

**GENOME-WIDE PATTERNS OF POLYMORPHISM IN AN INBRED LINE OF  
THE AFRICAN MALARIA MOSQUITO ANOPHELES GAMBIAE**

David A. Turissini<sup>1</sup>, Stephanie Gamez<sup>1</sup>, Bradley J. White<sup>1,2\*</sup>

<sup>1</sup>Department of Entomology, University of California, Riverside, CA, USA

<sup>2</sup>Center for Disease Vector Research, University of California, Riverside, CA, USA

\*Author for Correspondence: [bwhite@ucr.edu](mailto:bwhite@ucr.edu)

© The Author(s) 2014. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

## ABSTRACT

*Anopheles gambiae* is a major mosquito vector of malaria in Africa. While increased use of insecticide-based vector control tools have decreased malaria transmission, elimination is likely to require novel genetic control strategies. It can be argued that the absence of an *An. gambiae* inbred line has slowed progress towards genetic vector control. In order to empower genetic studies and enable precise and reproducible experimentation, we set out to create an inbred line of this species. We found that amenability to inbreeding varied between populations of *An. gambiae*. After full-sib inbreeding for 10 generations, we genotyped 112 individuals – 56 saved prior to inbreeding and 56 collected after inbreeding – at a genome wide panel of SNPs. While inbreeding dramatically reduced diversity across much of the genome, we discovered numerous, discrete genomic blocks that maintained high heterozygosity. For one large genomic region, we were able to definitively show that high diversity is due to the persistent polymorphism of a chromosomal inversion. Inbred lines in other eukaryotes often exhibit a qualitatively similar retention of polymorphism when typed at a small number of markers. Our whole genome SNP data provides the first strong, empirical evidence supporting associative overdominance as the mechanism maintaining higher than expected diversity in inbred lines. While creation of *An. gambiae* lines devoid of nearly all polymorphism may not be feasible, our results provide critical insights into how more fully isogenic lines can be created.

## INTRODUCTION

Model species are a powerful tool for investigating novel biological processes as they are easily maintained in a laboratory environment, allowing for precise genetic and experimental manipulation (Miklos and Rubin 1996; Rubin, et al. 2000). However, to be of practical benefit, insights obtained from model species need to be confirmed in non-model systems, an often difficult, time-consuming, and expensive task (Hunter 2008).

The African malaria mosquito *Anopheles gambiae* is a good example of an important non-model system with limited genetic resources (White, et al. 2011). Endemic to Sub-Saharan Africa, *An. gambiae* is a highly synanthropic species and one of the most efficient malaria vectors in the world (Collins and Paskewitz 1995; Coluzzi, et al. 2002). Since the publication of its genome in 2002 (Holt, et al. 2002), much effort has been devoted to understanding *An. gambiae* molecular biology with the end goal of devising novel genetic control strategies to disrupt disease transmission (Beaty, et al. 2009; Blandin, et al. 2009; Windbichler, et al. 2011). However, both genetic work within *An. gambiae* and the translation of discoveries from other invertebrates are impeded by the absence of an inbred line.

Due to the relative difficulty of rearing and maintaining *An. gambiae*, researchers use eggs from multiple females to establish colonies (Mpofu, et al. 1993; Benedict 1997; della Torre, et al. 1997; Benedict, et al. 2009). While some genetic variation is lost through bottlenecks associated with colonization (Norris, et al. 2001), even long-established *An. gambiae* colonies retain high diversity

(Morlais, et al. 2004; Lawniczak, et al. 2010; Neafsey, et al. 2013). Variation within colonies of *An. gambiae* hinders progress of genomic studies in two key ways. First, despite being the second arthropod to have its genome sequenced, the *An. gambiae* reference genomes remain poorly assembled (Holt, et al. 2002; Sharakhova, et al. 2007; Lawniczak, et al. 2010) due to high heterozygosity in the source DNA. Incomplete and improper reference genome assembly can cause spurious, deficient, or erroneous results for virtually any genomics-based study due to reliance on the reference genome for read mapping (Li and Durbin 2009), primer design (Rozen and Skaletsky 2000), gene ontology analysis (Consortium 2013), etc. In an attempt to minimize heterozygosity, more recent *Anopheles* genome projects obtained source DNA from the progeny of a single female (more extensive inbreeding was deemed infeasible). Unfortunately, genetic variation was still too high to produce assemblies that rival those of *Drosophila* despite similar genome sizes (Neafsey, et al. 2013). Second, inbred lines would improve the performance of quantitative genetic studies, which have the potential to identify genetic polymorphisms that control key phenotypes of medical or ecological importance (e.g. Riehle, et al. 2006; Blandin, et al. 2009). Such studies are empowered by high variation between, but not within, experimental lines (Mackay, et al. 2009).

Inbred lines in other Dipteran insects such as *Aedes* mosquitoes and *Drosophila* flies are common, relatively easy to produce, and highly useful in genetics and genomics studies (Nene, et al. 2007; Mackay, et al. 2012). However, no inbred lines for *An. gambiae* have ever been reported presumably

due to the actual/perceived difficulty in creating such lines. To fill this gap, we set out to create an inbred line of *An. gambiae* that would enable genetic studies and, ultimately, improved vector control. Here, we report the first inbred line of *An. gambiae* and quantify the extent to which polymorphism was reduced through ten generations of full-sib inbreeding. Interestingly, genome-wide analyses identify multiple, large genomic regions that maintained unexpectedly high levels of heterozygosity. Similar qualitative results have been reported after extensive inbreeding in diverse eukaryotes including insects (Munstermann 1994; Rumball, et al. 1994), mammals (Eriksson, et al. 1976; Connor and Belucci 1979; Mina, et al. 1991), mollusks (McGoldrick and Hedgecock 1997), and plants (Strauss 1986). In light of our data, we discuss the likely mechanism underlying the retention of diversity in inbred lines and plausible strategies for creating more highly inbred lines in *An. gambiae*.

## RESULTS

### *Inbreeding of Different Colonies*

We attempted to inbreed five different colonies of *An. gambiae* via full-sibling mating (see Methods) with variable results (Table 1). *Anopheles gambiae* is composed of two ecologically and genetically divergent forms termed M and S (Lehmann and Diabate 2008; Costantini, et al. 2009; Simard, et al. 2009;

Lawniczak, et al. 2010) . It has been proposed that these two forms should be elevated to species status (Coetzee, et al. 2013), although gene flow between them still occurs at non-negligible rates (Reidenbach, et al. 2012; Weetman, et al. 2012; Lee, et al. 2013). Interestingly, we found that the S Form of *An. gambiae* was unamenable to inbreeding. We were unable to get even a single female from PIMPERENA, a Malian S Form colony, to lay eggs when isolated, despite the fact that females from this colony lay hundreds to thousands of eggs when kept en masse in a single cage. While we did succeed in obtaining eggs from a single female of the S Form NDKO colony, none of her female progeny laid eggs resulting in the death of the line after just one generation.

