*SUPPLEMENT TO:*

# Genomic signatures of rapid adaptive divergence in a tropical montane species

Per G.P. Ericson [1 *], Martin Irestedt [1], Huishang She [2], Yanhua Qu [1, 2 *]

[1] *Department of Bioinformatics and Genetics, Swedish Museum of Natural History, PO Box 50007, SE-104 05, Stockholm, Sweden*
[2] *Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, 100101, China*

[*] Corresponding authors: per.ericson@nrm.se and quyh@ioz.ac.cn

Accepted for publication in *Biology Letters*.

*CONTENT*

1. MATERIAL AND METHODS (DETAILED DESCRIPTION)

*Samples information, extraction, sequencing, reference mapping and variant calling*

As a reference genome for subsequent mapping we built the *de novo* genome of a *Ailuroedus stonii* individual collected 28 May 1985 at Kerea, 50 km N of Port Moresby on the Vanapa River, 9°04'30''S, 147°10'30''E, Central Province, Papua New Guinea, (tissue sample Museum Victoria Z43608 = voucher specimen Australian National Wildlife Collection ANWC B24961). DNA was extracted from cryofrozen tissue using the KingFisher duo extraction robot and the KingFisher™ Cell and Tissue DNA Kit according to the manufacturer's instructions. A total of 217.65 Gb of high-quality sequence data was obtained by sequencing five DNA libraries (one short-insert-sized, paired-end [180 bp] library, two mate-pair [3 and 5-8 kb] libraries, one 10X Genomics Chromium Genome library, and one Hi-C library) on an Illumina HiSeq X platform at the National Genomics Institute. The Hi-C sequencing was made by the Science for Life Laboratory (National Genomics Institute, Stockholm). In the first step a library was prepared using the Omni-C (Dovetail Genomics) kit. This is a proximity-ligation protocol using a sequence-independent endonuclease, generating data for identification of topologically-associated functional domains and scaffolding. The library was sequenced by the means of restriction enzyme digestion, followed by proximity ligation to capturing genome organisation within a ligated molecule. Ligated library molecules were then sequenced as paired-end reads to reveal which of the genome were physically proximal in the nuclei at the time of sampling.

Low quality and duplicated reads were filtered out before an initial assembly done by the Science for Life Laboratory (National Genomics Institute, Stockholm) using the ALLPATHS_LG assembler (Butler et al. 2008). The final genome assembly was done using the HiRise pipeline (Dovetail Genomics, Putnam et al. 2016) and resulted in a genome length of 1,092 Mb (6,982X estimated physical coverage). The assembly resulted in 2,364 scaffolds covering 1,092 Mb with a scaffold N50 of 75 Mb and contig N50 of 436 kb. The 23 largest scaffolds cover 97% of the genome and we used these for the downstream analyses. They were annotated by blasting the window sequences to the chicken genome (Gallus.gallus.5.0.cds) using BLAST+ v2.6.0 (Camacho et al. 2009). For 22 scaffolds we found almost perfect matches with individual chicken chromosomes 1 to 20 (in two cases two scaffolds matched different parts of the same chicken chromosome). The remaining scaffold seemingly comprises genomic regions corresponding to chicken chromosomes 21, 22, 23, 24, 27, 28 and 33. We stress that when using chromosome numbers in the text and figures these are only tentative and refer to the chicken genome.

*Amblyornis papuensis* is a rare bird species, both in the wild and in museum collections. It occupies parts of the New Guinea Highlands (Central Range) and has two subspecies, *Amblyornis p. papuensis* in the west of the species' range in Indonesia (Wissel Lakes, in Weyland Mts area; Nassau Range and Oranje Mts; Bele R and L Habbema region), and *Amblyornis p. sanfordi* in the east of the range in Papua New Guinea (Mt Hagen, Mt Giluwe, Tari Gap, and S Karius Range) (Frith & Frith 2020). We are not aware of any fresh sample of the species, but by courtesy of American Museum of Natural History, New York, Australian National Wildlife Collection, Canberra, Natural History Museum, Tring, and Yale Peabody Museum, New Haven, we have obtained toe pads of eight museum study skins, amounting to four each of the two subspecies. DNA was extracted from the samples using the Qiagen QIAamp DNA Mini Kit following the protocol described in Irestedt et al. (2006). The sequencing libraries were prepared using the protocol published by Meyer and Kircher (2010). Each library was then amplified in four independent PCR reactions with unique indexes and these uniquely tagged products were pooled and cleaned. Finally, the samples were pooled and sequenced at equal molarity on one lane on the Illumina NovaSeq platform. The Illumina sequencing reads were processed using a custom-designed workflow available

at https://github.com/mozesblom to remove adapter contamination, low-quality bases and low-complexity reads. Overlapping read pairs were merged using PEAR (Zhang et al. 2014) and SuperDeduper (Petersen et al. 2015) was used to remove PCR duplicates. Trimming and adapter removal was done with Trimmomatic v0.32 (Bolger et al. 2014; default settings) and overall quality and length distribution of sequence reads were inspected with Fastqc v0.11.5 (Andrews 2010) before and after the cleaning. As erroneous DNA degradation patterns almost exclusively appear at the ends of sequence reads (Schubert et al. 2012), we shortened all reads obtained from museum study skins by deleting 5bp from both ends in order to reduce this "noise".

We used BWA mem v0.7.12 (Li & Durbin 2009) to map the polished reads against the 23 largest scaffolds (including 97% of the whole genome) of the *Ailuroedus stonii* genome. The mapping resulted in a mean coverage of 20.6X (range 15-31X) for the eight individuals. We chose to reduce the number of scaffolds in order to speed-up the variant calling. High-quality SNPs were called from the BAM-files using mpileup in Samtools v1.4 (Li et al. 2009) and including individual biallelic genotypes with a depth between 10 and 100 reads per individual and a quality of 10 or more. A total of 2,467,355 high-quality SNPs were used for downstream analyses.

