Appendix A

Lake Malawi genetic diversity

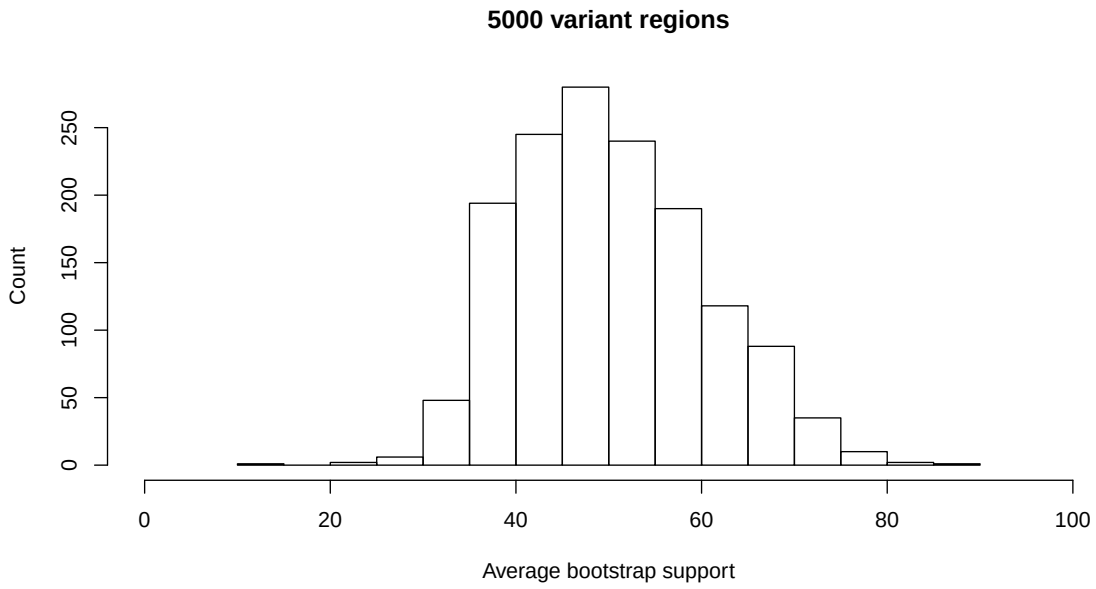

Fig. A.1 **Average bootstrap values for 1,460 phylogenies representing regions along the genome.**

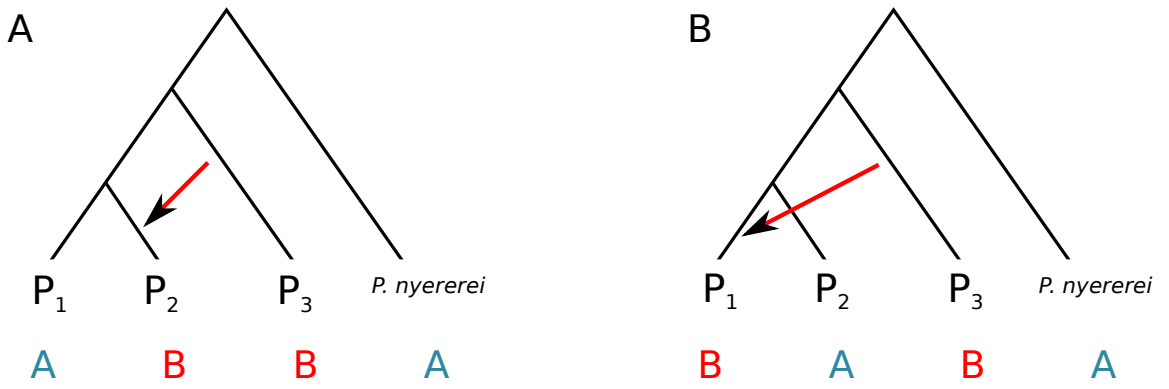

Fig. A.2 **The ABBA-BABA test for introgression.** The ancestral allele is denoted by the A character and the derived allele by the B character. **(A)** Differential gene flow P_3 to P_2 will result in an excess of shared derived alleles between them - the ABBA pattern. **(B)** Gene flow P_3 to P_1 will result in an excess of the BABA pattern.

Appendix B

Lake Massoko speciation

Table B.1 **Results of a survey of fish fauna in six crater lakes of Rungwe District, Tanzania.** Conducted in July and November 2011.

