DR. YING ZHEN (Orcid ID : 0000-0001-8635-9687)
DR. ERIC C ANDERSON (Orcid ID : 0000-0003-1326-0840)
MISS SIRENA LAO (Orcid ID : 0000-0003-2907-7266)

Article type : Original Article

Title: Genomic divergence across ecological gradients in the Central African rainforest songbird (*Andropadus virens*)

Authors:

Ying Zhen[a,b*], *Ryan J. Harrigan*[b], *Kristen C. Ruegg*[b,c], *Eric C. Anderson*[d], *Thomas C. Ng*[e], *Sirena Lao*[b], *Kirk E. Lohmueller*[a,f] *and Thomas B. Smith* [a,b]

Affiliations:

[a] Department of Ecology and Evolutionary Biology, University of California, Los Angeles, 610 Charles E Young Drive East, Los Angeles, CA 90095, USA.

[b] Center for Tropical Research, Institute of Environment and Sustainability, University of California, Los Angeles, La Kretz Hall, Suite 300, Los Angeles, CA 90095, USA.

[c] Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA 95060, USA.

[d] Fisheries Ecology Division, Southwest Fisheries Science Center, National Marine Fisheries Service, 110 McAllister Way, Santa Cruz, CA 95060, USA.

[e] Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95060, USA.

[f] Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, 90095, USA.

Keywords:
RADseq, genomic divergence, ecotone, rainforest, environmental gradient, adaptation

[*]**Corresponding author:**

Ying Zhen

Address: Center for Tropical Research, Institute of Environment and Sustainability,
University of California, Los Angeles, La Kretz Hall, Suite 300, Los Angeles, CA 90095,
USA.

Phone: 310-206-6234

Fax: 310-825-5446

Email: zhen@g.ucla.edu

Running title: Genomic divergence across ecological gradients

Abstract

The little greenbul, a common rainforest passerine from sub-Saharan Africa, has been the subject of long-term evolutionary studies to understand the mechanisms leading to rainforest speciation. Previous research found morphological and behavioral divergence across rainforest-savanna transition zones (ecotones), and a pattern of divergence with gene flow suggesting divergent natural selection has contributed to adaptive divergence and ecotones could be important areas for rainforests speciation. Recent advances in genomics and environmental modeling make it possible to examine patterns of genetic divergence in a more comprehensive fashion. To assess the extent to which natural selection may drive patterns of differentiation, here we investigate patterns of genomic differentiation among populations across environmental gradients and regions. We find compelling evidence that individuals form discrete genetic clusters corresponding to distinctive environmental characteristics and habitat types. Pairwise F_{ST} between populations in different habitats is significantly higher than within habitats, and this differentiation is greater than what is expected from geographic distance alone. Moreover, we identified 140 SNPs that showed extreme differentiation among populations through a genome-wide selection scan. These outliers were significantly enriched in exonic and coding regions, suggesting their functional importance. Environmental association analysis of SNP variation indicates that several environmental variables, including temperature and elevation, play important roles in driving the pattern of genomic diversification. Results lend important new genomic evidence for environmental gradients being important in population differentiation.

Introduction

Rainforests are heralded for their exceptionally high biological diversity, yet the evolutionary mechanisms for the generation and maintenance of this diversity have been debated for decades (Haffer 1969; Mayr & O'Hara 1986; Martin 1991; Smith *et al.* 1997; Schneider *et al.* 1999; Moritz *et al.* 2000; Ogden & Thorpe 2002; Price 2008; Schluter 2009; Hoorn *et al.* 2010; Ribas *et al.* 2011; Smith *et al.* 2014; Beheregaray *et al.* 2015). Models of rainforest speciation abound. Some emphasize the importance of neutral processes, for example genetic drift in allopatric populations isolated by historical refugia (Haffer 1969), some favor processes such as landscape change (Hoorn *et al.* 2010; Ribas *et al.* 2011) or dispersal (Smith *et al.* 2014), while others point toward a dominant role of divergent natural selection across ecological gradients and ecotones (Smith *et al.* 1997; Schneider *et al.* 1999; Ogden & Thorpe 2002; Smith *et al.* 2005; Schluter 2009; Smith *et al.* 2011; Beheregaray *et al.* 2015). Each process is expected to shape the genomes of natural populations in different ways, leaving a signal that provides insights into the evolutionary mechanisms that may have led to divergence. Such information is of importance not only to evolutionary geneticists interested in understanding the processes involved in speciation, but also to conservation decision makers, who are interested in preserving biodiversity and prioritizing new regions for protection in the face of rapid anthropogenic change and climate change.

In this study, we explore the roles that population-level processes play in shaping biodiversity in Central Africa by examining the genomic diversity in a common songbird, the little greenbul (*Andropadus virens*). The little greenbul provides a particularly useful taxon for this inquiry because it has a broad geographic distribution across sub-Saharan Africa where it occurs in ecologically diverse habitats, and has been the subject of long-term studies of intra-specific diversity and speciation. In the case of *A. virens*, as well as some other rainforest taxa, the rainforest-savanna transition zones (ecotones) have been shown to drive phenotypic divergence and likely speciation (Smith *et al.* 1997, 2005; Kirschel *et al.* 2009; Freedman *et al.* 2010; Mitchell *et al.* 2015; Nadis 2016). Compared to the central rainforest, ecotone habitats differ dramatically in numerous ways. For example, ecotones have less tree cover, lower levels of precipitation, and greater intra-annual variation in environmental variables. These ecological differences may lead to distinctive food resources, pathogens, acoustic environments and predation levels (Slabbekoorn & Smith 2002; Smith *et al.* 2005, 2013). Consequently, these differences in both abiotic and biotic environments are hypothesized to result in divergent selection in ecotone and rainforest populations, leading to locally adapted populations (Smith *et al.* 1997, 2005; Kirschel *et al.* 2009; Freedman *et al.* 2010; Sehgal *et al.* 2011; Kirschel *et al.* 2011). This hypothesis is supported by the fact that parapatric *A. virens* populations across rainforest-ecotone gradients have undergone significant divergence in morphological (i.e. body mass, wing, tail, tarsus and beak length) and vocal characteristics despite significant levels of gene flow (Smith *et al.* 1997; Slabbekoorn & Smith 2002; Smith *et al.* 2005; Kirschel *et al.* 2011; Smith *et al.* 2013). This pattern of divergence with gene flow and the role of ecotones in driving adaptive divergence is further supported by the fact that allopatric rainforest populations of *A. virens* that were geographically isolated in West and Central Africa for two million years, had much lower levels of phenotypic divergence in these traits compared to the level of divergence observed across a narrow (often 100km) rainforest-ecotone gradient (Smith *et al.* 2005). Together, results for *A. virens* and those from other species suggest that strong divergent natural selection across the rainforest-savanna ecotone transition contributes to adaptive phenotypic divergence despite high levels of ongoing gene-flow (Smith *et al.* 1997, 2001, 2005). Evidence for divergence with gene flow in *A. virens* is also consistent with models of ecological speciation where natural selection caused by shifts in ecology or invasions of new habitats can result in divergence in fitness related traits and might play a prominent role in speciation (Orr & Smith 1998; Schneider *et al.* 1999; Ogden & Thorpe 2002; Rundle & Nosil 2005; Schluter 2009; Beheregaray *et al.* 2015). Opportunities for this kind of divergence are possible across the little greenbul range, as they occur across a wide diversity of habitats, including mountains and islands, which are also known as hotspots of diversification and speciation (Darwin 1859; Myers *et al.* 2000; Orme *et al.* 2005). Previous research has found that, compared to *A. virens* populations in mainland rainforests, mountain populations and island populations also show significant divergence in morphological traits typically related to fitness in birds, including body mass, wing length, tail length, tarsus length and bill size (Smith *et al.* 2005). Moreover, both habitats have considerable gene flow with mainland rainforest populations in Lower Guinea (Smith *et al.* 2005), suggesting natural selection may play an important role in divergence of mountain and island populations in *A. virens.*

To date, the paucity of high-resolution genomic data for rainforest species such as *A. virens* hinders a full exploration of the evolutionary mechanisms that may be important for diversification. Previous genetic studies on *A. virens* population structure utilized a handful of mtDNA markers (Smith *et al.* 2001) and microsatellite loci (Smith *et al.* 2005). These limited resources were unable to differentiate ecotone and forest populations at genetic level, therefore debates still remain whether the observed phenotypic divergence might be the results of plasticity in traits in response to varying environmental conditions, or strictly genomic divergence between populations in ecotone and rainforest. Recent development of next generation sequencing techniques (NGS), especially restriction site associated DNA (RAD) sequencing, enables one to *de novo* assemble hundreds of thousands of RAD loci across the genome in hundreds of samples without a reference genome. This cost effective method to produce genomic-wide population data provides unprecedented opportunities to assess the patterns of diversity with much greater resolution, to find potential population structure and to identify candidate loci under local selection in non-model species such as *A. virens*.

