PLOS Genetics

Demography-informed selection signatures reveal different roles of two polymorphic inversions in local adaptation of Anopheles gambiae

--Manuscript Draft--

Manuscript Number:	PGENETICS-D-15-00120
Full Title:	Demography-informed selection signatures reveal different roles of two polymorphic inversions in local adaptation of Anopheles gambiae
Short Title:	Demography-informed selection signatures in inversions
Article Type:	Research Article
Section/Category:	Evolution
Keywords:	Inversions, selection, demography, Anopheles gambiae
Abstract:	Chromosomal inversions are important structural changes that may facilitate divergent selection when they capture co-adaptive loci. However, identifying selection targets within inversions can be challenging. The high degree of differentiation between heterokaryotypes as well as the differences in demographic histories of collinear regions compared to inverted ones reduce the power of detection with traditional outlier scans. Here, we developed a new approach that uses discriminant functions to classify loci that are under selection (or drift) informed from inversion-specific expectations for selected and neutral patterns as characterized by a diversity of summary statistics. We demonstrate the approach with an analysis of RAD sequencing data we collected in a classic Dipteran species with polymorphic inversion clines - Anopheles gambiae, a malaria vector species from sub-Saharan Africa. Contrary to minimal geographic structure among populations in collinear regions, individuals are clustered by SNPs from two polymorphic inversions with the 2La and 2Rb arrangements predominating in dry habitats compared with prevalence of the 2L+a and 2R+b arrangements in wet habitats. Despite of both being adaptive introgressions, estimates of the demographic histories, and hence expectations for patterns of neutral and selected within the two inversions differ. Specifically, the origin of 2La predated the divergence of two species. Nevertheless, with our approach we were able to detect putative regions subject to selection within both chromosomal inversions, with much higher power than traditional FST-outlier analysis. Moreover, the arrangement 2L+a that associates with wet habitat exhibits much more selection signatures than the arrangement 2La that associates with dry habitat. We discuss the implications of these results with respect to studies of rapid adaptation in these malaria vectors, and in particular, the insights our newly developed approach offers for identifying not only potential targets of selection but also the lineage that has undergone adaptive change.
Additional Information:	
Question	**Response**
Data Availability PLOS journals require authors to make all data underlying the findings described in their manuscript fully available, without restriction and from the time of publication, with only rare exceptions to address legal and ethical concerns (see the PLOS Data Policy and FAQ for further details). When submitting a manuscript, authors must provide a Data Availability Statement that describes where the data underlying their manuscript can be found. Your answers to the following constitute your statement about data availability and will be included with the article in the event of publication. **Please note that**	Yes - all data are fully available without restriction

simply stating 'data available on request from the author' is not acceptable. ***If, however, your data are only available upon request from the author(s), you must answer "No" to the first question below, and explain your exceptional situation in the text box provided.*** Do the authors confirm that all data underlying the findings described in their manuscript are fully available without restriction?	
Please describe where your data may be found, writing in full sentences. **Your answers should be entered into the box below and will be published in the form you provide them, if your manuscript is accepted.** If you are copying our sample text below, please ensure you replace any instances of XXX with the appropriate details. If your data are all contained within the paper and/or Supporting Information files, please state this in your answer below. For example, "All relevant data are within the paper and its Supporting Information files." If your data are held or will be held in a public repository, include URLs, accession numbers or DOIs. For example, "All XXX files are available from the XXX database (accession number(s) XXX, XXX)." If this information will only be available after acceptance, please indicate this by ticking the box below. If neither of these applies but you are able to provide details of access elsewhere, with or without limitations, please do so in the box below. For example: "Data are available from the XXX Institutional Data Access / Ethics Committee for researchers who meet the criteria for access to confidential data." "Data are from the XXX study whose authors may be contacted at XXX." * typeset	All relevant data are within the paper and its Supporting Information files. All customized scripts, setting files for programs and genomic data will be deposited on Dryad.
Additional data availability information:	Tick here if the URLs/accession numbers/DOIs will be available only after acceptance of the manuscript for publication so that we can ensure their inclusion before publication.

Dear *PloS Genetics* journal,

We are delighted to submit the manuscript "**Demography-informed selection signatures reveal different roles of two polymorphic inversions in local adaptation of *Anopheles gambiae***" as possible publication in *PloS Genetics*.

With the increasing availability of physical maps among different species, the important role of chromosomal inversions in maintaining adaptive divergence has been demonstrated in many systems. Because chromosomal inversions suppress recombination between heterokaryotypes, co-adaptive genotypes within inversions that confer local adaptations can be protected. However, a method that identifies selection signature of loci/region within inversions in an empirical system of adaptive divergence is lacking.

In this study, we develop a new approach for locating a selection signature when population divergence (or speciation) is promoted by inversions, a genomic architecture that limits the power of traditional tests for divergent selection (i.e., F_{ST}-outlier tests). Genomic outlier scans either have an inherent assumption for a simple demographic model which might not capture the characteristics of the species' history or treat the genome as a homogenous pool evolving neutrally under a single common demographic history. In addition, it depends on an expectation of the background level of divergence generated by genetic drift to be distinguishable from that of selected scenarios. Therefore, these are improper methods to be applied for regions with reduced recombination (i.e., high levels of background divergence) or evolved under different demographic histories than the rest of the genome, such as inversions. The approach we designed circumvents these problems, with the intent to lower false discovery rates as well as increase detection power. First, separate demographic histories are inferred for inversion and collinear regions to generate region-specific background levels of differentiation. Second, genetic measures other than differentiations, such as haplotype and genetic diversity, are also included in the analyses to increase the power of distinguishing adaptive alleles from neutral ones. Third, discriminant functions are used to predict selection instead of an outlier approach by using training datasets from neutral versus selection simulations.

We applied this approach to an analysis of two common inversions in a severe malaria vector species, *Anopheles gambiae*, which are associated with aridity adaptations. We found the signature of selection traces predominantly in the wet-adapted $2L+^{a}$ arrangement rather than the dry-adapted 2La arrangement, which was previously overlooked in studies of 2La inversions. We think this study is worthy of publishing in *PloS Genetics* because it is the first attempt to identify selection signature within inversions in an empirical system using inversion-specific coalescent expectations. It has a wide application as it directly tests whether inversions evolved under Kirkpatrick and Barton (2006)'s local adaptation theory as well as how many loci/regions were involved.

Best regards,

Qixin He and L. Lacey Knowles
Department of Ecology and Evolutionary Biology, University of Michigan

Title: Demography-informed selection signatures reveal different roles of two polymorphic inversions in local adaptation of *Anopheles gambiae*

Running Title: Demography-informed selection signatures in inversions

Type: Article

Authors: Qixin He and L. Lacey Knowles

Contact information:
Qixin He, Ph. D. candidate
Email**:** heqixin@umich.edu
Affiliation: Department of Ecology & Evolutionary Biology, 1109 Geddes Ave., Museum of Zoology, University of Michigan, Ann Arbor MI 48109-1079. phone (734) 763-7943, fax (734) 763-4080

L. Lacey Knowles, Professor and Curator
Email: knowlesl@umich.edu
Affiliation: Department of Ecology & Evolutionary Biology, 1109 Geddes Ave., Museum of Zoology, University of Michigan, Ann Arbor MI 48109-1079. phone (734) 763-7943, fax (734) 763-4080,
Corresponding Author: Qixin He.
Manuscript information: 4 Figs + 1 Table; Supplementary, 6 Figs + 2 Tables + 1 Text file
Data Archival Location: Dryad upon acceptance
Keywords: Inversions, selection, demography, *Anopheles gambiae*

Abstract

Chromosomal inversions are important structural changes that may facilitate divergent selection when they capture co-adaptive loci. However, identifying selection targets within inversions can be challenging. The high degree of differentiation between heterokaryotypes as well as the differences in demographic histories of collinear regions compared to inverted ones reduce the power of detection with traditional outlier scans. Here, we developed a new approach that uses discriminant functions to classify loci that are under selection (or drift) informed from inversion-specific expectations for selected and neutral patterns as characterized by a diversity of summary statistics. We demonstrate the approach with an analysis of RAD sequencing data we collected in a classic Dipteran species with polymorphic inversion clines - *Anopheles gambiae*, a malaria vector species from sub-Saharan Africa. Contrary to minimal geographic structure among populations in collinear regions, individuals are clustered by SNPs from two polymorphic inversions with the 2La and 2Rb arrangements predominating in dry habitats compared with prevalence of the $2L+^{a}$ and $2R+^{b}$ arrangements in wet habitats. Despite of both being adaptive introgressions, estimates of the demographic histories, and hence expectations for patterns of neutral and selected within the two inversions differ. Specifically, the origin of 2La predated the divergence of two species. Nevertheless, with our approach we were able to detect putative regions subject to selection within both chromosomal inversions, with much higher power than traditional F_{ST}-outlier analysis. Moreover, the arrangement $2L+^{a}$ that associates with wet habitat exhibits much more selection signatures than the arrangement 2La that associates with dry habitat. We discuss the implications of these results with respect to studies of rapid adaptation in these malaria vectors, and in particular, the insights our newly developed approach offers for identifying not only potential targets of selection but also the lineage that has undergone adaptive change.

Author summary

Evidence supporting the prevalent role of inversions in facilitating adaptive divergence is now common. Nonetheless, identifying the loci within inversions that confer local adaptation has been difficult. Here, we present a new approach to identifying targets of selection, and demonstrate the approach with an analysis of two common inversions in a malaria vector species, *Anopheles gambiae*. Specifically, discriminant functions were built to detect regions under

selection based on the comparison of the observed data to expectations generated from simulations of neutral versus selection under inversion specific demographic histories. With this new approach we found the signature of selection traces predominantly to the 2L+^{a} arrangement rather than the introgressed 2La arrangement, suggesting that the same region has involved in two episodes of adaptations - specifically to wet and dry habitats, respectively. Our approach can be widely applied for chromosomal regions with different histories, as well as non-model species with no genome reference or sparse genomic markers, and represents a significant improvement over traditional tests like F_{ST}-outlier analyses to detect targets of selection.

Introduction

Detecting the signature of natural selection from molecular data has been a central focus of geneticists, not only because of the deeper understanding of molecular evolution it brings, but also its potential in revealing important functional information [1,2]. With the advances in sequencing technologies, a plethora of new methods have been developed for identifying selected sites and regions of the genomes. These approaches capture different attributes the signature of selection might leave. These include tests for patterns of exceptionally long haplotypes [3-5], surges in linkage desequilibirum (LD) [6,7], and skewed site frequency spectrums [8-10]. However, when the goal is to identify selection under spatially divergent selection among populations, a predominant approach is to rely on F_{ST}-outlier tests to scan for regions with exceptionally high divergence between ecologically divergent populations [11-15]. The appeal of the approach also extends from its broad applicability, especially with the lack of genomic resources for detailed estimates of LD or haplotypes in most taxa. However, the power of such methods becomes inherently limited when adaptive mutations occur in regions with reduced recombination, such as with chromosomal inversions. This becomes particular problematic for studying adaptive divergence given the important role inversions play in maintaining co-adaptive genotypes [16,17].

Identifying selection targets within alternative inversions can be challenging because of two characteristics that distinguish them from collinear regions (i.e., genomic regions without inversions). First, in genomic regions with inversions there is an overall high divergence between the inverted and non-inverted (i.e., standard) chromosomal arrangement [e.g., 18], which itself diminishes the power to detect selection [19,20]. Second, as a consequence of the effect that inversions have on the mixing of alleles among individuals (i.e., alleles captured by inversions are protected from mixing with alleles on standard chromosomes), the genome is a mosaic, with the amount of gene flow differing across regions depending on whether it includes an inversion. This means that tests that rely upon a single neutral parameterization (either with simulations under island-model used in FDIST2 [21] or coancestry matrix in FLK [13] will necessarily be based on a mis-specification for expected patterns of divergence, which can then further exacerbate the difficulties with detecting targets of selection.