*Population structure and demographic history*

Population genetic structure was inferred from the full data set of 2,467,355 SNPs with a principal component analysis (PCA) using *smartpca* in EIGENSOFT v6.1.4 (Price et al. 2006). Figure 1*c* shows a striking difference in the variation of individual PCA scores within the *papuensis* and *sanfordi* populations, respectively. We believe this reflects the considerably larger variation in heterogeneity distributions among the *papuensis* individuals than among those of *sanfordi* (electronic supplementary material, figure S5).

We also used the coalescent-based program SNAPP v1.3.0 in BEAST2 v2.4.8 (Bouckaert et al. 2014; Bryant et al. 2012) to perform a Bayesian MCMC analysis of the SNP data. The analyses may be interpreted with some caution as SNAPP assumes a strict isolation model with constant population sizes, a condition that may not be met in the populations (see Results). To make the analysis computationally feasible we first randomly sampled 1% SNPs from the full data set and then filtered the non-biallelic variants. In the end, 80,157 biallelic SNPs shared by the two populations of *Amblyornis papuensis* and three outgroup species (*Amblyornis macgregoriae*, *A. subalaris* and *A. inornata*) were retained for analysis. All individuals were assigned to different "populations" to avoid inferring a certain topology on the analysis. We chose wide and uninformative distributions as priors of the model parameters. The forward and backward mutation rates were set to be estimated during the course of the MCMC chain, and the rate parameters were sampled from an inverse gamma distribution. For the Yule prior for the species tree the lambda parameter, that governs the rate of divergence, was uniformly distributed in the range of zero to one. We run the analyses for 2,000,000 iterations and assessed convergence of the MCMC chains by plotting likelihood scores against iterations and checking that all parameter ESS values were 200 or larger. Trees were sampled for every 1,000 iteration and we discarded the first 10% as burn-in. We ran the analysis three times to ascertain stability of the results. We plotted the distribution of species trees in the posterior sample using DensiTree v2.1.11 (Bouckaert 2010).

The standardized multilocus heterozygosity was calculated from the 2,467,355 SNPs data set using the R script inbreedR (https://rdrr.io/cran/inbreedR/src/R/).

We studied the demographic history of *Amblyornis papuensis* using different, complementary methods. PopSizeABC (Boitard et al. 2016), an approximate Bayesian computation pipeline, was used to estimate temporal variation in effective population size ($N_e$) in each lineage. PopSizeABC estimates variation in population size for a group of

individuals and traces the temporal variation up until a few thousand years ago. PopSizeABC uses the SNP data set (2,467,355 SNPs) to calculate summary statistics of the genome-wide site frequency spectrum (SFS) and the average zygotic linkage disequilibrium (LD) at specific time bins (Boitard et al. 2016). These statistics are first calculated for an empirical data set, and then compared with the corresponding statistics calculated from a large number of simulated data sets. The simulated data sets are obtained by cutting the empirical data set into segments of 2 million bp each and then randomly select 100 such segments. $N_e$ was estimated in 21 discrete time windows between 2,400 to 130,000 years BP. In the analyses we set the recombination rate to $1.0*10^{-8}$, and the genomic mutation rate per generation to $4.6*10^{-9}$ (Smeds et al. 2016). The general biology and life-history parameters for *Amblyornis papuensis* are largely unknown, and this is also true for the generation length (measured as the average age of parents of the population). The generation length depends not only on the age of first reproduction, but also on adult survival and maximum longevity. Bird et al. (2020) estimated that *Amblyornis papuensis* has a generation length of 5.35 years, which we have used herein. We compared the summary statistics for the empirical data sets with 400,000 simulated data sets to identify the simulations that are most similar. These were then selected by applying a simple rejection method with an acceptance (tolerance) rate of 0.01.

We also used Fastsimcoal v2.6 (Excoffier et al. 2013) to infer the demographic history of *Amblyornis papuensis*. We generated a two-dimensional, unfolded site frequency spectrum (SFS) using the doSaf and realSFS functions in ANGSD v0.917 (Korneliussen et al. 2014) for a 76 Mb region of scaffold 446 that does not contain any highly (top 5%) selected 50 kb windows. We compared four demographic models combining two hypothetical demographic events: a bottleneck (present or not) in the ancestral population and presence or not of bidirectional gene flow between the current populations. All parameters were selected from uniform distributions. For each demographic inference, we ran two separate analyses with 100 replicates each, and we set the number of coalescent simulations (-n) to 200,000. We used the Akaike (1974) information criterion (AIC) to evaluate which model had the higher likelihood. For this model, we run simulations in 1,000 replicates, each including 20 estimation loops with 200,000 coalescent simulations. To determine the best parameter estimates, we selected the 5% most likely replicate runs (that is, those with the smallest difference between the estimated and observed likelihood) and used this subset to calculate the mean for all demographic parameters, along with their 95% confidence intervals (based on 200,000 resamples). We re-calculated estimates of divergence time in units of years, effective population sizes, and migration rates by scaling with a neutral mutation rate of $4.6*10^{-9}$ (Smeds et al. 2016) and a generation time of 5.35 years (see above).

*Environmental heterogeneity analysis and selection*

To compare local climatic conditions for the two populations in New Guinea we used QGIS v3.10.6 to collect 19 bioclimatic variables (BIO1-BIO19, 5 minutes data) from the WorldClim v2.1 database for the period 1970-2000 (Fick & Hijmans 2017) for a large number of randomly sampled localities within each of the two population's core distribution. Totally 19 bioclimatic variables were scored for 1,939 (western distribution) and 2,938 (eastern distribution) localities situated between 2,600 and 2,800 m a.s.l. Although *Amblyornis papuensis* occurs at both lower and higher altitudes than this (Frith & Frith 2020), we chose to restrict the elevational range to get comparable climatic data from each area. The Z-transformed data for all localities in each core area was subjected to principal component analysis (*prcomp*) in R. Differences in component mean values were tested for statistical significance with two-tailed *t*-tests. To test the robustness of our results we repeated the analysis for a wider elevational span – for localities situated between 2,300-3,300 m a.s.l. The resulting plot of PC2 and PC3 are almost identical (see figure S2 below).