Here we take a population genomic approach leveraging single nucleotide polymorphism (SNP) data generated from RAD sequencing to survey the genome-wide diversity of *A. virens* across multiple ecological habitats in Cameroon and Equatorial Guinea, including rainforests, ectones, mountains, as well as island. Our specific objectives for this approach were to: 1) estimate overall levels of genetic diversity in *A. virens*; 2) determine population structure and differentiation across habitats; 3) identify candidate loci that are potential targets of selection; 4) understand the biological functions of these candidate loci using transcriptome data; and 5) characterize genetic turnover across environmental gradients.

Materials and methods

Sampling, DNA extraction and RADseq library construction

For RAD sequencing, blood samples from adult *A. virens* were collected in Central Africa and stored in Queens Lysis Buffer (Smith *et al.* 1997, 2005). Overall, 217 samples were collected from 15 geographically distant sampling sites (Figure 1A), representing 15 populations. Sampling sites were classified into one of four habitat types by researchers in the field and had previously been confirmed using remote sensing data (Slabbekoorn & Smith 2002; Smith *et al.* 2005, 2013). Low quality samples were removed by filtering, resulting in a total of 182 samples included in the final analysis (see *RADseq data bioinformatics processing* below). This included seven rainforest populations (83 samples), five ecotone populations (59 samples), two mountain populations (18 samples) and one population from the island of Bioko (12 samples). Each population was represented by 5-22 samples, with a mean representation of 12 samples (Table S1) (Willing *et al.* 2012; Nazareno *et al.* 2017).

RAD library preparation followed the methods for traditional RAD as described in Ali et al (2016) that were slightly modified from the original RAD protocol as described in Baird et al (2008). In short, genomic DNA (50 ng) for each sample was digested with 2.4 units of SbfI-HF (New England Biolabs, NEB, R3642L) at 37 °C for 1 hr in a 12 µl reaction volume buffered with 1X NEBuffer 4 (NEB, B7004S). Samples were heated to 65 °C for 20 minutes, and then 2 µl indexed SbfI P1 RAD adapter (10 nM) was added to each sample. Ligation of inline barcoded P1 adaptors was performed by combining 2 µl of the ligation mix (1.28 µl water, 0.4 µl NEBuffer 4, 0.16 µl rATP (100 mM, Fermentas R0441) with 0.16 µl T4 DNA Ligase (NEB, M0202M) and heating at 20 °C for 1 hr followed by incubation at 65 °C for 20 min. Following the ligation, half the per sample volume or 5 µl of each of the 96 samples were pooled into a single tube and cleaned using 1X Agencourt AMPure XP beads (Beckman Coµlter, A63881); the remainder of the sample was stored for use in an additional library preparation if needed. The pooled DNA was then re-suspended in 100 µl low TE and sheared to an average fragment size of 500 base pairs using a Bioruptor NGS sonicator (Diagenode). Sheared DNA was then concentrated to 55.5 µl using Ampure XP beads, and used as the template in the NEBNext Ultra DNA Library Prep Kit for Illumina (NEB E7370L; version 1.2). The NEBNext protocol for library prep was followed apart from the fact that we used custom P2 adaptors which were created by annealing an NEBNext Mµltiplex Oligo for Illumina (NEB, E7335L) to the oligo GATCGGAAGAGCACACGTCTGAACTCCAGTCACIIIIIIATCAGAACA*A (the * represents a Phosphorothioated DNA base). In addition, instead of the USER enzyme step, we used a universal P1 RAD primer (AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC*G) and a universal P2 RAD primer (CAAGCAGAAGACGGCATACG*A) during final amplification. The final RAD library was cleaned using AMPure XP beads and sequenced at the UC Berkeley QB3 Vincent J Coates Genome Sequencing Laboratory (GSL) on an Illumina HiSeq2000 (Illumina, San Diego, CA) using single-end 100 bp sequence reads.

RADseq data bioinformatics processing

We used FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to assess overall data quality of each RADseq sequencing run. To remove the lowest quality bases, we trimmed all raw sequencing reads (100bp) by 14 bp at the 3' end using *seqtk* (https://github.com/lh3/seqtk). We processed RADseq reads using the *Stacks* pipeline version 1.32 (Catchen *et al.* 2011, 2013) in the following manner. First, we demultiplexed the trimmed data by P1 barcodes and removed low-quality reads and those containing adapter sequences using *process_radtags.* After demultiplexing, reads were 80 bp in length (without barcodes) and data from different runs were combined together. These reads were used to *de novo* assemble RAD loci using *denovo_map.pl* (parameter settings: m=3 M=4 n=4). The parameters for *de novo* assembly were determined empirically to limit the impact of over-splitting of loci following methods described in Ilut *et al.* (2014) and

Harvey *et al.* (2015). Specifically, we chose one sample that had a depth of coverage close to the median depth coverage of all samples, and ran the *de novo* assembly over a wide range of values of M (1-8; n=M) using *ustacks*. The percentage of homozygous and heterozygous loci plateaued at M=4, suggesting this value appropriately minimized over-splitting of alleles (Figure S1). Thus, we used M=4 for the final run on all samples. *Stacks* implements a multinomial-based likelihood model for SNP calling, by estimating the likelihood of two most frequently observed genotypes at each site and performing a standard likelihood ratio test using a chi-square distribution (Hohenlohe *et al.* 2010; Catchen *et al.* 2011). We used the default alpha (chi-square significance level) of 0.05. Paralogous loci that were stacked together were identified and removed by later quality control steps built into *Stacks* (e.g. max number of stacks per loci = 3; Ilut *et al.* 2014; Harvey *et al.* 2015). After the first round of assembly using *denovo_map.pl*, we ran stacks' correction mode (*rxstacks-cstacks-sstacks*) using the bounded SNP model with a 0.05 upper bound for the error rate (bound_high = 0.05). The *rxstacks* program made corrections to genotype and haplotype calls based on population information, rebuilt the catalog loci, and filtered out loci with average log likelihood less than -8.0 (http://catchenlab.life.illinois.edu/stacks/).

We then identified a set of high-quality RAD loci for downstream population genetic analysis using the following steps: 1) We only kept RAD loci if they were present in at least 80% of all samples, because loci that only assembled in small subset of samples had limited utility in downstream analyses as well as possibly higher error rates. 2) We filtered out RAD loci that had more than 40 SNPs along the 80 bp RAD loci sequence, as these likely represented sequencing errors or over-clustering of paralogous loci. In the final dataset, the RAD locus that has the most SNPs possessed 25 SNPs. Because the alignments look reasonable for the RAD loci that have higher number of SNPs, we did not apply any additional filters to avoid introducing additional biases. 3) We mapped the RAD loci sequences from *A. virens* to the closest reference genome available, the zebra finch genome (version 3.24), using BLAT and removed RAD loci that mapped to multiple positions in the zebra finch genome. 4) We used BLAT to compare RAD loci sequences against each other, and removed ones that had a match. This step further removes over-splitting RAD loci.

Following these filters, we obtained our final consensus set of RAD loci (Table S2). Samples that were missing more than 20% of the final consensus RAD loci were identified in a preliminary run and were removed from final analysis because they likely had low quality DNA, low quality libraries, or low sequencing coverage. 182 samples were included in the final dataset (see above). Genotypes were called and filtered using methods implemented in the *Stacks* pipeline (Hohenlohe *et al.* 2010). We exported genotypes for the final consensus RAD loci in VCF format using stacks *populations* program (only bi-allelic SNPs). Additional filters based on genotype calls were performed in *vcftools* (https://vcftools.github.io/index.html) or using custom scripts, which includes removing SNPs from the last 7 bp of the RAD loci as this part of the locus was enriched for erroneous SNPs due to the lower sequencing quality at the 3' end of reads, and filtering sites that have genotyping rate less than 80% of all samples.

We used the resulting full SNP dataset with SNPs from all frequencies to estimate genetic diversity statistics such as number of segregating sites (S), average pairwise differences (π) and Waterson's θ (θ_w) in each population (Table S1). Rare SNPs that had a minor allele frequency (MAF) less than 2% in the whole sample set were subsequently removed using *vcftools*, and the remaining SNPs were used for downstream analyses such as PCA, pairwise F_{ST} calculations, *Bayescan* outlier analysis, *gradientForest* analysis.

RNA extraction, RNAseq library preparation, and *transcriptome de novo assembly*

A. virens lacks a reference genome. In order to help determine which of the RAD loci are transcribed, we collected RNAseq data and made a *de novo* assembly of *A. virens* transcriptome. Fresh tissue samples were collected from 10 live individuals in the field (five tissue types: blood, brain, breast tissue, heart and liver). Tissue samples were stored in either PAXgene (Blood RNA Tubes, PreAnalytiX/Qiagen, Switzerland) or Allprotect (Tissue Reagent, Qiagen, Germany) buffer and shipped to laboratory facilities at UCLA. RNA was extracted from each sample and tissue type separately using an RNeasy kit (Qiagen, Germany), and based on quality of extractions (both overall concentration and 260/280 ratio), three RNA samples from three tissue types (brain, heart and liver) were chosen to perform library preparations. RNAseq libraries were prepared using Illumina TruSeq RNA Library Prep Kit v.2 (Illumina, San Diego, California) following the manufacture's protocol, and libraries were indexed, normalized, pooled, and sequenced on a single lane on Illumina HiSeq 2500 (paired end 100 bp reads, Rapid Run mode) at GSL.