Lake	Species	Family	Tribe	Probable Status
Kingiri	*Astatotilapia* sp. 'kingiri black'	Cichlidae	Haplochromini	Endemic
	Rhamphochromis sp. 'kingiri dwarf'	Cichlidae	Haplochromini	Endemic
	Rhamphochromis sp. 'kingiri brevis'	Cichlidae	Haplochromini	Endemic
	Serranochromis robustus	Cichlidae	Haplochromini	Native
	Coptodon rendalli	Cichlidae	Tilapiini	Native
	Oreochromis shiranus	Cichlidae	Tilapiini	Native
	Oreochromis (Nyasalapia) squamipinnis	Cichlidae	Tilapiini	Native
	Clarias gariepinus	Clariidae		Native
	Micropanchax johnstoni	Poeciliidae		Native
	Barbus radiatus	Cyprinidae		Native
	Barbus trimaculatus	Cyprinidae		Native
Ilamba	*Astatotilapia* sp. 'ilamba black'	Cichlidae	Haplochromini	Endemic
	Otophraynx sp. 'Ilamba tetrastigma'	Cichlidae	Haplochromini	Endemic
	Oreochromis cf shiranus	Cichlidae	Tilapiini	Native/Endemic?
	Oreochromis cf squamipinnis	Cichlidae	Tilapiini	Native/Endemic?
	Clarias gariepinus	Clariidae		Native
	Mesobola cf. spinifer	Cyprinidae		Native
	Barbus paludinosos	Cyprinidae		Native
	Barbus trimaculatus	Cyprinidae		Native
	Barbus macrotaenia	Cyprinidae		Native
	Barbus radiatus	Cyprinidae		Native
Ikapu	*Astastotilapia* sp. 'ikapu dark'	Cichlidae	Haplochromini	Endemic
	Tilapia sparrmanii	Cichlidae	Tilapiini	Native/Introduced?
	Oreochromis 'golden chambo'	Cichlidae	Tilapiini	Endemic
	Clarias gariepinus	Clariidae		Native/Introduced?
Itamba	*Astatotilapia* sp. 'itamba dark'	Cichlidae	Haplochromini	Endemic
	Oreochromis cf. shiranus	Cichlidae	Tilapiini	Native/Endemic?
	Oreochromis (Nyasalapia) cf. karongae	Cichlidae	Tilapiini	Native/Endemic?
	Oreochromis niloticus	Cichlidae	Tilapiini	Introduced
Massoko	*Astatotilapia* sp. 'massoko benthic'	Cichlidae	Haplochromini	Endemic
	Astatotilapia sp. 'massoko littoral'	Cichlidae	Haplochromini	Endemic
	Coptodon rendalli	Cichlidae	Tilapiini	Introduced?
	Oreochromis (Nyasalapia) squamipinnis	Cichlidae	Tilapiini	Native/Endemic?
	Clarias gariepinus	Clariidae		Introduced?
Itende	*Astatotilapia* sp. 'itende'	Cichlidae	Haplochromini	Endemic
	Oreochromis (Nyasalapia) squamipinnis	Cichlidae	Tilapiini	Native/Endemic?

Table B.2 **An overview of *Astatotilapia* samples collected for RAD sequencing.**

Sampling location (ecomorph)	N	Sampling Dates	Collector(s)	Latitude S	Longitude E
Lake Massoko (benthic)	5	17/07/2011	MG, BN, GT, SM, AS	9°20'0	33°45'18
Lake Massoko (littoral)	3	17/07/2011	MG, BN, GT, SM, AS	9°20'0	33°45'18
Lake Massoko (small, unassigned)	4	17/07/2011	MG, BN, GT, SM, AS	9°20'0	33°45'18
Lake Itende	3	27/11/2011	MG, GT, AS	9°19'19	33°47'15
Lake Ikapu	3	20/07/2011	MG, BN, GT, SM, AS	9°22'12	33°48'25
Lake Itamba	2	19/07/2011	MG, BN, GT, SM, AS	9°21'04	33°50'39
Lake Ilamba	2	17/07/2011	MG, BN, GT, SM, AS	9°23'33	33°50'09
Lake Kingiri	8	15+21/07/2011	MG, BN, GT, SM, AS	9°25'08	33°51'29
Ruo river	2	22/05/2009	MG, AS, JS	15°50'77	35°11'69
Unaka lagoon	1	24/07/2011	MG, AS	12°23'59	34°05'17
Mbenji island	2	–/01/2011	MG (from imported wild stock)	13°26'	34°29'
Makanjila	2	18/01/2011	MG, PP, JB	13°41'35	34°50'51
Chisumulu island	1	–/01/2011	MG (from imported wild stock)	12°00'	34°37'
Mgowesa river	2	16/07/2011	MG, BN, GT, SM, AS	9°23'43	33°49'38
Astatotilapia tweddlei (Lake Chilwa)	1	19/05/2009	MG, AS, JS	15°22'18	35°35'20