To evaluate the contribution of divergent selection to the genetic differentiation between *papuensis* and *sanfordi* we calculated the Z-transformed $F_{ST}$ value Z($F_{ST}$) of the populations of each non-overlapping 50 kb window. Estimations of the patterns of linkage disequilibrium (LD) with PopLDdecay (Zhang et al. 2019) show that LD is not a problem when applying a window size of 50 kb. The 50 kb windows with Z($F_{ST}$) over 1.76 (the top 5%) were arbitrarily defined as outlier regions. We annotated these by blasting the window sequences to the chicken genome (Gallus.gallus.5.0.cds) using BLAST+ v2.6.0 (Camacho et al. 2009). Identified genes within the top 5% selected windows were then referred to gene ontology (GO) categories using Panther Classification System (Mi et al. 2019).

We localized targets of strong selective sweeps by analyzing SNP patterns using the site frequency spectrum (SFS). The selective sweeps should be complete to be efficiently detected, i.e., the beneficial mutation should be fixed in the population and not be fixed for too long. The sweeps should also be strong in relation to recombination as the method we used utilize the neutral genomic regions that are around the beneficial mutation. SweeD (Pavlidis et al. 2013) calculates SFS in a grid (i.e. windows) across each scaffold and implements a composite likelihood ratio (CLR) test based on the sweepfinder algorithm (Nielsen et al. 2005). The CLR uses the variation of the whole or derived SFS of a whole scaffold to compute the ratio of the likelihood of a selective sweep at a given position to the likelihood of a null model without a selective sweep (Badouin et al. 2017). The null hypothesis relies on the SFS of the whole-genome sequence rather than on a standard neutral model, which makes it more robust to demographic events such as population expansions (Nielsen et al. 2005; Pavlidis et al. 2010). CLRs were computed at 4kb equidistant points between the first and last SNP in every scaffold. Based on the results of the $F_{ST}$ analysis of selection we can safely assume that parts of the genomes have evolved under positive selection. This allows us to define a significance threshold from the data set itself. We compared the mean CLR across all points in the two populations, as well as the number of points with a CLR score larger than two times the standard deviation of the mean for the population.

References


Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.

Andrews, S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Badouin, H., Gladieux, P., Gouzy, J., Siguenza, S., Aguileta, G., Snirc, A., Le Prieur, S., Jeziorski, C., Branca, A. & Giraud, T. 2017. Widespread selective sweeps throughout the genome of model plant pathogenic fungi and identification of effector candidates. *Molecular Ecology, 26,* 2041-2062.

Boitard, S., Rodríguez, W., Jay, F., Mona, S. & Austerlitz, F. 2016. Inferring population size history from large samples of genome-wide molecular data - an approximate Bayesian computation approach. *PLoS Genetics, 12,* e1005877.

Bolger, A.M., Lohse, M. & Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30,* 2114-2120.

Bouckaert, R. 2010. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics, 26,* 1372-1373.

Bouckaert, R, Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M.A., Rambaut, A. & Drummond, A.J. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput. Biol., 10(4),* e1003537.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. & RoyChoudhury, A. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution, 29,* 1917-1932.

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., Nusbaum, C. & Jaffe, D.B. 2008. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Research, 18,* 810-820.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T.L. 2009. BLAST+: architecture and applications. *BMC Bioinformatics, 10,* 421.

Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V.C. & Foll, M. 2013. Robust demographic inference from genomic and SNP data. *PLoS Genetics, 9,* e1003905.

Fick, S.E. & Hijmans, R.J. 2017. WorldClim 2: new 1km spatial resolution climate surfaces for global land areas. *International Journal of Climatology, 37,* 4302-4315.

Frith, C.B., Frith, D.W. 2004. *The Bowerbirds.* Oxford: Oxford University Press.

Frith, C.B., Frith, D.W. 2020. Archbold's Bowerbird (*Archboldia papuensis*), version 1.0. In *Birds of the World* (J. del Hoyo, A. Elliott, J. Sargatal, D.A. Christie, and E. de Juana, Editors). Cornell Lab of Ornithology, Ithaca, NY, USA.

Irestedt, M., Ohlson, J.I., Zuccon, D., Källersjö, M. & Ericson, P.G.P. 2006. Nuclear DNA from old collections of avian study skins reveals the evolutionary history of the Old World suboscines (Aves, Passeriformes). *Zoologica Scripta, 35,* 567-580.

Korneliussen, T.S., Albrechtsen, A. & Nielsen, R. 2014. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics, 15,* 356.

Li H. & Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25,* 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25,* 2078-2079.

Meyer, M. & Kircher, M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc., 2010(6)*:prot5448.

Mi, H., Muruganujan, A., Huang, X., Ebert, D., Mills, C., Guo, X. & Thomas, P.D. 2019. Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nature Protocols, 14,* 703-721. https://doi.org/10.1038/s41596-019-0128-8

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M.J., Clark, A.G. & Bustamante, C. 2005. Genomic scans for selective sweeps using snp data. *Genome Research, 15,* 1566-1575. doi: 10.1101/gr.4252305.

Pavlidis, P., Jensen, J.D. & Stephan, W. 2010. Searching for footprints of positive selection in whole-genome snp data from nonequilibrium populations. *Genetics, 185,* 907-922.

Pavlidis, P., Zivkovic, D., Stamatakis, A. & Alachiotis, N. 2013. SweeD: likelihood-based detection of selective sweeps in thousands of genomes. *Molecular Biology and Evolution, 30,* 2224-2234.

Petersen, K.R., Street, D.A., Gerritsen, A.T., Hunter, S.S. & Settles M.L. 2015. Super deduper, fast PCR duplicate detection in fastq files. In: *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics.* p. 491-492.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. & Reich, D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics, 38,* 904-909.

Putnam, N.H., O'Connell, B.L., Stites, J.C., Rice, B.J., Blanchette, M., Calef, R., Troll, C.J., Fields, A., Hartley, P.D., Sugnet, C.W., Haussler, D., Rokhsar, D.S. & Green, R.E.

2016. Chromosome-scale shotgun assembly using an in vitro method for longrange linkage. *Genome Research, 26,* 342-350.
Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J. F., Al-Rasheid, K. A., Willerslev, E., Krogh, A., & Orlando, L. 2012. Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics, 13,* 178.
Smeds, L., Qvarnström, A. & Ellegren, H. 2016. Direct estimate of the rate of germline mutation in a bird. *Genome Research, 26,* 1211-1218.
Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M. & Yang, T.-L. 2019. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics, 35,* 1786-1788.
Zhang, J.J., Kobert, K., Flouri, T. & Stamatakis A. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics, 30,* 614-620.

