Supplementary Material

Full Title: Interspecific gene flow and the evolution of specialisation in black and white rhinoceros.

Short title: Pliocene gene flow in African Rhinoceros

Authors: Yoshan Moodley*, Michael V. Westbury, Isa-Rita M. Russo, Shyam Gopalakrishnan, Andrinajoro Rakotoarivelo, Remi-Andre Olsen, Stefan Prost, Tate Tunstall, Oliver A. Ryder, Love Dalén, Michael W. Bruford

*Corresponding author. Email: yoshan.moodley@univen.ac.za

This PDF file includes:

Tables S1 to S10
Figs. S1 to S6

Supplementary Methods

Additional tests of admixture

To further investigate putative signs of introgression among our African rhinoceros genomes, we implemented Treemix v 1.13 (Pickrell and Pritchard, 2012), specifying the Sumatran rhinoceros as the root and various numbers of migration edges (0-3) (Supplementary Figures S3-S6). First we created a PLINK file using the five rhinoceros genomes mapped to the Sumatran rhinoceros with ANGSD (-doplink 2), and the following parameters: -uniqueonly 1, -remove_bads 1, -minq 25, -minmapq 25, -dopost 1, -GL 1, -docounts 1, -domaf 2, -postcutoff 0.99, -SNP_pval 1e-6, -geno_minDepth 5, -domajorminor 1, and -minind 5. We used PLINK v1.90b3.42 (Purcell et al. 2007) to convert the ANGSD output PLINK file into a frequency file, which was then converted into a Treemix input file using plink2treemix.py (https://bitbucket.org/nygcresearch/treemix). We then ran four independent Treemix runs specifying different numbers of migration edges (-m 0-3), and the Sumatran rhinoceros as root. The tree and residues were produced in R using the plotting_funcs.R script available with Treemix.

In an attempt to identify relationships that were not well-modelled, potentially due to migration or over-parameterisation, we visualised the residual fits of the model to the data. When specifying two migration edges, we uncovered gene flow as we would expect based on geography, that is migration between the southern white and southern black rhinoceros and migration between the eastern black and the northern white rhinoceros. However, when also considering the residual fits, we noted that those relationships had the highest standard error (SE) which could indicate over-parameterisation. Moreover, when specifying zero migration edges, we identified the Sumatran/northern white rhinoceros to have the highest SE, suggesting some migration between these species. This, however, is highly unlikely given that the species are distributed on different continents and may reflect the limitations of Treemix when used on deeply diverging species level comparisons, as opposed to population level comparisons for which the software was intended. Furthermore, the ~18 Ma divergence between Sumatran and white rhinoceros (Margaryan *et al*. 2020) further highlights the unlikeliness of this migration event. A further limitation of Treemix analyses could arise if gene flow had occurred between the ancestral black and ancestral white rhinoceros lineages as suggested by the Dfoil results. This would be represented as reduced branch lengths as opposed to migration edges and would therefore be impossible to detect using this method. Therefore, due to the implausibility of migration between the Sumatran and northern white rhinoceros, and the relatively high SE between lineages that show gene flow when adding migration edges, we interpret zero migration edges as being the most plausible. To further investigate these results, we also implemented the threepop and fourpop tests, also known as the F3- and F4-statistics (Reich *et al*. 2009, Keinan *et al*. 2007). These tests look for the presence of gene flow between a defined triplet ((A, B), C) or quadruplet ((A, B), (C, D)), respectively, with only significantly negative statistics being informative of migration. For these analyses we specified 25,000 SNP windows to calculate the standard error (-k), resulting in 3,107 independent blocks, and the treemix.frq file from above as input. We then extracted the results obtained from phylogenetically correct combinations of the African rhinos.

We found no significantly negative statistics for any of the threepop combinations (Supplementary Table S6), suggesting no significant evidence for gene flow. However, we did find a significantly negative statistic in the fourpop test, suggesting some level of gene flow between the four African populations (Supplementary Table S7).

Literature cited

Keinan A, Mullikin JC, Patterson N and Reich D (2007) Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet*, 39(10):1251–5.

Margaryan A, Sinding MH, Liu S, Vieira FG, Chan YL, Nathan SK, Moodley Y, Bruford MW, Gilbert MT (2020) Recent mitochondrial lineage extinction in the critically endangered Javan rhinoceros. *Zoological Journal of the Linnean Society*.