In contrast, we found that the M Form of *An. gambiae* was relatively amenable to inbreeding. Of the three colonies tested, all were successfully inbred for at least 5 generations and two of the three were inbred for 10 generations or more. We had the greatest success with the NGS colony, which is still alive and currently on its 20<sup>th</sup> generation of full-sib mating. Our variable success with inbreeding is consistent with previous ecological analyses of niche partitioning between the two forms. Studies in Cameroon and Burkina Faso suggest that the M Form exploits marginal habitats (irrigated rice fields, polluted urban environments, and coastal areas) while the S Form dominates in canonical *An. gambiae* habitats (ephemeral puddles and pools)(Costantini, et al. 2009; Simard, et al. 2009; Gimonneau, et al. 2012; Kamdem, et al. 2012). Indeed, in our experiment the M Form was more robust and thrived in the relatively unsuitable (i.e. marginal) laboratory environment. Alternatively, the differential inbreeding

success between M and S may not be related to ecology; instead, it could be due to fewer recessive deleterious alleles segregating in populations of M Form relative to S Form. While intriguing, the differential inbreeding success of M and S should not be over-interpreted as the number of colonies tested was relatively low and variance could potentially be due to chance alone. Larger sample sizes of both forms are needed to confirm the hypothesis.

### *The Overall Impact of Inbreeding on Polymorphism*

To assess the decrease in polymorphism achieved by inbreeding the NGS colony, we genotyped 112 individual female mosquitoes – 56 saved prior to the start of inbreeding and 56 collected after 10 generations of inbreeding – at a genome-wide panel of single nucleotide polymorphisms (SNPs). To genotype individual mosquitoes in parallel we employed a double-digest restriction associated DNA (ddRAD) sequencing methodology (Peterson, et al. 2012). Briefly, DNA from individual mosquitoes was extracted, digested with two restriction enzymes, barcoded, size selected, and PCR amplified. The goal was to simultaneously sequence a representative and reproducible fraction of the genome from individual mosquitoes on the Illumina platform.

For the following analyses we only used SNPs genotyped in at least 16 individual mosquitoes from both the NGS and NGS INBRED groups leaving us with a total of 159,314 SNPs on the autosomes and 11,099 SNPs on the X. We observed a relatively even distribution of SNPs across the euchromatin of the autosomes, with a dip in SNP density at the heterochromatic centromeres

(Sharakhova, et al. 2010). Overall, the X chromosome exhibited a lower SNP frequency than the autosomes. The observed distribution of polymorphic sites is consistent with studies of nucleotide variation in natural populations (Lawniczak, et al. 2010; White, et al. 2010; Clarkson, et al. 2014) suggesting that the ddRAD approach worked well in generating an unbiased, genome-wide panel of markers (Figure 1).

Overall, 90,049 or 57% of the total autosomal SNPs went from polymorphic in NGS to monomorphic in NGS INBRED. Polymorphism was reduced to a greater extent on the autosomes than the X, where only 4,394 or 40% of the total SNPs went from polymorphic in NGS to monomorphic in NGS INBRED (Figures 2 and 3). Specifically, average values of  $\pi$  per SNP on the autosomes declined from 0.30 in NGS to 0.12 in NGS INBRED. On the X chromosome,  $\pi$  per SNP only declined from 0.25 in NGS to 0.16 in NGS INBRED. Figure 3 plots  $\pi$  for individual SNPs in NGS versus NGS INBRED. As expected and detailed above, most SNPs went from polymorphic in NGS to monomorphic in NGS INBRED. However, a cluster of SNPs in the upper right hand corner remained persistently polymorphic in both the NGS and NGS INBRED lines. SNPs in this category may exhibit strong heterosis, recessive lethality, or be linked with such sites, explaining their persistence at a frequency of ~50%. Alternatively, a small fraction of apparently heterozygous SNPs may actually be due to mapping errors where reads from two different genomic regions map to a single locus due to incomplete or improper reference genome assembly. Surprisingly, 7,155 SNPs appear monomorphic in NGS, but



polymorphic in NGS INBRED. These SNPs were likely at a low frequency in NGS, were not present in our sample of 56 individuals, and increased in frequency during the inbreeding process.

### *Spatial Heterogeneity in Diversity Reduction*

To compare polymorphism levels across the genome between NGS and NGS INBRED, we performed window-based analyses of  $\pi$  (Figure 4). While diversity is markedly reduced in NGS INBRED relative to NGS in many regions, polymorphism levels nearly equal to those observed prior to inbreeding exist over discrete, large genomic intervals. For example, a region on 2L extending from 16 Megabases (Mb) proximal of the centromere all the way to the telomere shows only slightly reduced diversity in NGS INBRED. Another ~25 Mb region around the centromere of chromosome 3 shows little, if any, reduction in polymorphism. Inbreeding did not reduce  $\pi$  substantially in other smaller regions, such as near the telomere of 2R. The X chromosome shows a more spatially homogenous, albeit less dramatic, drop in diversity than the autosomes.

With an initial frequency of 50%, only 1/1024 SNPs should remain polymorphic after ten generations of inbreeding. However, nearly half of all SNPs remained polymorphic in NGS INBRED. The persistently polymorphic sites are mostly found in large linkage blocks. While linkage disequilibrium decays rapidly in natural populations of *An. gambiae* (Harris, et al. 2010; Cheng, et al. 2012; Marsden, et al. 2014), within colonies longer haplotypes may exist which could plausibly explain the observed tracts of heterozygosity.

### *2La Inversion Analysis*

The large block of non-reduced polymorphism on 2L overlaps with the position of a known polymorphic chromosomal inversion in *An. gambiae*. The *2La* inversion extends from genomic position ~20,000,000 to 42,000,000 on 2L and the alternative arrangements are referred to as *2La* (ancestral) and *2L<sup>a</sup>* (derived) (Sharakhov, et al. 2006). The inversion segregates at a low frequency (~5% *2La*) in the Cameroonian region where the founder mosquitoes of the NGS colony were collected (Wondji, et al. 2005; Simard, et al. 2009). In natural populations, the inversion is always found at Hardy-Weinberg equilibrium (Toure, et al. 1998; Powell, et al. 1999; Costantini, et al. 2009; Simard, et al. 2009). To determine if polymorphism of the *2La* inversion was the cause of the large diversity block on 2L, we molecularly determined inversion status for the 56 NGS and 56 NGS INB that we sequenced. Prior to inbreeding the *2La* arrangement was segregating at a frequency of 45% (13 *2L<sup>a</sup>/+<sup>a</sup>*; 36 *2La/+<sup>a</sup>*; 7 *2La/a*) in the colony. After inbreeding the *2La* arrangement had increased in frequency to 68% (3 *2L<sup>a</sup>/+<sup>a</sup>*; 30 *2La/+<sup>a</sup>*; 23 *2La/a*), but not to fixation. We have previously shown high divergence between alternative arrangements of *2La* with *F<sub>ST</sub>* values routinely exceeding 0.5 (White, et al. 2007a; Cheng, et al. 2012). Thus, if persistently polymorphic in a population, the regions captured by the inversion should display high levels of polymorphism.

