Supplementary Information: Genome assembly and annotation of the gray mangrove *Avicennia marina*

Guillermo Friis, Joel Vizueta, Edward G. Smith, David R. Nelson, Basel Khraiwesh, Enas Qudeimat, Kourosh Salehi-Ashtiani, Alejandra Ortega, Alyssa Marshall, Carlos M. Duarte, John A. Burt

Sequencing and assembly of the *Avicennia marina* genome

Chicago library preparation and sequencing

A Chicago library was prepared as described previously (Putnam et al. 2016). Approximately 500ng of HMW gDNA (mean fragment length = 80 kb) was reconstituted into chromatin *in vitro* and fixed with formaldehyde. Fixed chromatin was digested with DpnII, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed, and the DNA purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to ~350 bp mean fragment size and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were sequenced on an Illumina HiSeqX to produce 235 million 2x150bp paired end reads, which provided 381.78X physical coverage of the genome (1-100 kb pairs).

Dovetail HiC library preparation and sequencing

A Dovetail HiC library was prepared in a similar manner as described previously (Lieberman-Aiden et al. 2009). For each library, chromatin was fixed in place with formaldehyde in the nucleus and then extracted Fixed chromatin was digested with DpnII, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed, and the DNA purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to ~350 bp mean fragment size and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were sequenced on an Illumina HiSeqX to produce 212 million 2x150bp paired end reads, which provided 23,864.92X physical coverage of the genome (10-10,000 kb pairs).

Scaffolding the assembly with HiRise

Chicago library reads and Dovetail HiC library reads were used as input data for HiRise (Putnam et al, 2016). An iterative analysis was conducted. First, Chicago library sequences need to be aligned to a draft genome, for which we used a previously released assembly of *Avicennia marina* (Lyu et al. 2018; Xu et al. 2017; assembly accession: GCA_900003535.1). HiRise does not use scaffolds under 1 Kb, and they were thereby excluded from the aligning. The aligning was conducted using a modified SNAP read mapper (http://snap.cs.berkeley.edu). The separations of Chicago read pairs mapped within draft scaffolds were analyzed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify and break putative miss-joins, to score prospective joins, and make joins above a

threshold. After aligning and scaffolding Chicago data, Dovetail HiC library sequences were aligned and scaffolded following the same method (Figure S1).

To explore synteny patterns and the improvement in the scaffolding level from the draft genome to the final HiRise assembly, we generated two synteny plots between the 32 putative chromosomes of the HiRise assembly and (i) the 196 scaffolds accounting for more than half of the draft genome (i.e. L50 = 196) and the (ii) 115 scaffolds accounting for more than 90% of the Chicago assembly (i.e. L90 = 115). Mapping coordinates were computed with blastn from the BLAST suite (Tatusova and Madden 1999), and the plots were produced with Circos (Krzywinski et al. 2009). A link density histogram showing the mapping positions of the read pairs is also provideded in Figure S2.

Genome annotation

Messenger RNA sequencing

Samples for RNA-seq were gathered on the coast of the Red Sea near Jeddah in the Kingdom of Saudi Arabia (22.324 °N, 39.100 °E). Gray mangrove samples from root, stem, leaf, flower, and seed were collected during the winter of 2012 and flash-frozen in liquid nitrogen and subsequently stored at -80°C for RNA extraction. For stem tissue, young or new growth was used. For roots, newly emerging vertical pneumatophores were used. For leaves, 15 leaves from different trees were divided into three groups; RNA was extracted from each group and pooled within each tissue for library preparation. Total RNA was isolated using TRIzol reagent (Invitrogen, USA) following the manufacturer's instructions. RNA was extracted in triplicates from each tissue, then pooled for library preparation. To improve completeness, an extra library

was prepared pooling isolated RNA from all the extracted tissues. The *A. marina* mRNA libraries were prepared using TruSeq RNA sample prep kit (Illumina, Inc.) following the manufacturer's instructions, with inserts that range in size from approximately 100-400 bp. Library quality control and quantification were performed with a Bioanalyzer Chip DNA 1000 series II (Agilent) and sequenced with a HiSeq2000 (Illumina, Inc.).