Pickrell J, Pritchard J. (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *Nature Precedings*. Mar 2:1-.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics. Sep 1;81(3):559-75.

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., and Singh, L. (2009) Reconstructing Indian population history. *Nature*, 461(7263):489–94.

Supplementary Tables

Table S1: Four-taxon ABBA-BABA tests for introgression among African rhinoceros lineages. Significant tests are shown, with exchanging pairs of taxa highlighted. See Figure S2 for a diagrammatic representation of these results.

H1	H2	H3	nABBA	nBABA	Dstat	SE	Z
Northern white	Southern white	Southern black	173235	143189	0.09495	0.00255	37.17
Northern white	Southern white	Eastern black	192098	169577	0.06227	0.00259	24.06
Southern black	Eastern black	Southern white	165945	151444	0.04569	0.00257	17.77
Southern black	Eastern black	Northern white	176803	149911	0.08231	0.00268	30.67

H1-3, correspond to positions one to three in the four taxon test; SE, standard error; Z, z-score must be greater than 3 or less than -3 to test to be significant.

Table S2. Five-taxon Dfoil tests for introgression among African rhinoceros lineages. The analysis considered introgressed segments of DNA in three independent size classes: 100, 50 and 20 kilobases.

Direction of gene flow	**# of windows**	**# of windows**	**# of windows**
	100kb windows	**50kb windows**	**20kb windows**
None	8265	15938	33546
na	955	2506	12762
Northern white into Southern black	1	2	4
Southern white into Southern black	0	2	1
Northern white into Eastern black	2	3	2
Eastern black into Southern white	0	1	1
Southern black into Northern white	0	0	2
Southern black into Southern white	7	2	2
Between Ancestral white and Eastern black	79	161	276
Between Ancestral white and Southern black	59	124	232
Eastern black to Northern white	3	3	3
Southern white into Eastern black	1	0	1

Table S3. Forty-seven loci inferred to have been exchanged between black and white rhinoceros using Dfoil. Genes were identified in admixture tracts inferred by Dfoil. Genes with the same name in the rhino annotation as the human annotation are designated with a -.

Dfoil admixture genes	Human anologs
ADNP	-
AKR1D1	-
AMOTL1	-
ARID5B	-
ASCC3	-
ASPH	-
AURKA	-
B3GNT5	-
CCL22	-
CLCN4	-
COLEC12	-
EMCN	-
GALNT7	-
GJB4	-
HAPLN1	-
IMMP2L	-
LMAN1	-
LOC100049798	1OKO_GORGO
LOC100061235	ZNF146
LOC100064538	Uncharacterised
LOC100065592	ZNF709
LOC102147400	Uncharacterised
LOC102147489	Uncharacterised
LOC102148198	Uncharacterised
LOC102149190	Uncharacterised
LOC102150619	Uncharacterised
LOC102150749	Uncharacterised
LOC106781365	Uncharacterised
LOC106782351	TIGD1
LOC111768373	Uncharacterised
LOC111768924	Uncharacterised
LOC111769246	Uncharacterised
LOC111771019	Uncharacterised
LOC111772601	Uncharacterised
LOC111772794	Uncharacterised
LOC111772923	Uncharacterised
LOC111772963	Uncharacterised
LOC111774282	Uncharacterised
LSAMP	-
OLFM4	-
RHPN2	-
RIN2	-
SLC6A9	-
STX12	-
TUBB1	-
WFDC3	-
XPO1	-

Table S4. Number of consecutive windows showing any signal of introgression, regardless of direction, based on three different window sizes taken from the Dfoil analysis (Table S2). Genes with the same name in the rhino annotation as the human annotation are designated with a -.

Number of consecutive windows	100kb windows	50kb windows	20kb windows
1	135	262	463
2	6	8	19
3	1	2	3
4	1	0	2
5	0	1	0
6	0	0	1

Table S5. D3 statistics results based on two different non-overlapping sliding windows (1Mb, and 100kb). P-values >0.05 indicate no significant differential gene flow among taxa within each triplet. SW = Southern white, NW = Northern white, SB = Southern black, EB = Eastern black.

	1MB windows			100kb windows		
Triplet Compared	Mean	Standard deviation	p-value	Mean	Standard deviation	p-value
(SW,NW)EB)	-0.0002	0.0025	0.4233	-0.0003	0.0072	0.4701
(SW,NW)SB)	0.0000	0.0025	0.4988	-0.0001	0.0071	0.4939
(EB,SB),SW)	0.0010	0.0026	0.2149	0.0011	0.0097	0.4124
(EB,SB),NW)	0.0013	0.0023	0.1373	0.0013	0.0088	0.3831

Table S6. Results of the three-population (F3) test within the Treemix package. Positive values indicate no evidence for gene flow within each triplet.