To determine the effect of the segregating inversion on diversity we performed window-based analyses of  $\pi$  for the three different karyotypes present

in NGS and NGS INBRED (Figure 5). As expected diversity within homokaryotypes ( $2L^{+a}$  or  $2La$ ) is near zero in NGS INBRED, but heterokaryotypes have high diversity. Thus, we conclude that persistence of the  $2La$  inversion explains the large block of high diversity on the left arm of chromosomes 2. Interestingly, when we perform window-based analyses of  $\pi$  on NGS individuals  $2L^{+a}$  homokaryotypes show markedly higher diversity than  $2La$  homokaryotypes in the rearranged region (Figure 6). This pattern likely results from the colony being founded from eggs laid by multiple female mosquitoes from a population with a high frequency of the  $2L^{+a}$  arrangement. In contrast, the other blocks of extended heterozygosity do not overlap with known inversions. However, the second largest diversity block straddles the centromere of chromosome 3, a region of presumed low recombination that may functionally segregate as a single locus. Similar to the case for the  $2La$  inversion, two divergent 'centromere haplotypes' may be segregating in NGS and neither was lost through inbreeding.

## DISCUSSION

We successfully subjected a colony of *An. gambiae* to full-sib mating for 10 generations, dramatically reducing diversity across the genome. However, discrete blocks of the genome retained high diversity. Under neutrality, the vast majority of polymorphisms should have been eliminated during inbreeding: in the absence of linkage and selection, a SNP with an initial frequency of 50% should become monomorphic >99.9% of the time after ten generations of full-sib mating.

### *Associative Overdominance*

Previous studies of allozyme markers in inbred lines of a variety of organisms including *Aedes* mosquitoes (Matthews and Craig Jr 1987; Munstermann 1994), *Culex* mosquitoes (Knop, et al. 1987), *Drosophila* (Sing, et al. 1973; Frankham, et al. 1993; Rumball, et al. 1994), chicken (Mina, et al. 1991), mouse (Connor and Belucci 1979), oyster (McGoldrick and Hedgecock 1997), pine tree (Strauss 1986), rat (Eriksson, et al. 1976), and *Tribolium* beetles (Yeh 1972) reported patterns similar to ours. In these studies, heterozygosity at a portion of protein markers persisted at much higher rates than expected during inbreeding.

Associative overdominance has often been invoked to explain the discrepancy between the expected and actual efficiency of inbreeding (Rumball, et al. 1994; Wang and Hill 1999). During inbreeding individuals will necessarily differ in their degree of homozygosity across the genome (Weir and Cockerham 1973). At some loci, natural selection will strongly favor heterozygotes due the presence of either recessive deleterious mutations or overdominant polymorphisms (Ohta 1971; Ohta and Cockerham 1974; Strobeck 1979; Charlesworth 1991; Wang and Hill 1999). Loci where natural selection favors heterozygotes will thus remain polymorphic even in the face of inbreeding. Additionally, neutral loci linked to the polymorphism(s) under balancing selection will also exhibit elevated polymorphism (Santiago and Caballero 1995; Nordborg, et al. 1996). Tighter linkage will cause longer persistence of polymorphism due to the rarity of recombination between the neutral loci and overdominant polymorphism.

Computational simulations of full-sib inbreeding show that observed diversity at allozyme loci in *D. melanogaster* inbred lines is consistent with the expected effects of associative overdominance (Rumball, et al. 1994; Wang and Hill 1999). However, because these previous studies only employed a handful of markers definitive evidence for associative overdominance was lacking.

A dense set of genome-wide markers allows us to clearly implicate associative overdominance as the mechanism underlying the retention of diversity in the NGS INBRED line of *An. gambiae*. Associative overdominance theory makes two clear predictions regarding the spatial patterns of diversity across the genome of an inbred line. First, heterozygous sites should be clustered in linked, genomic intervals. Second, the size of the high-diversity blocks should negatively correlate with recombination rate (Ohta 1971; Ohta and Cockerham 1974; Wang and Hill 1999). Our analysis strongly supports both predictions. Genome-wide analyses revealed heterogeneity in diversity levels across the genome with discrete chromosomal intervals of high polymorphism. The two largest blocks of high diversity are located in regions of low recombination. One is ~25 Mb region roughly centered on the centromere of chromosome 3 – pericentromeric regions in *An. gambiae* have highly reduced recombination rates (Pombi, et al. 2006; Slotman, et al. 2006; Stump, et al. 2007). The second large region overlaps the 2La inversion. While recombination in inversion homokaryotypes is not reduced, gene flux between alternative arrangements is highly suppressed (Stump, et al. 2007). Consistent with associative overdominance, our analysis demonstrated that the presence of

alternative karyotypes is responsible for elevated polymorphism in NGS INBRED on chromosome 2L (Strobeck 1983).

However, our results do not meet all theoretical expectations of associative overdominance. Simulations also predict that heterogametic sex chromosomes should homogenize faster than the autosomes. Our results displayed the opposite pattern; the autosomes showed larger declines in diversity than the X chromosome. However, it should be noted that the predicted differences in homogenization rates are small and stochastic forces can cause results from a single study to deviate massively from the prediction (Wang and Hill 1999).

#### *Creating more homozygous lines*

Overall, our results suggest that producing isogenic or nearly isogenic lines of *An. gambiae* will be difficult, but certain steps can be taken to maximize the effectiveness of inbreeding. First, inbreeding for many generations will reduce the size of heterozygous linkage blocks by providing more opportunity for recombination to break the association between neutral and overdominant loci (Latter, et al. 1995). Identifying inversions polymorphisms and selecting for homokaryotypic lines prior to inbreeding will also aid recombination in breaking up linkage blocks. Finally, our results from the S-Form colonies PIMPERENA and NDKO suggest the possibility of social cues playing a role in oviposition. When caged en masse, females from these colonies will readily lay eggs. However, when isolated we were generally unable to induce females from either colony to

oviposit. While the literature does not provide any evidence of social cues in oviposition, it would be worthwhile to place lone gravid females into cages with males and sterilized (irradiated) females and determine if the oviposition rate increased. Overall, there does not appear to be any intrinsic barriers to successfully inbreeding *An. gambiae*. Although we lost 4 out of the 5 inbred lines we started, this rate is not qualitatively different from that observed in large-scale *Drosophila melanogaster* inbreeding experiments (Garcia, et al. 1994; Rumball, et al. 1994). While an 80% loss rate is quite tolerable in *Drosophila* due to the ease of maintenance, the large amount of labor involved in rearing *An. gambiae* makes this loss rate more burdensome. In summary, while NGS INBRED contains more polymorphism than would be ideal, its lowered diversity makes it an important tool for vector biologists. Further improvement to the line can be achieved through continued inbreeding and, if possible, direct selection for a homokaryotypic arrangement of the 2La inversion.

