Supplementary B

Supplementary Material for: Integrating kelp genomic analyses and geological data to reveal ancient earthquake impacts

FULL GENETIC METHODS

DNA extraction and GBS

DNA was extracted and purified following the same method described in Peters et al. (2020), except that we used the updated Qiagen DNeasy Plant Pro DNA extraction kit. The kit protocol was followed, except that the lysis step was extended to 24 hours, and immediately after lysis, samples were treated with 100 µl isopropanol and incubated at 65°C for 30 minutes, vortexing every 15 minutes. A new GBS library including 90 *D. antarctica* samples was generated for this study (Supplementary Table B2). Another 126 samples sequenced across four previously GBS libraries (Parvizi, Dutoit, Fraser, & Waters, 2022; Vaux, Craw, Fraser, & Waters, 2021; Vaux, Parvizi, Craw, Fraser, & Waters, 2022) were also used for this study (Supplementary A; Supplementary Table B1). DNA was digested using the *PstI-HF* enzyme, following the GBS protocol described by Elshire et al. (2011), with the same modifications described by Peters et al. (2020). The size selection varied between 200 – 600 bp (Supplementary A; Supplementary Table B2). The five libraries were sequenced on five separate runs using mid output flow cells on the Illumina NextSeq 500 platform (75 bp paired-end; Supplementary A; Supplementary Table B2).

Processing of GBS data

The *process_radtags* component of STACKS 2.53 (Rochette, Rivera-Colón, & Catchen, 2019) was used to demultiplex all samples into paired forward and reverse reads per individual, using inline barcodes. The *process_radtags* component removed low quality reads and reads with missing barcodes or *PstI* restriction sites (-c -q). This process included the rescue barcode and RADtag parameter (-r) to retrieve additional reads, and reads were

truncated to 68 bp (-t 68). For the new GBS library, a total of 317,501,416 reads (87.3%) were retained after this initial filtering (Supplementary Table B2).

Demultiplexed pairs of reads were assembled into loci without a reference genome using the *de_novo* pipeline in STACKS. In *ustacks*, the minimum depth of coverage used to create a stack was two (-m 3), the maximum distance (in nucleotides) allowed between stacks was two (-M 2), and the maximum distance allowed to align secondary reads to primary stacks was four (-N 4). A bounded SNP model was applied with the error rate not being allowed to exceed 5% (--bound_high 0.05). In *cstacks*, the number of mismatches allowed between sample loci when building a catalog was two (-n 2). For the *populations* component of STACKS, samples were organised into a single, panmictic population. Each locus was required to be present in 70% of individuals within the population (-p 1 -r 0.7). A minimum minor allele frequency of 5% was enforced for loci (--min_maf 0.05). Only the first SNP of each locus was used (--write_single_snp) and all SNPs were processed as biallelic and assumed to represent nuclear loci.

The settings in STACKS listed above were selected after iteratively modifying parameters in STACKS (Supplementary Table B3), as recommended by Mastretta-Yanes et al. (2015). We aimed to maximise the number of variant loci, while paying attention to coverage depth, missing data per sample and per locus, and the risk of erroneously combining too many reads. Overall, these preliminary investigations of revealed the data to be highly consistent across parameter changes, and relative to the other iterations, the final selected parameter settings yielded a medium number of loci with low missing data.

Filtering loci

The initial output loci from STACKS were filtered. Loci estimated to be in linkage disequilibrium (LD) within the panmictic population were identified using PLINK 1.9 with a cut-off of 0.8 (Purcell et al., 2007). One locus from each pair of loci estimated to be in LD was removed at random. Using VCFTOOLS 0.1.16 (Danecek et al., 2011), genotypes for a locus were removed from individuals if they had a coverage depth below 8 reads (--minDP 8), and then after recoding loci, sites with ≥50% missing data among all individuals were removed (--max-missing 0.5). The loci estimated to be in LD or to have low genotype coverage depth were organised into a list and excluded (-B; Catchen et al. 2013) and the *populations* component of was re-run (same settings as above).

Variation in coverage depth per locus was investigated in the subsequent dataset using VCFR 1.13 (Knaus & Grünwald, 2017), and loci that were outliers for mean coverage depth and the standard deviation of coverage depth were identified. The outlier range was 1.5 times the interquartile range, above the upper quartile and below the lower quartile. These coverage depth outlier loci were added to an updated, second list of excluded loci and the *populations* component of STACKS was re-run to produce the final filtered dataset.

Missing data per sample and variation in loci coverage depth (same methods as described above) was assessed for the filtered datasets using VCFR. Some output files from STACKS were converted to different file formats in PGDSPIDER 2.1.1.3 (Lischer & Excoffier, 2012) for some downstream analyses. For phylogenetic reconstructions, VCF files were converted to the phylip format, with loci filtered to require at least four samples per locus (-m 4), using VCF2PHYLIP 2.0 (Ortiz, 2019).

Genotypic analyses

Phylogenetic relationships among samples were inferred by constructing unrooted maximum likelihood (ML) and neighbour-joining (NJ) trees with free rates using IQTREE 1.6.12 (Nguyen, Schmidt, Von Haeseler, & Minh, 2015) and VCF-KIT 0.1.6 (Cook & Andersen, 2017) respectively. The analysis in IQTREE was conducted with 10,000 ultrafast bootstrap replicates and the implementation of the *modelfinder* algorithm. Trees were visualised in FIGTREE 1.4.4 (FigTree, 2018).

