome Res. 2005; 15:1468. [PubMed: 16251456]
22. Akey JM, et al. Proc Natl Acad Sci U S A. 2010; 107:1160. [PubMed: 20080661]

23. Du W, et al. Insect Mol Biol. 2005; 14:179. [PubMed: 15796751]

24. Mehren JE. Curr Biol. 2007; 17:R240. [PubMed: 17407751]

25. Neafsey DE, et al. Science. 2010 **in review**.

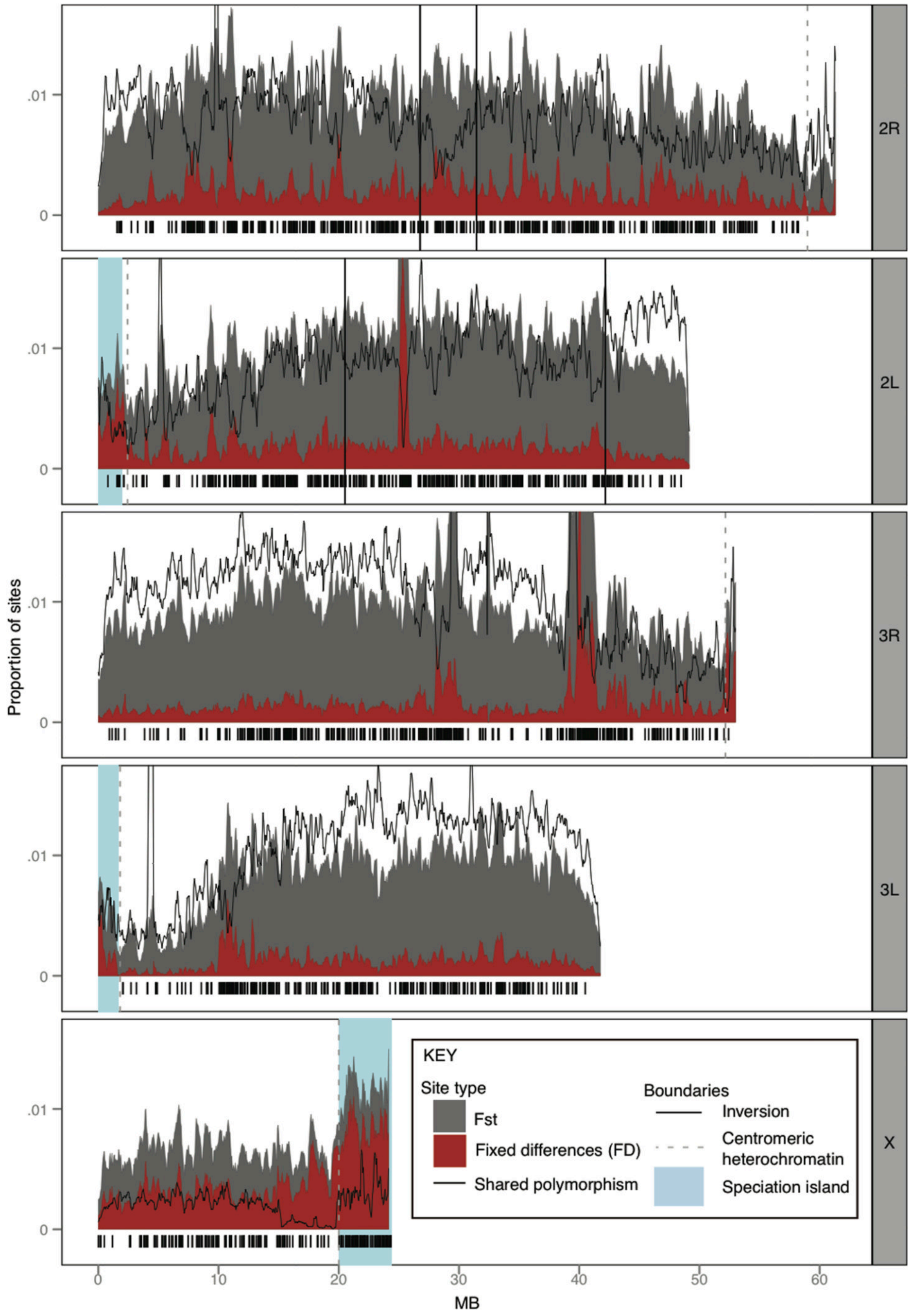

Fig. 1.

Sliding window analysis of polymorphism and divergence in M and S based on 250kb windows with 50kb steps. Approximate boundaries of chromosomal rearrangements differing between M and S colonies (2Rc and 2La) are indicated by solid black vertical lines. “Speciation islands” *sensu* (14, 15) are shaded in blue for reference. “F_{ST}” refers to the mean per-site estimate (19). Underneath the X axis, vertical black bars mark the approximate location of 1kb windows whose divergence values fall in the top percentile of the distribution across autosomes (or the X chromosome, calculated separately). For both 250kb and 1kb windows, only windows meeting coverage and quality restrictions (19) are plotted.