### *Sequencing Strategies for Evolutionary Genetics*

In our experiment, reduced-representation sequencing was a cost-effective alternative to whole genome sequencing (WGS); indeed, it is very unlikely that WGS data would have provided any additional insights. Furthermore, it is worth highlighting the benefit of using barcodes to sequence individual mosquitoes rather than pooled DNA ('pool-seq,' e.g. Zhu, et al. 2012). To this point, we could have sequenced two pools of individual mosquitoes: NGS and NGS INBRED. While we would have obtained similar insights into the levels and spatial

heterogeneity of polymorphism with pooled data, we would have been unable to retroactively determine the 2La karyotype of each individual. Evolutionary genomicists should weigh the costs and benefits of sequencing individuals versus performing pooled sequencing in the future as individual sequencing provides flexibility when addressing unexpected results.

## METHODS

### *Mosquito Inbreeding*

Mosquitoes were maintained in the University of California Riverside insectaries under standard conditions (White, et al. 2013). Prior to inbreeding the colony was maintained in our insectaries for six months with an approximate size of 600 individuals per generation with roughly equal number of males and females. Inbreeding of the NGS colony was accomplished via a full-sibling mating



scheme. To initiate the inbred line, NGS mosquitoes were allowed to emerge en masse and mate for seven days after which female mosquitoes were offered a blood meal. Polyandry is very rare in natural populations of *An. gambiae* (Tripet, et al. 2003), but can be marginally higher in laboratory colonies (Gomulski 1990). On the next day, five blood fed females were removed from the colony and placed into individual 250ml cups. To induce oviposition, each cup had ~50 milliliters of ddH<sub>2</sub>O and was lined with a strip of wet-strength filter paper. After two nights, the cups were checked and eggs were synchronously hatched. Larvae from each female were reared in 1 liter of ddH<sub>2</sub>O under standard conditions (White, et al. 2013). Only the isofemale line that produced the most pupae was retained. The same inbreeding process was reiterated each subsequent generation on the adult progeny of one female. Inbreeding of other colonies followed the same protocol, except for PIMPERENA where additional females (15 total) were isolated in an attempt to start the inbreeding process.

### *Library Preparation*

Our genotyping-by-sequencing (GBS) protocol is a slightly modified version of the ddRAD protocol designed by Peterson and colleagues (Peterson, et al. 2012). The goal is to simultaneously sequence a reduced, yet highly reproducible, portion of the genome from individual mosquitoes. First, mosquito DNA is extracted in 96-well plate format using the ZR-96 Quick-gDNA kit (Zymo Corporation, San Diego, CA). DNA is eluted in 35ul ddH<sub>2</sub>O, 5ul (~50 nanograms) of which is subsequently used for GBS library preparation. DNA from individual

mosquitoes is then digested at 37°C for three hours with two restriction enzymes, *MluC1* and *NlaIII* (New England Biolabs, Ipswich, MA). The digested DNA is purified using Ampure magnetic beads (Beckman Coulter; Brea, CA). For each individual, a barcoded adapter (1 out of 48) is ligated to the sticky end produced by *NlaIII*, while a universal adapter is ligated to the sticky end produced by *MluC1*. Adapters are ligated to DNA fragments with T4 DNA ligase (NEB) for 2 hours at 22°C, heat killed at 65°C, and allowed to cool to room temperature at a rate of 1°/min. Adapter sequences are given in Table S1. Ligation reactions from up to 48 individuals with unique barcodes are then pooled together. To concentrate and clean the pooled DNA for size selection, it is subjected to two rounds of Ampure magnetic bead purification resulting in a final volume of 30ul. Next, the pooled DNA is loaded onto a Blue Pippin (Sage Science; Beverly, MA) and narrow size selection is performed with a target of 400bp. Size-selected DNA is then directly used as a template for PCR amplification. Each library contains up to 48 individuals and can be amplified with a unique, indexed primer allowing for multiple libraries to be pooled in a single lane. PCR reactions consisted of 1x NEB Q5 PCR Buffer, 10mM each DNTP, 20pmol of both the universal and indexed primer, 0.25 U NEB Q5 polymerase, and 4ul of size-selected template. Reaction conditions were 98°C for 1 minute, 10 cycles of 98°C for 8 sec, 68°C for 20sec, and 72°C for 20sec, followed by a final extension at 72°C for 2 min. Eight separate PCR reactions were run for each library and pooled to avoid bias in the final library. Primer sequences are given in Table S2. A detailed protocol for library preparation can be found at [mosquitogenomics.org/protocols](http://mosquitogenomics.org/protocols).

For this experiment, single-end, 100 base pair (bp) sequencing was performed on 48 individuals from each group (NGS and NGS INB) using the Illumina HiSeq2500 at the UCR Genomics Core (a total of 288 individuals were multiplexed in this lane). The remaining eight individuals from each group were subjected to paired-end, 100bp sequencing on a HiSeq2500 (a total of 96 individuals were multiplexed in this lane). Since our pipeline filters out any SNPs found in less than 16 individuals from both lines (see below), none of the second-pair reads were retained for analysis. The additional mosquitoes sequenced in the aforementioned lanes were part of a different project.

#### *Data Processing and Analysis*

Each barcode/adaptor is a minimum of three mutational steps away from any other barcode/adaptor. Thus, sequencing reads were assigned to individuals based on adaptor and barcode sequence, allowing up to one mismatch in each sequence. After assigning reads to individuals, reads were mapped against the *An. gambiae* PEST reference genome using BWA (Li and Durbin 2009) with  $n=8$  and otherwise default parameters. After mapping, BAM files for all NGS and NGS INBRED mosquitoes were merged and genotypes were called in parallel for individuals using the GATK program with default parameters (McKenna, et al. 2010). A custom perl script extracted individual genotype information from the resulting vcf file and calculated per site  $\pi$ . Indels and multi-allelic ( $>2$ ) SNPs were excluded from polymorphism analysis due to difficulties in reliably calling genotypes from short read sequencing data. After filtering, mean coverage per

individual was 5.10x for NGS INB and 4.21x for NGS with an average of 2.86 SNPs per ddRAD fragment. Spatial patterns of diversity were investigated by averaging  $\pi$  for each SNP in non-overlapping 1 Mb windows across the genome. Results were robust to varying window sizes (data not shown).

### *2La Inversion Typing*

The *2La* inversion karyotype for individual mosquitoes was determined in two different ways. First, each individual was karyotyped using a molecular PCR diagnostic that spans the inversion breakpoints (White, et al. 2007b). Second, STRUCTURE (Hubisz, et al. 2009; Pritchard, et al. 2000) was run with  $k=2$  using genotype information from each individual for all SNPs between the *2La* breakpoints. STRUCTURE uses a Bayesian approach to assign a probability that each individual belongs to certain populations ( $k$ ) defined by the user. Groups correlating with *2La* homokaryotypes (~100% likelihood of assignment to population 1), *2L*<sup>a</sup> homokaryotypes (~100% likelihood of assignment to population 2), and heterokaryotypes (~50% likelihood of assignment to both populations) were readily identified. The PCR karyotype and the STRUCTURE karyotype agreed for all individual mosquitoes except two. For these two individuals, the STRUCTURE karyotype was used for downstream analysis as the PCR reaction has previously produced occasional false-calls (Obbard, et al. 2009; White, et al. 2007b).