Population structure was assessed using principal components analysis (PCA) implemented in ADEGENET 2.13 (Jombart, 2008; Jombart & Ahmed, 2011). The maximum number of 'meaningful' principal components (PCs) to interpret was determined by comparing PC Eigen values. Population structure and admixture among samples was further assessed using LEA 2.8 (Frichot & François, 2015), which analysed 10 values of K. The LEA analysis was conducted with default settings.

FULL GENETIC RESULTS

Genotypic analyses

After filtering, the dataset contained 4,269 loci. A total of 243 loci were excluded for being in LD or for having low or outlying coverage depth (Supplementary Table B4). Mean missing data per sample was relatively low (Supplementary Table 4), and missing data was consistent among most samples (Supplementary Figures B1 and B2). After filtering, the coverage depth for loci was well constrained (Supplementary Figure B3).

The phylogenetic reconstructions for ML and NJ trees were similar (Figure 3, Supplementary Figures B4 and B5). Samples clustered into geographic groupings, with the Cape Campbell and Ward Beach sites forming a northern group, Wharanui and Waipapa Bay forming a central group, and the Kaikōura Peninsula, Southern Kaikōura and Hurunui

forming a southern group (Figure 3, Supplementary Figures B4 and B5). Samples from Rarangi and Banks Peninsula consistently formed a distinct, basal clade relative the other South Island samples, when the trees were rooted using the North Island sampling (Supplementary Figures B4 and B5).

For PCA, the broken-stick test indicated that only the first two PCs should be retained (Supplementary Figure B6). PC1 (13.9% of variation) and PC2 (9.2%) revealed the same clusters identified in the phylogenetic reconstructions (Figure 4a). Specifically, Ward Beach and Cape Campbell were clustered together, as were Wharanui and Waipapa Bay, and samples from the Kaikōura Peninsula, Southern Kaikōura and Hurunui formed another grouping (Figure 3). Samples from the North Island, Rarangi and Banks Peninsula were separated (Figure 3). In both the phylogenetic reconstructions and the PCA (Figures 3 and 4a, Supplementary Figures B4 and B5), 10 samples clustered unexpectedly with individuals from geographically distant populations – potentially indicating low levels of dispersal and admixture.

According to cross entropy values, the highest values of K (i.e. 9 or 10) were favoured using LEA admixture analysis (Supplementary Figure B7). However, clustering among samples was clearly hierarchical and population structure was highly consistent across all values of K (i.e. 2 – 10; Supplementary Figure 8). For example, samples from Wharanui and Waipapa Bay were consistently clustered together from K = 2 to K = 7, and from K = 8 onwards Wharanui was distinguished as a separate group (Supplementary Figure B8). Some clusters distinguished a small number of individuals or contributed to clustering uncertainty for a small proportion of samples. These clusters potentially indicate uncertainty due to genuine admixture among some locations, and they may have been influenced by the remaining missing data. Nonetheless, the large clusters consistently identified across most values of K in both datasets were highly concordant with the phylogenetic and PCA results.

Notably, for K = 9, the following groups were distinguished: the North Island, Rarangi and Banks Peninsula, Cape Campbell and Ward Beach, Wharanui, Waipapa Bay, Kaikōura Peninsula, and Southern Kaikōura and Hurunui (Figure 4b, Supplementary Figure B8).

SUPPLEMENTARY B REFERENCES

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140. doi: 10.1111/mec.12354

Cook, D. E., & Andersen, E. C. (2017). VCF-kit: Assorted utilities for the variant call format. *Bioinformatics*, *33*(10), 1581–1582. doi: 10.1093/bioinformatics/btx011

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. doi: 10.1093/bioinformatics/btr330

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, *1*, 269–271.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, *6*(5), 1–10. doi: 10.1371/journal.pone.0019379

Felicísimo, Á. M., Muñoz, J., & González-Solis, J. (2008). Ocean surface winds drive dynamics of transoceanic aerial movements. *PLoS ONE*, *3*(8), 1–7. doi: 10.1371/journal.pone.0002928

Fernández-López, J., & Schliep, K. (2019). rWind: download, edit and include wind data in ecological and evolutionary analysis. *Ecography*, *42*(4), 804–810. doi: 10.1111/ecog.03730

FigTree. (2018). FigTree 1.4.4. Retrieved from http://tree.bio.ed.ac.uk/

Frichot, E., & François, O. (2015). LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution*, *6*(8), 925–929. doi: 10.1111/2041-210X.12382

Goudet, J. (2005). HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, *5*(1), 184–186. doi: 10.1111/j.1471-8286.2004.00828.x

Gruber, B., Unmack, P. J., Berry, O. F., & Georges, A. (2018). DARTR: An R package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Molecular Ecology Resources*, *18*(3), 691–699. doi: 10.1111/1755-0998.12745

Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). Circlize implements and enhances circular visualization in R. *Bioinformatics*, *30*(19), 2811–2812. doi: 10.1093/bioinformatics/btu393

Jombart, T. (2008). Adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405. doi: 10.1093/bioinformatics/btn129

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics*, *27*(21), 3070–3071. doi: 10.1093/bioinformatics/btr521

Knaus, B. J., & Grünwald, N. J. (2017). VCFR: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources*, *17*(1), 44–53. doi: 10.1111/1755-0998.12549