Triplet	F3 statistic	SE	Z-score
((Northern white;Southern white), Eastern Black)	0.00328898	7.33E-05	44.8633
((Northern white;Southern white), Southern black)	0.0034585	7.45E-05	46.4355
((Eastern black;Southern black), Northern white)	0.00191813	7.51E-05	25.5253
((Eastern black;Southern black), Southern white)	0.00208766	7.57E-05	27.5961

SE, standard error; Z, z-score must be less than -3 to be significant for gene flow.

Table S7. Results of the four-population (F4) test within the Treemix package. A significant negative value indicates gene flow occurred somewhere within the defined quadruplet.

Quadruplet	f4	SE	Z-score
((Southern white,Northern white); (Eastern black,Southern black))	-0.000169524	6.41E-06	-26.443

SE, standard error; Z, z-score must be less than -3 to be significant for gene flow.

Table S8. D-statistics results from simulated data based on our hypothesised model of ancient migration between the ancestral black and white lineages and no recent gene flow. The migration rate used was 0.5. We performed 20,000 replicates, of which 19,398 independent replicates contained enough informative information and were used to acquire the standard deviation (SD), standard error (SE), and Z score. Z score must be greater than 3 or less than -3 to test to be significant.

Topology	Mean D-score	SD	SE	Z
(((Southern White, Northern white), Eastern black), Sumatran)	0.01782	0.56873	0.00408	4.36362
(((Southern White, Northern white), Southern black), Sumatran)	0.01790	0.56752	0.00408	4.39314
(((Southern black, Eastern black), Southern white), Sumatran)	-0.00641	0.56646	0.00407	-1.57501
(((Southern black, Eastern black), Northern white), Sumatran)	-0.00479	0.56509	0.00406	-1.18110

Table S9. D-statistics results from simulated data based on our hypothesised model of ancient migration between the ancestral black and white lineages and no recent gene flow. The migration rate used was 1. We performed 20,000 replicates, of which 19,774 independent replicates contained enough informative information and were used to acquire the standard deviation (SD), standard error (SE), and Z score. Z score must be greater than 3 or less than -3 to test to be significant.

Topology	**Mean D-score**	**SD**	**SE**	**Z**
(((Southern White, Northern white), Eastern black), Sumatran)	-0.00517	0.52028	0.00370	-1.39761
(((Southern White, Northern white), Southern black), Sumatran)	-0.00751	0.52053	0.00370	-2.02748
(((Southern black, Eastern black), Southern white), Sumatran)	-0.00514	0.51149	0.00364	-1.41190
(((Southern black, Eastern black), Northern white), Sumatran)	-0.00307	0.51056	0.00363	-0.84678

Table S10. D-statistics results from simulated data based on our hypothesised model of ancient migration between the ancestral black and white lineages and no recent gene flow. Th migration rate used was 2. We performed 20,000 replicates, of which 19,918 independent replicates contained enough informative information and were used to acquire the standard deviation (SD), standard error (SE), and Z score. Z score must be greater than 3 or less than -3 to test to be significant.

Topology	**Mean D-score**	**SD**	**SE**	**Z**
(((Southern White, Northern white), Eastern black), Sumatran)	-0.00866	0.47500	0.00337	-2.57206
(((Southern White, Northern white), Southern black), Sumatran)	-0.00868	0.47458	0.00336	-2.58023
(((Southern black, Eastern black), Southern white), Sumatran)	-0.02066	0.46785	0.00332	-6.23071
(((Southern black, Eastern black), Northern white), Sumatran)	-0.01977	0.46715	0.00331	-5.97258

Supplementary Figures

Figure S1. PSMC plot of the southern black rhinoceros, originally a ~35-fold coverage genome (black trajectory), here downsampled to 16-fold coverage for comparison with Figure 2B (orange trajectory). Differences between the plots were negligible.