MG = Martin Genner, GT = George Turner, BN = Benjamin Ngatunga, SM = Semvua Mzighani, AS = Alan Smith, JS = Jennifer Swanstrom, PP = Paul Parsons, JB = Jon Bridle.

Table B.3 **The migration parameter `M` and the probability of migration** At any point, the coalescent simulation is tracking n_1 alleles from population 1 and n_2 alleles from population 2, denoted as (n_1,n_2). Going back in time, 'events' occur with probability $\mathbb{P}(event)$ in which either: a) a common ancestor of two lineages is found within a population; or b) a lineage migrates to the other population. This table gives probabilities that the event is a migration, for a range of parameters `M`, and numbers of tracked lineages (n_1,n_2). The calculation is based Hudson [253]:

$$\mathbb{P}(migration|event) = \frac{n_1 \frac{M}{2}}{\binom{n_1}{2}+\binom{n_2}{2}+(n_1+n_2)\frac{M}{2}} + \frac{n_2 \frac{M}{2}}{\binom{n_1}{2}+\binom{n_2}{2}+(n_1+n_2)\frac{M}{2}}$$

M	Lineages tracked (n_1,n_2)		
	(74,64)	(37,32)	(18,16)
5	6.82%	12.93%	23.74%
10	12.76%	22.89%	38.37%
20	22.63%	37.26%	55.46%
40	36.91%	54.29%	71.35%

Table B.4 **Location and lengths of highly diverged regions (HDRs)**

scaffold	start coordinate	end coordinate	length (bp)	scaffold	start coordinate	end coordinate	length (bp)
0	10512411	10559800	47389	51	1450783	1493272	42489
0	10570498	10598504	28006	55	3423595	3500130	76535
0	11529594	11540402	10808	57	46109	77869	31760
0	11994849	12015103	20254	57	1615373	1638983	23610
0	14003832	14040483	36651	64	55966	175700	119734
0	18256071	18263999	7928	74	591451	600724	9273
5	1920004	1936600	16596	77	2089974	2164351	74377
6	2399603	2417150	17547	78	6039	59940	53901
11	5426321	5452278	25957	82	2236206	2273645	37439
12	3879628	3890173	10545	83	1379873	1425116	45243
14	2810211	2821278	11067	84	2355047	2388352	33305
14	3582492	3609841	27349	84	2399084	2517997	118913
14	3661853	3697260	35407	88	819852	845401	25549
15	1934238	1967068	32830	88	1194601	1316288	121687
15	1981201	2033263	52062	88	1372483	1527476	154993
15	2912637	2961336	48699	88	1732907	1868455	135548
15	3390209	3412823	22614	88	1908746	1943289	34543
15	4641580	4880808	239228	88	2418799	2435992	17193
15	4907565	5049805	142240	91	129230	153938	24708
15	5452492	5474330	21838	92	296055	342364	46309
15	6705210	6818468	113258	93	1295671	1314656	18985
15	7208463	7304325	95862	95	1001404	1044619	43215
15	7317980	7335547	17567	97	298706	313982	15276
15	7507678	7592890	85212	97	2158978	2172667	13689
15	7682086	7700855	18769	97	2188270	2212097	23827
18	2797554	2832090	34536	99	330458	339693	9235
18	6702155	6728393	26238	99	355072	639642	284570
26	5297874	5318387	20513	108	572788	738675	165887
26	5517553	5550216	32663	108	814090	941030	126940
30	183937	257768	73831	108	963573	1023559	59986
30	797497	844062	46565	112	1966090	2014162	48072
30	3978481	4066243	87762	113	1062779	1122847	60068
31	4041909	4078026	36117	114	505730	526658	20928
32	4518067	4589712	71645	114	1902474	2005892	103418
32	4616851	4625659	8808	120	918534	962612	44078
32	4886114	4908718	22604	121	1894243	1983613	89370
33	4163850	4200587	36737	126	420889	439080	18191
36	1010120	1029957	19837	143	1376555	1380941	4386
39	465841	688825	222984	148	1458116	1644136	186020
39	1004596	1047034	42438	148	1669247	1754172	84925
39	1207716	1236548	28832	162	1227615	1263777	36162
39	2153726	2185052	31326	164	0	113596	113596
39	2245171	2269911	24740	164	196412	276496	80084
39	2294746	2311616	16870	186	693304	711017	17713
39	2323506	2340670	17164	190	805453	832175	26722
40	1569517	1608679	39162	206	177202	290266	113064
40	1870518	1891875	21357	229	252128	283116	30988
43	3655049	3696771	41722	229	470627	578767	108140
45	2785077	2828731	43654	304	0	70114	70114