Lischer, H. E. L., & Excoffier, L. (2012). PGDSpider: An automated data conversion tool for

connecting population genetics and genomics programs. *Bioinformatics*, *28*(2), 298–299. doi: 10.1093/bioinformatics/btr642

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., & Emerson, B. C. (2015). Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference. *Molecular Ecology Resources*, *15*(1), 28–41. doi: 10.1111/1755-0998.12291

Metservice. (2021). *Moana Backbone Model: A 25-year Hydrodynamic Hindcast Model of New Zealand Waters*. doi: 10.5281/zenodo.5895265

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. doi: 10.1093/molbev/msu300

Ortiz, E. M. (2019). *vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis*. 10.5281/zenodo.2540861. doi: 10.5281/zenodo.2540861

Parvizi, E., Dutoit, L., Fraser, C., & Waters, J. (2022). Concordant phylogeographic responses to large-scale coastal disturbance in intertidal macroalgae and their epibiota. *Molecular Ecology*, *31*(2), 646–657. doi: 10.1111/mec.16245

Parvizi, E., Fraser, C. I., Dutoit, L., Craw, D., & Waters, J. M. (2020). The genomic footprint of coastal earthquake uplift. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1930), 2–8. doi: 10.1098/rspb.2020.0712rspb20200712

Peters, J. C., Waters, J. M., Dutoit, L., & Fraser, C. I. (2020). SNP analyses reveal a diverse pool of potential colonists to earthquake-uplifted coastlines. *Molecular Ecology*, *29*(1), 149–159. doi: 10.1111/mec.15303

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. doi: 10.1086/519795

Rochette, N. C., Rivera-Colón, A. G., & Catchen, J. M. (2019). Stacks 2: analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular Ecology*, *28*(21), 4737–4754. doi: 10.1111/mec.15253

Sjoberg, D. (2022). *ggsankey: Sankey, Alluvial and Sankey Bump Plots*. doi: 10.1177/0048393103262550

van Etten, J. (2017). R package gdistance: Distances and routes on geographical grids. *Journal of Statistical Software*, *76*(1). doi: 10.18637/jss.v076.i13

Vaux, F., Craw, D., Fraser, C. I., & Waters, J. M. (2021). Northward range extension for *Durvillaea poha* bull kelp: response to tectonic disturbance? *Journal of Phycology*, *57*(5), 1411–1418. doi: 10.1111/jpy.13179

Vaux, F., Parvizi, E., Craw, D., Fraser, C. I., & Waters, J. M. (2022). Parallel recolonisations generate distinct genomic sectors in kelp following high magnitude earthquake disturbance. *Molecular Ecology*. doi: 10.1111/mec.16535

Waples, R. S., Waples, R. K., & Ward, E. J. (2021). Pseudoreplication in genomics-scale datasets. *Molecular Ecology Resources*, *22*(1), 503–518. doi: 10.1101/2020.11.12.380410

Wickham, H. (2016). *Elegant Graphics for Data Analysis: ggplot2*. New York: Springer-Verlag.

SUPPLEMENTARY TABLES

SUPPLEMENTARY TABLE B1
Sampling of *D. antarctica* 'NZ North' from 17 locations.

Sample group	Location	Collection date(s)	n Total
North Island	Boom Rock	29/07/20	14
	Orongorongo Beach	02/12/19	
	Cod Rocks	04/12/19	
Rarangi		17/11/20	9
Cape Campbell		20/11/16 15/11/20	13
Ward Beach		19/11/16 15/09/18 17/11/20	38
Wharanui		19/11/16 15/09/18 18/09/19 16/11/20	28
Waipapa Bay		19/11/16 14/09/18 16/11/20	50
Kaikōura Peninsula		20/11/16 28/04/17 14/11/20	33
Southern Kaikōura	Rakanui	20/11/16	12
	Raramai Tunnels	31/03/07	
	Oaro	13/11/20	
Hurunui	Elliots Garden, Gore Bay	13/11/20	11
	Napenape	18/11/20	
Banks Peninsula	Le Bons Bay	11/12/08	8
	Peraki Bay	08/03/09	
	Te Oka Bay	07/03/09	
		Totals	**216**

SUPPLEMENTARY TABLE B2

Sequencing information and results the five GBS libraries used to sequence the 216 *D. antarctica* samples for this study. A new GBS library, FV3, was generated for this study (highlighted in yellow), whereas the other four libraries were produced for previous studies (Parvizi, Fraser, Dutoit, Craw, & Waters, 2020; Vaux et al., 2021, 2022). The table includes the number of reads removed during quality control using the *process_radtags* component of STACKS 2.53.