## 2. DATA REPOSITORIES

Raw Illumina sequences are deposited in Sequence Reads Archive, National Center for Biotechnology Information, SRA accession [NCBI does not allow a temporary link to the data itself, only to metadata: ftp://ftp-trace.ncbi.nlm.nih.gov/sra/review/SRP242614_20210119_164248_5db09cb32d8e4e9ff6be1f76a085f9ce].
The *Ailuroedus stonii* genome assembly and data related to the analyses are deposited in Dryad [temporary link during the review process: https://datadryad.org/stash/share/WUc1va-3KDbDtUlhKeF5H05snl8ZauRh-5jiOekyNFc].

## 3. SUPPLEMENTARY FIGURES

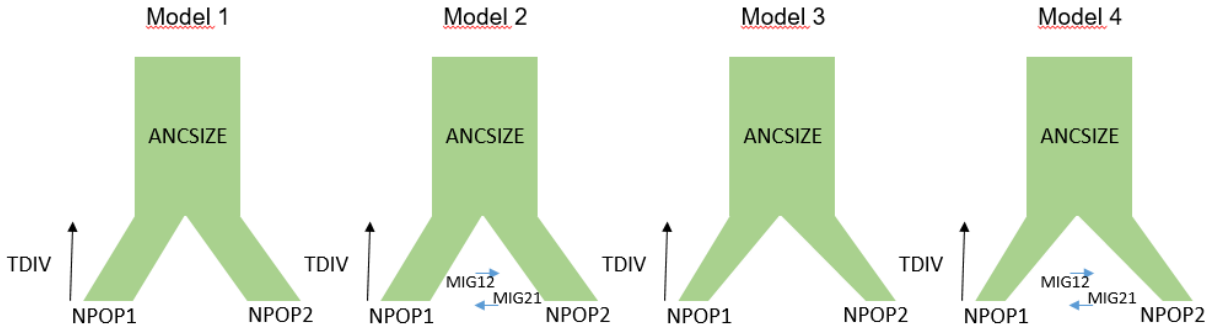


*Figure S1:* The observed site frequency spectrum (SFS) was compared with simulations using four models of hypothesized demographic events. All models postulate a split of an ancestral population into two daughter populations. The models differ in that two (M3 and M4) hypothesize a decrease of the effective population size after the split, and two (M1 and M2) hypothesize occurrence of gene flow between the daughter populations.

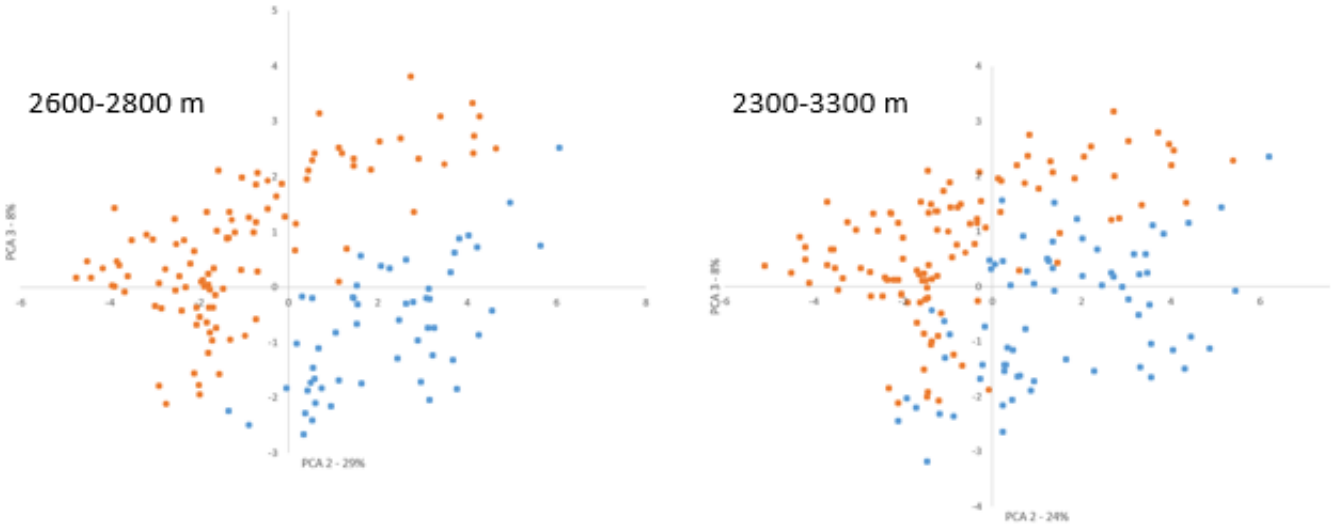


*Figure S2:* Results of principal component analyses of 19 bioclimatic variables (BIO1-BIO19, 5 minutes data) obtained for a large number of random localities within each of the two populations' core-areas in the Central Range in New Guinea. Orange dots represents localities within the western Central Range (*papuensis* core-distribution) and blue dots localities within the eastern Central Range (*sanfordi* core-distribution). The left figure shows the result when restricting the localities to those at elevations between 2,600-2,800 m a.s.l. The right figure shows the results for a wider elevational range, 2,300-3,300 m a.s.l. The two analyses yield highly similar results and we have based our results on the more narrow elevational range where the vast majority of the individuals in each population lives.