We obtained one lane of paired-end RNAseq data pooled from three tissue types. We first removed bases with quality scores lower than 20 and minimum sequence length of 30bp using *trim_galore* (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). We then pooled the remaining paired-end reads from different tissues together for a *de novo* assembly of the transcriptome using the *Trinity* pipeline (Grabherr *et al.* 2011). We assessed the quality of the assembly using scripts provided in the *Trinity* package, and predicted the coding regions in the assembled transcriptome using *TransDecoder* in *Trinity*.

Detecting population structure using genomic data

To detect the underlying population structure among samples, we performed a Principal Component Analysis (PCA) using the bioconductor package *SNPrelate* (Zheng *et al.* 2012). 47,482 SNPs with MAF >=2% were used in PCA. The first six principal components were visually examined to identify clustering patterns of samples and to determine whether these genetic clusters tend to segregate with ecological factors or geography. We also used the program ADMIXTURE (Alexander *et al.* 2009) to estimate the ancestry of individual genotypes, using only the first SNP of each RAD loci to limit the impact of linkage disequilibrium. The analysis was run for K = 1-15.

To quantify pairwise population differentiation, we calculated pairwise F_{ST} between populations using *SNPrelate*. The correlation of population genetic differentiation (pairwise F_{ST}) and geographic distance, in other words, the presence of isolation by distance (IBD), was estimated by a simple Mantel test with 999,999 permutations using vegan2.2-1 in R (Mantel 1967; Oksanen *et al.* 2017). Mantel tests are reported using both raw F_{ST} and F_{ST} /(1- F_{ST}), as well as both raw Euclidian geographic distance and log-transformed distances (Slatkin 1995; Rousset 1997).

Moreover, we found that pairwise F_{ST} between populations from different habitats were higher than pairwise F_{ST} computed between populations from the same habitat (see Results). In principle, this pattern could solely be driven by isolation by distance as populations from the same habitat tend to be located closer geographically to each other compared to population from different habitats. To determine whether the elevated F_{ST}

between populations from different habitats (compared to F_{ST} between populations from the same habitat) could be explained simply by the differences in geographic distance, we performed permutation tests that accounted for the fact that populations from different habitats tend to be further apart. Specifically, we divided the population pairs into five bins based upon their geographic distances from each other (i.e. <200km, 200-400km, 400-600km, 600-800km, >800km). Then within each bin, we permutated whether the pairwise F_{ST} values are from a within-habitat comparison or a between-habitat comparison. We generated 10,000 such permutations and, for each permutation, performed a t-test on whether the F_{ST} values for between-habitat comparisons were higher than those for within-habitat comparisons. From the permuted data, we built a null distribution of t-statics, which accounted for the effect of geography. Our final empirical p-value for the observed data was calculated as the percentage of permutated datasets that had a t-statistic as large or larger than the one seen in the original data. Similar permutation analyses were applied to the dataset including all habitats as well as to a dataset that only considered rainforest and ecotone populations. In the null distribution of t-statistics for the test of whether F_{ST} is higher between than within habitats, we found that none of the 10,000 permutated datasets had a t-statistic of F_{ST} as large or larger than the one seen in the original data, suggesting a p-value < 1e-04. However, for the null distribution of t-statistics for the test of whether distance is higher between or within habitats, 1581 of the 10,000 permutated datasets had a t-statistic as large or larger than the one seen in the original data, suggesting a p-value = 0.158. This suggests that our null distribution of t-statistics accounts for the fact that populations from similar habitats tend to be closer together geographically.

As an alternative method to test whether habitat contributed to the observed pattern of population differentiation above and beyond geographic distance, we created a binary matrix that indicated whether a pair of populations was from the same habitat or not. We tested the correlation of genetic distance matrix and this matrix while controlling for geographic distance using a partial Mantel test using vegan 2.2-1 in R (Mantel 1967; Oksanen *et al.* 2017). Partial Mantel tests are performed using both raw F_{ST} and F_{ST} /(1-F_{ST}), as well as both raw Euclidian geographic distance and log-transformed distances (Slatkin 1995; Rousset 1997).

Identifying outlier SNPs under selection

We used *BayeScan2.1* (Foll & Gaggiotti 2008) to identify highly differentiated SNPs that are candidates to be under natural selection. This program takes a Bayesian approach to search for SNPs exhibiting extreme F_{ST} values that could be due to local adaptation. Outlier SNPs were identified using SNPs with MAF >= 2% across all samples, specifying all 15 populations or four habitats (see supplemental Notes). We ran 5000 iterations using prior odds of 10 and assessed the statistical significance of a locus being an outlier using a false discovery rate (FDR) of 5%.

To explore the spatial patterns of population differentiation across chromosomes, we mapped the consensus RAD loci to the zebra finch genome using BLAT with default parameters. For the uniquely mapped RAD loci, we plotted the F_{ST} of each SNP by genome coordinates to examine spatial patterns of outlier SNPs. To interpret the potential biological function of the outlier SNPs identified by Bayescan analysis, we used a zebra finch genome annotation (v3.2.4.78) to identify outlier SNPs mapped to annotated genic regions.

We further examined whether candidate loci under selection were enriched in exonic (transcribed) or coding regions. To do this, we mapped RAD loci to the *de novo* assembled *A. virens* transcriptome using BLAT with default parameters. Any RAD locus that mapped to the transcriptome was considered from exonic regions of the genome, and the remaining RAD loci were labeled “non-transcribed” regions of the genome. Similarly, we mapped RAD loci to predicted coding sequences and categorized them into coding and non-coding RAD loci. We then used a one-sided Fisher’s exact test to examine whether there was significant enrichment of outlier loci in exon or coding regions of the genome. Finally, we cross-checked these outliers to see if there was any significant associations with environment using Latent Factor Mixed Models (Frichot *et al.* 2013) (see Supplemental Notes for more details).

Detecting environmental drivers of genomic variation

In addition to population structure, we also tested whether allele frequencies in different populations were associated with environmental variables across the range of *A. virens* using the package *gradientForest* (Ellis *et al.* 2012) in the R statistical framework

(R Working Group, 2014). Gradient forests are an extension of random forests (Breiman 2001) that treat response variables (in this case, individual SNP minor allele frequencies within each population) as members of a larger community (the total genome), and provides summary statistics based on ensembled forest runs to indicate an overall association of changes in allele frequency to particular environmental variables (Ellis *et al.* 2012; Fitzpatrick & Keller 2015). Gradient forests were run using the following changes to the default settings: number of trees run for each environmental variable = 500, number of SNPs included in each bin = 1000. Allelic frequencies across the genome were predicted for unsampled geographic locations by generating a random set of 100,000 points across the range of *A. virens*. Then we used our final gradient forest model to predict allele frequencies at each of those points, given their environmental characteristics. Ordinary Kriging (Oliver & Webster 1990) was then used within ArcMap (ESRI, Redlands, CA) to generate a continuous surface across the known range of *A. virens* in Cameroon.

We used a suite of 17 environment variables (Table S6), including bioclim measures of temperature and precipitation (n=9; any variables showing a Pearson's correlation coefficient > 0.7 were removed) downloaded from the Worldclim database (www.worldclim.org), measures of vegetation and tree cover captured using the NASA MODerate-resolution Imaging Spectroradiometer (MODIS, n=4), elevation and slope captured via the Shuttle Radar Topography Mission (n=2), and surface moisture estimates measures using the Quick Scatterometer (QuikSCAT, n=2). In addition to these variables, and to account for purely geographic associations, we also included Euclidean distances (measured as Latitude and Longitude) as predictor variables in all model.

Results

SNP discovery and overall genetic diversity

We used RAD sequencing to survey the genome-wide diversity of *A. virens*. The final sample set included 15 populations from four different habitats, including rainforests, ecotones, mountains, and an island (Figure 1A; Table S1). After removing low quality reads and samples, we obtained a total of 916 million reads for 182 *A. virens* samples (SRA:xxxxxx). The number of raw sequence reads per sample ranged from 1.60 to 20.73 million. On average, 99.2% of these reads were utilized in the *de novo* assembly of the

RAD loci. The mean coverage depth ranged from 16× to 136× per sample (mean = 38×, median = 32×, Figure S2). Using this dataset, we assembled and identified 34,657 high quality RAD loci that passed our quality filters and were genotyped in more than 80% of all final samples. On these 34,657 consensus RAD loci, there were a total of 255,290 SNPs. The median number of SNPs per RAD locus is seven. With a minimum minor allele frequency filter of 2%, we retained 47,482 SNPs that were present on 23,882 RAD tags (Table S2; Supplemental Notes).