GBS library name	CF1	CF2	FV1	FV2	FV3	Total
Total number of multiplexed samples	192	216	96	203	192	899
Multiplexed *D. antarctica* samples for this study	27	12	71	16	90	216
Number of adapter plates	2	3	1	3	2	
Size selection (bp)	250 - 450	300 - 600	200 - 600	200 - 600	200 - 500	
Sequencing platform	Illumina NextSeq 500	Illumina NextSeq 500	Illumina NextSeq 500	Illumina NextSeq 500	Illumina NextSeq 500	
Read length (bp)	75	75	75	75	75	
Paired reads?	Yes	Yes	Yes	Yes	Yes	
PhiX spike-in (approx.)	5%	5%	5%	5%	5%	
Total Sequences	307,797,516	310,206,086	315,678,078	330,627,614	363,842,806	1,628,152,100
Barcode Not Found	152,618,572	37,478,802	26,414,122	45,595,966	42,773,066	304,880,528
Low Quality	79,148	80,258	125,261	71,969	265,853	622,489
RAD Cutsite Not Found	2,868,624	6,485,392	3,029,044	2,753,042	3,302,471	18,438,573
Retained Reads	152,231,172	266,161,634	286,109,651	282,206,637	317,501,416	1,304,210,510
Percentage retained reads	49.5%	85.8%	90.6%	85.4%	87.3%	80.1%

SUPPLEMENTARY TABLE B3

Results of varying certain parameters in *ustacks*, *cstacks* and *populations* components of STACKS 2.53. Complete lists of parameters used for each component are provided in the Methods. The STACKS runs selected for analysis are highlighted in green.

N	ustacks	cstacks	# pops	populations	# variant and invariant loci	# variant loci	Mean missing data per sample
216	-m 3 -M 2 -N 4	-n 2	1	-p 1 -r 0.90	6,177	2,255	5.7%
216	-m 3 -M 2 -N 4	-n 2	1	-p 1 -r 0.80	8,995	3,833	9.4%
216	-m 3 -M 2 -N 4	-n 2	1	-p 1 -r 0.70	10,629	4,872	12.9%
216	-m 3 -M 2 -N 4	-n 2	1	-p 1 -r 0.60	12,035	5,919	17.0%
216	-m 3 -M 2 -N 4	-n 2	1	-p 1 -r 0.50	13,440	7,092	21.8%
216	-m 4 -M 2 -N 4	-n 2	1	-p 1 -r 0.60	11,252	5,549	16.6%
216	-m 3 -M 2 -N 4	-n 1	1	-p 1 -r 0.60	11,976	5,944	17.2%
216	-m 3 -M 2 -N 4	-n 3	1	-p 1 -r 0.60	12,015	5,954	17.1%
216	-m 3 -M 3 -N 5	-n 2	1	-p 1 -r 0.60	11,657	5,763	17.1%

SUPPLEMENTARY TABLE B4

Details for final filtered loci dataset. The dataset was filtered loci for linkage disequilibrium (LD), genotype coverage depth, and locus coverage depth. The final column lists the mean missing data per sample for the dataset, as estimated by VCFR.

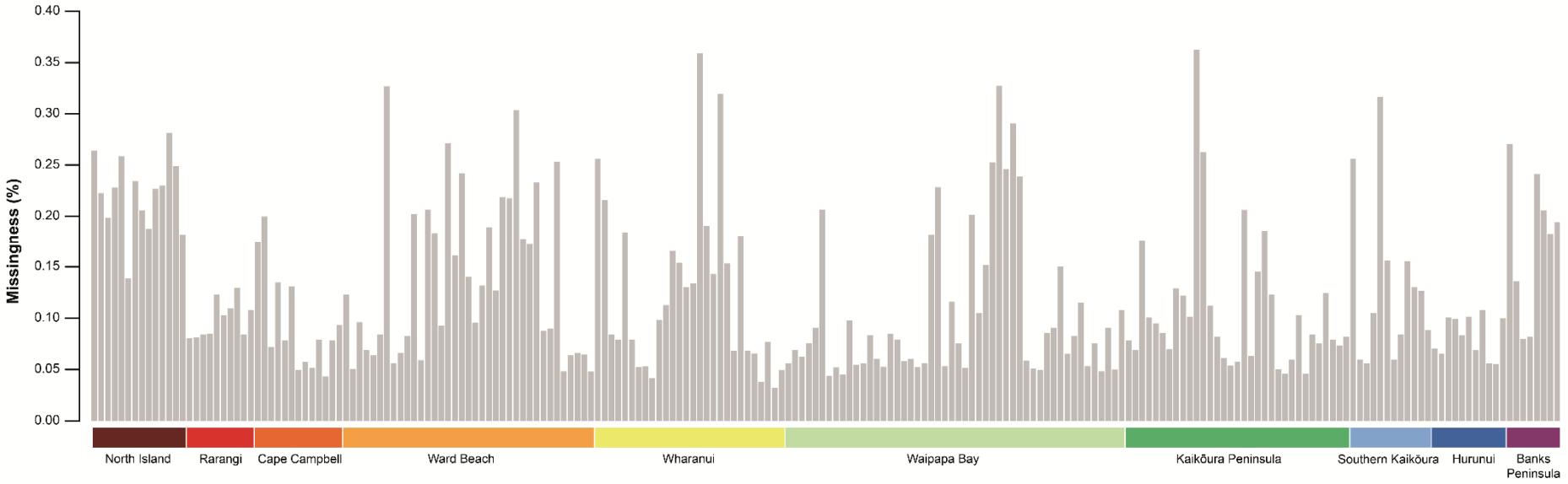
Locus representation			Filtering						
n	**# groups**	***populations* settings**	**# LD**	**# Genotype depth <8**	**# Loci depth outliers**	**# Highly correlated loci**	**Total excluded loci**	**# final variant loci (SNPs)**	**Mean missing data per sample**
216	1	-p 1 -r 0.70	42	153	55	N/A	243	4,629	12.4%

20 SUPPLEMENTARY B FIGURES

21 SUPPLEMENTARY FIGURE B1

22 The percentage of missing data (missingness, estimated in VCFR) per sample in the GBS dataset (4,629 loci). Sample groups are labelled under
23 samples.