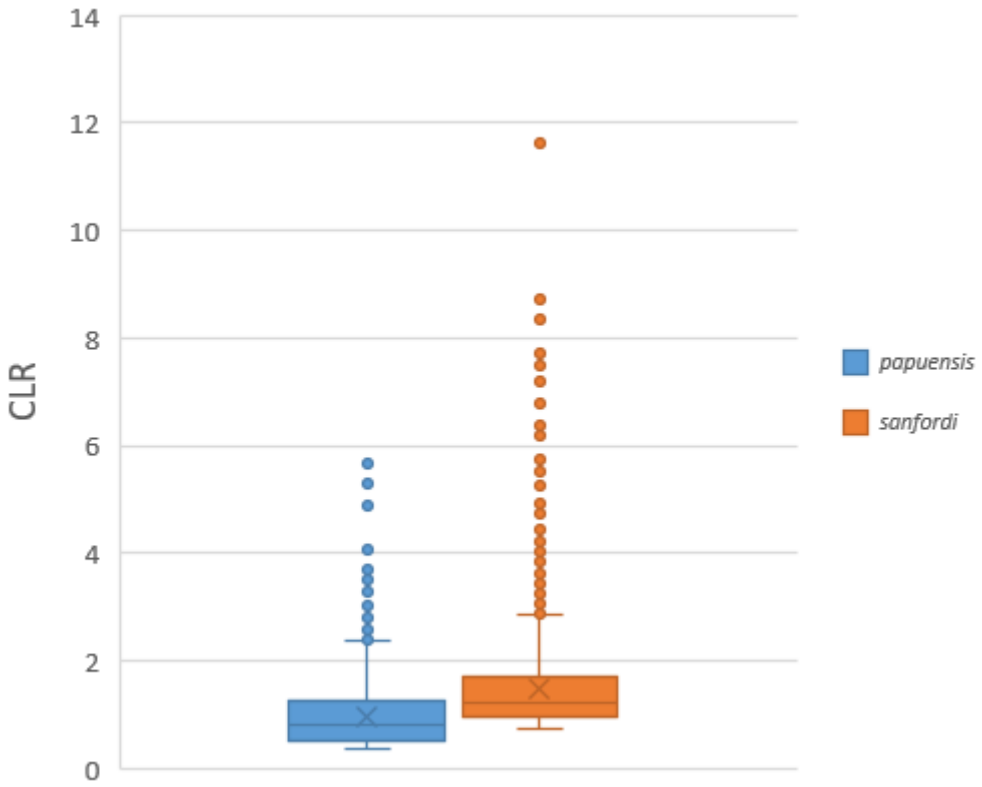


*Figure S3:* Box-plots of CLR (composite likelihood ratio) values for regions across the whole genome exhibiting signs of strong selective sweeps (CLR > 2*s.d.) in populations *papuensis* (n = 2,184) and *sanfordi* (n = 4,104), respectively. The average CLR value across the genome is almost three times larger in the *sanfordi* population than in *papuensis* (0.11 vs. 0.04, Wilcoxon, p <0.0001).

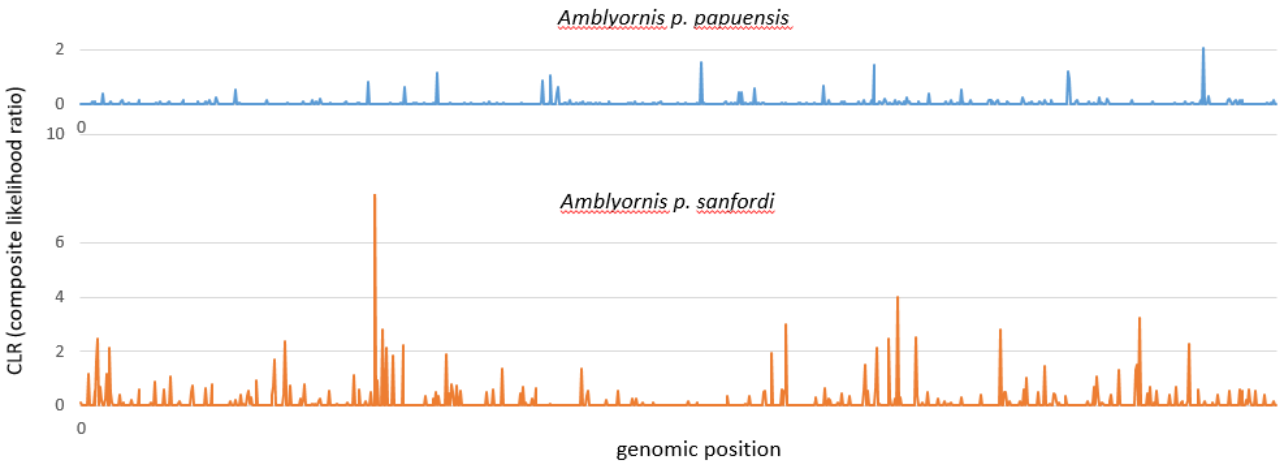


*Figure S4:* Genome-wide analyses of selective sweeps with SweeD shows that the number of windows exhibiting signs of selective sweeps are more common in *sanfordi*, and that the CLR (composite likelihood ratio) values are also higher in this population. The average CLR per window is 0.04 for *papuensis* and 0.11 for *sanfordi* (the difference is not statistically significant) and the number of regions with a CLR larger than 2*s.d. are 2,184 (2.4%) in *papuensis* and 4,104 (4.4%) in *sanfordi* ($\chi^2$ 607.00, $p < 0.0001$). The figure shows a genomic region that constitutes ca. 10% of the total genome.

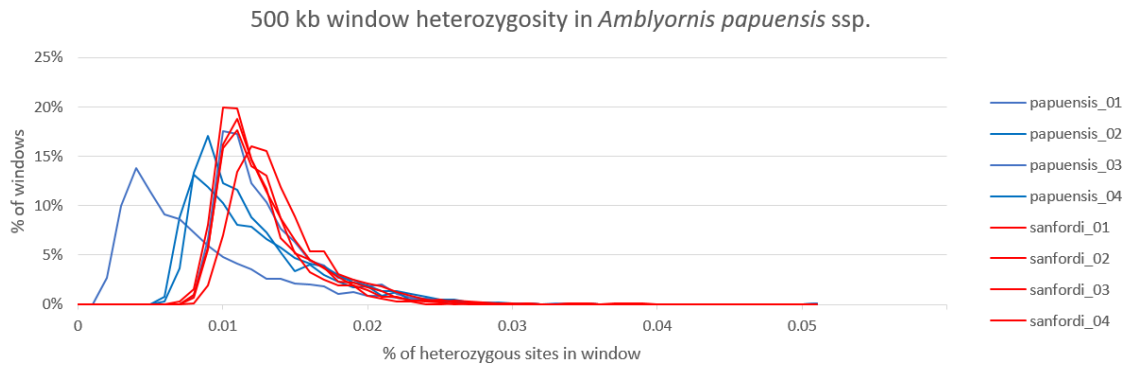


*Figure S5:* Comparison of heterogeneity distributions in the eight individuals studied. The individuals of the *sanfordi* population shows a larger degree of similarity than those of the *papuensis* population.