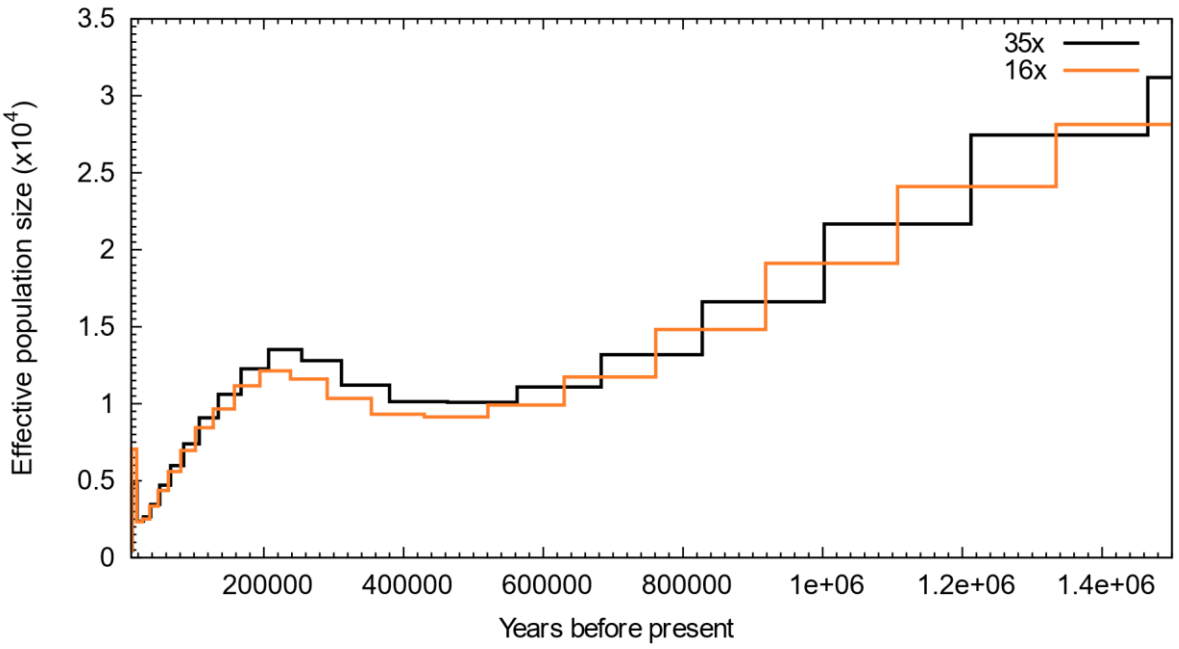

Figure S2. D-statistics results. Black circles show the D-statistic. Error bars represent three standard errors on either side of the D-statistic. Topologies excluding the outgroup (H1, H2), H3) are shown on the y axis. A positive D-statistic represents gene flow between the H2 and H3 (ABBA) individuals.