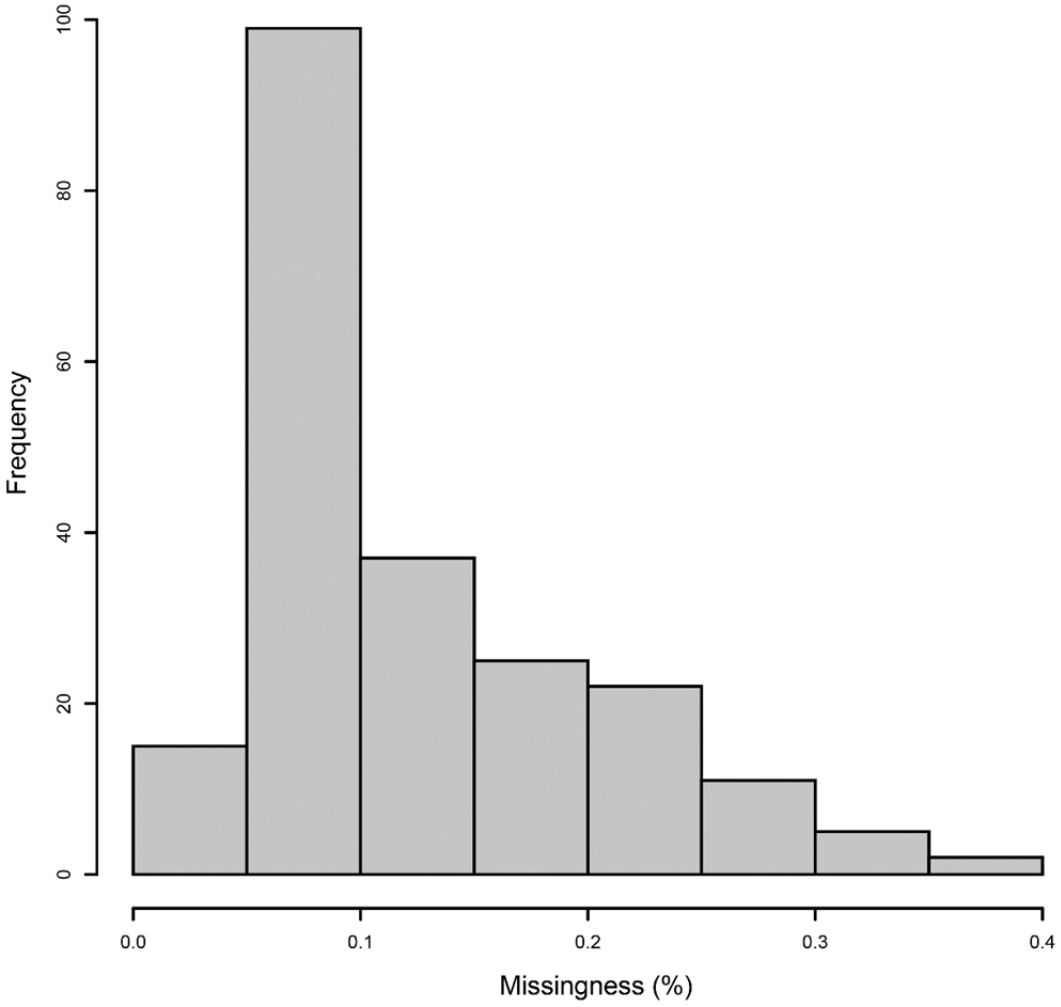
24



25

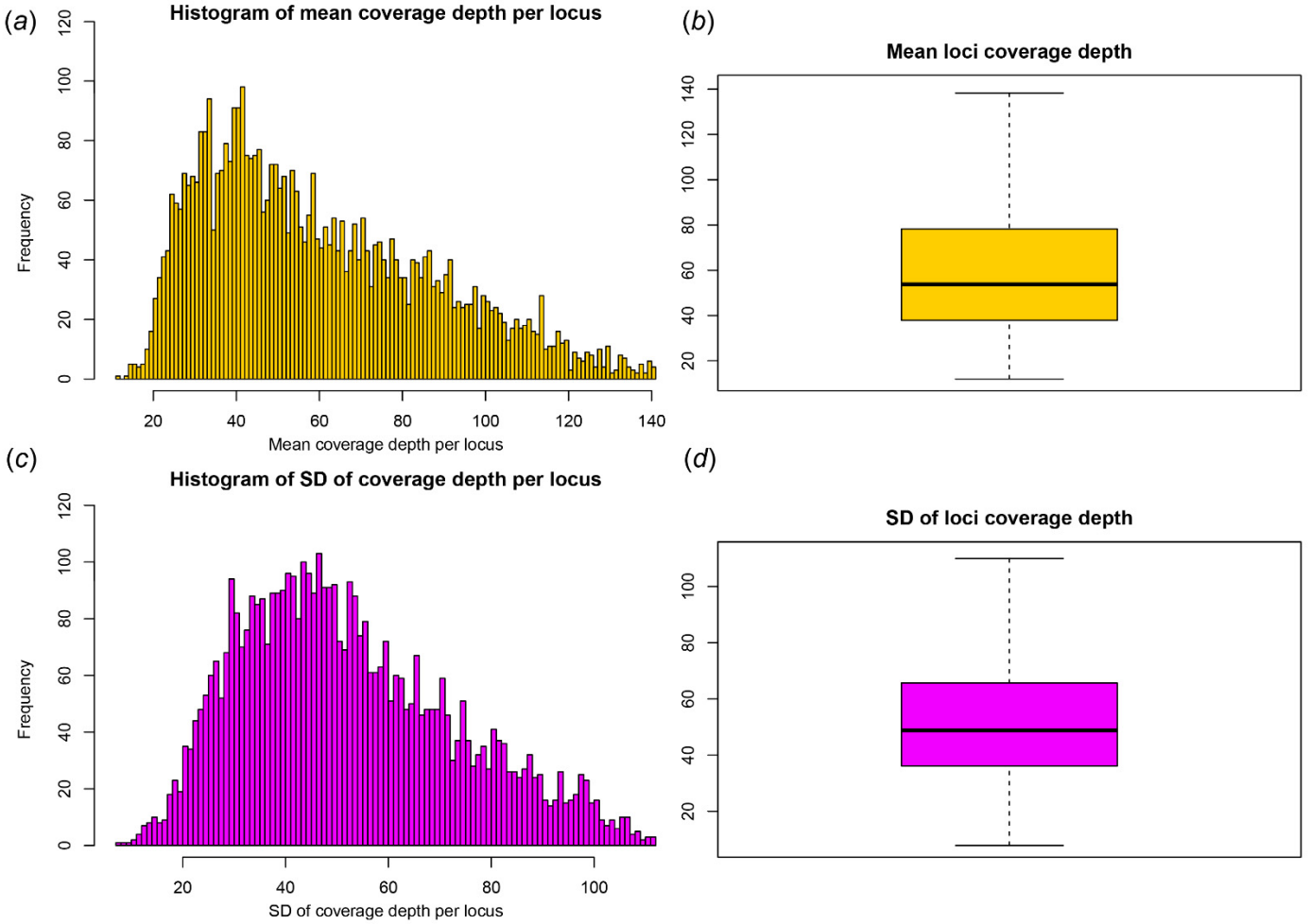
SUPPLEMENTARY FIGURE B2

Histograms of missing data (missingness, estimated in VCFR) per sample in the GBS dataset (4,629 loci).