## 4. SUPPLEMENTARY TABLES

*Table S1:* Samples sequenced in the study. Abbreviations: AMNH, American Museum of Natural History, New York; ANWC, Australian National Wildlife Collection, Canberra; BMNH, Natural History Museum, London; YPM, Peabody Museum, Yale University, New Haven.

| Taxon | Sample ID | Locality | Date | Sex | Material | Coverage |
|---|---|---|---|---|---|---|
| *Amblyornis papuensis papuensis* | AMNH 342259 | Indonesia, Papua Prov., Bele River, 2200 m | 4 Dec 1938 | male | toe pad | 15.3 |
| *Amblyornis papuensis papuensis* | AMNH 342261 | Indonesia, Papua Prov., Bele River, 2200 m | 23 Nov 1938 | female | toe pad | 20.4 |
| *Amblyornis papuensis papuensis* | YPM 75437 | Indonesia, Papua Prov., Ilaga | 6 Sep 1960 | male | toe pad | 27.4 |
| *Amblyornis papuensis papuensis* | YPM 75438 | Indonesia, Papua Prov., Ilaga | 9 Sep 1960 | male | toe pad | 31.1 |
| *Amblyornis papuensis sanfordi* | BMNH 195317217 | Papua New Guinea, Mount Giluwe | 1953 | male | toe pad | 19.5 |
| *Amblyornis papuensis sanfordi* | ANWC B01296 | Papua New Guinea, Southern Highlands, Kare, Mendi, -6.0684, 143.7019 | 5 Sep 1961 | female | toe pad | 20.7 |
| *Amblyornis papuensis sanfordi* | AMNH 705703 | Papua New Guinea, Western Highlands prov., Base Camp, Tomba, Mt. Hagen, 9-10,000 ft | 12 Jul 1950 | male | toe pad | 14.8 |
| *Amblyornis papuensis sanfordi* | AMNH 705710 | Papua New Guinea, Western Highlands prov., Base Camp, Tomba, Mt. Hagen, 8,700 ft | 20 Jul 1950 | female | toe pad | 15.5 |

*Table S2:* The 23 largest scaffolds were analysed, corresponding to 97% of the total genome length. Blasting to the chicken genome showed that most scaffolds match chicken chromosomes.

| scaffold | length (bp) | inferred chicken chromosome |
|---|---|---|
| 1 | 31'822'539 | 8 |
| 2 | 20'509'599 | 4 |
| 37 | 21'332'042 | 10 |
| 40 | 11'197'321 | 19 |
| 80 | 11'492'157 | 17 |
| 81 | 35'485'005 | 6 |
| 130 | 12'831'607 | 18 |
| 131 | 16'864'701 | 14 |
| 166 | 25'960'477 | 9 |
| 229 | 15'015'738 | 20 |
| 284 | 65'164'962 | 5 |
| 285 | 56'874'832 | 21, 22, 23, 24, 27, 28, 33 [3] |
| 331 | 116'393'903 | 3 |
| 346 | 38'874'525 | 7 |
| 370 | 19'267'510 | 13 |
| 405 | 73'802'003 | 4 [1] |
| 445 | 80'441'600 | 10 |
| 499 | 21'120'569 | 12 |
| 527 | 75'446'723 | 1 [12] |
| 541 | 20'950'921 | 11 |
| 605 | 14'570'870 | 15 |
| 631 | 155'504'331 | 2 [19, 20] |
| 1402 | 119'797'195 | 1 [8, 9, 23] |
| Sum | 1'060'721'130 | |

*Table S3:* The demographic models tested using Fastsimcoal. The models differ in their combination of a hypothesized population change after the split between the populations, and the presence of gene flow (electronic supplementary material, figure S1). Model M4, inferring size changes of the populations and presence of gene flow after they split, received the highest likelihood by AIC comparison.

| | Holocene population change | Recent migration | log-likelihood | Parameters | AIC | DeltaAIC | Weight |
|---|---|---|---|---|---|---|---|
| **M4** | **yes** | **yes** | **-24802625.61** | **12** | **49605275.22** | **0** | **1** |
| M2 | no | yes | -24806934.34 | 7 | 49613882.68 | 8607 | 0 |
| M3 | yes | no | -24819587.29 | 9 | 49639192.59 | 33917 | 0 |
| M1 | no | no | -24820216.90 | 5 | 49640443.79 | 35169 | 0 |

*Table S4:* Inferred demographic parameters under model M4 (Figure 1d). The parameter TDIV is scaled by a generation length of 5.35 years.

| Parameters | Point estimation | 95% C.I. | |
|---|---|---|---|
| | | lower bound | upper bound |
| Size of ancestral population (ANCSIZE), number of individuals | 3870 | 3554 | 4187 |
| Effective population size *papuensis* (NPOP1), number of individuals | 1645 | 1444 | 1846 |
| Effective population size *sanfordi* (NPOP2), number of individuals | 182 | 162 | 211 |
| Divergence time between *papuensis* and *sanfordi* (TDIV), years | 11834 | 10174 | 12457 |
| Geneflow between *papuensis* and *sanfordi*, number of individuals/generation (MIG12) | 0.005 | 0.005 | 0.007 |
| Geneflow between *sanfordi* and *papuensis*, number of individuals/generation (MIG21) | 0.012 | 0.011 | 0.014 |

*Table S5:* Test of means of 19 bioclimatic variables observed at randomly sampled localities within the distributions (between 2,600-2,800 m a.s.l.) of populations *papuensis* (western New Guinea, 1939 localities) and *sanfordi* (eastern New Guinea, 2938 localities), respectively.