### *Acknowledgements*

Funds for this project were provided to BJW by the University of California Riverside, Mosquito Research Foundation Grant #2013-19, and NIH R01AI113248. We acknowledge Nora Besansky, Carlo Costantini, Paul Howell, Marcy Kern, Tovi Lehmann, and Isabelle Morlais for establishing colonies and providing mosquito eggs. We thank James Ricci and Raissa Green for assistance with mosquito rearing. We appreciate the work of John Weger and Glenn Hicks in optimizing on-machine conditions for running ddRAD libraries on the HiSeq2500. Insightful comments from four anonymous reviewers substantially improved the manuscript. The following reagents were obtained through the MR4 as part of the BEI Resources Repository, NIAID, NIH: *Anopheles gambiae* Mali-NIH M form, MRA-860, deposited by Tovi Lehmann >> Nora Besansky and *Anopheles gambiae* Pimperena S form, MRA-861, deposited by Nora Besansky.

## LITERATURE CITED

- Beaty BJ, et al. 2009. From Tucson to genomics and transgenics: The Vector Biology Network and the emergence of modern vector biology. *PLoS neglected tropical diseases* 3: e343.
- Benedict MQ. 1997. Care and maintenance of anopheline mosquito colonies. In. *The Molecular Biology of Insect Disease Vectors*: Springer. p. 3-12.
- Benedict MQ, et al. 2009. Colonisation and mass rearing: learning from others. *Malar J* 8: S4.
- Blandin SA, et al. 2009. Dissecting the Genetic Basis of Resistance to Malaria Parasites in *Anopheles gambiae*. *Science* 326:
- Charlesworth D 1991. The apparent selection on neutral marker loci in partially inbreeding populations. *Genet Res* 57: 159-175.
- Cheng C, et al. 2012. Ecological genomics of *Anopheles gambiae* along a latitudinal cline: a population-resequencing approach. *Genetics* 190: 1417-1432.
- Clarkson CS, et al. 2014. Adaptive introgression between *Anopheles* sibling species eliminates a major genomic island but not reproductive isolation. *Nature communications* 5.
- Coetzee M, et al. 2013. *Anopheles coluzzii* and *Anopheles amharicus*, new members of the *Anopheles gambiae* complex. *Zootaxa* 3619: 246-274.
- Collins FH, Paskewitz SM 1995. Malaria: current and future prospects for control. *Annual Review of Entomology* 40: 195-219.

- Coluzzi M, Sabatini A, Della Torre A, Di Deco MA, Petrarca V 2002. A polytene chromosome analysis of the *Anopheles gambiae* species complex. *Science* 298: 1415-1418.
- Connor JL, Belucci MJ 1979. Natural selection resisting inbreeding depression in captive wild housemice (*Mus musculus*). *Evolution*: 929-940.
- Consortium GO 2013. Gene Ontology annotations and resources. *Nucleic acids research* 41: D530-D535.
- Costantini C, et al. 2009. Living at the edge: biogeographic patterns of habitat segregation conform to speciation by niche expansion in *Anopheles gambiae*. *BMC Ecology* 9: 16.
- della Torre A, Merzagora L, Powell JR, Coluzzi M 1997. Selective introgression of paracentric inversions between two sibling species of the *Anopheles gambiae* complex. *Genetics* 146: 239-244.
- Eriksson K, Halkka O, Lokki J, Saura A 1976. Enzyme polymorphism in feral, outbred and inbred rats (*Rattus norvegicus*). *Heredity* 37: 341-349.
- Frankham R, Smith G, Briscoe D 1993. Effects on heterozygosity and reproductive fitness of inbreeding with and without selection on fitness in *Drosophila melanogaster*. *Theoretical and Applied Genetics* 86: 1023-1027.
- Garcia N, Lopez-Fanjul C, Garcia-Dorado A 1994. The genetics of viability in *Drosophila melanogaster*: effects of inbreeding and artificial selection. *Evolution*: 1277-1285.

- Gimonneau G, et al. 2012. Larval habitat segregation between the molecular forms of the mosquito *Anopheles gambiae* in a rice field area of Burkina Faso, West Africa. *Medical and Veterinary Entomology* 26: 9-17.
- Gomulski L 1990. Polyandry in nulliparous *Anopheles gambiae* mosquitoes (Diptera: Culicidae). *Bulletin of Entomological Research* 80: 393-396.
- Harris C, Rousset F, Morlais I, Fontenille D, Cohuet A 2010. Low linkage disequilibrium in wild *Anopheles gambiae* s.l. populations. *BMC Genet* 11: 81.
- Holt RA, et al. 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* 298: 129-149.
- Hubisz MJ, Falush D, Stephens M, Pritchard JK 2009. Inferring weak population structure with the assistance of sample group information. *Molecular ecology resources* 9: 1322-1332.
- Hunter P 2008. The paradox of model organisms. *EMBO reports* 9: 717-720.
- Kamdem C, et al. 2012. Anthropogenic habitat disturbance and ecological divergence between incipient species of the malaria mosquito *Anopheles gambiae*. *PLoS One* 7: e39453.
- Knop NF, Asman SM, Reisen WK, Milby MM 1987. Changes in the biology of *Culex tarsalis* (Diptera: Culicidae) associated with colonization under contrasting regimes. *Environmental entomology* 16: 405-414.
- Latter B, Mulley J, Reid D, Pascoe L 1995. Reduced genetic load revealed by slow inbreeding in *Drosophila melanogaster*. *Genetics* 139: 287-297.



- Lawniczak MK, et al. 2010. Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science* 330: 512-514.
- Lee Y, et al. 2013. Spatiotemporal dynamics of gene flow and hybrid fitness between the M and S forms of the malaria mosquito, *Anopheles gambiae*. *Proceedings of the National Academy of Sciences of the United States of America* 110: 19854-19859.
- Lehmann T, Diabate A 2008. The molecular forms of *Anopheles gambiae*: a phenotypic perspective. *Infect Genet Evol* 8: 737-746.
- Li H, Durbin R 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25: 1754-1760.
- Mackay TF, et al. 2012. The *Drosophila melanogaster* genetic reference panel. *Nature* 482: 173-178.
- Mackay TF, Stone EA, Ayroles JF 2009. The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* 10: 565-577.
- Marsden CD, et al. 2014. Diversity, Differentiation, and Linkage Disequilibrium: Prospects for Association Mapping in the Malaria Vector *Anopheles arabiensis*. *G3: Genes| Genomes| Genetics* 4: 121-131.
- Matthews T, Craig Jr G 1987. Heterozygosity in inbred strains of the tree-hole mosquito *Aedes triseriatus*. *Biochemical genetics* 25: 647-655.
- McGoldrick DJ, Hedgecock D 1997. Fixation, segregation and linkage of allozyme loci in inbred families of the Pacific oyster *Crassostrea gigas*

(Thunberg): implications for the causes of inbreeding depression. *Genetics* 146: 321-334.