In this study, we develop a new approach for locating a selection signature when population divergence (or speciation) is promoted by inversions, a genomic architecture that limits the power of traditional tests for selection, as described above. By first estimating region specific demographic histories, we are able to generate region specific neutral expectations for demography-adjusted selection tests [see also 22]. We then built a discriminant function using combinations of summary statistics from sequences simulated under neutral and selected scenarios, which can then be used to assign empirical sites into neutral or selection classes based on predictions from the discriminant function. We apply the newly developed approach to *Anopheles gambiae*, which like other Dipeteran species, has a long history of research on inversion polymorphisms [23]. As a widespread species with large population size, it typically lacks significant population structure, except for the geographic structure observed at genomic regions characterized by seven commonly segregating inversions on the chromosome 2 [24-26]. Here we focus on two large inversions, 2La and 2Rb (21.4Mb and 7 Mb, respectively), that exhibit stable clines, but which defy detection of putatively selected regions/genes within the inversions because of high LD and F_{ST} across the inversions [27]. The high frequency of 2La in dry geographic areas (savannas) compared to wet areas (forests) [28], a predictable cycle in its frequency during dry and wet seasons, and higher resistance to dessication in experiments [29], identify its role in adaptive divergence, as do similar trends observed in 2Rb [30]. In addition to the general difficulties with identifying the targets of selection within inverted regions discussed above, another potential complication is associated with the origins of the two inversions. Both 2La and 2Rb are examples of adaptive introgression from another species in the same species complex, *An. arabiensis* [31], which is sympatric with *An. gambiae* in arid savanna areas [32,33]. This suggests that the region contained in the inversions might be old enough to be highly differentiated from the standard chromosome in *An. gambiae*, which exacerbates the aforementioned high divergence problem between heterokaryotypes.

Nevertheless, despite these challenges, as our analyses demonstrate, we are able to not only identify genomic areas associated with targets of selection, but we are also able to make statistical statements about what lineage underwent selective divergence with respect to the ancestral state. We highlight how these new findings have interesting implications for how we think about rapid divergence in the malaria vector *An. gambiae*. We also discuss the applicability of our procedure for tests of selection in other taxa more generally, including those where

specific genomic regions have vastly different history than the rest of the genome because of the mosaic nature.

Results

We collected genomic data in 259 *An. gambiae* individuals and 8 *An. arabiensis* from six sites along a gradient of wet to dry habitats in Cameroon (see supplementary text S1 for specimen identification results; Table S1 for sampling information). From individually-barcoded double digest genomic Radseq libraries [34], and two lanes of 100bp paired-end sequencing on Illumina HiSeq2000 platform, 25,966 loci (i.e., RADtags from the genomic preparation) were mapped onto Chromosome 2, 3 and the X, after filtering for ambiguously mapped reads and loci with low coverage per sample or low presence across samples. Although we recognize that this represents a small proportion of the genome (about 1%), the goal of this manuscript is to demonstrate the promise of the approach for detecting approximate targets of selection, as opposed to providing a full analysis of the proportion of sites under selection within inversions contributing to adaptive divergence, which would require additional sequencing that is beyond the scope of this study.

Our new approach of targeting specific selection within inversion involves several procedures (summarized in Fig. 1a). First we estimated divergence time and introgression rate between the two species from collinear regions (Fig. 1b) using site frequency spectrum [35]. We then estimated inversion specific parameters, such as gene flux rate between alternative inversions and the age of inversion mutations (Fig. 1c). This framework was then used for tests of selection against neutral expectations that are region specific based upon discriminant functions derived from summary statistics estimated from empirical versus simulated data under the inferred demographic history.

Establishing a demographic null model for tests of selection

Principal Components Analysis (PCA) and Discriminant Analysis of Principal Components (DAPC) analyses of SNP data showed a lack of population genetic structure in the species in the collinear regions. Specifically, no distinctive geographic structure among the six sampled populations was apparent from the PCA (Fig. S3) and one group ($K = 1$) received the highest

support in the DAPC analysis (see supplementary text for details about tests of geographic structure). Therefore individuals were pooled across populations to estimate a demographic history that could be used to establish a null expectation for patterns of genomic divergence under a model of drift.

Specific demographic histories were inferred for different genomic regions from the region-specific site frequency spectrum (SFS) under a composite-likelihood approach implemented in fastsimcoal2 (for details see methods) [35]. Estimations showed that the two species diverged fairly recently around 87K years ago (~1.4Ne generations ago, T_{div} in Table 1; assuming 12 generations a year [26]) with small but constant genetic exchange (on the order of 1E-7 to 1E-8, which is about exchanging 0.01 to 0.2 individuals per generation; *m* in Table 1).

In contrast to the collinear regions, SNPs from 2La (2L: 20524058-42165532) and 2Rb (2R: 19023925-26758676) were distributed into one of three clusters in the PCAs (first PC explaining 20.3% and 11.9% of the total variance in 2La and 2Rb, respectively; Fig. S4 a,d). The results from the DAPC also supported $K = 3$ as the most likely number of genetic clusters (Fig. S4b, e). Individuals that form the three clusters identified from these analyses correspond to the three genotypes associated with the inverted genomic region: namely, the inverted homokaryotypes (referred to as *I/I* hereafter), the heterokaryotypes (*I/S*), and the standard (i.e., non-inverted) homokaryotypes (*S/S*) based on comparison with molecular karyotyping results (Fig. S4c, f). Separate demographic histories were therefore inferred for the different genomic arrangements (i.e., inverted versus standard chromosomal regions) for both the 2La and 2Rb regions, with a recombination parameter to accommodate occasional gene flux between alternative karyotypes (Fig. 1c).

Inversions were modeled as an introgressed between *An. arabiensis* and *An. gambiae* after the species' divergence time (see demographic model in Fig. 1c). The age of inversion mutation and their introgression time will strongly influence the baseline divergence expected for detecting selection. Interestingly, the coalescent time of *S* and *I* (T_{IS}, Fig. 1c) of 2La was around 3Ne generations ago (Table 1), which is much more ancient than species divergence time (T_{div}). In contrast, T_{IS} of 2Rb is similar to T_{div}. Despite its ancient origin (which predates the divergence of *An. arabiensis* and *An. gambiae*), the introgression event of 2La was found to have occurred more recently (~0.5*Ne*) than 2Rb (~0.8*Ne*) (Table 1).

Signature of selection within inversions

A modified F_{ST}-outlier scan based on inversion specific demographic history was first performed. Due to structural constraints of inversions, double recombination rates are higher in the center of an inversion versus breaking points. We therefore divided inversion regions into 150kb segments and adjusted recombination rates from our average estimates based on local 2Mb F_{ST} estimations (for details see methods) and then simulated 1000 neutral cases per region. We found that because the divergence of *I* and *S* of 2La was quite old, F_{ST} estimates between I and S for neutrally evolved DNA segments tended to be very high (e.g., the 95 percentile and 99 percentile of simulation estimates generally exceed 0.8; see Fig. 2a). Consequently, the power of differentiating selection from neutral genes based on F_{ST} estimates alone is severely limited within the 2La inversion (~0.3 and 0.1 with 0.05 and 0.01 false-positive rates respectively; Fig 4), in contrast to collinear regions where most F_{ST} estimates are concentrated around 0 (Fig. 2c). This is less exacerbated in the younger 2Rb (Fig. 2b; ~0.42 and 0.3 with 0.05 and 0.01 false-positive rates respectively; Fig 4). Yet, the empirical F_{ST} distribution has much longer tail than the collinear F_{ST} distribution (Fig. 2d), again highlighting the challenge with detecting putatively selected regions given the elevated F_{ST}-values that reduces detection power with very high false-discovery rates (Fig. 4; see Material and Methods for the distinction between false-positive rates and false-discovery rates).

Since outlier analysis based solely on F_{ST} measures have limited power to differentiate selection from drift (see Fig. 4), we designed a new approach to make use of a diversity of summary statistics. 1000 simulations of 50kb sequences containing selected locus were ran for each of the three scenarios: 1) neutral; 2) selection occurs on sites associated with *S* in *Anopheles gambiae*; 3) selection occurs on sites associated with *I* in *An. gambiae* and *An. arabiensis* (Fig. 1a). Discriminant functions (DAPC) were built to differentiate the three scenarios based on the values observed across the summary statistics of simulated RADtags (Fig. 3a). This approach yielded a significant gain in power (Fig. 4a-d) and a reduction in false-discovery rates (Fig. 4e-h). In order to minimize random effects from individual RADtags, average summary statistics of random loci sampled across the simulated 50kb regions at a similar density as the empirical data were calculated to build the discriminant function which classify regions as containing selected locus or not. The power of assigning regions correctly to three scenarios is therefore above 0.9 (Table

S2) with a comparable false-positive rates (~0.06) as the outlier analysis with a top 5% cut-off (Fig. 4). Moreover, in contrast to the pervasive problem with high false-discovery rates based on F_{ST}-values alone (i.e., falsely identified targets of selection), such cases significantly decreased when multiple summary statistics are used (Fig. 4e-h; Table S2). Interestingly, heterozygosity (H) and θ_π contributed more of the variation used in differentiating each scenario compared to F_{ST}.

After empirical estimates of summary statistics were transformed into discriminant scores, regions were assigned into neutral or selection classes based on predictions from the discriminant function estimated from the simulations (see Fig 3a). We found that on 2Rb, four regions contained selected sites associated with I (red bars on Fig 3b), while four contain selection associated with S (blue bars in Fig. 3b). These regions span some loci that were identified as F_{ST} outliers, but not all of them. In 2La, we recovered a much higher number of regions identified as being selected in S compared to those associated with I (Fig. 3).

Discussion

Our study is the first attempt to identify selection signature within inversions in an empirical system using inversion-specific coalescent expectations. By modeling the unique origin and introgression histories of each inversion compared to collinear regions, our approach allows for identifying targets of selection through discriminant function analyses trained with a suite of genetic measures. Compared to traditional F_{ST}-outlier analysis, simulation-trained discriminant function classification analysis has a much higher detection power and reduced false discovery rates (Fig. 4). When the F_{ST}-outlier analysis was applied to *Anopheles gambiae* populations, older inversions (i.e., 2La) suffered more from reduced power of detection because of the increasing baseline of divergence estimators for neutral sequences in older inversions (see Fig. 2, 4). In contrast, the simulation-trained discriminant function classification analysis achieved similar detection power and false discovery rates for both 2La and 2Rb (Fig. 4). In addition to the higher detection power, the new approach can also decipher the specific branch where selection occurred. In particular, selection was found to have occurred predominantly on *Anopheles gambiae* standard arrangement ($2L+^{a}$) rather than the introgressed arrangement (2La), suggesting that the same region has involved in two episodes of adaptations - specifically to wet and dry habitats, respectively. Our study highlighted the importance of adaptive introgression via inversions as a mechanism for maintaining adaptive divergence in species with high gene flow among populations inhabiting heterogeneous environments.

Detecting selection in regions with reduced recombination rates

The traditional indexes (F_{ST}, D_{XY}) in genomic scan to quantify inter-population differentiation are still the most effective way to detect targets of divergent selection in most cases (e.g., Soria-Carrasco et al. 2014). With the availability of population genomic data, such genomic scan studies become a common practice (reviewed in [36]) and many studies reported regions of elevated divergence, termed genomic islands (e.g., [37-39]; but see [40]. While common practices do not accommodate regions with reduced recombination or different demographic histories, we tried to circumvent the overall high divergence problem in inversions by generating null expectations from neutral evolution in selection identification. However, this is still not ideal because the detection power is highly diminished when variability of divergence estimators

among neutral sequences increases with the age of divergence (see overall low power and high false discovery rates in F_{ST}-outlier results in Fig. 4, S5-6).