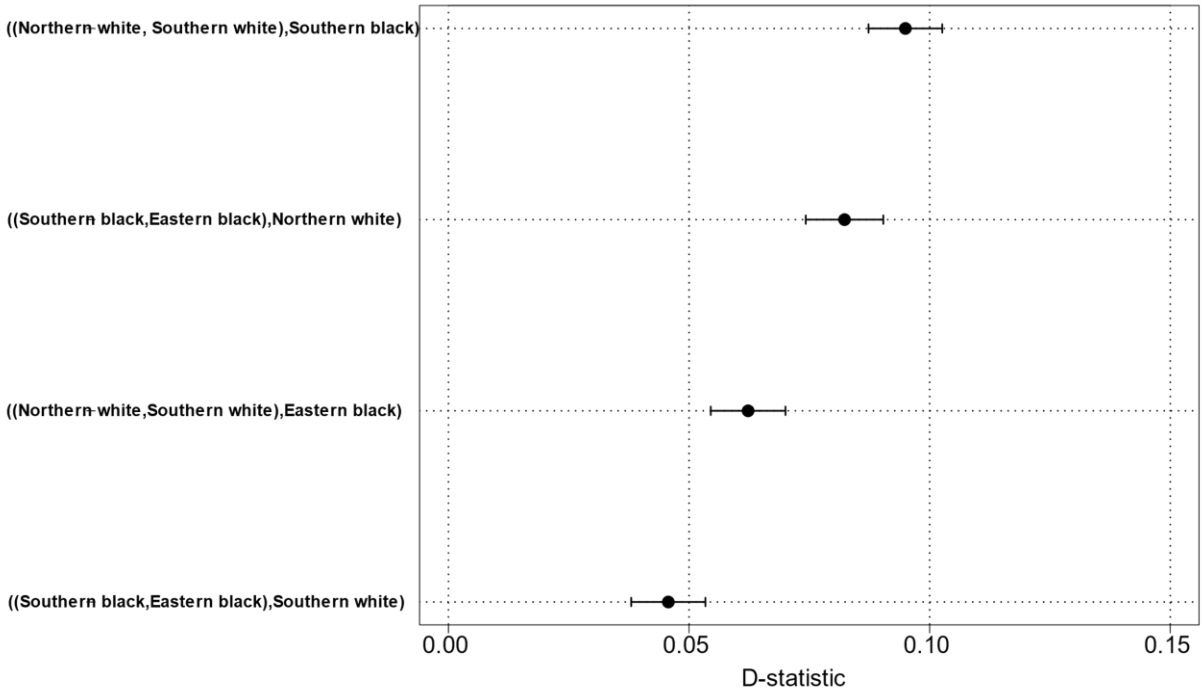

Figure S3. Treemix graph showing the maximum likelihood phylogenetic reconstruction among African rhinoceros species (left) and the residual fit for the proposed model (right). In this analysis, the number of migration edges was set to 0; that is, no assumed migration.

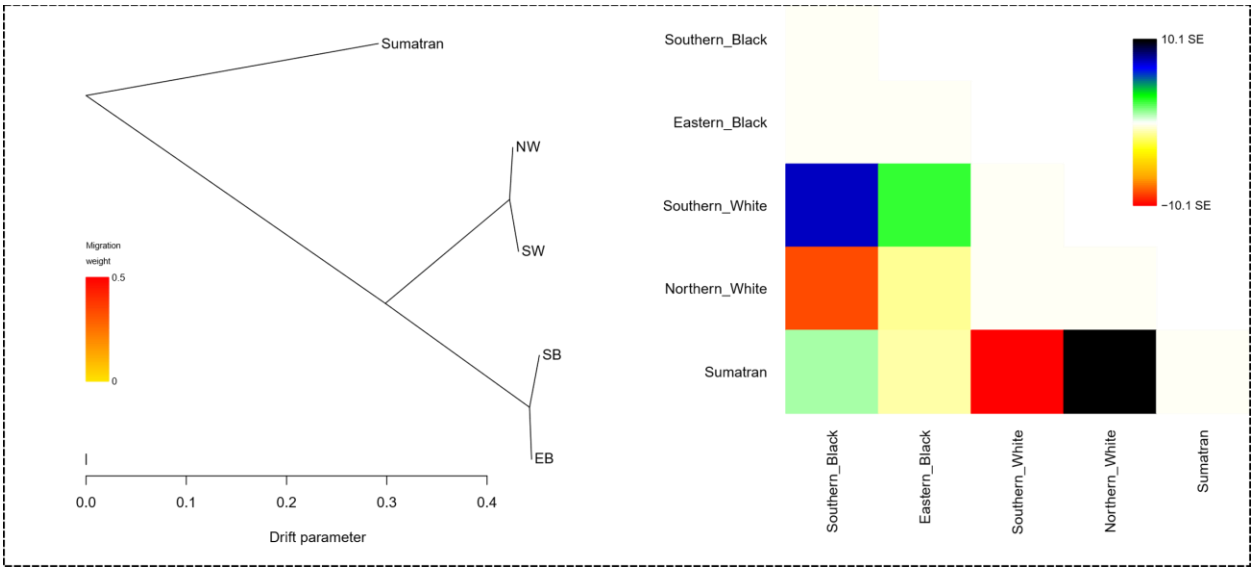

Figure S4. Treemix graph showing the maximum likelihood phylogenetic reconstruction among African rhinoceros species (left) and the residual fit for the proposed model (right). In this analysis, the number of migration edges was set to 1.