Adaptive variability analysis based on genome resequencing data

Genome resequencing and variant calling

Whole genome resequencing was carried out for the 60 individuals from 6 complementary populations around the Arabian Peninsula (Figure 1 of the manuscript; Table S1) at Novogene facilities. Illumina paired-end 150 bp libraries with insert size equal to 350 pb were prepared and sequenced in a Novaseq platform. A total of 2.4G reads were produced resulting in a mean coverage per site and sample of 85X before filtering. Read quality was evaluated using FASTQC (Andrews 2010) after sorting reads by individual with AXE (Murray and Borevitz 2017). Trimming and quality filtering treatment was conducted using Trim Galore (Krueger 2015), resulting in a set of reads ranging between 90 and 138 bp long. Reads were then mapped against the *A. marina* reference genome using the mem algorithm in the Burrows-Wheeler Aligner (BWA; Li and Durbin 2009). Read groups were assigned and BAM files generated with Picard Tools version 1.126 (http://broadinstitute.github.io/picard). We used the HaplotypeCaller + GenotypeGVCFs tools from the Genome Analysis Toolkit (GATK; McKenna et al. 2010) version 4.1.8.1 to produce a set of single nucleotide polymorphisms (SNPs) in the variant call format (*vcf*). Four samples presenting more than a 25% of missing data were discarded at this

point. Using vcftools (Danecek et al. 2011), we retained biallelic SNPs excluding those out of a range of coverage between 4 and 50 or with a genotyping phred quality score below 40. A threshold for SNPs showing highly significant deviations from Hardy-Weinberg equilibrium (HWE) with a p-value of 10^{-4} was also implemented to filter out false variants arisen by the alignment of paralogous loci. We then applied GATK generic hard-filtering recommendations consisting on QualByDepth (QD) > 2.0; FisherStrand (FS) < 60.0; RMSMappingQuality (MQ) > 40; MappingQualityRankSumTest (MQRankSum) > -12.5; ReadPosRankSum (RPRK) < -8.0; and StrandOddsRatio (SOR) > 3.0 (GATK Best Practices; See Appendix I; Auwera et al. 2013; DePristo et al. 2011). To produce the final SNP matrix used in genome scans, positions for which one or more samples were not genotyped were removed, along with those presenting a minor allele count (MAC) below 3 and only the SNPs from the 32 major scaffolds were retained. Final dataset consisted on 56 samples and 538,185 SNPs.

F_{ST} scan and t-SNE analysis

A Weir and Cockerham (1984) F_{ST} analysis was conducted using vcftools on sliding windows of 20 Kb for the SNP dataset. A Manhattan plot was produced with the R-package qqman (Turner 2016). We identify 200 outlier loci using the R function boxplots.stats and implementing a restrictive coefficient of 1.5, i.e. a value 1.5 times higher than the length of the third and fourth interquartile range for the F_{ST} distributions. Genes overlapping 123 F_{ST} outliers were then recovered resulting in a list of 109 candidate genes (Table S3). Associated GO terms are summarized in a WEGO plot (Figure S5; Ye et al. 2006)

To explore patterns of variation in the functional, putatively under selection loci identified in the F_{ST} scan, we conducted a t-distributed stochastic neighbor embedding (t-SNE) analysis, a powerful method for analyzing high-dimensional data. To do so, we extracted the variable positions from the outlier loci and translated them into counts of the non-reference allele for each position with vcftools, resulting in a 613 SNP matrix. We then used Rtsne R-package (Krijthe et al. 2018) to perform the analysis with a perplexity value of 18, as recommended by the authors for our number of points, and dimensionality equal to 2. Linear regressions were run for each one of the recovered t-SNE axes with the annual range of sea surface temperature (SST) for each one of the mangrove populations. Environmental georeferenced data was obtained from GMED (http://gmed.auckland.ac.nz/, Basher et al. 2018). T-SNE axes and fitted values representing the correlation with the gradient of temperature range were plotted for visual interpretation.

Table S1. Resequenced samples of *Avicennia marina*. Seq. depth refers to the mean sequence coverage of the called SNP set for each individual.