McKenna A, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20: 1297-1303.

Miklos GLG, Rubin GM 1996. The role of the genome project in determining gene function: insights from model organisms. *Cell* 86: 521-529.

Mina N, Sheldon B, Yoo B, Frankham R 1991. Heterozygosity at protein loci in inbred and outbred lines of chickens. *Poultry Science* 70: 1864-1872.

Morlais I, Ponçon N, Simard F, Cohuet A, Fontenille D 2004. Intraspecific nucleotide variation in *Anopheles gambiae*: new insights into the biology of malaria vectors. *The American Journal of Tropical Medicine and Hygiene* 71: 795-802.

Mpofu S, Masendu H, Kanyimo K, Mtetwa C 1993. Laboratory colonization of *Anopheles quadriannulatus* from sympatry with other sibling species of the *Anopheles gambiae* complex in Zimbabwe. *Medical and Veterinary Entomology* 7: 122-126.

(Munstermann LE 1994. Unexpected genetic consequences of colonization and inbreeding: allozyme tracking in *Culicidae* (Diptera). *Annals of the Entomological Society of America* 87: 157-164.

Neafsey DE, et al. 2013. The evolution of the *Anopheles* 16 genomes project. *G3: Genes| Genomes| Genetics* 3: 1191-1194.

- Nene V, et al. 2007. Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science* 316: 1718-1723.
- Nordborg M, Charlesworth B, Charlesworth D 1996. The effect of recombination on background selection. *Genet Res* 67: 159-174.
- Norris DE, Shurtleff AC, Toure YT, Lanzaro GC 2001. Microsatellite DNA polymorphism and heterozygosity among field and laboratory populations of *Anopheles gambiae* ss (Diptera: Culicidae). *J Med Entomol* 38: 336-340.
- Obbard DJ, Welch JJ, Little TJ 2009. Inferring selection in the *Anopheles gambiae* species complex: an example from immune-related serine protease inhibitors. *Malar J* 8: 117.
- Ohta T 1971. Associative overdominance caused by linked detrimental mutations. *Genet Res* 18: 277-286.
- Ohta T, Cockerham CC 1974. Detrimental genes with partial selfing and effects on a neutral locus. *Genet Res* 23: 191-200.
- Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One* 7: e37135.
- Pombi M, Stump AD, Della Torre A, Besansky NJ 2006. Variation in recombination rate across the X chromosome of *Anopheles gambiae*. *Am J Trop Med Hyg* 75: 901-903.
- Powell JR, Petrarca V, della Torre A, Caccone A, Coluzzi M 1999. Population structure, speciation, and introgression in the *Anopheles gambiae* complex. *Parassitologia* 41: 101-113.

- Pritchard JK, Stephens M, Donnelly P 2000. Inference of population structure using multilocus genotype data. *Genetics* 155: 945-959.
- Reidenbach KR, et al. 2012. Patterns of genomic differentiation between ecologically differentiated M and S forms of *Anopheles gambiae* in West and Central Africa. *Genome biology and evolution* 4: 1202-1212.
- Riehle MM, et al. 2006. Natural malaria infection in *Anopheles gambiae* is regulated by a single genomic control region. *Science* 312: 577-579.
- Rozen S, Skaletsky HJ. 2000. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S, editors. *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Totowa, NJ: Humana Press. p. 365-386.
- Rubin GM, et al. 2000. Comparative genomics of the eukaryotes. *Science* 287: 2204-2215.
- Rumball W, Franklin I, Frankham R, Sheldon B 1994. Decline in heterozygosity under full-sib and double first-cousin inbreeding in *Drosophila melanogaster*. *Genetics* 136: 1039-1049.
- Santiago E, Caballero A 1995. Effective size of populations under selection. *Genetics* 139: 1013-1030.
- Sharakhov IV, et al. 2006. Breakpoint structure reveals the unique origin of an interspecific chromosomal inversion (2La) in the *Anopheles gambiae* complex. *Proc Natl Acad Sci U S A* 103: 6258-6262.

- Sharakhova MV, et al. 2010. Genome mapping and characterization of the *Anopheles gambiae* heterochromatin. *BMC Genomics* 11.
- Sharakhova MV, et al. 2007. Update of the *Anopheles gambiae* PEST genome assembly. *Genome biology* 8: R5.
- Simard F, et al. 2009. Ecological niche partitioning between the M and S molecular forms of *Anopheles gambiae* in Cameroon: the ecological side of speciation. *BMC Ecology* 9: 17.
- Sing CF, Brewer GJ, Thirtle B 1973. Inherited biochemical variation in *Drosophila melanogaster*: noise or signal? I. Single-locus analyses. *Genetics* 75: 381-404.
- Slotman MA, et al. 2006. Reduced recombination rate and genetic differentiation between the M and S forms of *Anopheles gambiae* s.s. *Genetics* 174: 2081-2093.
- Strauss SH 1986. Heterosis at allozyme loci under inbreeding and crossbreeding in *Pinus attenuata*. *Genetics* 113: 115-134.
- Strobeck C 1983. Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. *Genetics* 103: 545-555.
- Strobeck C 1979. Partial selfing and linkage: the effect of a heterotic locus on a neutral locus. *Genetics* 92: 305-315.
- Stump AD, et al. 2007. Genetic exchange in 2La inversion heterokaryotypes of *Anopheles gambiae*. *Insect Molecular Biology* 16: 703-709.

- Toure YT, et al. 1998. The distribution and inversion polymorphism of chromosomally recognized taxa of the *Anopheles gambiae* complex in Mali, West Africa. *Parassitologia* 40: 477-511.
- Tripet F, Toure YT, Dolo G, Lanzaro GC 2003. Frequency of multiple inseminations in field-collected *Anopheles gambiae* females revealed by DNA analysis of transferred sperm. *The American Journal of Tropical Medicine and Hygiene* 68: 1-5.
- Wang J, Hill WG 1999. Effect of selection against deleterious mutations on the decline in heterozygosity at neutral loci in closely inbreeding populations. *Genetics* 153: 1475-1489.
- Weetman D, Wilding CS, Steen K, Pinto J, Donnelly MJ 2012. Gene flow–dependent genomic divergence between *Anopheles gambiae* M and S forms. *Molecular Biology and Evolution* 29: 279-291.
- Weir B, Cockerham CC 1973. Mixed self and random mating at two loci. *Genet Res* 21: 247-262.
- White BJ, Cheng C, Simard F, Costantini C, Besansky NJ 2010. Genetic association of physically unlinked islands of genomic divergence in incipient species of *Anopheles gambiae*. *Mol Ecol* 19: 925-939.
- White BJ, Collins FH, Besansky NJ 2011. Evolution of *Anopheles gambiae* in Relation to Humans and Malaria. *Annual Review of Ecology, Evolution, and Systematics*, Vol 42 42: 111-132.
- White BJ, et al. 2007a. Localization of candidate regions maintaining a common polymorphic inversion (2La) in *Anopheles gambiae*. *PLoS Genet* 3: e217.