Table B.5 **GO enrichment terms in candidate 'islands of speciation' ±50kb**

GO ID	Term	Total	Found	Expected	p-value
GO:0005813	centrosome	59	4	0.66	0.0041
GO:0009798	axis specification	31	3	0.33	0.0043
GO:0044877	macromolecular complex binding	233	7	2.4	0.0098
GO:0005874	microtubule	77	4	0.86	0.0105
GO:0046530	photoreceptor cell differentiation	44	3	0.47	0.0116
GO:1903034	regulation of response to wounding	17	2	0.18	0.014
GO:0030916	otic vesicle formation	19	2	0.2	0.0174
GO:0043484	regulation of RNA splicing	20	2	0.22	0.0192
GO:0022037	metencephalon development	23	2	0.25	0.025
GO:0006414	translational elongation	25	2	0.27	0.0293
GO:0004812	aminoacyl-tRNA ligase activity	27	2	0.28	0.0311
GO:0006418	tRNA aminoacylation for protein translation	26	2	0.28	0.0315
GO:0015629	actin cytoskeleton	63	3	0.7	0.033
GO:0004879	ligand-activated sequence-specific DNA binding RNA polymerase II transcription factor activity	31	2	0.32	0.0402
GO:0014070	response to organic cyclic compound	94	4	1.01	0.0472
GO:0032101	regulation of response to external stimulus	33	2	0.36	0.0488

Table B.6 **GO terms significantly enriched in all HDRs ±10kb**

GO ID	Term	Total	Found	Expected	p-value
GO:0005874	microtubule	77	4	0.69	0.0049
GO:0005200	structural constituent of cytoskeleton	12	2	0.11	0.0052
GO:0043401	steroid hormone mediated signaling pathway	46	3	0.42	0.0085
GO:0003707	steroid hormone receptor activity	47	3	0.43	0.009
GO:0005057	receptor signaling protein activity	58	3	0.53	0.0159
GO:0006418	tRNA aminoacylation for protein translation	26	2	0.24	0.0235
GO:0004812	aminoacyl-tRNA ligase activity	27	2	0.25	0.0251
GO:0001755	neural crest cell migration	28	2	0.26	0.027
GO:0045454	cell redox homeostasis	29	2	0.27	0.0288
GO:0051216	cartilage development	74	3	0.68	0.0303
GO:0004879	ligand-activated sequence-specific DNA binding RNA polymerase II transcription factor activity	31	2	0.28	0.0325
GO:0048675	axon extension	37	2	0.34	0.0451