| | western New Guinea | | eastern New Guinea | | | | |
|---|---|---|---|---|---|---|---|
| | mean | s.d. | mean | s.d. | *t* | p (two-tailed) | Bonferroni corrected |
| BIO1: Annual Mean Temperature | 13.13 | 1.711 | 13.00 | 0.892 | 0.92 | 0.3587 | n.s. |
| BIO2: Mean Diurnal Range | 10.77 | 0.289 | 11.02 | 0.337 | -9.11 | 0.0000 | p < .00001 |
| BIO3: Isothermality | 91.33 | 0.760 | 91.44 | 0.555 | -1.85 | 0.0655 | n.s. |
| BIO4: Temperature Seasonality | 31.78 | 2.663 | 37.81 | 3.137 | -23.13 | 0.0000 | p < .00001 |
| BIO5: Max Temperature of the Warmest Month | 19.09 | 1.685 | 18.94 | 0.960 | 1.09 | 0.2772 | n.s. |
| BIO6: Min Temperature of the Warmest Month | 7.30 | 1.701 | 6.89 | 0.882 | 3.30 | 0.0010 | n.s. |
| BIO7: Temperature Annual Range | 11.79 | 0.330 | 12.05 | 0.359 | -8.58 | 0.0000 | p < .00001 |
| BIO8: Mean Temperature of the Wettest Quarter | 13.19 | 1.705 | 13.34 | 0.900 | -1.31 | 0.1904 | n.s. |
| BIO9: Mean Temperature of the Driest Quarter | 12.98 | 1.684 | 12.53 | 0.890 | 3.60 | 0.0003 | p < .01 |
| BIO10: Mean Temperature of the Warmest Quarter | 13.43 | 1.717 | 13.39 | 0.894 | 0.23 | 0.8176 | n.s. |
| BIO11: Mean Temperature of the Coldest Quarter | 12.72 | 1.722 | 12.53 | 0.889 | 1.55 | 0.1217 | n.s. |
| BIO12: Annual Precipitation | 3018.87 | 184.026 | 2789.25 | 164.736 | 14.70 | 0.0000 | p < .00001 |
| BIO13: Precipitation of the Wettest Month | 309.54 | 14.761 | 320.11 | 12.333 | -8.69 | 0.0000 | p < .00001 |
| BIO14: Precipitation of the Driest Month | 221.48 | 18.862 | 150.18 | 21.160 | 39.77 | 0.0000 | p < .00001 |
| BIO15: Precipitation Seasonality | 11.73 | 1.923 | 24.13 | 3.616 | -47.87 | 0.0000 | p < .00001 |
| BIO16: Precipitation of the Wettest Quarter | 881.64 | 36.784 | 875.61 | 29.034 | 2.04 | 0.0422 | n.s. |
| BIO17: Precipitation of the Driest Quarter | 679.20 | 50.741 | 481.24 | 63.722 | 38.43 | 0.0000 | p < .00001 |
| BIO18: Precipitation of the Warmest Quarter | 801.98 | 35.816 | 799.40 | 55.669 | 0.62 | 0.5374 | n.s. |
| BIO19: Precipitation of the Coldest Quarter | 703.40 | 60.569 | 486.47 | 68.856 | 37.40 | 0.0000 | p < .00001 |

*Table S6:* Principal component analysis of 19 bioclimatic variables observed at random localities within each population's core distribution area. For a detailed description of methods, see the text in electronic supplementary material.

| | PCA_1 | PCA_2 | PCA_3 | PCA_4 | PCA_5 | PCA_6 | PCA_7 | PCA_8 | PCA_9 | PCA_10 | PCA_11 | PCA_12 | PCA_13 | PCA_14 | PCA_15 | PCA_16 | PCA_17 | PCA_18 | PCA_19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BIO1: Annual Mean Temperature | 0.292 | 0.178 | | | 0.106 | | | | | | | 0.116 | 0.13 | 0.194 | | 0.367 | 0.783 | 0.137 | |
| BIO2: Mean Diurnal Range | 0.145 | -0.249 | -0.431 | -0.292 | 0.336 | | 0.18 | | | | | | | | | | -0.102 | 0.698 | |
| BIO3: Isothermality | | | 0.2 | -0.93 | | | -0.23 | | | | | | | | | | | -0.153 | |
| BIO4: Temperature Seasonality | | -0.326 | 0.143 | 0.167 | 0.439 | 0.317 | -0.727 | | | | | | | | | | | | |
| BIO5: Max Temperature of the Warmest Month | 0.298 | 0.157 | | | 0.171 | | | | | 0.161 | 0.253 | -0.329 | -0.33 | | -0.129 | | | -0.112 | -0.7 |
| BIO6: Min Temperature of the Warmest Month | 0.265 | 0.238 | 0.139 | | | | | | -0.102 | 0.163 | 0.26 | -0.339 | -0.34 | | -0.135 | | | | 0.685 |
| BIO7: Temperature Annual Range | 0.134 | -0.26 | -0.476 | | 0.357 | | 0.229 | | | | | | | | | | | -0.668 | 0.203 |
| BIO8: Mean Temperature of the Wettest Quarter | 0.302 | 0.139 | 0.128 | | 0.126 | | | | 0.266 | | | 0.148 | 0.445 | -0.734 | | -0.126 | | | |
| BIO9: Mean Temperature of the Driest Quarter | 0.274 | 0.222 | | | 0.122 | | | | -0.348 | -0.518 | -0.562 | | -0.245 | -0.136 | 0.263 | | | | |
| BIO10: Mean Temperature of the Warmest Quarter | 0.294 | 0.167 | 0.112 | | 0.117 | | | | 0.109 | | | 0.2 | 0.172 | 0.508 | | -0.705 | | | |
| BIO11: Mean Temperature of the Coldest Quarter | 0.288 | 0.191 | | | | | | | | | | 0.168 | 0.17 | 0.318 | | 0.568 | -0.595 | | |
| BIO12: Annual Precipitation | -0.263 | 0.225 | | | 0.219 | 0.194 | 0.114 | -0.105 | 0.655 | -0.466 | | -0.305 | | | | | | | |
| BIO13: Precipitation of the Wettest Month | -0.226 | -0.115 | 0.393 | | 0.329 | 0.406 | 0.4 | -0.478 | -0.27 | 0.199 | | | | | | | | | |
| BIO14: Precipitation of the Driest Month | -0.193 | 0.323 | -0.21 | | 0.105 | | -0.15 | | -0.136 | 0.33 | -0.526 | -0.327 | 0.299 | | -0.408 | | | | |
| BIO15: Precipitation Seasonality | 0.146 | -0.358 | 0.244 | | | | 0.126 | 0.208 | | -0.14 | | -0.642 | 0.449 | 0.129 | 0.254 | | | | |
| BIO16: Precipitation of the Wettest Quarter | -0.297 | | 0.222 | | 0.136 | 0.163 | 0.271 | 0.825 | | | | 0.214 | | | | | | | |
| BIO17: Precipitation of the Driest Quarter | -0.189 | 0.33 | -0.195 | | 0.112 | | | | 0.109 | 0.346 | | -0.104 | | | 0.795 | -0.123 | | | |
| BIO18: Precipitation of the Warmest Quarter | -0.194 | | 0.282 | | 0.483 | -0.801 | | | | | | | | | | | | | |
| BIO19: Precipitation of the Coldest Quarter | -0.185 | 0.331 | -0.182 | | 0.152 | 0.101 | -0.126 | | -0.463 | -0.394 | 0.515 | | 0.36 | | | | | | |
| % explained variation | 49.4% | 28.7% | 7.5% | 5.4% | 5.1% | 2.5% | 0.9% | 0.4% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% |
| cumulative % explained variation | 49.4% | 78.1% | 85.6% | 91.0% | 96.1% | 98.5% | 99.4% | 99.9% | 99.9% | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% |