The number of segregating sites ranges from 25,936 to 70,598 per population. Waterson's θ (θ_w) was estimated to be 0.0049 - 0.0076/bp (mean = 0.0064/bp), and π ranges from 0.0034 - 0.0037/bp (mean = 0.0036/bp) (Table S1), which is comparable to π estimated from other bird species (Nadachowska-Brzyska *et al.* 2013; Romiguier *et al.* 2014). Overall levels of genetic diversity are comparable in each habitat, including the island population (Table S1). The finding that θ_w is larger than π indicates an excess of low-frequency variants relative to the standard neutral model which could be driven by recent population expansions.

Transcriptome assembly and annotation

Transcriptome assembly was performed using 169 million paired-end RNAseq data from three different tissue types. The assembled transcriptome had a GC content of 45%. The average contig length was 815 bp and N50 was 1,619 bp. In total, trinity produced 237,226 genes and 286,494 transcripts, and predicted 81,018 coding sequences from these transcripts (Table S5). Of the 34,657 RAD loci we genotyped, 8412 RAD loci (24.2%) were mapped to the *de novo* assembled *A. virens* transcriptome, and 3,618 RAD loci (10.4%) were mapped to the predicted coding sequences (Figure S3). The RAD tags overlapping coding sequence tend to have fewer SNPs than those that do not overlap with coding sequences (Figure S4), consistent with the fact that the coding regions are likely under stronger selective constraint.

Population structure

We used PCA to identify population structure in little greenbuls. The first two PCs revealed a clear clustering pattern of individuals from the same habitats (Figure 1B). Populations from island, mountains, rainforests, and ecotones formed four discrete clusters, suggesting genomic divergence across ecological gradients and habitats. Island and mountain populations were most distinct (Figure 1B), however samples from all four habitats separated on PC1, including those from ecotone and rainforest habitats. PC2 further separated island and mountain samples from rainforest and ecotone samples. Remarkably, results suggest that, within rainforest and ecotone habitats, individual populations could be distinguished solely on the basis of genomic markers, mostly by PC1, with individuals from the same sampling sites clustering together (Figure 1C). The level of separation of ecotone populations from rainforest populations along PC1 roughly followed a latitudinal gradient, corresponding to environmental and rainfall gradients that distinguish ecotone in the north from rainforest in the south of Cameroon (see environmental analyses below). Specifically, samples collected from sites Wakwa and Ngaoundaba Ranch, toward the more extreme edge of the ecotone habitat and had the most extreme ecotone environmental conditions, formed clusters that were most distant from the rainforest samples, while samples collected from Betare Oya, at lower ecotone that was closest to the central rainforests, formed a cluster closest to the rainforest (Figure 1C). The pattern of genomic differentiation across habitats was confirmed using the program *Admixture* (Figure S5).

Pairwise F_{ST} between populations ranged from 0.017 to 0.078 (mean = 0.038; Figure 2A-B; Table S3), indicating low overall levels of genomic differentiation across populations. There was significant correlation between pairwise F_{ST} and geographic distance between the populations (Mantel r = 0.34; mantel simulated p-value = 0.003), suggesting isolation by distance contributes to population differentiation. However, pairwise F_{ST} between populations from different habitats was significantly higher than pairwise F_{ST} between populations within the same habitat (one tailed t test, p-value = 1.36e-10; Figure 2A). Pairwise geographic distances between populations from different habitats were also significantly higher than pairwise distances between populations within the same habitat (one tailed t-test, p-value = 1.015e-06). To account for the fact that populations from different habitats were also geographically further apart, we performed a permutation test, where we randomized whether a population pair was from the same or different habitats in different bins stratified by their geographic distance. Using permutated datasets, we built a null distribution of these t-statistics (that already includes the effect of geographic distance), which we used to evaluate the significance of our observed value. The higher F_{ST} value for between-habitat comparison was highly significant when compared to this improved null distribution (p-value < 1e-04, Figure 2C and Figure S6), suggesting that isolation by distance alone cannot explain the higher F_{ST} between habitats than within habitats. Similarly, only considering rainforest and ecotone populations, pairwise F_{ST} was significantly higher between habitats as compared to within habitats (one tailed t-test, p-value = 9.793e-05). Application of the same permutation test shows the higher F_{ST} between ecotone and rainforest populations (p-value=0.0055) cannot be explained by geographic distance alone (Figure 2D and Figure S6).

To confirm this finding using an alternative statistical approach, we used partial Mantel tests to determine the contribution of habitat types of population pairs to their genetic differentiation (F_{ST}), controlling for geographic distance. We found a highly significant and positive correlation between genetic distance and whether a pair of population comes from the same habitat, and greater genetic differentiation (higher F_{ST}) from between-habitat populations compared to within-habitat populations, while controlling for geographic distance (Table 1). Taken together these results suggest that factors other than geographic location, such as local adaptation significantly contribute to population differentiation between habitats.

In addition, mountain and island populations were more diverged from other populations (Figure 2B). Interestingly, F_{ST} between two mountain populations were exceptionally high (F_{ST}= 0.060) compared to other within habitat pairwise F_{ST} (ranging from 0.017 to 0.040, Figure 2A), despite the fact that the two mountain populations were geographically very close to each other. The F_{ST} values between mountain populations and lowland forest/ecotone populations were larger than pairwise F_{ST} values between lowland populations, suggesting mountain populations are highly differentiated both from one another and from lowland populations.

Candidate loci under selection

To further explore potential candidate loci under selection, we identified SNPs with extreme allele frequency differences across populations, which should be enriched by targets of local adaptation. We identified 140 outlier SNPs across all populations with a

False Discovery Rate of 5% using *Bayescan*. These candidate SNPs are potential targets of divergent selection across different sampling sites (Figure S7). The 140 outlier SNPs reside in 119 loci, and 40 of these loci mapped to the zebra finch genome (Figure S9). Of these, 36 mapped to main scaffolds of known chromosomes and four mapped to the Z chromosome. Only 13 of these outlier loci mapped to annotated genic regions on the zebra finch genome and nine mapped to genes with functional annotations (Table S4).

In order to uncover the functional significance of outlier loci, we used the *de novo* assembly of greenbul transcriptome to partition the RAD loci and SNPs into different categories depending on whether they mapped to coding regions or transcribed (exonic) regions (Figure S10). This enabled us to test for enrichment of outlier SNPs in putatively functional regions. Of the 47,482 SNPs, 9677 mapped to the transcriptome, and 42 were outliers based on a Bayescan analysis. Using a one-sided Fisher's exact test, we detected significant enrichment of outlier loci in exonic regions of the genome (p=0.0044; Table S2; Figure S10). Using the predicted coding sequence from the transcriptome, 3,602 SNPs mapped to the predicted coding sequences and 21 of these were outliers. We again detected a significant enrichment of outlier SNPs in protein-coding sequences (p=0.002; Table S2; Figure S10). Taken together, these enrichment results provide additional confidence that the outlier loci found using Bayescan captured functionally important, biologically relevant genetic variants, which were not merely loci that fell within the tail of a neutral distribution.

Genomic Turnover Across Environments

Because some environmental adaptation may involve shifts in allele frequency at many loci across the genome (e.g., polygenic selection involving many genes of small effect), we used the *gradientForest* approach to look for correlations in allele frequencies associated with environmental variables. A total of 7238 SNPs, ~15% of all SNPs, had R^2 values above 0 (0.0073-0.83) when testing for a correlation between frequency and an environmental variable. Of the 19 environmental and geographical variables included in models (Table S6), variables capturing temperature variation (Min Temp: minimum temperature of the coldest month, Temp Range: mean diurnal temperature range, and Mean temp: mean annual temperature) and elevation were most important in explaining environmentally associated variation in SNPs (Figure 3A). In some cases, measures of surface moisture or tree cover were also important, but axes for these variables largely overlapped with temperature or elevation measures along PC plot (likely the result of co-linearity in environmental variables) (Figure 3B). Results from LFMM analyses indicated these same variables were also associated with differentiation observed at hundreds of individual loci, although exact functions of these regions remain unknown (see Supplemental Notes).

Geographic variables alone were not as important in explaining variation in allele frequency, again suggesting that geographic distance cannot fully account for all variation in SNP frequencies across the range of little greenbuls. Above and beyond neutral processes, selective pressures imposed by differences in these environments best explains the observed genomic patterns of variation. Predictions across Cameroon suggest strong genomic turnover (defined as coordinated shifts in allele frequencies across the genome) throughout the forest, savannah, and ecotone regions, with diagnostic genomic variation occurring in each of these habitats (Figure 3). Distinct SNP frequencies at high elevations (Figure 3B-C), and the fact that elevation explains a large proportion of variation of allele frequencies in the greenbul genome (largely allied with PC1, Figure 3B) also suggest a unique genetic signature in populations at elevation.