Seq. ID	Locality	Field code	Country	LAT	LONG	Collection date	Seq. Depth
NRS01	North Red Sea	M1-2017-1	Saudi Arabia	26.91522	36.01074	01/05/2017	16.09
NRS02	North Red Sea	M1-2017-2	Saudi Arabia	26.91522	36.01074	01/05/2017	16.20
NRS03	North Red Sea	M1-2017-3	Saudi Arabia	26.91522	36.01074	01/05/2017	15.77
NRS04	North Red Sea	M1-2017-4	Saudi Arabia	26.91522	36.01074	01/05/2017	19.00
NRS06	North Red Sea	M1-2017-6	Saudi Arabia	26.91522	36.01074	01/05/2017	15.23
NRS07	North Red Sea	M1-2017-7	Saudi Arabia	26.91522	36.01074	01/05/2017	17.73
NRS08	North Red Sea	M1-2017-8	Saudi Arabia	26.91522	36.01074	01/05/2017	15.43
NRS10	North Red Sea	M1-2017-10	Saudi Arabia	26.91522	36.01074	01/05/2017	19.05
SND01	Bahrain	19-156	Bahrain	26.15218	50.59536	02/03/2019	23.15
SND02	Bahrain	19-157	Bahrain	26.15170	50.59485	02/03/2019	17.46
SND03	Bahrain	19-158	Bahrain	26.15200	50.59450	02/03/2019	14.26
SND04	Bahrain	19-159	Bahrain	26.15210	50.59441	02/03/2019	20.20
SND05	Bahrain	19-160	Bahrain	26.15221	50.59428	02/03/2019	21.12
SND06	Bahrain	19-161	Bahrain	26.15231	50.59414	02/03/2019	21.89
SND07	Bahrain	19-162	Bahrain	26.15244	50.59400	02/03/2019	17.61
SND08	Bahrain	19-167	Bahrain	26.15310	50.59331	02/03/2019	22.58
SND09	Bahrain	19-168	Bahrain	26.15327	50.59318	02/03/2019	16.18
SND10	Bahrain	19-170	Bahrain	26.15355	50.59288	02/03/2019	26.67
RGB02	Ras Ghurab	18-008	U.A.E.	24.60115	54.56653	29/11/2018	20.38
RGB03	Ras Ghurab	18-009	U.A.E.	24.60137	54.56772	29/11/2018	22.75
RGB04	Ras Ghurab	18-010	U.A.E.	24.6013	54.56824	29/11/2018	21.64
RGB05	Ras Ghurab	18-011	U.A.E.	24.60143	54.5686	29/11/2018	15.71
RGB06	Ras Ghurab	18-012	U.A.E.	24.60164	54.56905	29/11/2018	20.44
RGB07	Ras Ghurab	18-013	U.A.E.	24.60157	54.56933	29/11/2018	24.30
RGB08	Ras Ghurab	18-014	U.A.E.	24.60177	54.57045	29/11/2018	17.10
RGB09	Ras Ghurab	18-015	U.A.E.	24.60061	54.56556	29/11/2018	16.25
RGB10	Ras Ghurab	18-016	U.A.E.	24.60081	54.56593	29/11/2018	16.35
QRM01	Qurm	19-076	Oman	23.62594	58.48225	27/01/2019	16.00
QRM02	Qurm	19-077	Oman	23.62552	58.48292	27/01/2019	22.32
QRM03	Qurm	19-078	Oman	23.62501	58.48163	27/01/2019	20.94
QRM04	Qurm	19-079	Oman	23.62476	58.48149	27/01/2019	17.26
QRM05	Qurm	19-080	Oman	23.62470	58.48146	27/01/2019	19.64
QRM06	Qurm	19-082	Oman	23.62430	58.48133	27/01/2019	15.65
QRM07	Qurm	19-083	Oman	23.62415	58.48132	27/01/2019	20.14
QRM08	Qurm	19-084	Oman	23.62387	58.48127	27/01/2019	15.96
QRM10	Qurm	19-086	Oman	23.62355	58.48118	27/01/2019	16.79
FLM01	Filim	19-106	Oman	20.60761	58.18702	28/01/2019	20.21

FLM02	Filim	19-107	Oman	20.60719	58.18694	28/01/2019	21.66
FLM03	Filim	19-108	Oman	20.60707	58.18681	28/01/2019	23.36
FLM04	Filim	19-110	Oman	20.60645	58.18640	28/01/2019	15.20
FLM05	Filim	19-111	Oman	20.60630	58.18618	28/01/2019	19.37
FLM06	Filim	19-112	Oman	20.60603	58.18598	28/01/2019	17.22
FLM07	Filim	19-115	Oman	20.60589	58.18464	28/01/2019	24.22
FLM08	Filim	19-118	Oman	20.60527	58.18408	28/01/2019	14.69
FLM09	Filim	19-119	Oman	20.60534	58.18392	28/01/2019	17.64
FLM10	Filim	19-120	Oman	20.60520	58.18368	28/01/2019	18.14
ESL01	Salalah	19-146	Oman	17.02167	54.23963	30/01/2019	17.76
ESL02	Salalah	19-147	Oman	17.02135	54.23957	30/01/2019	19.00
ESL03	Salalah	19-148	Oman	17.02202	54.23946	30/01/2019	19.38
ESL04	Salalah	19-149	Oman	17.02221	54.23940	30/01/2019	19.28
ESL05	Salalah	19-150	Oman	17.02240	54.23938	30/01/2019	18.84
ESL06	Salalah	19-151	Oman	17.02251	54.23947	30/01/2019	21.60
ESL07	Salalah	19-152	Oman	17.02258	54.23955	30/01/2019	19.05
ESL08	Salalah	19-153	Oman	17.02280	54.23968	30/01/2019	23.83
ESL09	Salalah	19-154	Oman	17.02295	54.23976	30/01/2019	24.04
ESL10	Salalah	19-155	Oman	17.02680	54.24016	30/01/2019	19.61