Table B.7 **GO terms significantly enriched in all HDRs ±50kb**

GO ID	Term	Total	Found	Expected	p-value
GO:0071407	cellular response to organic cyclic compound	79	8	1.51	0.00012
GO:0003707	steroid hormone receptor activity	47	5	0.9	0.0019
GO:0004879	ligand-activated sequence-specific DNA binding RNA polymerase II transcription factor activity	31	4	0.59	0.0027
GO:0090101	negative regulation of transmembrane receptor protein serine/threonine kinase signaling pathway	17	3	0.32	0.00381
GO:0005813	centrosome	59	5	1.13	0.0051
GO:0030916	otic vesicle formation	19	3	0.36	0.00528
GO:0046530	photoreceptor cell differentiation	44	4	0.84	0.00958
GO:0031012	extracellular matrix	70	5	1.34	0.0105
GO:0006418	tRNA aminoacylation for protein translation	26	3	0.5	0.01286
GO:0007051	spindle organization	26	3	0.5	0.01286
GO:0004812	aminoacyl-tRNA ligase activity	27	3	0.51	0.0142
GO:0005874	microtubule	77	5	1.47	0.0155
GO:0004519	endonuclease activity	52	4	0.99	0.0168
GO:0009798	axis specification	31	3	0.59	0.02075
GO:0005200	structural constituent of cytoskeleton	12	2	0.23	0.021
GO:0035141	medial fin morphogenesis	12	2	0.23	0.0211
GO:0021984	adenohypophysis development	13	2	0.25	0.02462
GO:0042802	identical protein binding	60	4	1.14	0.027
GO:0060037	pharyngeal system development	14	2	0.27	0.02837
GO:0022626	cytosolic ribosome	15	2	0.29	0.0323
GO:0045446	endothelial cell differentiation	16	2	0.31	0.0365
GO:1903034	regulation of response to wounding	17	2	0.32	0.04086
GO:0015698	inorganic anion transport	41	3	0.78	0.04291
GO:0032403	protein complex binding	135	6	2.57	0.0437
GO:0035088	establishment or maintenance of apical/basal cell polarity	18	2	0.34	0.0454