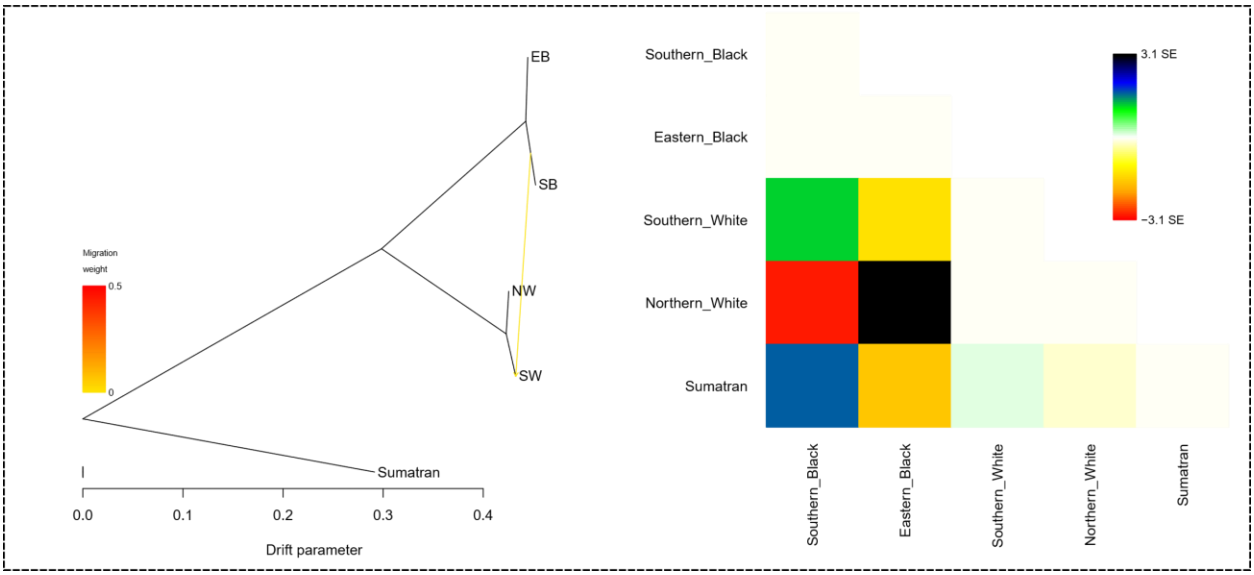

Figure S5. Treemix graph showing the maximum likelihood phylogenetic reconstruction among African rhinoceros species (left) and the residual fit for the proposed model (right). In this analysis, the number of migration edges was set to 2.

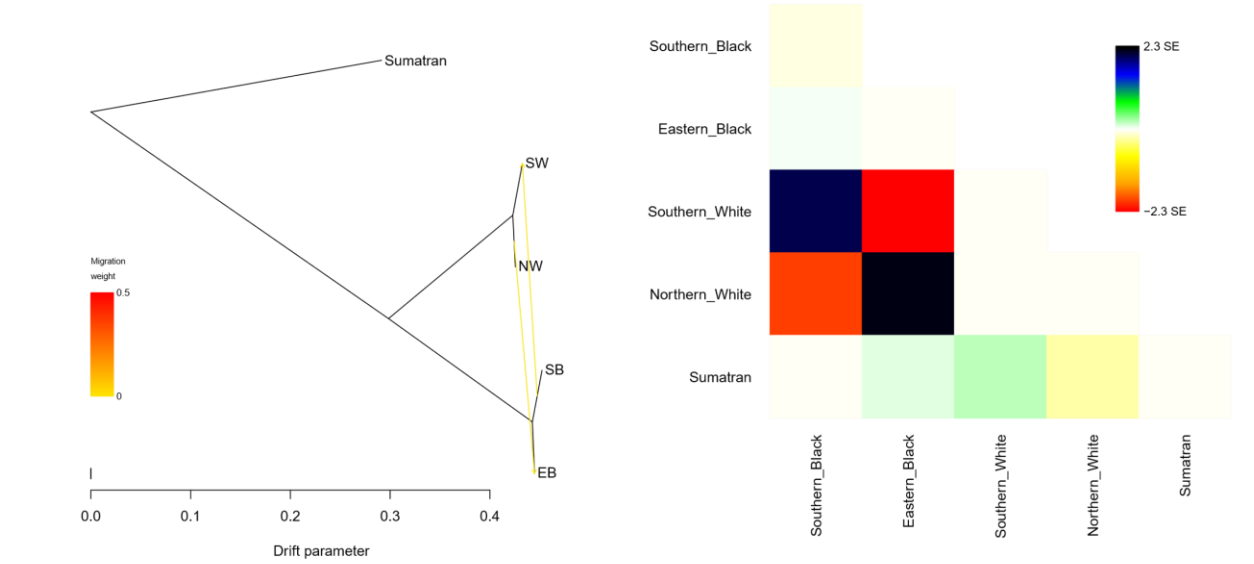

Figure S6. Treemix graph showing the maximum likelihood phylogenetic reconstruction among African rhinoceros species (left) and the residual fit for the proposed model (right). In this analysis, the number of migration edges was set to 3.