The simulation-trained discriminant function classification analysis in this study is similar to an ABC parameter estimation process, in which simulations close to observations are retained, and linear functions are built to estimate parameters from PCs transformed from summary statistics [41,42]. The difference is that we are choosing models (neutral or under selection) instead of estimating parameters because our simulations are fixed with specific parameters estimated from reconstructed demographic histories. In addition, the increased power of applying multiple summary statistics, converting them into PCs and transforming into discriminant functions is a good way to detect selection when expectations for a single statistic do not differ sufficiently under selection scenarios to provide discriminatory power. This approach has the great advantage of jointly considering all summary statistics together across different arrangements/species instead of relying on F_{ST} measures alone. Since RADtag sequences are relatively short (~100bp), simulated sequences still show a great degree of variation in summary statistics for each scenario so that discriminant functions do not differentiate them completely (Fig. 4). Nevertheless, true positive rate are ~0.8 in assigning to the correct scenarios. However, these rates can be improved significantly when we identify regions under selection using average summary statistics across a 50kb window encompassing randomly distributed short Radtags. Using regional average of statistics greatly reduced stochastic variability of each scenario so that correct assignment rates are above 90% in all scenarios (Fig. 4, Table S2).

Due to the rapid decay of linkage disequilibrium in the *Anopheles gambiae* complex (r^2 close to 0 between SNPs 1Kb apart; see [43]), local changes in linkage disequilibrium inside inversions are hard to detect, especially with our sparsely sampled genomic dataset. Yet, LD or extended haplotype tests might be informative within individuals of the same arrangements to detect selection signatures [e.g., 44] because although recombination between heterokaryotypes is highly reduced, it is not reduced within same karyotypes (e.g., LDhat [45] did not find significant changes in recombination rates within same karyotypes in our data). Nevertheless, our demography-informed discriminant analysis is still powerful for tests of selection in genomic regions with different evolution histories. These results highlight the general utility of this approach in non-model species, especially with sparse sampling of the genome or no reference

genome for applying methods for detecting changes in LD within the same karyotype from individually-barcoded whole genome sequences.

Age of inversions and its implications on adaptive introgression

Our estimation of the older divergence of 2L+^{a} from 2La corroborates recent karyotype and genomic phylogenies for the *Anopheles gambiae* complex [43,46], which predicted that 2L+^{a} evolved in the ancestor of *An. gambiae* and *An. arabiensis* and alternative karyotypes got fixed in the two species after their divergence. The coalescent time of 2L+^{a} and 2La is estimated to be around 3Ne, which was predicted to be a good age to detect selection from estimating coalescent time between alternative arrangements [47]. However, our F_{ST}-outlier analysis showed very limited power in 2La region, with a slightly higher power in the younger 2Rb region (see Fig. 4). The reason might rise from the fact that the introgression age of 2La is too short compared to its origin time so that gene flux between heterokaryotypes is not frequent enough to reduce the divergence between neutrally evolving regions inside inversions. 2Rb, in which the background divergence is not significantly older than the time of introgression, showed clearer patterns of regions with significantly higher divergence.

With the new approach, we gained equal power of detecting selection in both of the inversions and identified more loci/regions under selection in 2La than 2Rb (Fig. 3). More interestingly, the signature overwhelmingly showed more prevalent selection in sites associated with *S* rather than *I* in 2La, which has never been identified before. This might not be surprising given 2La's origin history. Past molecular studies have shown that 2La is the ancestral arrangement [48], from which 2L+^{a} arose and got fixed in *Anopheles gambiae* [31]. The new phylogenomic study on the seven species in the complex even challenged the long held belief of direction of introgression, proposing the possibility that 2La might have been a retained polymorphism in in *An. gambiae* and introgressed into *An. arabiensis* [43]. Although our study does not provide evidence for the direction of introgression, the fact that the younger 2L+^{a} arrangement shows stronger selection signal highlights the important fitness advantage associated in wet forest environments, which was previously overlooked in studies of 2La inversions[49]. Therefore, our data suggest that the alternative arrangements have each facilitated either wet or dry habitat adaptation at different stages of the species history. Additional genomic study will be needed to pinpoint the exact genes that are under selection given such inferences are beyond our dataset given the relatively low density of Radtag markers (on average ~10kb between adjacent markers).

Adaptation from mosaic genotypes

With the increasing availability of physical maps among different species, the important role of chromosomal inversions in maintaining adaptive divergence has been demonstrated in many systems, such as controlling flowering differences in *Mimulus guttatus* between different ecotypes [50] and wing patters that form Batesian mimicry in *Heliconius numata* [51]. The unique aspect of adaptation via polymorphic inversions in Sub-Saharan mosquito species *Anopheles gambiae* is the prevalent introgression and sharing of inversions among sibling species [31,49], which has posed challenges in recovering phylogenetic relationships within the *Anopheles gambiae* complex [31,52-54].

Our study showed how different sets of adaptive loci for different habitat (e.g., dry vs. wet) can be maintained in alternative rearrangements throughout a widespread species. The fact that the species complex does not have complete reproductive isolation and that they "borrow" pre-adapted inversions from each other while exploring new environments provides an interesting example of how adaptation leads to mosaic genotypes instead of new species, especially for species with big populations and high connectivity. This mode of adaptation coincides with recent theories that predict that mechanisms that reduce or suppress recombination, or increase linkage between co-adaptive and maladapted genotypes, will be advantageous because the gene complex can avoid being swamped when gene flow persists between populations that are under divergent selection [16,55-57].

Material and Methods

Sample collection and DNA extraction

Mosquitoes were collected indoors at each site using either aspirators or insecticide spray and preserved individually in 0.5ml tubes containing 100% ethanol. Morphologies of each sample were examined according to Gillies and Meillon [58] and Gillies and Coetzee [59] under dissecting microscope before DNA extraction with QIAamp DNA Mini Kit, whose yield ranged from 10-200 ng per sample.

Molecular identification of species and karyotypes

The species status of each sample (i.e., *gambiae, coluzzi,* or *arabiensis*) was determined by a PCR-RFLP method following Fanello et al. [60]. Briefly, the species was identified by the difference in the number of bands and/or fragment length after *HhaI* digestion of part of the integenic spacer (IGS) of the ribosomal DNA PCR products. The presence of inversions was determined by PCR of unique breakpoint regions of alternative arrangements. For 2La, primers were chosen to amplify a 492 bp region of 2La distal breakpoint and a 207 bp product from $2L+^{a}$ proximal breakpoint [28]. For *2Rb,* three primers amplify a 429 bp fragment on 2Rb breakpoint and a 630bp fragment on $2R+^{b}$ breakpoint [61]. If only one of the two PCR bands is present on a gel electrophoresis, then the sample is considered to be homokaryotype of one arrangement; alternatively, if both bands are present with similar brightness, the sample is considered to be heterokaryotype.

ddRAD library preparation and sequence analysis

Genomic DNA from each sample was individually barcoded and used in a reduced complexity library for Illumina sequencing using a double digestion Restriction Associated DNA sequencing procedure (ddRADseq; for details see [34]. Briefly, DNA was digested with the two most frequent restriction enzymes, MluCI and MseI, to maximize the number of unique short fragments. The digested products were then ligated by part of an Illumina adaptor sequence and a unique barcode. Ligation products were pooled among samples and size-selected between 340 and 420 base pairs (excluding adaptor lengths) using a Pippin Prep (Sage Science) machine. The targeted-size ligation products were amplified by iProof™ High-Fidelity DNA Polymerase (BIO-RAD) with 12 cycles. The library was sequenced in two lanes on the Illumina HiSeq2000

platform to generate paired-end 100 base pair reads. Sequences were identified to each sample based on the barcodes. Only reads with an average quality score of at least 30 (Phred) and an unambiguous barcode and restriction cut site were retained.

After filtering, sequences were mapped to the *Anopheles gambiae* reference genome AGam30 [62] using BWA-MEM algorithm in bwa with default settings [63]; mappings with quality scores above 10 using SAMTOOLS [64] were retained. SNPs were called from mapped contigs and genotypes were assigned using a maximum-likelihood statistical model [65,66] with the STACKS v1.03 pipeline [67]; default settings were used except where noted below. Specifically, loci (termed as "stacks" in the program) were identified from genomic locations with mappings of at least 5 copies of RAD sequences in each individual using the PSTACKS program to ensure credible calling of heterozygous SNPs in an individual [67]. A catalog of loci was built with the CSTACKS program from the PSTACKS output files across individuals to check the presence or absence of a particular locus at a genomic location. We retained loci that were present in at least 50% of all the samples and no more than two haplotypes per locus within each sample. All customized scripts, setting files for programs and genomic data are available on Dryad under doi: xxxxxxxxxx.

Population genetic structure of collinear and inverted regions

Geographic structure of *An. gambiae* populations were examined by measuring population divergence and performing principal component analysis (PCA). We performed separate analyses on collinear regions and inverted regions of each chromosome. Weir and Cockerham's F_{ST} (1984) and nucleotide diversity (π) were estimated on a per-site basis as well as for a window of 150kb (50kb steps were used to slide along the chromosome) using the POPULATIONS program in the STACKS pipeline [67]. SNPs were thinned to be at least 1000bp apart on the genome and imported into adegenet 1.4-1 package [68] in R [69] for PCA analyses. Only SNPs that are present in all populations and at least 80% of all individuals were included in the study. Missing genotypes were represented by the average value for the PCA analyses. Based on PCA results, discriminant analysis of principal components (DAPC; [70]) were run to determine genetic clusters within the species without prior assumptions on the model of population subdivision. DAPC runs K-means clustering on the transformed PCs to identify groups of individuals that maximize between-group genetic variation while minimizing within-group variation. The best

supported number of clusters is then determined by the comparison of model likelihoods through Bayesian Information Criterion (BIC) similar to the program STRUCTURE [71,72]. The agreement between genetic clusters inferred from inverted regions and molecular karyotyping assignment was also checked to assess the reliability of PCR identification methods.

Demographic history of collinear and inverted regions

Inference of demographic history implemented in FASTSIMCOAL2 [35] calculates composite-likelihood of joint-SFS across populations under user-specified demographic scenario by parameterized simulations sampled from priors and optimizes the parameter estimation through a conditional maximization algorithm (ECM). The derived and ancestral states of the SNPs were inferred from the comparison of four species in the *An. gambiae* complex: *An. gambiae* s.s., *An. arabiensis*, *An. quadriannulatus*, and *An. merus*. Whole genome scaffolds of the latter three species were mapped to *gambiae* genome using MUMMER 3.23 [73,74] and unique alignments were kept. Majority state of the diallelic SNP among four species was considered ancestral and multi-states SNPs were filtered. In order to maximize the number of SNPs included and ensure reasonable running time for each scenario, we excluded individuals with less complete sequencing coverage and subsampled SNPs in each case to infer region specific demographic histories. Point estimates were obtained from the run with the best maximum likelihood out of 50 realizations. Confidence intervals of parameter estimates were then obtained by 100 parametric bootstrapping runs from the point estimates.

Coalescent history between A. gambiae and A. arabiensis in collinear regions

We first estimated the divergence time (T_{div}), introgression rate (m) and recent population expansion (N_{cur}, T_{exp}) (Fig. 1b) of *An. gambiae* and *An. arabiensis* using joint-SFS built from Chromosome 3 as proxies for collinear regions (X chromosome has a different effective population size and different selection regime). In order to include more SNPs with missing data while obtaining an accurate estimation of the site frequency in the population, all SNPs were subsampled to 40 copies in *An. gambiae* and 6 in *An. arabiensis*. One variable SNP per RADtag were chosen to build the site frequency spectrum to ensure that SNPs are not linked. We fixed the population size (Ne) of *An. gambiae* to estimate other free parameters (see also [35]) (Fig. 1b, c); this parameter was set to ~Ne = 750,000 using a mutation rate of 3.5E-9 per base per generation (estimation from Drosophila resequencing, [75]) given that *An. gambiae* was

estimated to have a nucleotide diversity (π) of 0.01024 ± 4.0E-5 from all Radtags. Due to pervasive introgression between the *An. arabiensis* and *An. gambiae* (Besansky et al. 2003), we estimated the rate of gene flow between the two species after the divergence from their common ancestor (Fig. 1b).