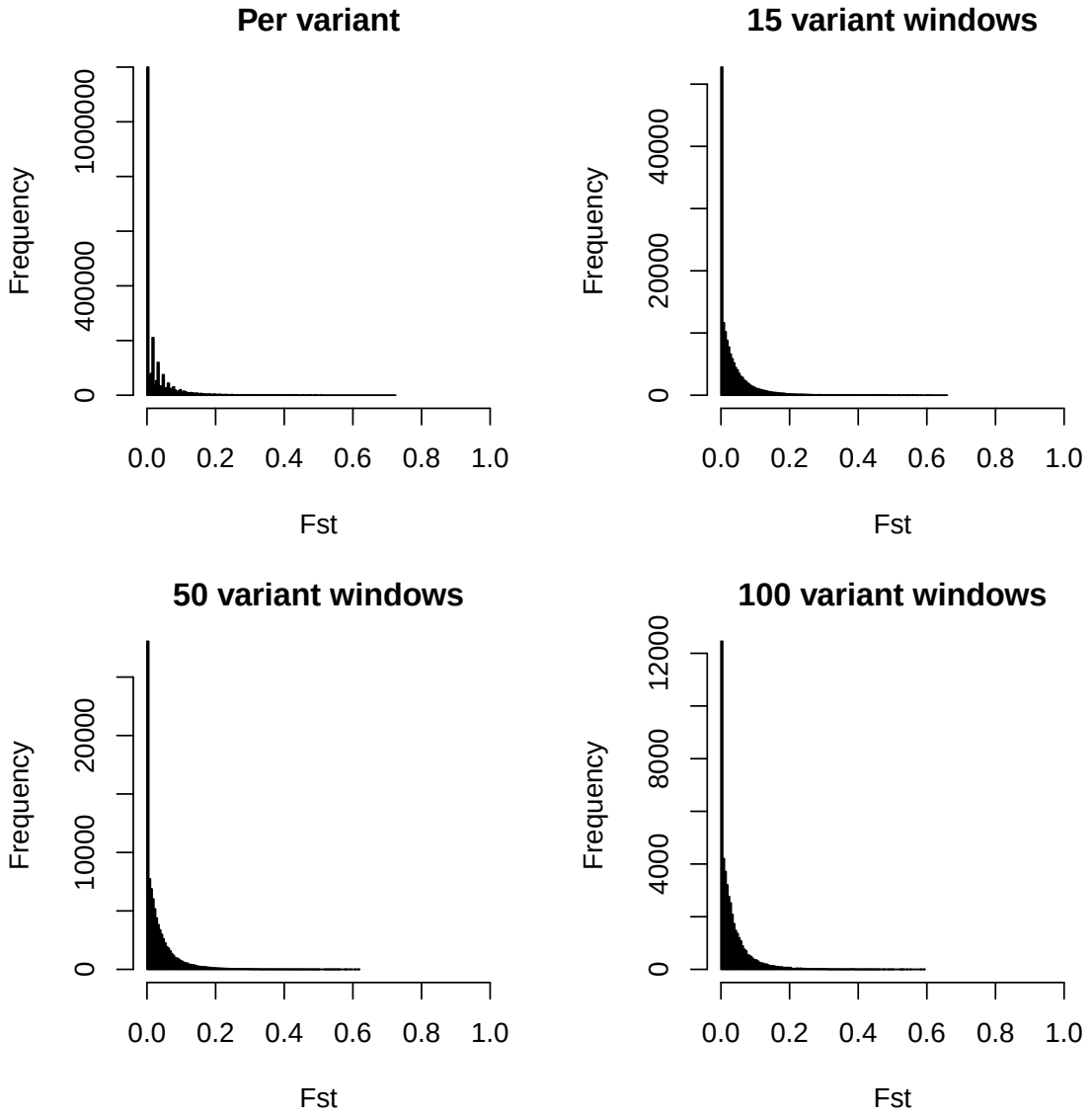

Fig. B.1 **Statistical distribution of within-Massoko F_{ST} divergence, per variant and in sliding windows of varying sizes** The distribution has a sharp L like shape, largely independent of the window size used, consistent with theoretical predictions about early stages of speciation with gene-flow [71].