Discussion

In this study, we used genome-wide RADseq SNPs to characterize the overall level of genetic diversity in *A. virens* populations across four different habitats. We found evidence of population structure of *A. virens* consistent with habitat type and previously observed phenotypic divergence. We demonstrated that population differentiation across habitats cannot be explained solely by isolation-by-distance, suggesting local adaptation further contributes to genomic divergence among habitats. We identified 140 outlier SNPs that are potential targets of selection and the fact that they are significantly enriched in exonic and coding regions suggests they are functionally important. Environmental association analysis further supports this conclusion and shows environmental variables, including temperature and elevation, are highly associated with patterns of genomic variation across the range of the little greenbul.

In addition to the differences between rainforest and ecotone populations, other habitats were found to harbor distinct patterns of genetic variation. The population from Bioko island formed a distinctive genetic cluster based on PCA, and also was identified as distinct in environmental association models (Figure 1B and 3), consistent with previous studies (Smith *et al.* 2005). Bioko island is 32 km off the coast of Africa, separated from mainland ~10,000 years ago and has an area of 2,017 km^2. Island populations and species may have smaller effective population sizes than mainland populations or sister taxa (Robinson *et al.* 2016), due to possible population bottleneck and considerably smaller ranges. As a result, island populations may have lower genetic diversity compared to their mainland counterparts (Frankham 1997). However, several recent empirical studies suggested this may not always be the case, particularly in birds (Francisco *et al.* 2015; James *et al.* 2016). In our study, the estimates of genetic variation using genome wide SNP markers in greenbul population on Bioko island are comparable to those from mainland populations (Table S1). This is consistent with the recent findings that island populations do not always have lower genetic diversity (Francisco *et al.* 2015; James *et al.* 2016), and the fact that Bioko island is a large island that only recently separated from the mainland.

Tropical mountains are well known to support disproportionally high biodiversity and are thought to be hotspots for avian speciation (Roy 1997; Myers *et al.* 2000; Smith *et al.* 2000; Orme *et al.* 2005; Fjeldså *et al.* 2007, 2012; Drovetski *et al.* 2013). Little greenbuls are found at elevations up to 2400 m, where environmental variables, particularly temperature and vegetation, change rapidly along altitudinal gradients. Our two mountain populations have high F_{ST} despite being geographically close and from same habitat type. Although the Euclidean distance between these two mountain populations is short, the environmental changes along altitudinal gradients are steep, causing isolation between populations from different mountains and forming “sky islands”, between which the level of gene flow probably is much lower than among lowland populations. Moreover, we found that the two different mountain populations exhibited the lowest within-population genetic variation among all sampled populations (Table S1). While this difference was not statistically significant (likely due to small sample sizes), this decreased level variation can inflate $F_{ST,}$ the relative measurement of population differentiation. It also suggests that mountain populations may have overall smaller effective population sizes (consistent with presumably smaller suitable habitat size for mountain populations) and/or have experienced serial bottleneck/founder effects as range expansions occurred. These processes can further contributed to divergence due to stronger genetic drift within each subpopulation leading to faster changes in allele frequency. The idea that elevation can drive genomic changes is supported by previous estimates of morphological divergence (Smith *et al.* 2005), and emphasizes the importance of preserving elevational gradients in tropical ecosystems in general (Thomassen *et al.* 2011).

Most of the genes containing outlier SNPs only have annotations predicted from human homologs, except two that have annotations from bird species. Both of these two genes are of particular interest. One outlier locus mapped to the 5UTR/coding junction of *EDIL3*, a calcium-binding protein that has been found to function in avian eggshell biomineralization (Marie *et al.* 2015). Avian eggshells protect the developing embryo and keep the egg free from pathogens. Environmental factors such as temperature, humidity, and partial oxygen pressure have been reported to affect avian eggshell structure, and previous studies documented rapid evolution of eggshell structure in response to colonization of novel environments in the house finch (Stein & Badyaev 2011). The second outlier locus mapped to *MLXIPL* (MLX interacting protein-like), which is co-activator of the carbohydrate response element binding protein that has been correlated with fat deposition in caged chicken (Proszkowiec-Weglarz *et al.* 2008; Li *et al.* 2015). Interestingly, seven more genes that contain outlier SNPs have annotation linked with metabolic traits or diseases in humans. For example, outlier SNPs were found in *RGS6* (Sibbel *et al.* 2011 p. 6), *CSAD* (Comuzzie *et al.* 2012) and *UNC13B* intron (Trégouet *et al.* 2008), which were associated with dietary fat intake, food preference, adiposity/obesity and diabetes in humans. Although metabolic traits were not measured, adult little greenbuls from the rainforest have significantly smaller body mass and body size compared to ecotone, mountain, and island populations (Smith *et al.* 1997, 2005), which could be the result of divergent selection of these genes associated with metabolic traits. Several recent studies have discussed the limitations of identifying F_{ST} outlier as loci under divergent selection, and suggest results should be interpreted carefully, because many other factors, including demographic history, recombination rate heterogeneity, and background selection may also create F_{ST} outliers (Roesti *et al.* 2012; Lotterhos & Whitlock 2014;

Cruickshank & Hahn 2014). Current work to model the demographic history of *A. virens* should help examine these various possibilities in greater detail.

Numerous hypotheses have been proposed for how biodiversity is generated in rainforests (Mayr & O'Hara 1986; Moritz *et al.* 2000). With the rapid advances in genomics and environmental modeling in the last decade, it is now possible to examine these mechanisms in greater depth. Using more powerful genome-wide data, we have shown, for the first time, strong patterns of population structure and genomic differentiation between rainforest and ecotone habitats in *A. virens* (Figure 1). Previously, no genetic differentiation was found between morphologically divergent populations in rainforest and ecotone habitats, leaving open the possibility that the observed morphological difference could simply be the result of a homogenized meta-population that differentially responds to environmental gradients. Although identifying the underlying genetic basis of morphological traits that differ between rainforest and ecotone populations was beyond the scope of this study, our results complement previous work by demonstrating that populations along the rainforest - ecotone gradient are diverging at the genomic level, and raise the possibility that local adaptation could account for the patterns of morphological variation previously observed across ecotone-rainforest gradients. Results also complement past research on reproductive behavior, which found differences in song characteristics along the forest-ecotone gradient, and showed experimentally that singing males respond more aggressively to male songs from their own habitat, suggesting incipient reproductive isolation driven by habitat (Slabbekoorn & Smith 2002; Kirschel *et al.* 2011; Smith *et al.* 2013). These patterns of differentiation are consistent with models of ecological speciation, where natural selection caused by shifts in ecology can promote speciation (Orr & Smith 1998; Schneider *et al.* 1999; Schluter 2000; Ogden & Thorpe 2002; Rundle & Nosil 2005; Price 2008; Räsänen & Hendry 2008; Schluter 2009; Beheregaray *et al.* 2015; Hanson *et al.* 2016). However, further research is necessary to more fully understand the evolutionary significance of divergence across ecological gradients and ecotones. In particular, studies investigating the underlying genetic basis of phenotypic differentiation and mate choice experiments would provide additional insights into their importance in divergence and speciation.

Author Contributions

T.B.S. and K.E.L conceived of and supervised the project. S.L. and R.J.H conducted the laboratory work. Sequence assemblies, population structure and outlier analysis was primarily carried out by Y.Z. with assistance from T.N., K.R., E.C.A. and K.E.L. Environmental association analysis was performed by R.J.H. The manuscript was written by Y.Z., R.J.H., K.R., K.E.L., and T.B.S., with input from all authors.

Acknowledgments

We thank University of California, Berkeley's Vincent J. Coates Genomic Sequencing Laboratory (GSL) for sequencing services. We acknowledge support from a QCB Collaboratory Postdoctoral Fellowship to Y.Z. and the QCB Collaboratory Community directed by Matteo Pellegrini. This research was supported by a Searle Scholars grant to K.E.L., National Science Foundation grant #PD-08-1269 to R.J.H., National Geographic grants and National Science Foundation grants IIA PIRE #1243524, DEB-9726425 and IRCEB9977072 to T.B.S.

References

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655–1664.

Beheregaray LB, Cooke GM, Chao NL, Landguth EL (2015) Ecological speciation in the tropics: insights from comparative genetic studies in Amazonia. *Evolutionary and Population Genetics*, **5**, 477.

Breiman L (2001) Random Forests. *Machine Learning*, **45**, 5–32.

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.

Catchen J, Hohenloh PA, Bassham SL, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular ecology*, **22**, 3124–3140.

Comuzzie AG, Cole SA, Laston SL *et al.* (2012) Novel Genetic Loci Identified for the Pathophysiology of Childhood Obesity in the Hispanic Population. *PLoS ONE*, **7**.

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

Darwin C (1859) *On the Origin of Species by Means of Natural Selection, Or, The Preservation of Favoured Races in the Struggle for Life*. J. Murray.