Coalescent history between A. gambiae and A. arabiensis in regions with inversion polymorphisms

We estimated the divergence time between standard chromosomes and inverted chromosomes (T_{IS}), time of introgression for inversions (T_{int}), and recombination rates between alternative rearrangements (r) using joint-SFS built from 2La or 2Rb regions while fixing other parameters that were estimated from the collinear region. SNPs were subsampled to 20 copies in *I* and *S*, and 6 in *An. arabiens*is. We fixed the parameters that have been estimated from collinear region models and focused on inversion specific parameters because collinear region has a larger SNP dataset. *I* and *S* of *An. gambiae* were treated as separate populations, but connected by a severely reduced recombination (r) (Fig. 1c). The same introgression rate estimated from collinear regions was applied to that between the inverted copy *I* in *An. gambiae* and *An. Arabiensis*, while the introgression rate between *S* in *An. gambiae* and *I* in *An. Arabiensis* was set to zero.

Selection signature in inversion regions

Based on reconstructed demographic histories for inversion regions, we obtained neutral expectations of population genomic measures through simulations. These measures can then be compared against empirical data to detect selection. We first applied F_{ST}-outlier analyses similar to traditional approaches. One difficulty lies in the heterogeneity of recombination rates inside inversions between heterokaryotypes (i.e., recombination rates are higher in the center and decrease sharply towards the breaking point region [76]), whereas the recombination rates estimated from SFS demographic modeling was an average. In order to make neutral simulations more realistic, we adjusted the recombination rate to be higher in the center region and lower on the two sides. First, empirical 2Mb F_{ST}-windows in 150kb-step were calculated and fed into a smooth spline function in R to get fitted values for each 150kb segment. Recombination rate was adjusted to the value that generated the empirical mean F_{ST}-value for the segment. 1000 demographic simulations were carried out using estimated parameters for each segment to generate 100bp DNA sequences. The population divergence measures, F_{ST}, between *I* and *S* were

estimated for the sequences. Lastly, empirical divergence measures for each locus were compared against the range of values from simulations to identify outliers.

Our second approach utilized sets of summary statistics to detect selection through simulation-trained discriminant function classification analysis. Three scenarios were run for 1000 replications under the current demographic model using MSMS [77] : a) pure neutral evolution; b) selection occurred on the branch of *An. arabiensis* and inverted chromosomes (I); c) selection occurred on the branch of *An. gambiae* and continued in standard chromosomes. Selection started from when two species diverged with selection coefficient ranging from 0.01 to 0.0001 (Fig. 2c). Selected locus is located in the center of a 50kb-long simulated region. We tested two ways of building discriminant functions: 1) based on summary statistics of one neutral locus that is close to the selected locus; 2) based on an average of summary statistics across several neutral loci of a region that contains selected locus. For the first approach, we sampled 100bp long sequences that are located 5Kb away from the selected locus (an average distance between empirical adjacent Radtags in our data). 9 summary statistics, including heterozygosity (H), θ_π of each population, and population pairwise-F_{ST}, were calculated for each simulated sequence. For the second approach, we sampled random sets of 100bp short sequences across the entire 50kb according empirical Radtag distributions and calculated an average of these summary statistics. In both cases, discriminant functions using principle components transformed from summary statistics (DAPC; [70]) were built based on the simulated training sets to differentiate the three scenarios. We then used the discriminant function to predict which scenario each empirical locus/region belonged to based on their estimated summary statistics.

We compared the performance of the outlier test with the discriminant function classification analysis by evaluating their false-positive rate, detection power (true positive rate) and false discovery rates. We define these rates, according to Lotterhos and Whitlock [20], as follows: false-positive rate is ratio of the number of significant neutral loci and the total number of neutral loci; the detection power is the ratio of identified selected loci and the total number of selected loci; the false-discovery rate is the number of significant neutral loci divided by the total number of significant loci. Three levels of selected loci were tested in simulated datasets to calculate power and false-discovery rates. Specifically, out of 1000 otherwise neutral loci, 1%, 5%, or 10% are simulated under selection. In our first approach, cut-off values for loci to be considered

significant were 95% and 99% of neutral distributions. Hence, we set the false-positive rates to be compared in our first approach as 5% and 1% (see Fig. 4). For the second approach, the cut-off values to be classified as an outlier (i.e., false positive rate) are not arbitrarily decided. Instead, the percentage of correct assignment from discriminant functions represent how likely different scenarios can be differentiated. Hence, the error rate in assigning neutral cases to selection cases is the false-positive rates.

Acknowledgments

We appreciate useful discussions with Anthony J. Cornel, Yooksook Jin, Huateng Huang, Mark Kirkpatrick, and Mark Christie. We would also like to thank Seraphin Menzepoh and Kevin Njabo for tremendous help in the field work. Financial support was provided by a NSF Doctoral Dissertation Improvement Grant (DEB-1210359) to Q. He and L. L. Knowles.

References

1. Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. Nature Reviews Genetics 8: 857-868.
2. Hohenlohe PA, Phillips PC, Cresko WA (2010) Using population genomics to detect selection in natural populations: key concepts and methodological considerations. International Journal of Plant Sciences 171: 1059-1071.
3. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832-837.
4. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913-918.
5. Voight BF, Kudaravalli S, Wen XQ, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biology 4: 446-458.
6. Kim Y, Nielsen R (2004) Linkage disequilibrium as a signature of selective sweeps. Genetics 167: 1513-1524.
7. Wang ET, Kodama G, Baidi P, Moyzis RK (2006) Global landscape of recent inferred Darwinian selection for *Homo sapiens*. Proceedings of the National Academy of Sciences of the United States of America 103: 135-140.
8. Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, et al. (2005) Genomic regions exhibiting positive selection identified from dense genotype data. Genome research 15: 1553-1565.
9. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, et al. (2005) Genomic scans for selective sweeps using SNP data. Genome Research 15: 1566-1575.
10. Ronen R, Udpa N, Halperin E, Bafna V (2013) Learning natural selection from the site frequency spectrum. Genetics 195: 181-193.
11. Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G (2008) LOSITAN: A workbench to detect molecular adaptation based on a F(st)-outlier method. BMC Bioinformatics 9.
12. Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. Molecular Ecology 13: 969-980.
13. Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, et al. (2010) Detecting selection in population trees: The Lewontin and Krakauer test extended. Genetics 186: 241-U406.
14. Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics 180: 977-993.
15. Gunther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. Genetics 195: 205-220.
16. Kirkpatrick M, Barton N (2006) Chromosome inversions, local adaptation and speciation. Genetics 173: 419-434.
17. Yeaman S (2013) Genomic rearrangements and the evolution of clusters of locally adaptive loci. Proceedings of the National Academy of Sciences 110: E1743-E1751.
18. Cheng C, White BJ, Kamdem C, Mockaitis K, Costantini C, et al. (2012) Ecological Genomics of Anopheles gambiae Along a Latitudinal Cline: A Population-Resequencing Approach. Genetics 190: 1417-1432.
19. Beaumont MA (2005) Adaptation and speciation: what can F_{ST} tell us? Trends in Ecology & Evolution 20: 435-440.
20. Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. Molecular ecology 23: 2178-2192.
21. Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. Proceedings of the Royal Society of London Series B: Biological Sciences 263: 1619-1626.

22. Rafajlović M, Klassmann A, Eriksson A, Wiehe T, Mehlig B (2014) Demography-adjusted tests of neutrality based on genome-wide SNP data. Theoretical Population Biology.
23. Coluzzi M, Sabatini A, Petrarca V, Dideco MA (1979) Chromosomal differentiation and adaptation to human environments in the *Anopheles gambiae* complex. Transactions of the Royal Society of Tropical Medicine and Hygiene 73: 483-497.
24. Czeher C, Labbo R, Vieville G, Arzika I, Bogreau H, et al. (2010) Population Genetic Structure of Anopheles gambiae and Anopheles arabiensis in Niger. Journal of Medical Entomology 47: 355-366.
25. Lanzaro GC, Toure YT, Carnahan J, Zheng LB, Dolo G, et al. (1998) Complexities in the genetic structure of Anopheles gambiae populations in west Africa as revealed by microsatellite DNA analysis. Proceedings of the National Academy of Sciences of the United States of America 95: 14260-14265.
26. Lehmann T, Hawley WA, Grebert H, Collins FH (1998) The effective population size of Anopheles gambiae in Kenya: Implications for population structure. Molecular Biology and Evolution 15: 264-276.
27. White BJ, Hahn MW, Pombi M, Cassone BJ, Lobo NF, et al. (2007) Localization of candidate regions maintaining a common polymorphic inversion (2La) in *Anopheles gambiae*. PLoS Genet 3: e217.
28. White BJ, Santolamazza F, Kamau L, Pombi M, Grushko O, et al. (2007) Molecular karyotyping of the 2LA inversion in Anopheles gambiae. American Journal of Tropical Medicine and Hygiene 76: 334-339.
29. Gray EM, Rocca KAC, Costantini C, Besansky NJ (2009) Inversion 2La is associated with enhanced desiccation resistance in Anopheles gambiae. Malaria Journal 8.
30. Simard F, Ayala D, Kamdem G, Pombi M, Etouna J, et al. (2009) Ecological niche partitioning between Anopheles gambiae molecular forms in Cameroon: the ecological side of speciation. BMC Ecology 9: 17.
31. Besansky NJ, Krzywinski J, Lehmann T, Simard F, Kern M, et al. (2003) Semipermeable species boundaries between *Anopheles gambiae* and *Anopheles arabiensis*: Evidence from multilocus DNA sequence variation. Proceedings of the National Academy of Sciences of the United States of America 100: 10818-10823.
32. Neafsey D, Lawniczak M, Park D, Redmond S, Coulibaly M, et al. (2010) SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes. Science 330: 514-517.
33. White BJ, Cheng C, Sangaré D, Lobo NF, Collins FH, et al. (2009) The population genomics of trans-specific inversion polymorphisms in Anopheles gambiae. Genetics 183: 275-288.
34. 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.
35. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. PLoS genetics 9: e1003905.
36. Nosil P, Feder JL (2012) Genomic divergence during speciation: causes and consequences Introduction. Royal Society Philosophical Transactions Biological Sciences 367: 332-342.
37. Harr B (2006) Genomic islands of differentiation between house mouse subspecies. Genome Research 16: 730-737.
38. Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, et al. (2012) Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. Philosophical Transactions of the Royal Society B: Biological Sciences 367: 343-353.
39. Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic islands of speciation in Anopheles gambiae. Plos Biology 3: 1572-1578.
40. Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Molecular ecology.
41. Beaumont MA, Zhang WY, Balding DJ (2002) Approximate Bayesian computation in population genetics. Genetics 162: 2025-2035.