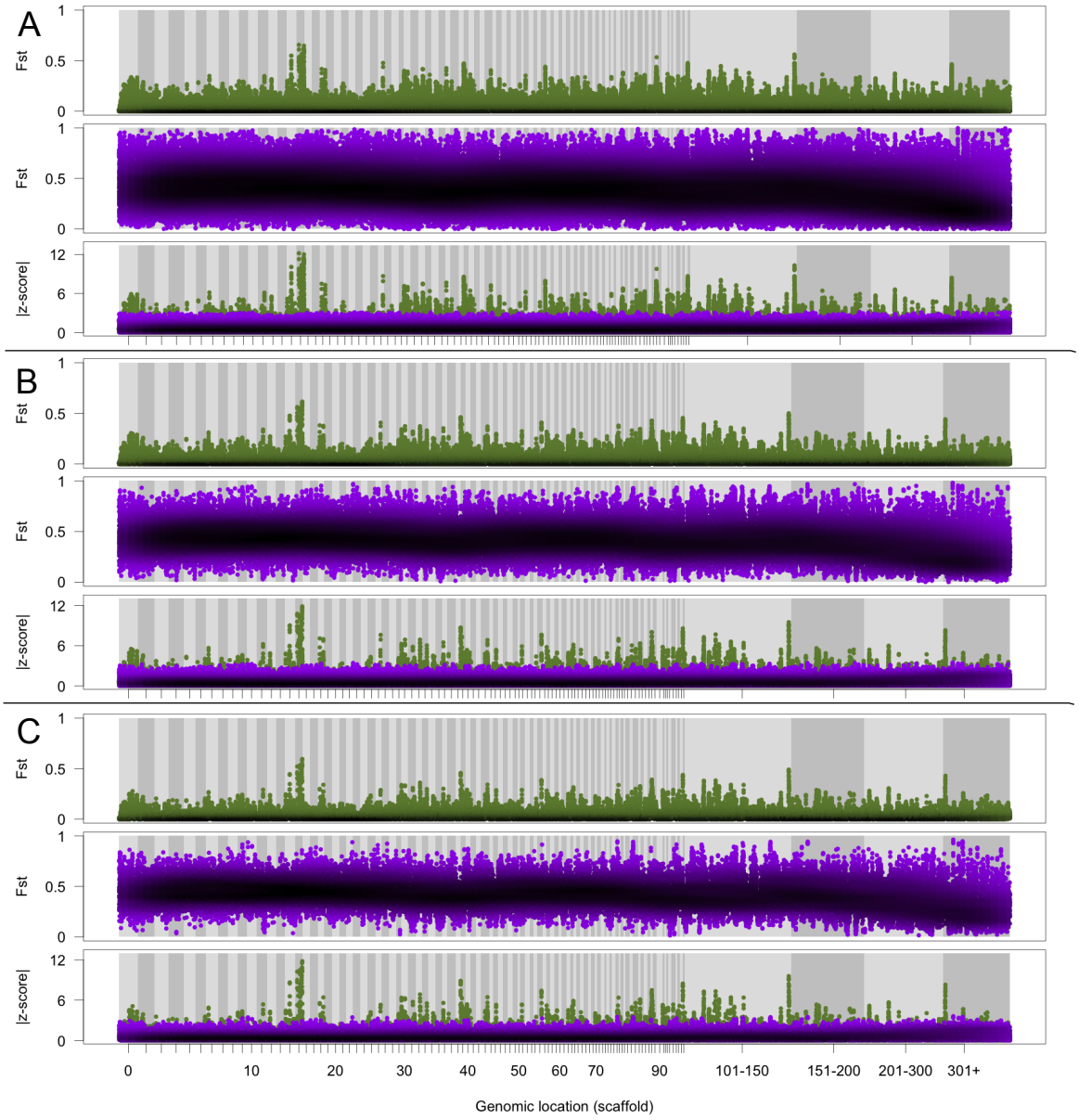

Fig. B.2 **Genome-wide pattern of F_{ST} divergence using sliding windows or varying sizes** The overall pattern of "genomic islands" raising above low background divergence is unaffected by varying the window size. Each figure shows genome-wide pattern of F_{ST} between *Massoko benthic* and *Massoko littoral* (green), and between combined Massoko and Itamba populations (purple), and absolute z-scores of Massoko-Itamba divergence (purple) and within-Massoko divergence (green). **(A)** Window size=15 variants; **(B)** Window size=50 variants; **(C)** Window size=100 variants.

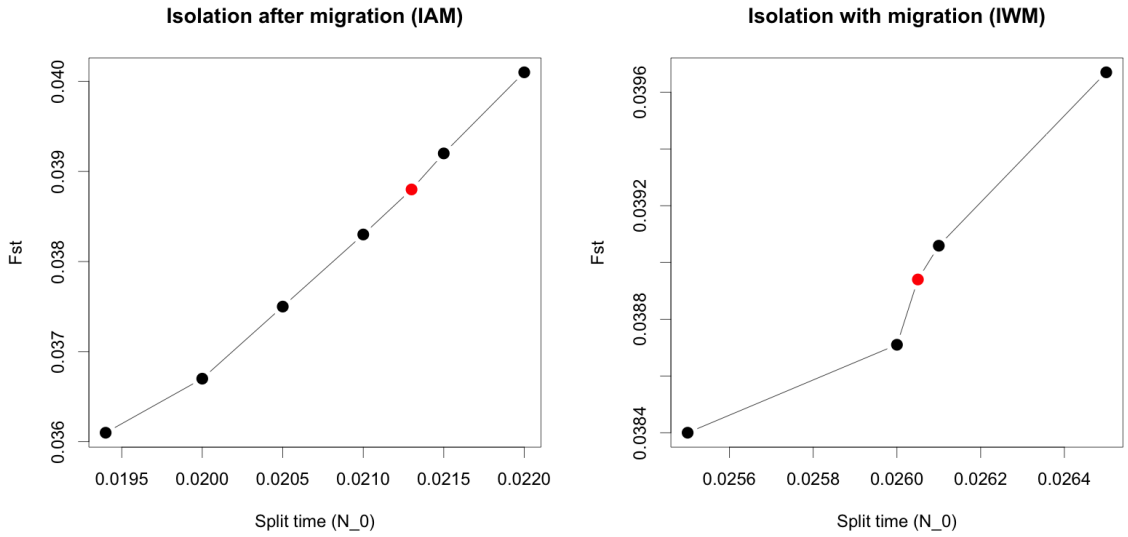

Fig. B.3 **Neutral simulations - fitting split time to match Massoko littoral-benthic F_{ST} divergence** The split time parameters were adjusted for simulations under both models of species formation to match the overall F_{ST} divergence observed between the *Massoko benthic* and *Massoko littoral* forms (0.0389). Several runs were tried (black points) until the optimal value was discovered for each model (red points).

AApos:	162	166	169	297	298	299
Ref:	V	S	T	G	A	A
H4:	L	A	A	S	S	S
H5:	.	.	A	.	.	S

Fig. B.4 **Amino-acid differences between the two haplotypes of the rhodopsin (*rho*) found in Lake Massoko *Astatotilapia*.** The differences are present at amino-acid positions 162, 166, 297, and 298. There are two additional amino-acid positions (169, 299) where both ecomorphs differ from the *M. zebra* reference.