Drovetski SV, Semenov G, Drovetskaya SS *et al.* (2013) Geographic mode of speciation in a mountain specialist Avian family endemic to the Palearctic. *Ecology and Evolution*, **3**, 1518–1528.

Ellis N, Smith SJ, Pitcher CR (2012) Gradient forests: calculating importance gradients on physical predictors. *Ecology*, **93**, 156–168.

Fitzpatrick MC, Keller SR (2015) Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. *Ecology Letters*, **18**, 1–16.

Fjeldså J, Bowie RCK, Rahbek C (2012) The Role of Mountain Ranges in the Diversification of Birds. *Annual Review of Ecology, Evolution, and Systematics*, **43**, 249–265.

Fjeldså J, Johansson US, Lokugalappatti LGS, Bowie RCK (2007) Diversification of African greenbuls in space and time: linking ecological and historical processes. *Journal of Ornithology*, **148**, 359–367.

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.

Francisco FO, Santiago LR, Mizusawa YM, Oldroyd BP, Arias MC (2015) Genetic structure of island and mainland populations of a Neotropical bumble bee species. *bioRxiv*, 27813.

Frankham R (1997) Heredity - Abstract of article: Do island populations have less genetic variation than mainland populations? *Heredity*, **78**, 311–327.

Freedman AH, Thomassen HA, Buermann W, Smith TB (2010) Genomic signals of diversification along ecological gradients in a tropical lizard. *Molecular Ecology*, **19**, 3773–3788.

Frichot E, Schoville SD, Bouchard G, François O (2013) Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. *Molecular Biology and Evolution*, **30**, 1687–1699.

Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature biotechnology*, **29**, 644–652.

Haffer J (1969) Speciation in Amazonian Forest Birds. *Science*, **165**, 131–137.

Hanson D, Moore J-S, Taylor EB, Barrett RDH, Hendry AP (2016) Assessing reproductive isolation using a contact zone between parapatric lake-stream stickleback ecotypes. *Journal of Evolutionary Biology*, **29**, 2491–2501.

Harvey MG, Judy CD, Seeholzer GF *et al.* (2015) Similarity thresholds used in DNA sequence assembly from short reads can reduce the comparability of population histories across species. *PeerJ*, **3**, e895.

Hohenlohe PA, Bassham S, Etter PD *et al.* (2010) Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. *PLoS Genet*, **6**, e1000862.

Hoorn C, Wesselingh FP, Steege H ter *et al.* (2010) Amazonia Through Time: Andean Uplift, Climate Change, Landscape Evolution, and Biodiversity. *Science*, **330**, 927–931.

Ilut DC, Nydam ML, Hare MP *et al.* (2014) Defining Loci in Restriction-Based Reduced Representation Genomic Data from Nonmodel Species: Sources of Bias and Diagnostics for Optimal Clustering. *BioMed Research International, BioMed Research International*, **2014**, **2014**, e675158.

James JE, Lanfear R, Eyre-Walker A (2016) Molecular Evolutionary Consequences of Island Colonization. *Genome Biology and Evolution*, **8**, 1876–1888.

Kirschel ANG, Blumstein DT, Smith TB (2009) Character displacement of song and morphology in African tinkerbirds. *Proceedings of the National Academy of Sciences*, **106**, 8256–8261.

Kirschel ANG, Slabbekoorn H, Blumstein DT *et al.* (2011) Testing Alternative Hypotheses for Evolutionary Diversification in an African Songbird: Rainforest Refugia Versus Ecological Gradients. *Evolution*, **65**, 3162–3174.

Li Q, Zhao XL, Gilbert ER *et al.* (2015) Confined housing system increased abdominal and subcutaneous fat deposition and gene expressions of carbohydrate response element-binding protein and sterol regulatory element-binding protein 1 in chicken. *Genetics and molecular research: GMR*, **14**, 1220–1228.

Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Molecular Ecology*, **23**, 2178–2192.

Mantel N (1967) The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research*, **27**, 209–220.

Marie P, Labas V, Brionne A *et al.* (2015) Quantitative proteomics and bioinformatic analysis provide new insight into protein function during avian eggshell biomineralization. *Journal of Proteomics*, **113**, 178–193.

Martin C (1991) *The Rainforests of West Africa: Ecology, Threats, and Protection*. Birkhauser, Basel ; Boston.

Mayr E, O'Hara RJ (1986) The biogeographic evidence supporting the Pleistocene forest refuge hypothesis. *Evolution*, **40**, 55–67.

Mitchell MW, Locatelli S, Sesink Clee PR, Thomassen HA, Gonder MK (2015) Environmental variation and rivers govern the structure of chimpanzee genetic diversity in a biodiversity hotspot. *BMC Evolutionary Biology*, **15**, 1.

Moritz C, Patton JL, Schneider CJ, Smith TB (2000) Diversification of Rainforest Faunas: An Integrated Molecular Approach. *Annual Review of Ecology and Systematics*, **31**, 533–563.

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J (2000) Biodiversity hotspots for conservation priorities. *Nature*, **403**, 853–858.

Nadachowska-Brzyska K, Burri R, Olason PI *et al.* (2013) Demographic Divergence History of Pied Flycatcher and Collared Flycatcher Inferred from Whole-Genome Re-sequencing Data. *PLoS Genet*, **9**, e1003942.

Nadis S (2016) Life on the edge: Saving the world's hotbeds of evolution | New Scientist.

Nazareno AG, Bemmels JB, Dick CW, Lohmann LG (2017) Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species. *Molecular Ecology Resources*, n/a-n/a.

Ogden R, Thorpe RS (2002) Molecular evidence for ecological speciation in tropical habitats. *Proceedings of the National Academy of Sciences*, **99**, 13612–13615.

Oksanen J, Blanchet FG, Friendly M *et al.* (2017) *vegan: Community Ecology Package*.

Oliver MA, Webster R (1990) Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information Systems*, **4**, 313–332.

Orme CDL, Davies RG, Burgess M *et al.* (2005) Global hotspots of species richness are not congruent with endemism or threat. *Nature*, **436**, 1016–1019.

Orr MR, Smith TB (1998) Ecology and speciation. *Trends in Ecology & Evolution*, **13**, 502–506.

Price T (2008) *Speciation in birds*. Roberts and Co., Greenwood Village, Colo.

Proszkowiec-Weglarz M, Humphrey BD, Richards MP (2008) Molecular cloning and expression of chicken carbohydrate response element binding protein and Max-like protein X gene homologues. *Molecular and Cellular Biochemistry*, **312**, 167–184.

Räsänen K, Hendry AP (2008) Disentangling interactions between adaptive divergence and gene flow when ecology drives diversification. *Ecology Letters*, **11**, 624–636.

Ribas CC, Aleixo A, Nogueira ACR, Miyaki CY, Cracraft J (2011) A palaeobiogeographic model for biotic diversification within Amazonia over the past three million years. *Proceedings of the Royal Society of London B: Biological Sciences*, rspb20111120.

Robinson JA, Vecchyo DO-D, Fan Z *et al.* (2016) Genomic Flatlining in the Endangered Island Fox. *Current Biology*, **26**, 1183–1189.

Roesti M, Salzburger W, Berner D (2012) Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evolutionary Biology*, **12**, 94.

Romiguier J, Gayral P, Ballenghien M *et al.* (2014) Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, **515**, 261–263.

Rousset F (1997) Genetic Differentiation and Estimation of Gene Flow from F-Statistics Under Isolation by Distance. *Genetics*, **145**, 1219–1228.

Roy MS (1997) Recent diversification in African greenbuls (Pycnonotidae: Andropadus) supports a montane speciation model. *Proceedings of the Royal Society B: Biological Sciences*, **264**, 1337–1344.

Rundle HD, Nosil P (2005) Ecological speciation. *Ecology Letters*, **8**, 336–352.

Schluter D (2000) *The ecology of adaptive radiation*. Oxford University Press, Oxford.

Schluter D (2009) Evidence for Ecological Speciation and Its Alternative. *Science*, **323**, 737–741.

Schneider CJ, Smith TB, Larison B, Moritz C (1999) A test of alternative models of diversification in tropical rainforests: Ecological gradients vs. rainforest refugia. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 13869–13873.

Sehgal RNM, Buermann W, Harrigan RJ *et al.* (2011) Spatially explicit predictions of blood parasites in a widely distributed African rainforest bird. *Proceedings of the Royal Society of London B: Biological Sciences*, **278**, 1025–1033.

Sibbel SP, Talbert ME, Bowden DW *et al.* (2011) RGS6 variants are associated with dietary fat intake in Hispanics: the IRAS Family Study. *Obesity (Silver Spring, Md.)*, **19**, 1433–1438.

Slabbekoorn H, Smith TB (2002) Habitat-Dependent Song Divergence in the Little Greenbul: An Analysis of Environmental Selection Pressures on Acoustic Signals. *Evolution*, **56**, 1849–1858.

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, **139**, 457–462.

Smith TB, Calsbeek R, Wayne RK *et al.* (2005) Testing alternative mechanisms of evolutionary divergence in an African rain forest passerine bird. *Journal of Evolutionary Biology*, **18**, 257–268.