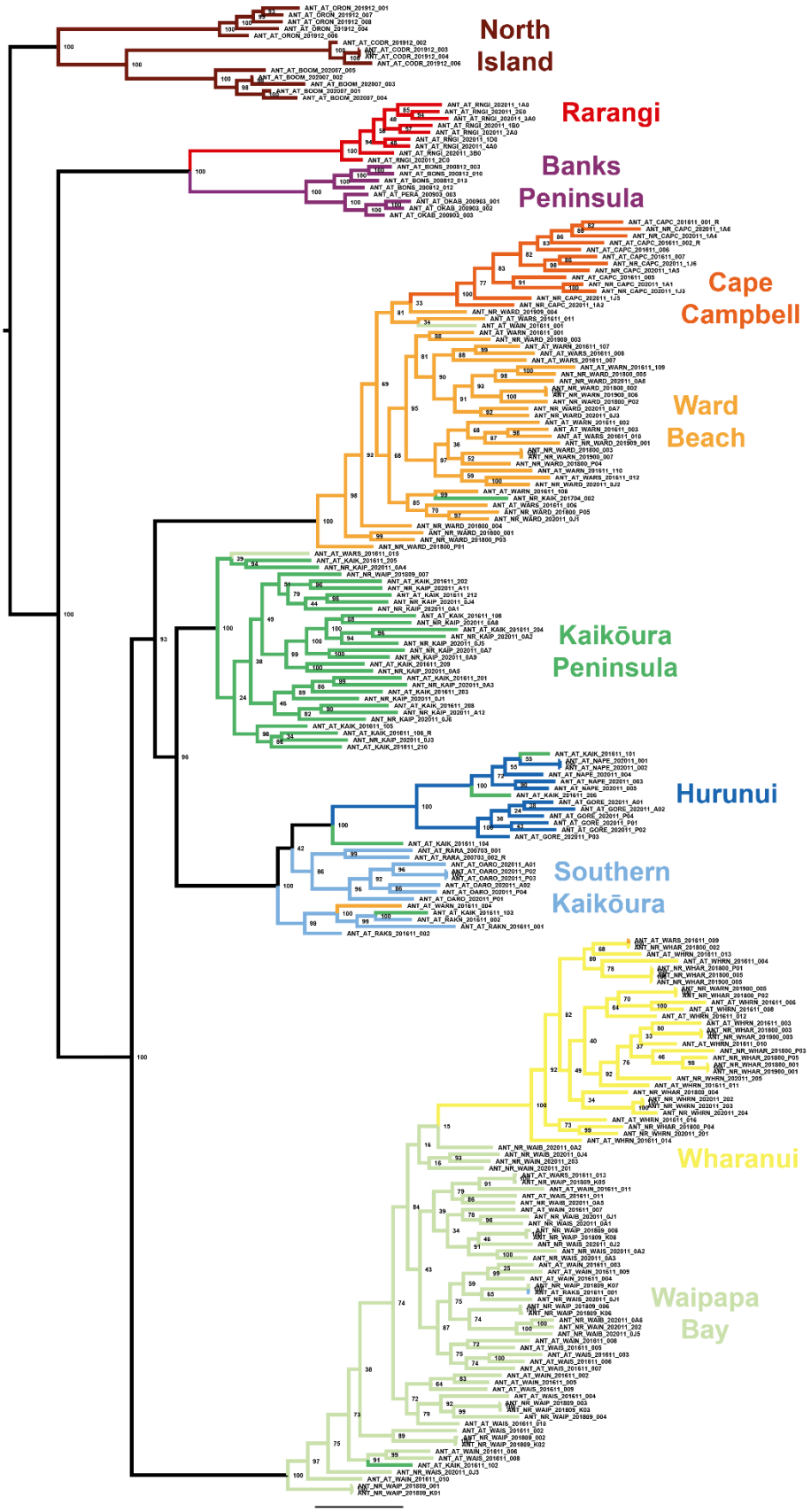
SUPPLEMENTARY FIGURE B3

Coverage depth per locus in the GBS dataset, as estimated in VCFR, including highly correlated loci. (*a*) a histogram of mean coverage depth per locus, (*b*) a box and whisker plot of mean coverage depth per locus, (*c*) a histogram of the