42. Wegmann D, Leuenberger C, Excoffier L (2009) Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics 182: 1207-1218.
43. Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, et al. (2015) Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science 347.
44. Lang M, Murat S, Clark AG, Gouppil G, Blais C, et al. (2012) Mutations in the neverland gene turned Drosophila pachea into an obligate specialist species. Science 337: 1658-1661.
45. Auton A, McVean G (2007) Recombination rate estimation in the presence of hotspots. Genome research 17: 1219-1227.
46. Kamali M, Xia A, Tu Z, Sharakhov IV (2012) A new chromosomal phylogeny supports the repeated origin of vectorial capacity in malaria mosquitoes of the Anopheles gambiae complex. PLoS pathogens 8: e1002960.
47. Guerrero RF, Rousset F, Kirkpatrick M (2012) Coalescent patterns for chromosomal inversions in divergent populations. Philosophical Transactions of the Royal Society B: Biological Sciences 367: 430-438.
48. Sharakhov IV, White BJ, Sharakhova MV, Kayondo J, Lobo NF, et al. (2006) Breakpoint structure reveals the unique origin of an interspecific chromosomal inversion (2La) in the Anopheles gambiae complex. Proceedings of the National Academy of Sciences of the United States of America 103: 6258-6262.
49. 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.
50. Lowry DB, Willis JH (2010) A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. Plos Biology 8: 14.
51. Joron M, Frezal L, Jones RT, Chamberlain NL, Lee SF, et al. (2011) Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. Nature 477: 203-206.
52. Besansky NJ, Powell JR, Caccone A, Hamm DM, Scott JA, et al. (1994) Molecular phylogeny of the *Anopheles gambiae* complex suggests genetic introgression between principal malaria vectors. Proceedings of the National Academy of Sciences of the United States of America 91: 6885-6888.
53. Bhutkar A, Gelbart WM, Smith TF (2007) Inferring genome-scale rearrangement phylogeny and ancestral gene order: a Drosophila case study. Genome Biology 8.
54. White BJ, Collins FH, Besansky NJ (2011) Evolution of Anopheles gambiae in relation to humans and malaria. Annual Review of Ecology, Evolution, and Systematics 42: 111-132.
55. Aeschbacher S, Bürger R (2014) The effect of linkage on establishment and survival of locally beneficial mutations. Genetics 197: 317-336.
56. Barton N (1995) A general model for the evolution of recombination. Genetical research 65: 123-144.
57. Yeaman S, Whitlock MC (2011) The genetic architecture of adaptation under migration-selection balance. Evolution 65: 1897-1911.
58. Gillies MT, Meillon Bd (1968) The Anophelinae of Africa South Or the Sahara (Ethiopian Zoogeographical Region): South African Institute of Medical Research Johannesburg.
59. Gillies M, Coetzee M (1987) A Supplement to the Anophelinae of Africa South of the Sahara. Publications of the South African Institute for Medical Research 55: 1-143.
60. Fanello C, Santolamazza F, Della Torre A (2002) Simultaneous identification of species and molecular forms of the Anopheles gambiae complex by PCR-RFLP. Medical and veterinary entomology 16: 461-464.
61. Lobo NF, Sangare DM, Regier AA, Reidenbach KR, Bretz DA, et al. (2010) Breakpoint structure of the Anopheles gambiae 2Rb chromosomal inversion. Malaria Journal 9.
62. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, et al. (2002) The genome sequence of the malaria mosquito Anopheles gambiae. Science 298: 129-+.
63. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754-1760.
64. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078-2079.

65. Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. G3: Genes, Genomes, Genetics 1: 171-182.
66. Hohenlohe PA, Catchen J, Cresko WA (2012) Population genomic analysis of model and nonmodel organisms using sequenced RAD tags. Data Production and Analysis in Population Genomics: Springer. pp. 235-260.
67. Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. Molecular ecology 22: 3124-3140.
68. Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24: 1403-1405.
69. R Core Team (2012) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
70. Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC genetics 11: 94.
71. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567-1587.
72. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959.
73. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, et al. (1999) Alignment of whole genomes. Nucleic Acids Research 27: 2369-2376.
74. Delcher AL, Phillippy A, Carlton J, Salzberg SL (2002) Fast algorithms for large-scale genome alignment and comparison. Nucleic acids research 30: 2478-2483.
75. Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, et al. (2009) Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. Genome research: gr. 091231.091109.
76. Navarro A, Betran E, Barbadilla A, Ruiz A (1997) Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. Genetics 146: 695-709.
77. Ewing G, Hermisson J (2010) MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics 26: 2064-2065.

Figure Legends

Figure 1. Schematic illustration of study design and population demographic scenarios of *Anopheles gambiae* and *An. arabiensis*. a) procedures involved in the detection of targets of selection; b) in collinear regions, *An. gambiae* and *An. arabiensis* diverged from a common ancestor at time T_{div} with a population size *Ne*. They experienced recent population expension at time T_{exp} to the current population size N_{cur}. The two species have constant gene flow since divergence; c) in regions with alternative arrangements, the arrangement, *S*, split with the alternative arrangement, *I*, at time T_{IS}; at time T_{int}, introgression occurred so that *I* split into two lineages, *An. gambiae* and *An. Arabiensis*, respectively. Alternative arrangements within *An. gambiae* have reduced recombination rate (*r*), while within the same arrangements *I*, the introgression rate (*m*) is the same as that in collinear regions. Parameters that maintain the same in inverted and collinear regions are shown in gray, while inversion specific parameters are shown in black.

Figure 2. Outlier analysis of inverted region scan. a) and b), dots represent F_{ST} measures of each RADtag locus between *S* and *I* chromosomes in *An. gambiae* populations along the region. Shades of yellow show quantiles of 25%, 50%, 75%, 95%, 99% respectively, of simulated values of divergence measures under reconstructed demographic histories with region-adjusted recombination rate. c) and d) empirical distributions of F_{ST} between individuals with *S* and *I* chromosomes on inverted region (green) or collinear region (red). a) and c), 2La regions; b) and d), 2Rb regions;

Figure 3. Candidate loci and regions under selection. All the dots represent RADtag loci that are classified to have experienced selection in the *S* lineage (red) or *I* lineage (blue). Y-axis shows the posterior probability of such classifications. Dots with solid color are the ones with larger than 0.9 in posterior probabilities of assignment. Bars represent regions that have been classified as either experienced selection in the *S* lineage (red) or *I* lineage (blue). Width of each bar corresponds to 50kb in our analysis.

Figure 4. Comparison of detection power and false-discovery rates between two selection detection methods. Upper panel, 2La region; lower panel, 2Rb region. a), c), e), and g) are the results for the scenario in which selection occurs on the branch of *S*. b, d), f), and h) are the results for selection on the branch of *I*. The *x*-axis presents cases where the proportion of selected loci consists of 1%, 5%, and 10% out of all simulated loci (i.e., rest of the loci are simulated under neutral scenario). All selected loci are generated with $s = 0.01$. See Fig. S5&6 for cases where s = 0.001 & 0.0001. See methods for the details of how rates are calculated. The inset shows the false-positive rates for each method.

Supporting Information Captions

S1 Text. Detailed results on population genetic structures of collinear and inverted regions.

Figure S1. Sampling locations and species composition of *Anopheles gambiae* species complex. The area of each pie chart correspond to the sample size. Map color from blue to red stands for humid to dry areas.

Figure S2. Principle component analyses using SNPs in different collinear genomic regions. Color of the dots represent different populations. Red dots are individuals of *An. arabiensis*.

Figure S3. Principle component analyses of *An. gambiae* using SNPs in different collinear genomic regions. Color of the dots represent different populations.

Figure S4. Principle component analyses of *Anopheles gambiae* using SNPs from 2La and 2Rb. Left and Right panels are the result for 2La and 2Rb, respectively. Top panel is the result for PCA clustering of individuals from different populations. Middle panel finds the best number of clusters based on BIC scores. The bottom panel shows how divergent each cluster is from each other on the discriminant function space.

Figure S5. Comparison of detection power and false-discovery rates between two selection detection methods. Upper panel, 2La region; lower panel, 2Rb region. a), c), e), and g) are the results for the scenario in which selection occurs on the branch of *S*. b, d), f), and h) are the results for selection on the branch of *I*. The *x*-axis presents cases where the proportion of selected loci consists of 1%, 5%, and 10% out of all simulated loci (i.e., rest of the loci are simulated under neutral scenario). All selected loci are generated with $s = 0.001$. See methods for the details of how rates are calculated. The inset shows the false-positive rates for each method.

Figure S6. Comparison of detection power and false-discovery rates between two selection detection methods. Upper panel, 2La region; lower panel, 2Rb region. a), c), e), and g) are the results for the scenario in which selection occurs on the branch of *S*. b, d), f), and h) are the results for selection on the branch of *I*. The *x*-axis presents cases where the proportion of selected loci consists of 1%, 5%, and 10% out of all simulated loci (i.e., rest of the loci are simulated under neutral scenario). All selected loci are generated with $s = 0.0001$. See methods for the details of how rates are calculated. The inset shows the false-positive rates for each method.

S1 Table. Collection sites, coordinates and sampling sizes of *Anopheles gambiae*.

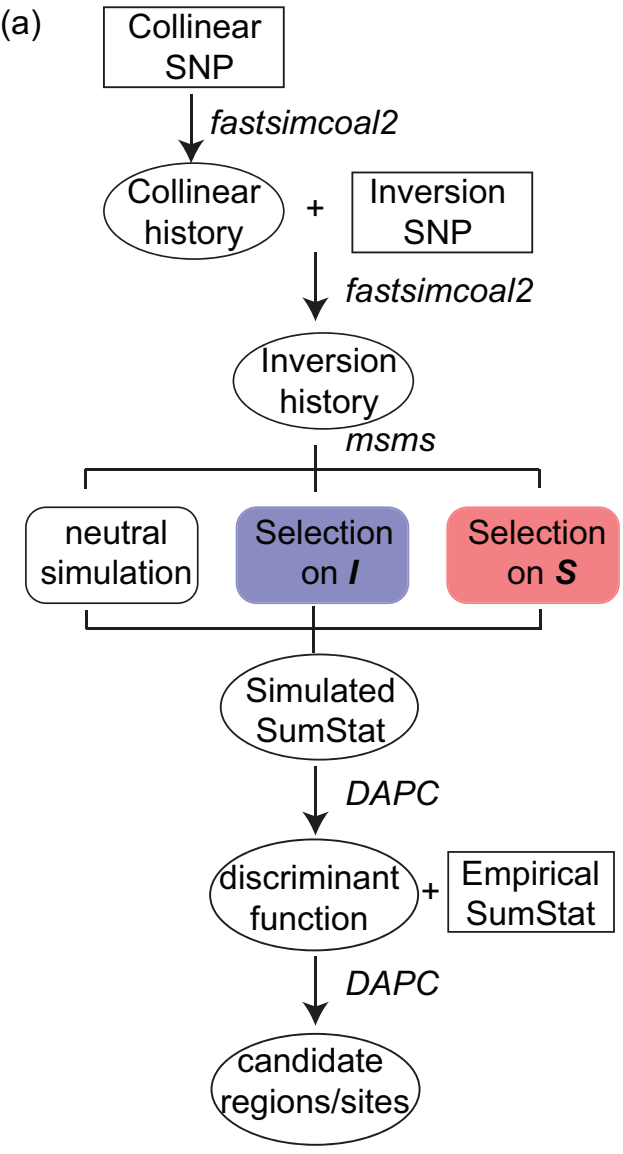
S2 Table. Power of differentiating selection form drift using average summary statistics of loci in a 5kb segment with/without selected locus inside.

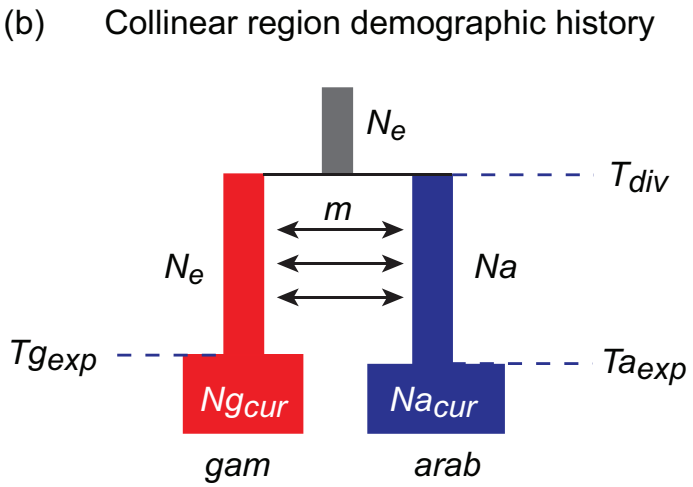
Tables

Table 1. Estimations of population genetic and demographic parameters using region-specific SFS implemented in FASTSIMCOAL2. All parameters were estimated relative to a fixed prior of *An. gambiae* (*Ne* = 750,000; see Material and Methods for details). See Fig. 1b, c for explanations of parameters. Note that the estimations of time are in the unit of generations.