- White BJ, et al. 2013. Dose and developmental responses of *Anopheles merus* larvae to salinity. *Journal of Experimental Biology* 216: 3433-3441.
- White BJ, et al. 2007b. Molecular karyotyping of the 2La inversion in *Anopheles gambiae*. *Am J Trop Med Hyg* 76: 334-339.
- Wondji C, et al. 2005. Species and populations of the *Anopheles gambiae* complex in Cameroon with special emphasis on chromosomal and molecular forms of *Anopheles gambiae* s.s. *J Med Entomol* 42: 998-1005.
- Yeh F. 1972. Inbreeding as a Control of Biochemical Polymorphism in Populations of *Tribolium Castaneum*. *Canadian Journal of Genetics and Cytology*. 14(3): 741.
- Zhu Y, Bergland AO, González J, Petrov DA 2012. Empirical validation of pooled whole genome population re-sequencing in *Drosophila melanogaster*. *PLoS One* 7: e41901.

## FIGURE LEGENDS

**Figure 1.** The ddRADseq approach successfully generated a genome-wide panel of SNPs. The number of SNPs genotyped in at least 16 individuals from both NGS and NGS INBRED are plotted in 1 Megabase (Mb) bins across the genome.

**Figure 2.** Dramatic reductions in polymorphism are observed in the NGS INBRED line. The distribution of  $\pi$  for individual SNPs is plotted for the (A) NGS autosomes, (B) NGS X chromosome, (C) NGS INBRED autosomes, and (D) NGS INBRED X chromosome. Red lines denote the mean value of  $\pi$ .

**Figure 3.** Most individual SNPs exhibit reduced polymorphism in NGS INBRED although both persistently polymorphic SNPs and SNPs with increased polymorphism exist.  $\pi$  values for individual SNPs in the NGS and NGS INBRED are plotted against each other for (A) the autosomes and (B) the X chromosome. The number of SNPs falling into each bin is denoted by the logarithmic color scale shown on the right.

**Figure 4.** Reductions in polymorphism are spatially heterogeneous across the genome. Black and red lines denote mean  $\pi$  for all SNPs in 1 Mb non-overlapping windows across the genome for NGS and NGS INBRED, respectively. Boundaries of the 2La inversion are marked with a shaded box.



Large, discrete blocks of both low and high nucleotide diversity in NGS INBRED are evident.

**Figure 5.** Polymorphism of the *2La* inversion is responsible for the large block of polymorphism on NGS INBRED chromosome 2L. Different colored lines denote mean  $\pi$  for all SNPs in 1 Mb stepping windows across the genome for the three different karyotypes present in NGS INBRED. Boundaries of the *2La* inversion are marked with a shaded box. Note the dramatic difference in polymorphism between heterokaryotypes and homokaryotypes.

**Figure 6.** The NSG colony was founded from mosquitoes mostly carrying the *2L+<sup>a</sup>* arrangement. Different color lines denote mean  $\pi$  for all SNPs in 1 Mb stepping windows across the genome for the three different karyotypes present in NGS. Boundaries of the *2La* inversion are marked with a shaded box. Note the clear difference in polymorphism between alternative homokaryotypes.

TABLES

Table 1. Results of attempted inbreeding for five *An. gambiae* colonies

Colony	Form	Origin	Established	Generation	Issue
<b>NDKO</b>	S	Ndkayo, Cameroon	2007	1	Stopped laying eggs
<b>PIMPERENA</b>	S	Pimperena, Mali	2005	0	Would not lay eggs
<b>YAOUNDE</b>	M	Yaounde, Cameroon	1991	10	Stopped bloodfeeding
<b>NGS</b>	M	N'gouso, Cameroon	2007	20	N/A
<b>MALI-NIH</b>	M	Niono, Mali	2005	5	Stopped laying eggs

FIGURE 1.

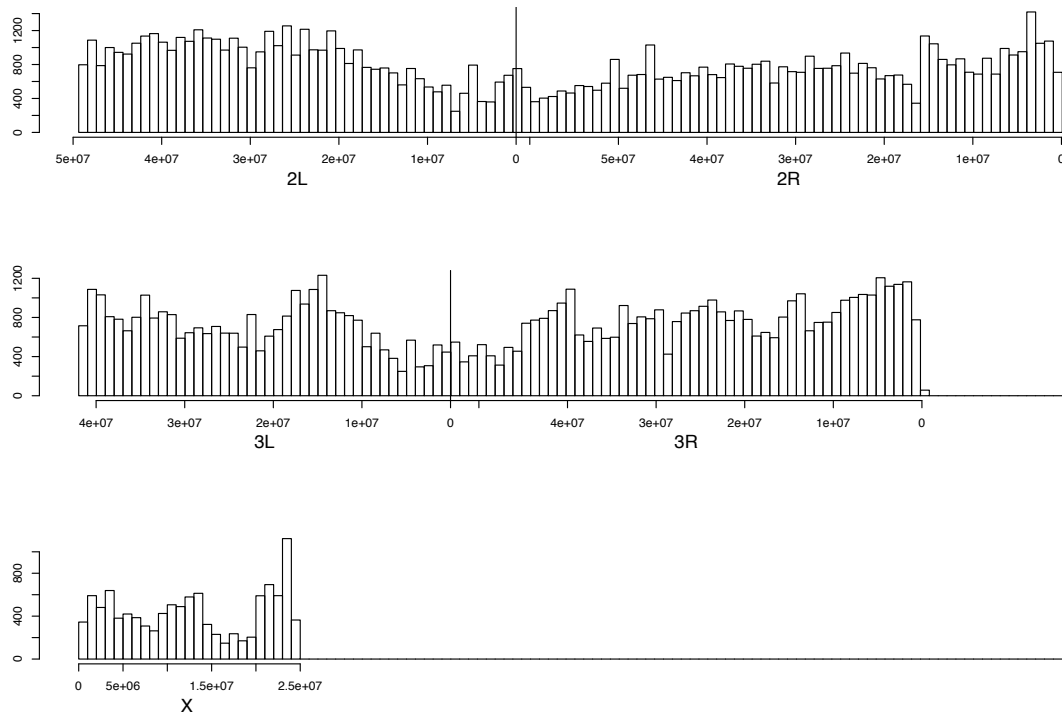


FIGURE 2.