standard deviation (SD) in coverage depth per locus, (*d*) a box and whisker plot of the standard deviation in coverage depth per locus.



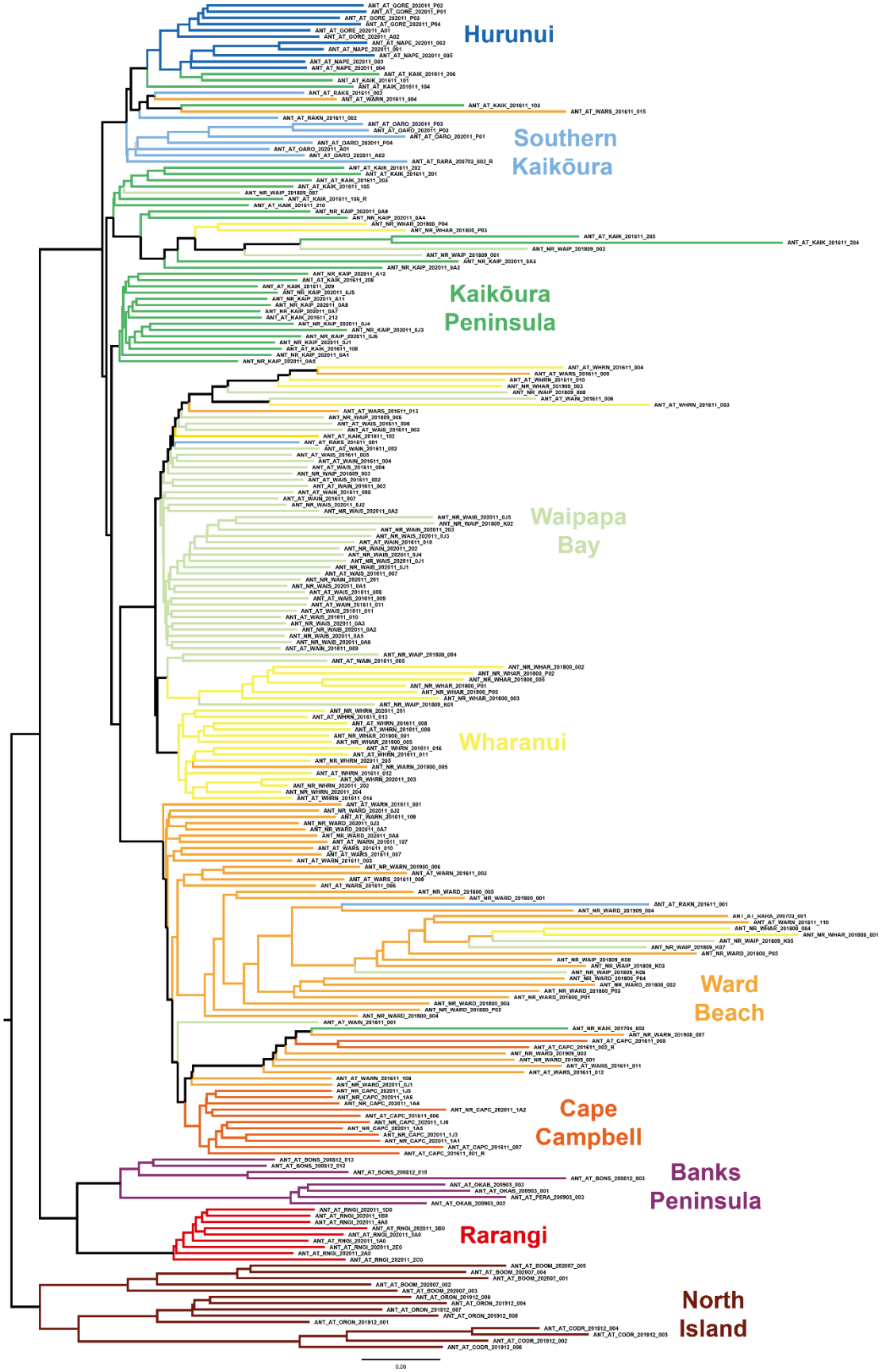
SUPPLEMENTARY FIGURE B4

Unrooted maximum likelihood (ML) phylogeny produced by IQTREE for the GBS dataset (4,629 loci).