Regions	Parameters	Point estimates	Relative to Ne	95% Confidence Interval	
Collinear	Na_{cur}	2,101,300	2.80	1,126,650	33,035,200
	Ng_{cur}	5,374,400	7.17	4,578,413	10,547,442
	Na	338,900	0.45	304,350	406,750
	m	1.75E-07		1.52E-08	2.70E-07
	Ta_{exp}	186,800	0.25	136,900	228,700
	Tg_{exp}	278,500	0.37	256,600	290,200
	T_{div}	1,052,500	1.40	824,600	1,240,400
2La	r	7.99E-08		9.25E-08	1.08E-07
	T_{int}	340,200	0.45	320,700	352,000
	T_{IS}	2,353,800	3.14	2236100	2568100
2Rb	r	4.79E-07		4.64E-07	5.15E-07
	T_{int}	591,800	0.79	570,100	625,200
	T_{IS}	1091300	1.46	916,600	1,140,500

Fig. 1
Click here to download Figure: Fig1.eps





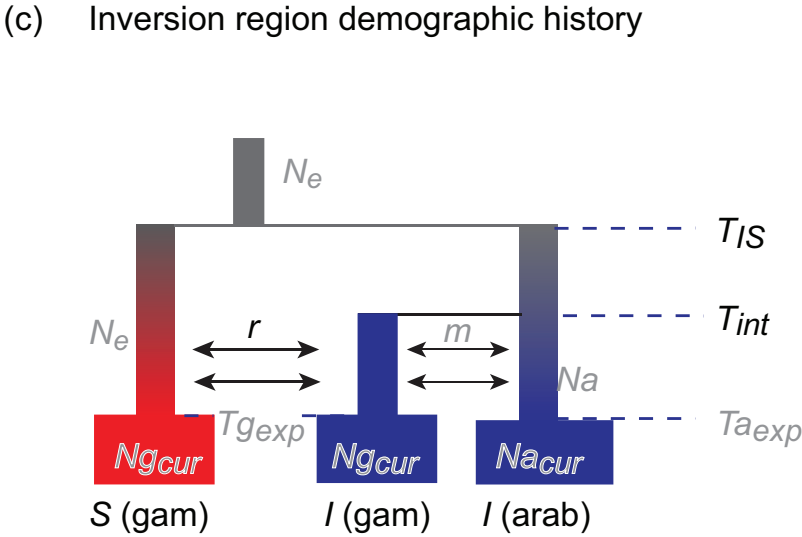


Fig.2
Click here to download Figure: Fig2.eps

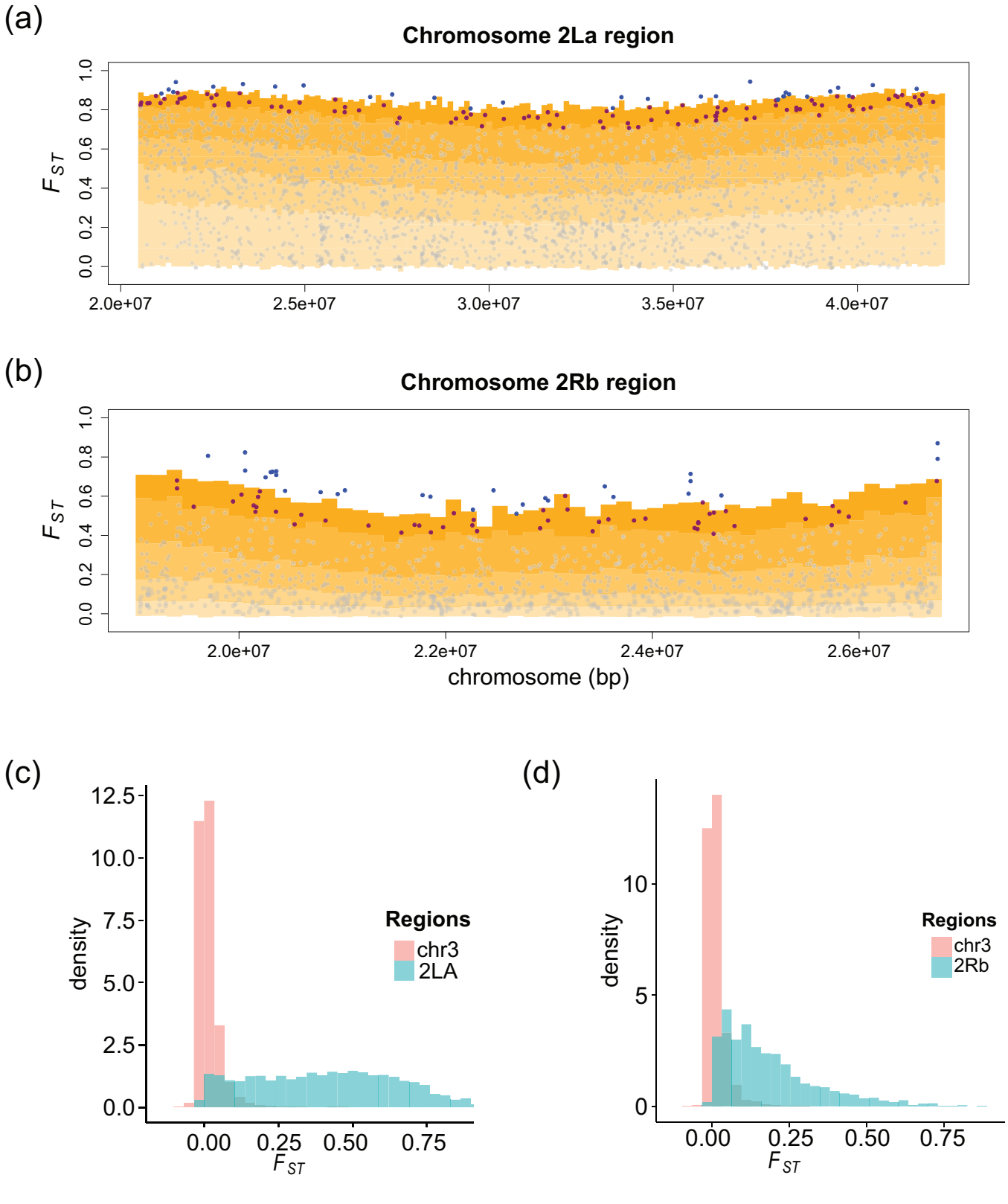


Fig.3
Click here to download Figure: Fig3.eps

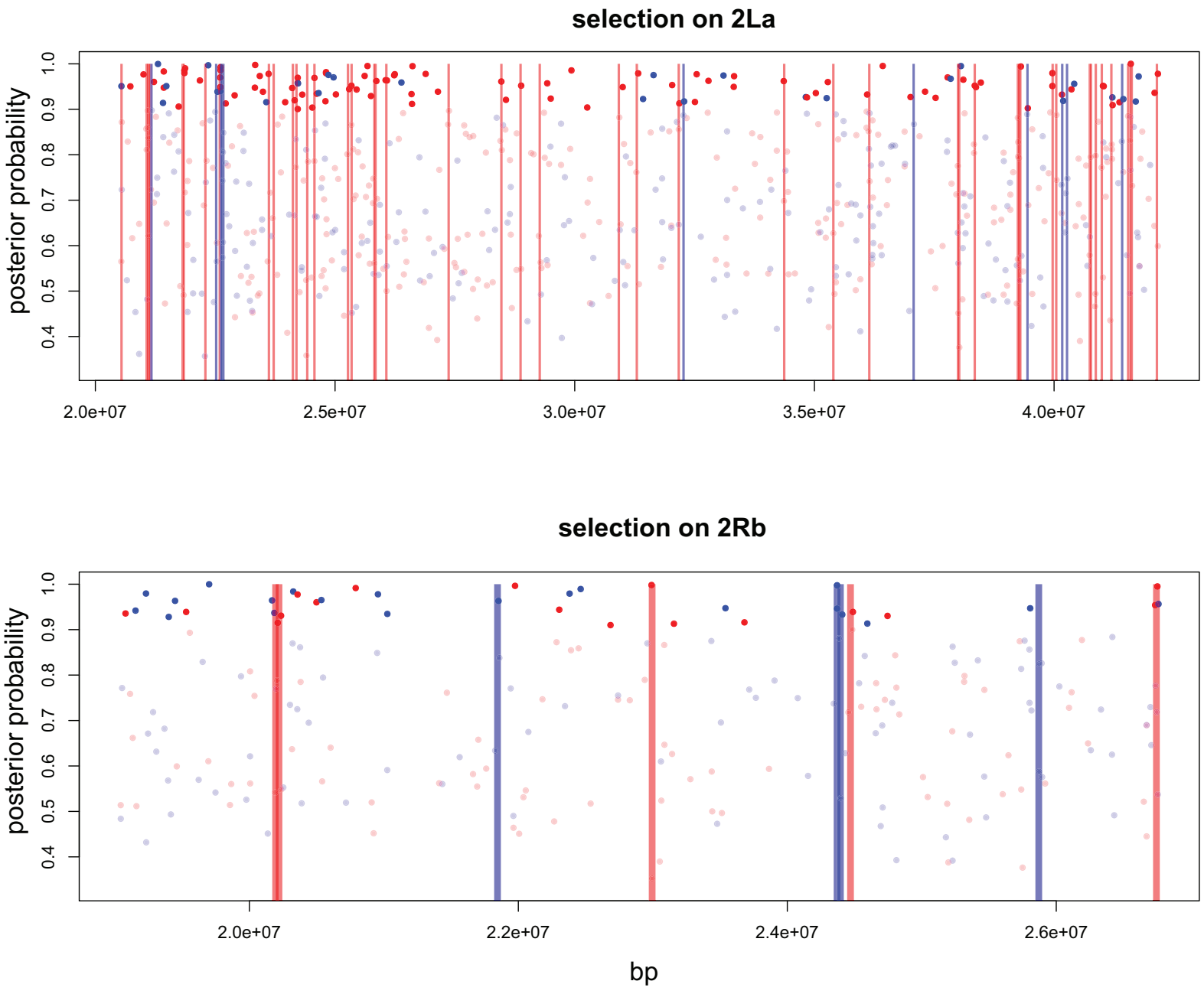


Fig.4

Click here to download Figure: fig4.eps

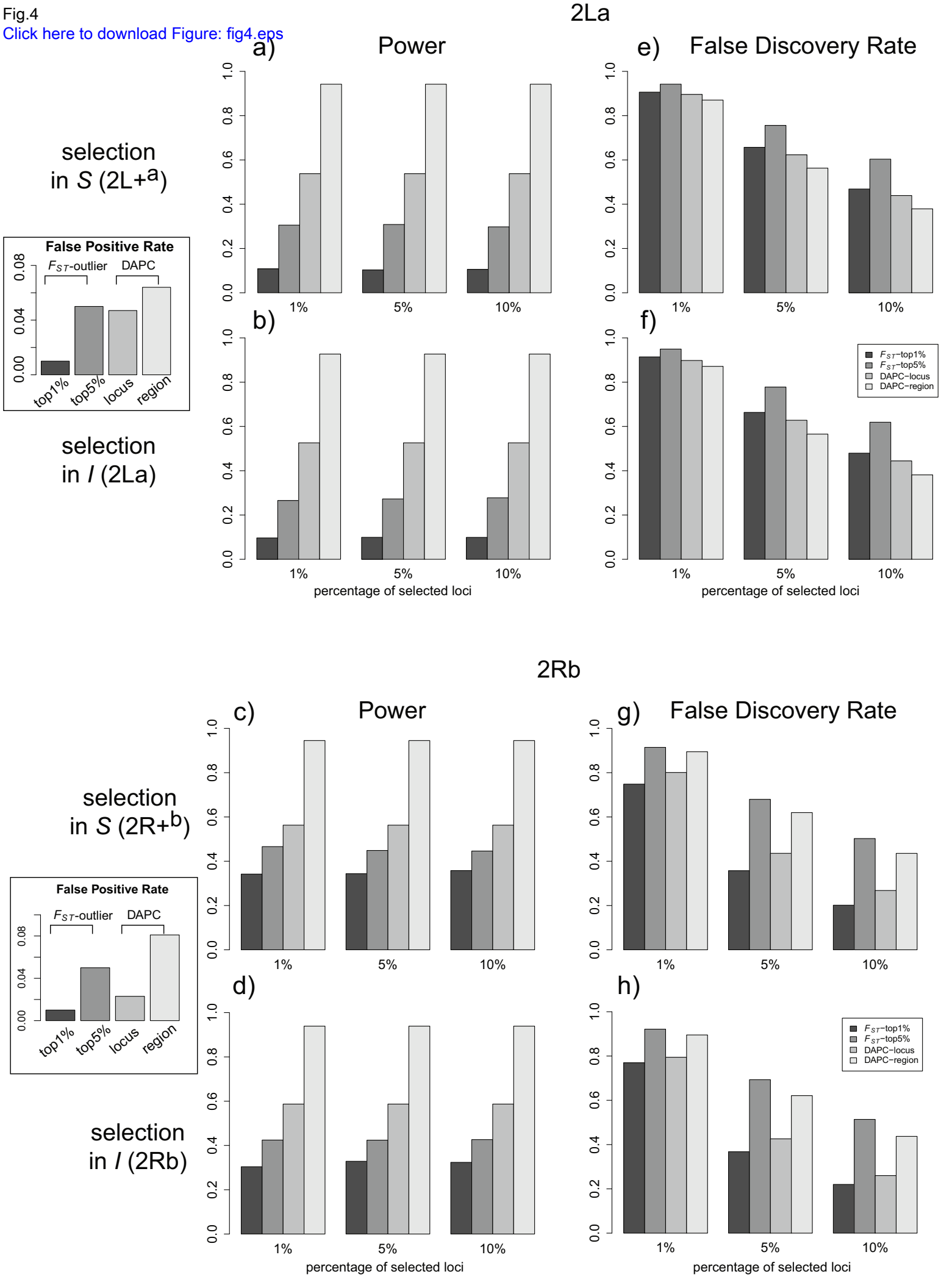


Fig. S1

Click here to download Figure: FigS1.eps

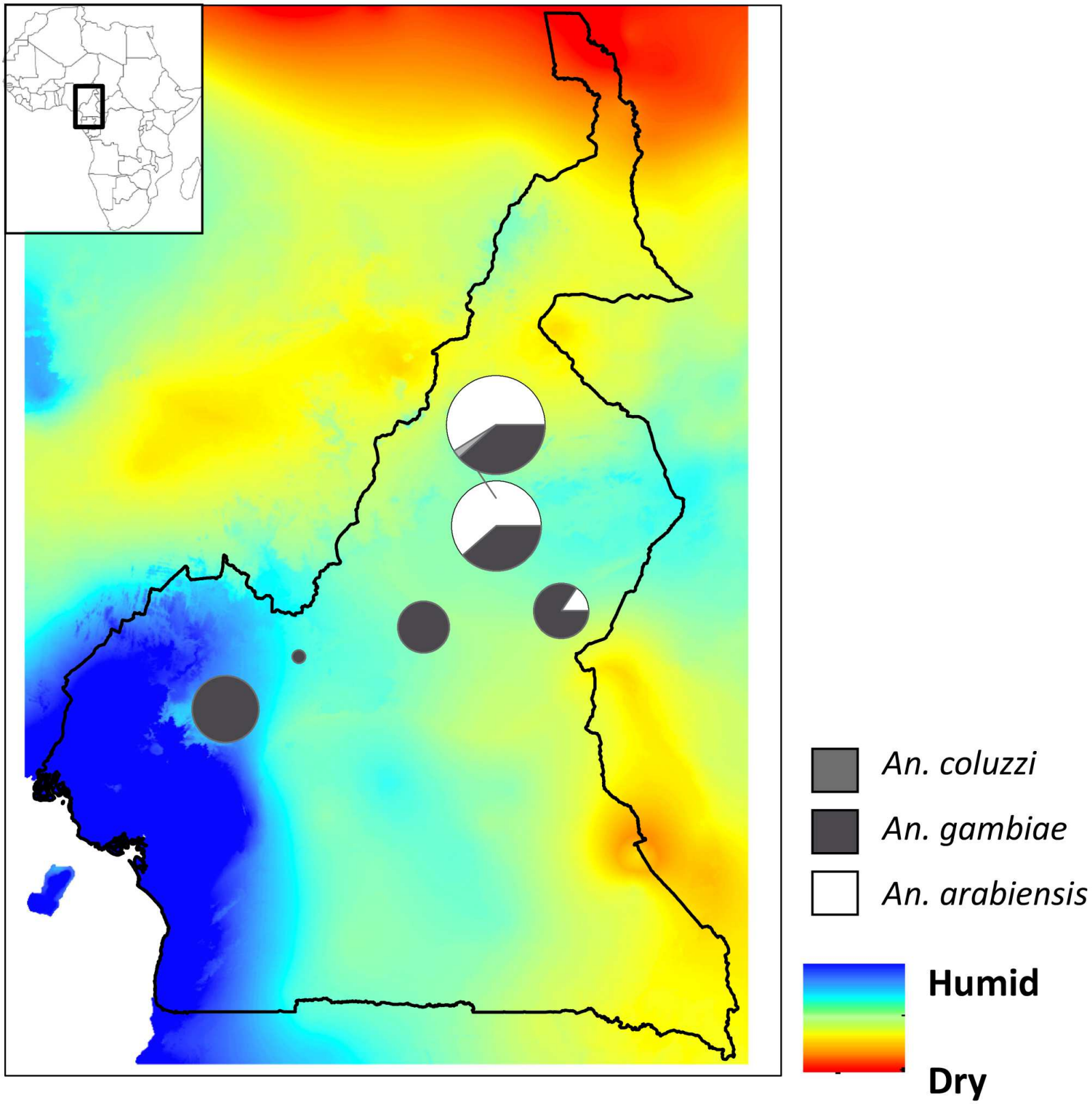
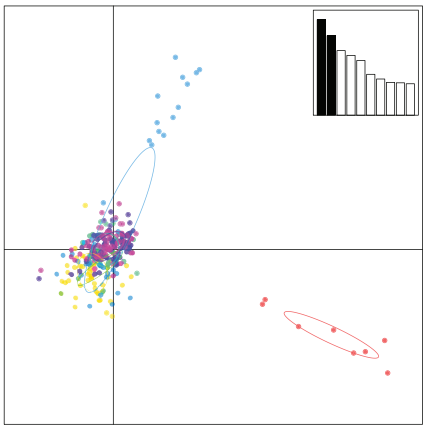
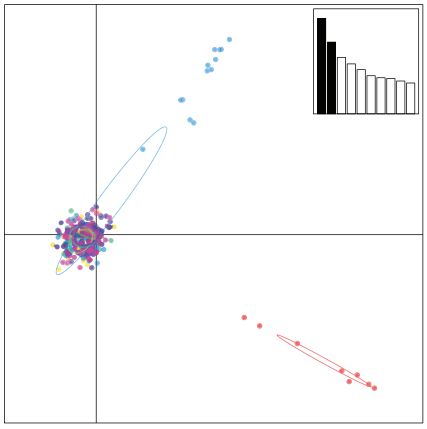