Smith TB, Harrigan RJ, Kirschel ANG *et al.* (2013) Predicting bird song from space. *Evolutionary Applications*, **6**, 865–874.

Smith TB, Holder K, Girman D *et al.* (2000) Comparative avian phylogeography of Cameroon and Equatorial Guinea mountains: implications for conservation. *Molecular Ecology*, **9**, 1505–1516.

Smith BT, McCormack JE, Cuervo AM *et al.* (2014) The drivers of tropical speciation. *Nature*, **515**, 406–409.

Smith TB, Schneider CJ, Holder K (2001) Refugial isolation versus ecological gradients. *Genetica*, **112–113**, 383–398.

Smith TB, Thomassen HA, Freedman AH *et al.* (2011) Patterns of divergence in the olive sunbird Cyanomitra olivacea (Aves: Nectariniidae) across the African rainforest–savanna ecotone. *Biological Journal of the Linnean Society*, **103**, 821–835.

Smith TB, Wayne RK, Girman DJ, Bruford MW (1997) A Role for Ecotones in Generating Rainforest Biodiversity. *Science*, **276**, 1855–1857.

Stein LR, Badyaev AV (2011) Evolution of eggshell structure during rapid range expansion in a passerine bird. *Functional Ecology*, **25**, 1215–1222.

Thomassen HA, Fuller T, Buermann W *et al.* (2011) Mapping evolutionary process: a multi-taxa approach to conservation prioritization. *Evolutionary Applications*, **4**, 397–413.

Trégouet D-A, Groop P-H, McGinn S *et al.* (2008) G/T Substitution in Intron 1 of the UNC13B Gene Is Associated With Increased Risk of Nephropathy in Patients With Type 1 Diabetes. *Diabetes*, **57**, 2843–2850.

Willing E-M, Dreyer C, Oosterhout C van (2012) Estimates of Genetic Differentiation Measured by FST Do Not Necessarily Require Large Sample Sizes When Using Many SNP Markers. *PLOS ONE*, **7**, e42649.

Zheng X, Levine D, Shen J *et al.* (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, **28**, 3326–3328.

Data Accessibility:

- RADseq data: NCBI SRA database BioProject ID PRJNA390986
- RNAseq data: NCBI SRA database BioProject ID PRJNA390772
- Data files including RAD loci consensus sequences, VCF file and sample information available at Dryad doi:10.5061/dryad.8n8t0

Table 1. Simple Mantel test for IBD (isolation-by-distance) and partial Mantel test for the effect of habitat.

Simple Mantel test: test for IBD

correlation between	Mantel r	p

F_{ST}	non-transformed distance	0.34	0.003
F_{ST}	log-transformed distance	0.29	0.008
$F_{ST}/(1- F_{ST})$	log-transformed distance	0.28	0.007

Partial Mantel test: test for the effect of habitat while controlling for IBD

correlation between		control for	Mantel r	p
F_{ST}	same habitat or not	non-transformed distance	0.48	9.00E-06
F_{ST}	same habitat or not	log-transformed distance	0.50	1.00E-06
$F_{ST}/(1- F_{ST})$	same habitat or not	log-transformed distance	0.50	3.00E-06

P-values were generated by 999,999 permutations.

Here “distance” refers to the geographic distance separating the pair of populations on which the F_{ST} value was computed.**Figure Legends**

Figure 1. Sampling and population structure. (A), Sampling locations. Each point is a sampling location, and habitat types are indicated by color same to (B). (B-C), PCA using SNPs that have a minor allele frequency higher than 2%. Each point presents a sample, and samples are colored by their habitat types (B) and by populations (C).

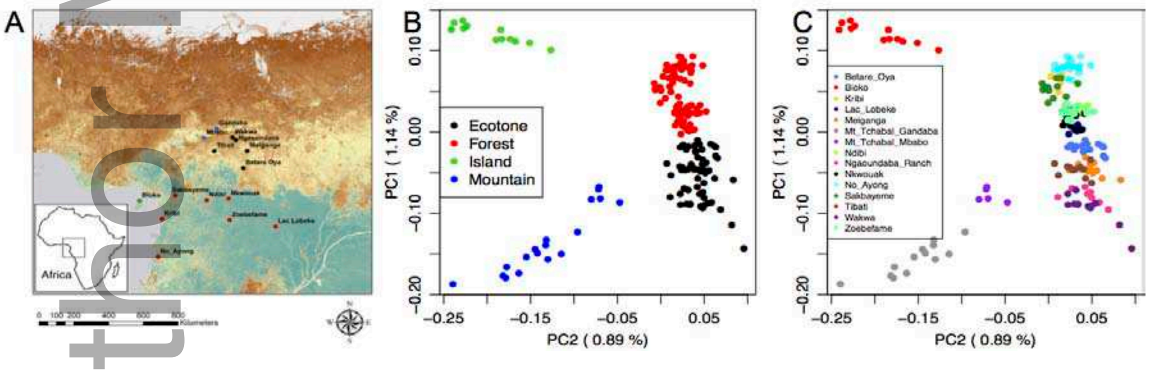


Figure 2. Pairwise population differentiation. (A), Pairwise F_{ST} between populations correlates with pairwise geographic distance between populations. Empty circles denote pairs of populations from the same type of habitat (shown by the color of the circle). Solid circles are pairs of populations from different types of habitats (shown by colors of the circle and inside). (B), Heat map of pairwise F_{ST}. Sampling locations are grouped by habitat type in both axes. (C) and (D), The pairwise F_{ST} of populations from different

habitats are greater than the pairwise F_{ST} of populations from the same habitat, even at the same geographic distance. (C) includes all populations from four habitats, and (D) includes only rainforest and ecotone populations. Histogram shows the null distribution of t-statistics generated by 10000 permutations of habitats within different bins of geographic distance (see Methods). Red dot shows the observed value.

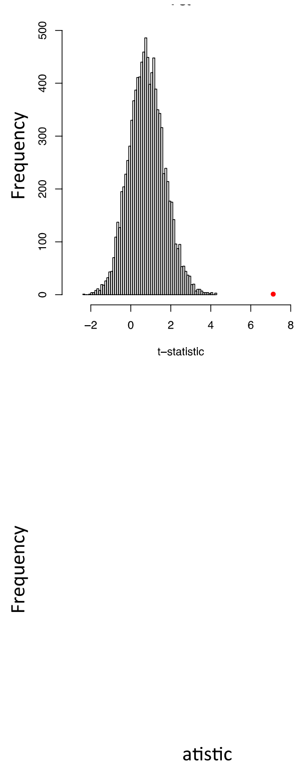


Figure 3. **Environmental drivers of genomic variation.** (A), Environmental and geographical variables ranked by their importance in explaining SNP allele frequency variation. (B), PC plot indicates the contribution of the environmental variables to the predicted patterns of frequency differentiation, with labeled vectors indicating the direction and magnitude of environmental gradients with greatest contribution. (C), Predicted spatial variation in population-level genetic composition from SNPs. Red points in (C) are locations where actual samples were collected in this study. Colors in (B) and (C) represent gradients in genomic turnover derived from transformed environmental predictors.

Locations with similar colors are expected to harbor populations with similar genetic composition.

Table 1. Simple Mantel test for IBD (isolation-by-distance) and partial Mantel test for the effect of habitat.

Simple Mantel test: test for IBD

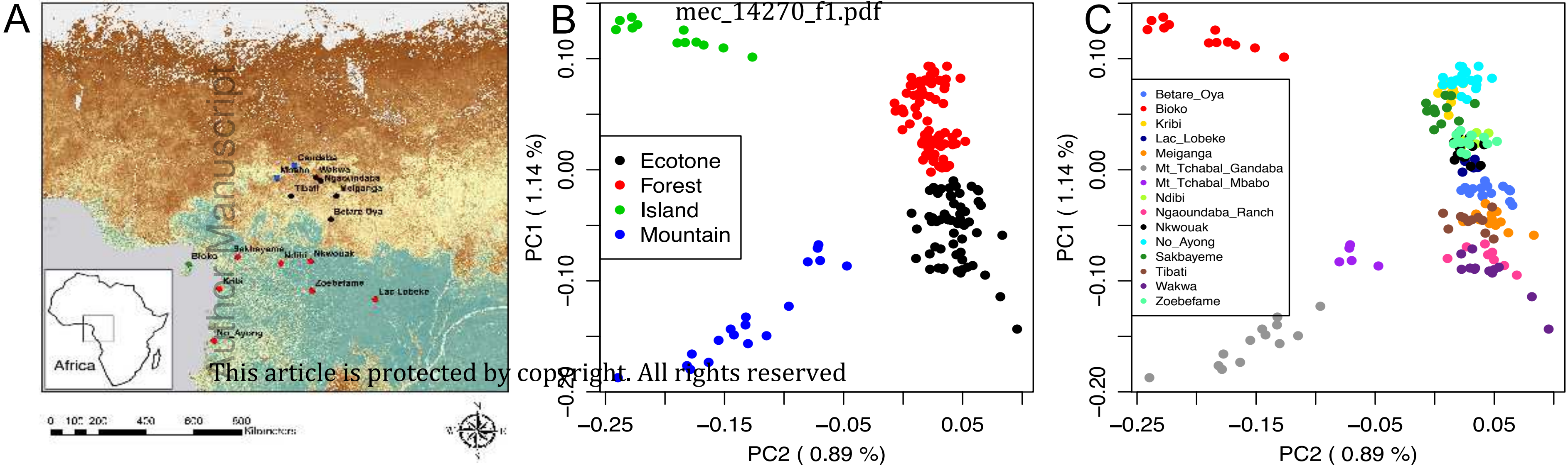
correlation between		Mantel r	p
F_{ST}	non-transformed distance	0.34	0.003
F_{ST}	log-transformed distance	0.29	0.008
$F_{ST}/(1- F_{ST})$	log-transformed distance	0.28	0.007