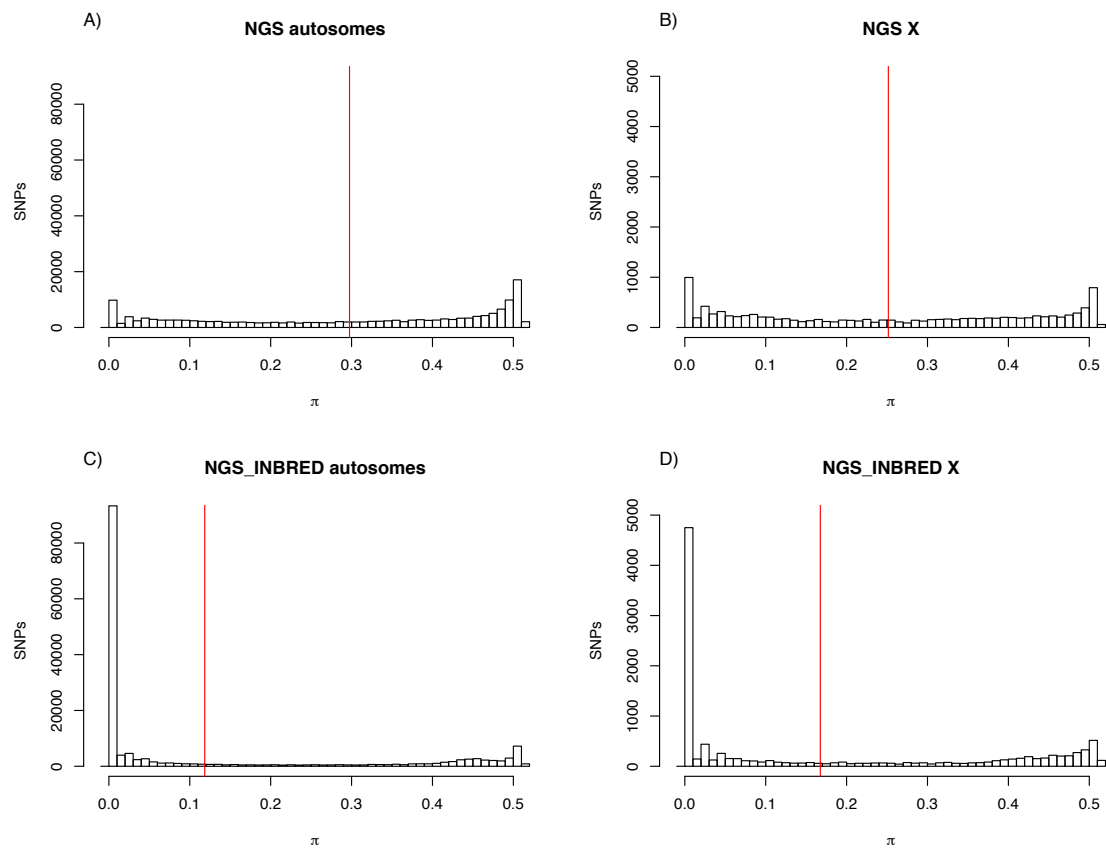


FIGURE 3.

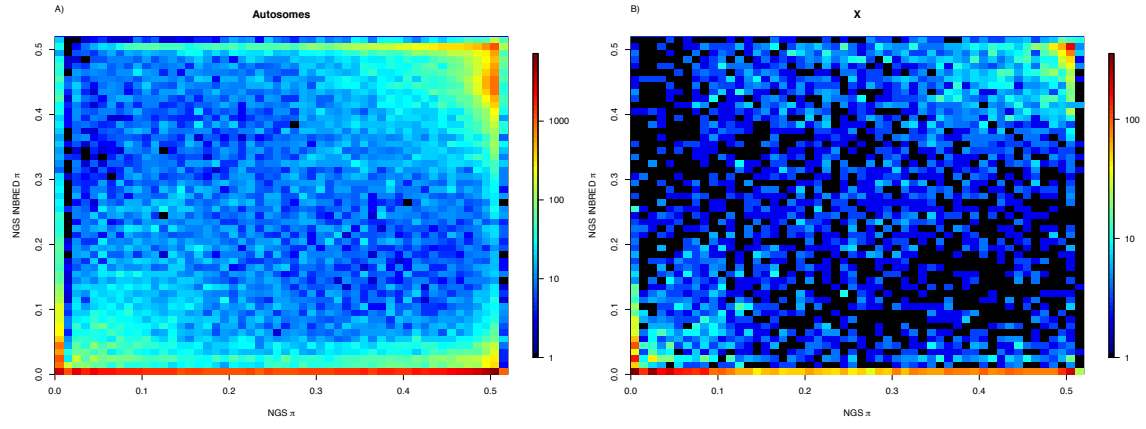


FIGURE 4.

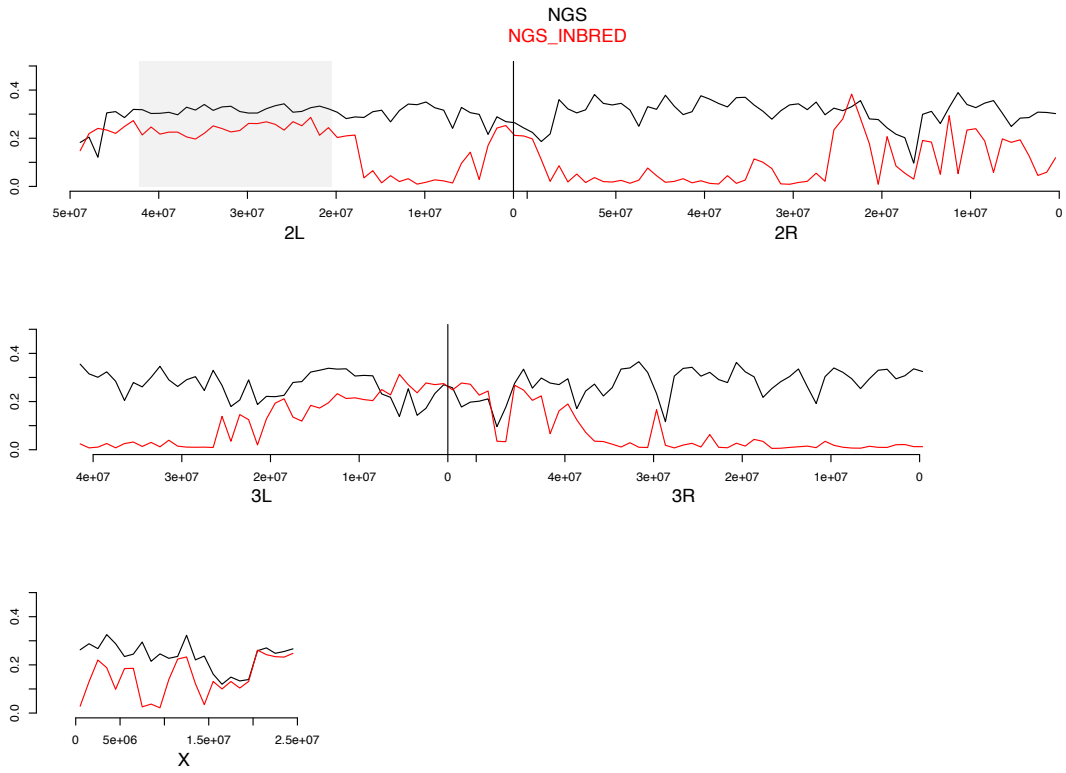


FIGURE 5.



FIGURE 6.