Fig. S2
Click here to download Figure: FigS2.eps

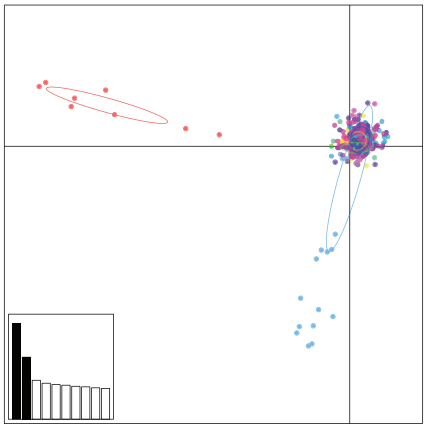
(a) 2L-collinear



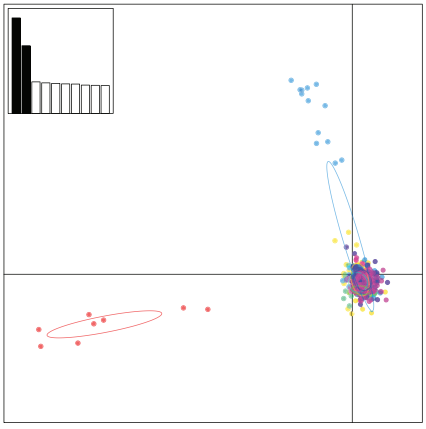
(b) 2R-collinear



(c) 3L



(d) 3R



(e) X

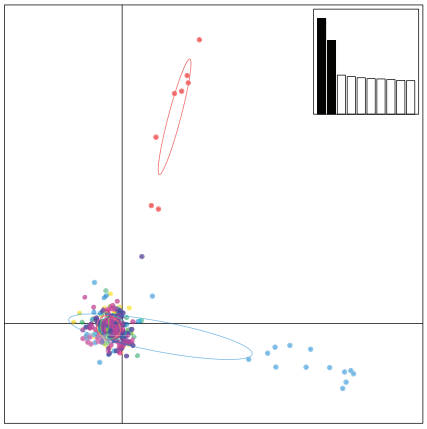


Fig. S3
Click here to download Figure: FigS3.eps

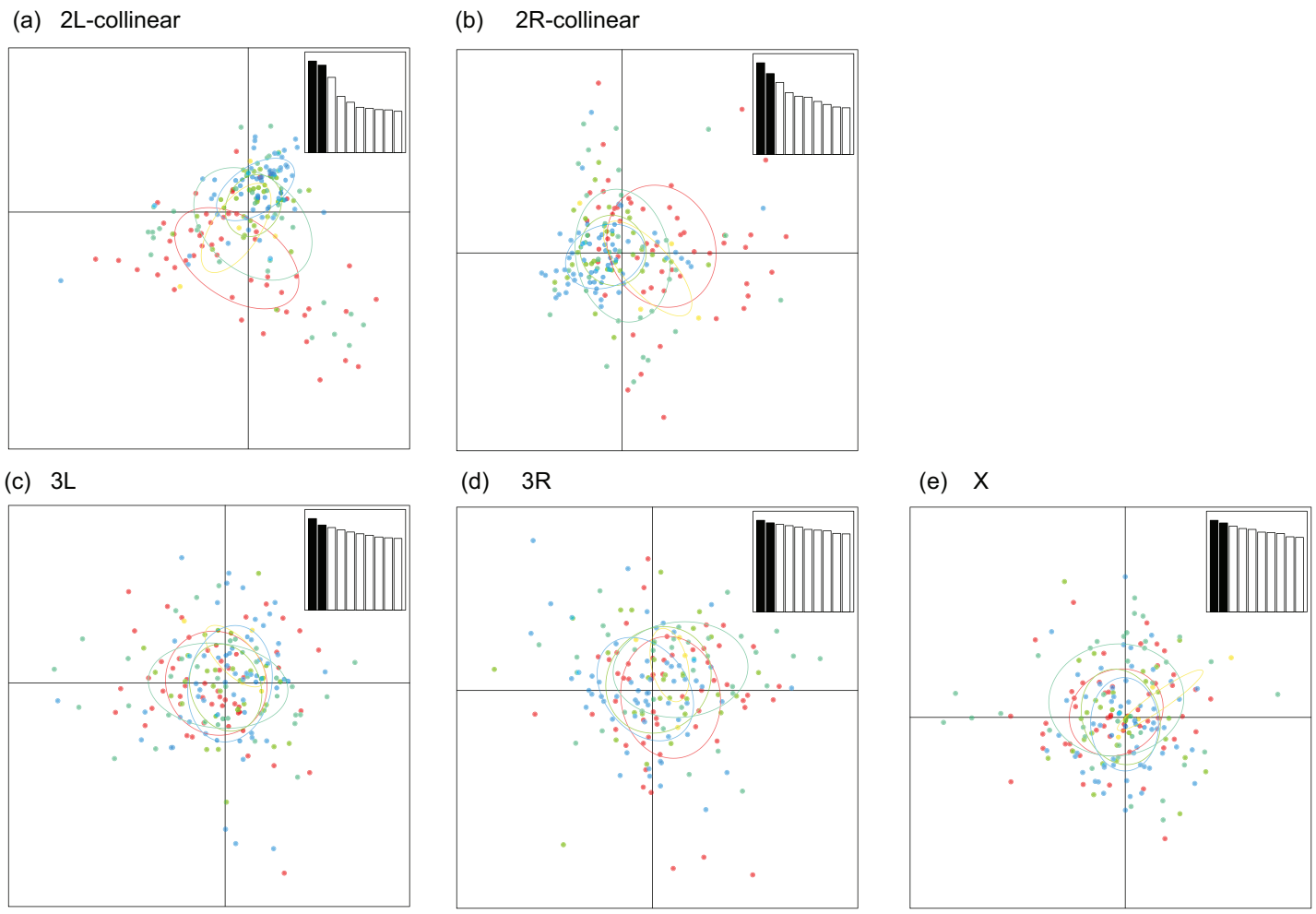


Fig. S4

Click here to download Figure: FigS4.eps

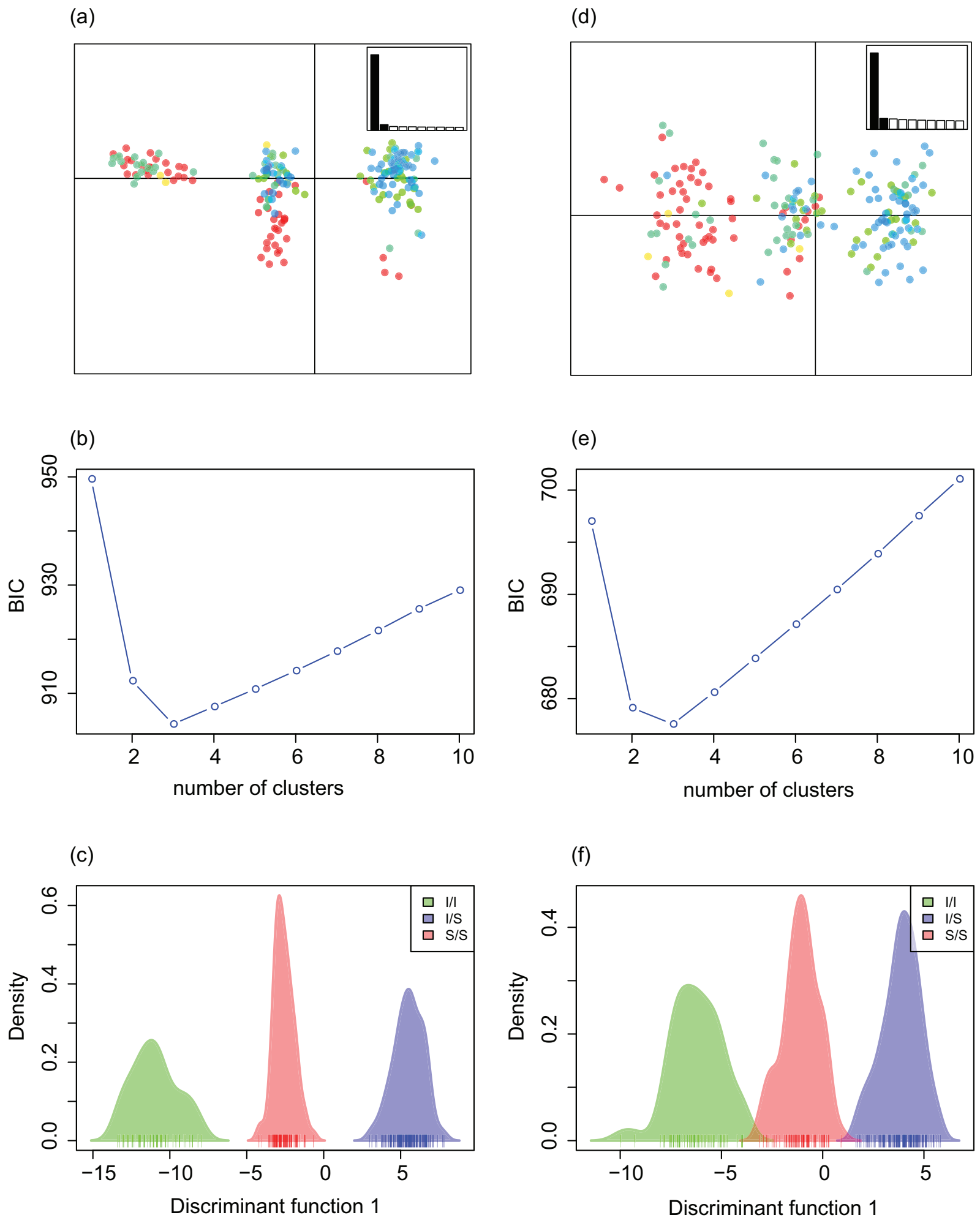


Fig. S5

Click here to download Figure: figS5.eps

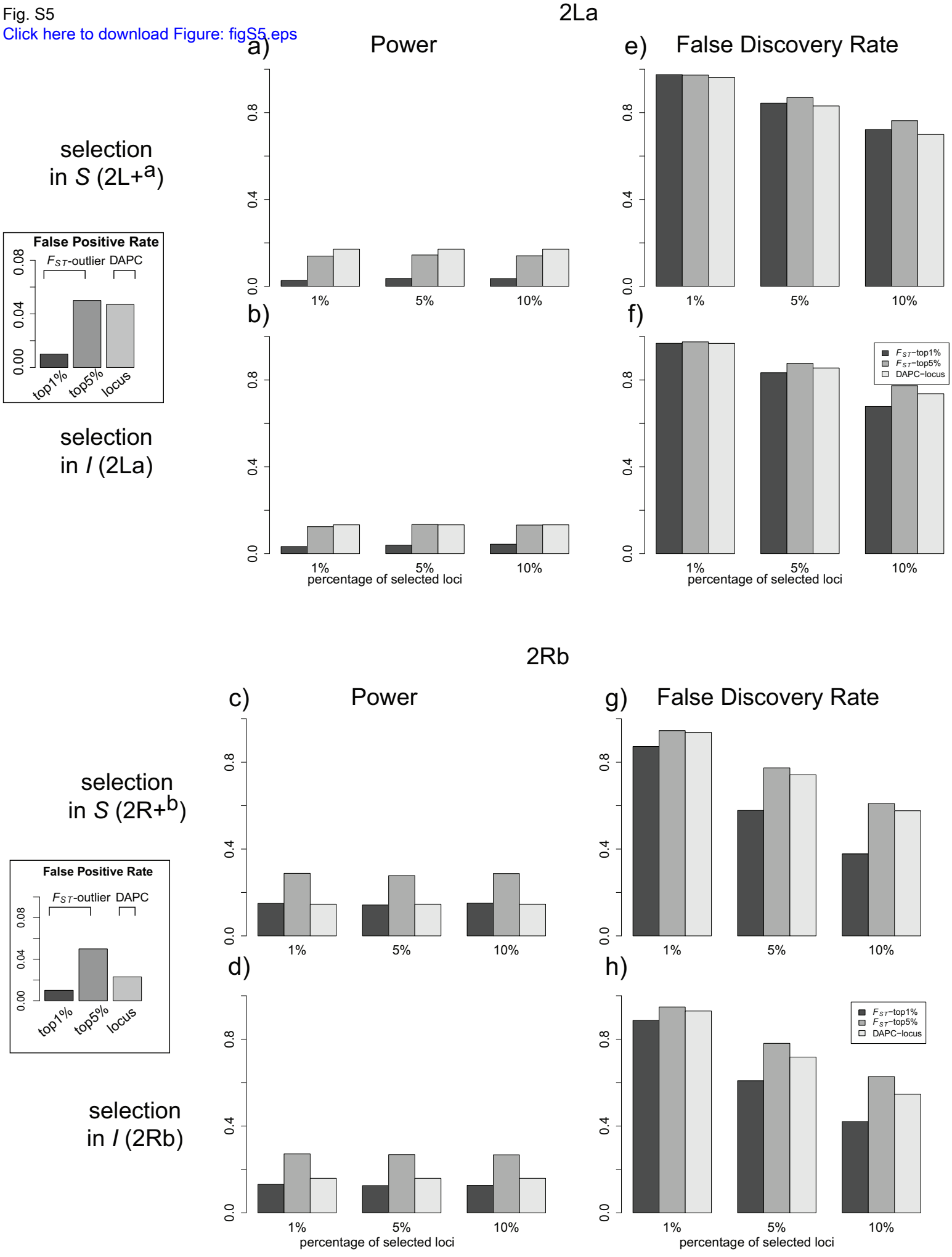
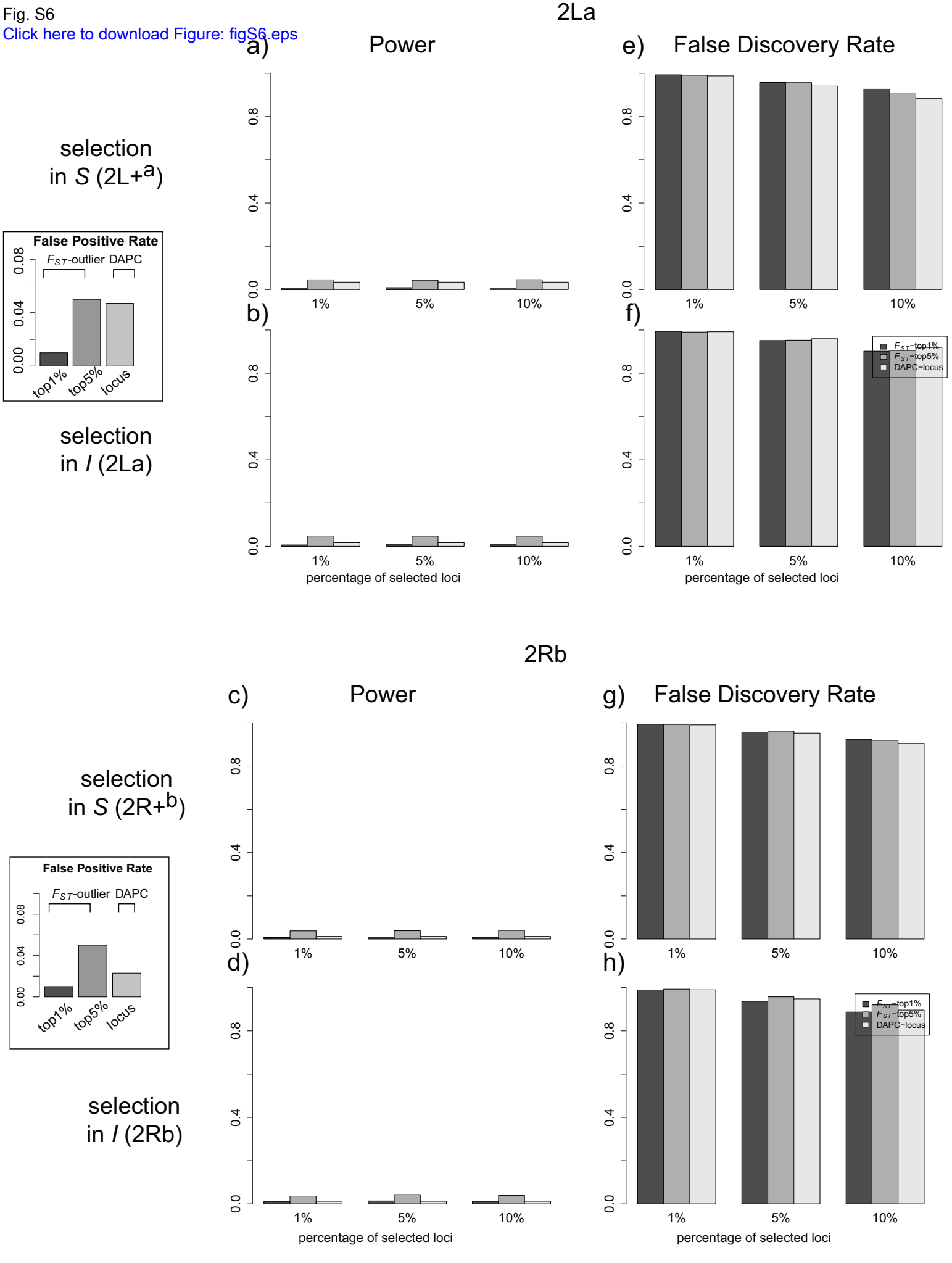


Fig. S6

Click here to download Figure: figS6.eps



Supporting Information

Click here to download Supporting Information: QXH&LLK_2014_SuppleText.docx