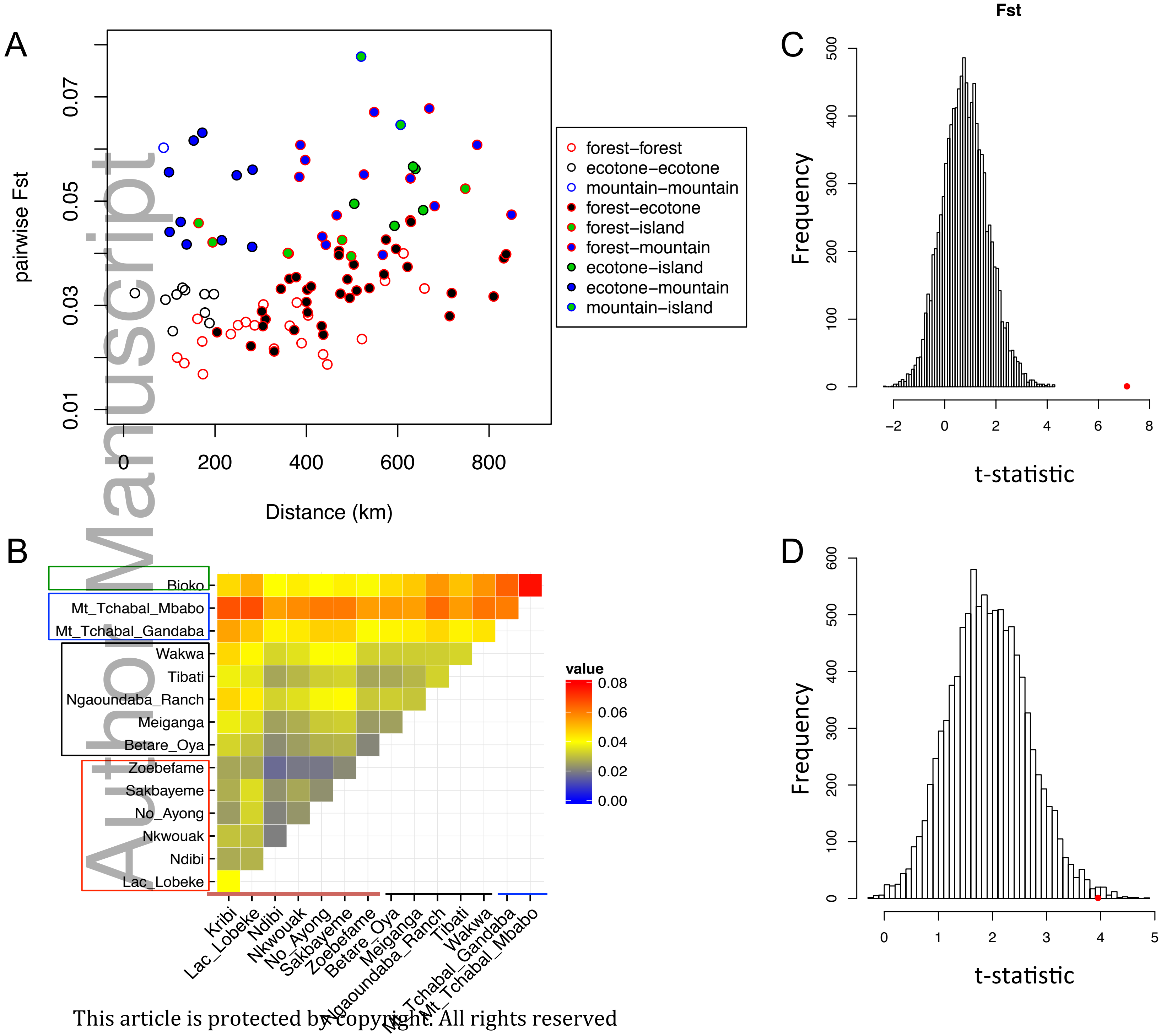
Partial Mantel test: test for the effect of habitat while controlling for IBD

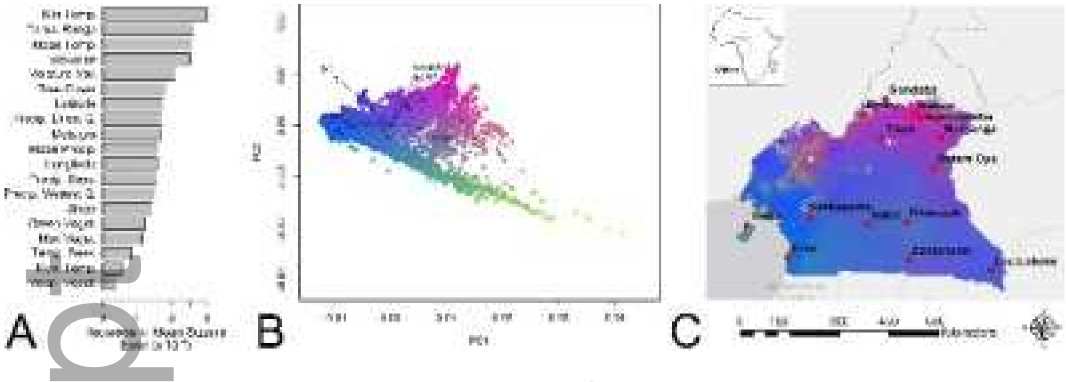
correlation between		control for	Mantel r	p
F_{ST}	same habitat or not	non-transformed distance	0.48	9.00E-06
F_{ST}	same habitat or not	log-transformed distance	0.50	1.00E-06
$F_{ST}/(1- F_{ST})$	same habitat or not	log-transformed distance	0.50	3.00E-06

P-values were generated by 999,999 permutations.

Here "distance" refers to the geographic distance separating the pair of populations on which the F_{ST} value was computed.


A
Gandaba
Mbabo
Wakwa
Ngaoundaba
Tibati
Meiganga
Betare Oya
Sakbayeme
Bioko
Ndibi
Nkwouak
Kribi
Zoebefame
Lac Lobeke
No_Ayong
Africa
0 100 200 400 600 800 Kilometers
N
W
E
S
B
Ecotone
Forest
Island
Mountain
PC1 (1.14 %)
0.10
0.00
−0.10
−0.20
PC2 (0.89 %)
−0.25
−0.15
−0.05
0.05
C
Betare_Oya
Bioko
Kribi
Lac_Lobeke
Meiganga
Mt_Tchabal_Gandaba
Mt_Tchabal_Mbabo
Ndibi
Ngaoundaba_Ranch
Nkwouak
No_Ayong
Sakbayeme
Tibati
Wakwa
Zoebefame
PC1 (1.14 %)
0.10
0.00
−0.10
−0.20
PC2 (0.89 %)
−0.25
−0.15
−0.05
0.05


A
pairwise Fst
0.01
0.03
0.05
0.07
0
200
400
600
800
Distance (km)
○ forest–forest
○ ecotone–ecotone
○ mountain–mountain
● forest–ecotone
● forest–island
● forest–mountain
● ecotone–island
● ecotone–mountain
● mountain–island
B
Bioko
Mt_Tchabal_Mbabo
Mt_Tchabal_Gandaba
Wakwa
Tibati
Ngaoundaba_Ranch
Meiganga
Betare_Oya
Zoebefame
Sakbayeme
No_Ayong
Nkwouak
Ndibi
Lac_Lobeke
Kribi
Lac_Lobeke
Ndibi
Nkwouak
No_Ayong
Sakbayeme
Zoebefame
Betare_Oya
Meiganga
Ngaoundaba_Ranch
Tibati
Wakwa
Mt_Tchabal_Gandaba
Mt_Tchabal_Mbabo
value
0.08
0.06
0.04
0.02
0.00
Fst
C
Frequency
0
100
200
300
400
500
−2
0
2
4
6
8
t-statistic
D
Frequency
0
100
200
300
400
500
600
0
1
2
3
4
5
t-statistic



mec_14270_f3.jpg