*Table S7:* Size comparisons between adult individuals of the populations in the western (*papuensis*) and eastern (*sanfordi*) parts of Central Range in New Guinea. This summary statistics were extracted from Frith & Frith (2004) and it has not been possible to analyze the data statistically at an individual level. See also figure 1f.

| | *A. p. papuensis* | | | | *A. p. sanfordi* | | | |
|---|---|---|---|---|---|---|---|---|
| | n | mean | min | max | n | mean | min | max |
| Males, adults | | | | | | | | |
| wing length | 2 | 159 | 157 | 160 | 21 | 168 | 161 | 174 |
| tail length | 2 | 152 | 148 | 156 | 20 | 173 | 155 | 189 |
| bill length | 2 | 33.0 | 32.8 | 33.1 | 20 | 33.5 | 31.8 | 35.9 |
| tarsus length | 2 | 39.6 | 38.9 | 40.2 | 21 | 43.4 | 39.4 | 45.9 |
| weight | 2 | 173.3 | 170 | 175 | 6 | 184 | 180 | 190 |
| Females, adults | | | | | | | | |
| wing length | 5 | 148 | 144 | 155 | 9 | 158 | 148 | 163 |
| tail length | 5 | 126 | 122 | 130 | 9 | 140 | 132 | 149 |
| bill length | 5 | 33.4 | 31.0 | 35.3 | 9 | 33.5 | 31.5 | 35.6 |
| tarsus length | 5 | 38.1 | 36.6 | 40.7 | 9 | 40.4 | 37.1 | 42.7 |
| weight | | | | | 4 | 176 | 163 | 185 |

*Table S8:* Significantly enriched Gene Ontology (GO) terms of biological processes among the 495 genes identified in the 50kb windows that showed the top 5% strongest signatures of divergent selection (measured as Z($F_{ST}$)) between the populations *papuensis* and *sanfordi*. The table shows the GO terms assigned by Panther, the number of genes in each category and the p-value (adjusted for false discovery rate) for the enrichment of the term. Note that a single gene may be associated with more than one biological process.

| GO term | Biological process | No. of genes | FDR corrected p-value |
|---|---|---|---|
| GO:0050804 | modulation of chemical synaptic transmission | 23 | 1.51E-02 |
| GO:0099177 | regulation of trans-synaptic signaling | 23 | 1.20E-02 |
| GO:0051239 | regulation of multicellular organismal process | 89 | 2.09E-03 |
| GO:0050793 | regulation of developmental process | 74 | 2.49E-02 |
| GO:0032879 | regulation of localization | 75 | 2.57E-02 |
| GO:0065008 | regulation of biological quality | 114 | 7.06E-04 |
| GO:0009987 | cellular process | 354 | 4.13E-02 |
| GO:0061448 | connective tissue development | 15 | 2.42E-02 |
| GO:0048856 | anatomical structure development | 127 | 1.28E-02 |
| GO:0032502 | developmental process | 132 | 2.81E-02 |
| GO:0048731 | system development | 107 | 4.02E-02 |
| GO:0007275 | multicellular organism development | 116 | 2.80E-02 |
| GO:0006816 | calcium ion transport | 16 | 4.27E-02 |
| GO:0051179 | localization | 143 | 1.96E-02 |
| GO:0030001 | metal ion transport | 28 | 4.70E-02 |
| GO:0032501 | multicellular organismal process | 144 | 4.17E-02 |
| GO:0007611 | memory | 9 | 4.38E-02 |

*Table S9:* Significantly enriched Gene Ontology (GO) terms of biological processes among the genes identified in the 50kb windows that showed the top 5% strongest signatures of divergent selection (measured as $D_{XY}$) between the populations *papuensis* and *sanfordi*. The table shows the GO terms assigned by Panther, the number of genes in each biological process, the p-value (adjusted for false discovery rate) for the enrichment of the term, and to which higher GO category the process belongs. Note that a single gene may be associated with more than one biological process.

| GO term | Biological process | No. of genes | FDR corrected p-value | Higher GO category |
|---|---|---|---|---|
| GO:0000904 | cell morphogenesis involved in differentiation | 29 | 1.27E-02 | developmental process GO:0032502 |
| GO:0009653 | anatomical structure morphogenesis | 67 | 2.96E-02 | developmental process GO:0032502 |
| GO:0048869 | cellular developmental process | 87 | 3.76E-02 | developmental process GO:0032502 |
| GO:0065007 | biological regulation | 275 | 1.62E-02 | biological regulationGO:0065007 |