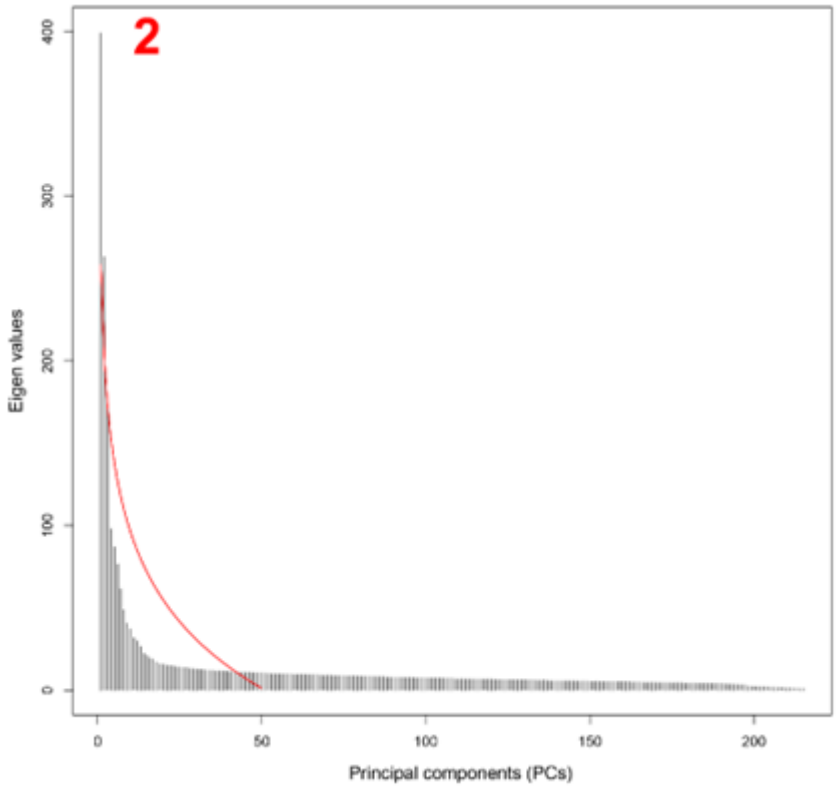
SUPPLEMENTARY FIGURE B5

Unrooted neighbour-joining (NJ) phylogeny produced by VCF-KIT for the GBS dataset (4,629 loci).



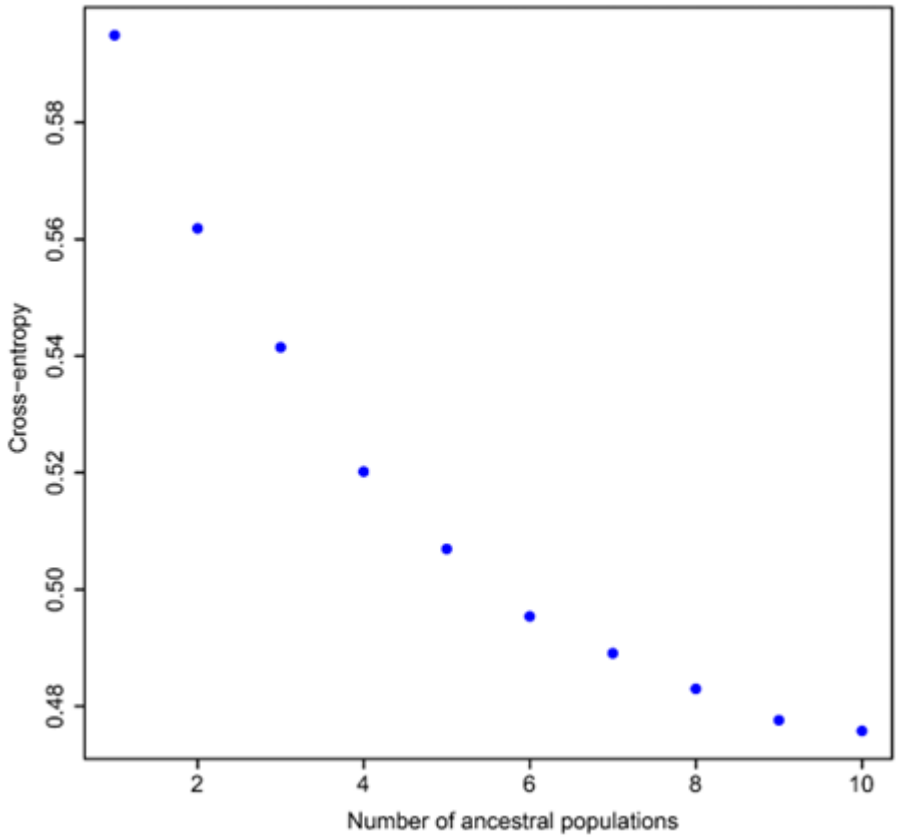
SUPPLEMENTARY FIGURE B6

The selection of retained principal components (PCs) for principal components analysis. Each graph shows the Eigen values for all PCs in the GBS dataset (4,629 loci). The red line shows the broken-stick test, the number of PCs above the broken-stick line that were retained for analysis are labelled in red.



SUPPLEMENTARY FIGURE B7

Cross entropy values for each iteration of K (1 – 10) applied in LEA for the GBS dataset (2,851 loci).



SUPPLEMENTARY FIGURE B8

Ancestry matrices (K = 2 – 10) generated by LEA for the GBS dataset (4,629 loci).