Table S2. Results of the repetitive elements annotation conducted with RepeatModeler v2.0.1 and RepeatMasker 4.0.9 (Flynn et al. 2019; Smit et al. 2015). Abbreviated types of elements correspond to short and long interspersed nuclear elements (SINEs and LINEs, respectively) and long terminal repeats (LTRs).

Type of element	Number	Length	Percentage
SINEs	21	2,883 bp	0.00%
LINEs	4077	24,404,03 bp	0.53%
LTRs	90146	91,329,761 bp	20.00%
DNA elements	28205	13,507,948 bp	2.96%
Unclassified	287652	76,246,996 bp	16.70%
Total interspersed repeats	183,527,991 bp		40.20%

Table S3. Genes linked to F_{ST} outliers and annotated orthologues.

Gene	Homolog	Organisms	Swissprot/Interpro annotation
jg9568	4CLL9	ARATH	4-coumarate--CoA ligase-like 9
jg27559	A0A178UJI8	ARATH	SWIM zinc finger
jg11657	A0A654EFS6	ARATH	Uncharacterized protein
jg36564	A0A654FUN5	ARATH	Uncharacterized protein
jg2194	A0A654GAV7	ARATH	CCT domain-containing protein
jg12875	AAE16	ARATH	Probable acyl-activating enzyme 16, chloroplastic
jg33974	AB9B	ARATH	ABC transporter B family member 9
jg9382	AMPD	ARATH	AMP deaminase
jg14653	ARI8	ARATH	Probable E3 ubiquitin-protein ligase ARI8
jg6141	ASG2	ARATH	Protein ALTERED SEED GERMINATION
jg26318	BAK1	ARATH	BRASSINOSTEROID INSENSITIVE 1-associated receptor kinase 1
jg26874	BAM7	ARATH	Beta-amylase 7
jg18084	BC10	ORYSJ	Glycosyltransferase BC10
jg22688	BH074	ARATH	Transcription factor bHLH74
jg18088	BH087	ARATH	Transcription factor bHLH87
jg22570	CATA3	NICPL	Catalase isozyme 3
jg12068	CCA11	ARATH	Cyclin-A1-1
jg15517	CFXQ	CYAM1	Ribulose bisphosphate carboxylase/oxygenase activase, chloroplastic
jg39383	CLPT1	HUMAN	Cleft lip and palate transmembrane protein
jg38448	CNBL3	ORYSJ	Calcineurin B-like protein
jg35228	COL13	ARATH	Zinc finger protein
jg28754	COPG2	ORYSJ	Coatomer subunit gamma-2
jg14275	CPR49	ARATH	GDSL esterase/lipase
jg5911	CSTR3	ARATH	CMP-sialic acid transporter 3
jg5910	CTL1	ARATH	Chitinase-like protein 1
jg22216	CTSL2	DICDI	CTD small phosphatase-like protein 2
jg5740	CYSK	SOLTU	Cysteine synthase
jg9428	DNJ16	ARATH	Chaperone protein dnaJ 16
jg20117	DPEP	SOLTU	4-alpha-glucanotransferase, chloroplastic/amyloplastic
jg27780	EAF6	XENTR	Chromatin modification-related protein
jg20333	ENDO4	ARATH	Endonuclease 4
jg28338	F4JQ74	ARATH	ATPase family associated with various cellular activities
jg19098	F4KCX5	ARATH	Uncharacterized protein
jg14025	FB345	ARATH	F-box protein
jg14023	FIL1	ANTMA	Stamen-specific protein
jg15317	FMO1	ARATH	Probable flavin-containing monooxygenase 1
jg26370	FOLT1	ARATH	Folate transporter 1, chloroplastic
jg38098	FRS5	ARATH	Protein FAR1-RELATED SEQUENCE 5
jg15505	FRS5	ARATH	Protein FAR1-RELATED SEQUENCE 5

jg26704	FURH	VIBFR	Furcatin hydrolase
jg26531	GATL7	ARATH	Probable galacturonosyltransferase-like 7
jg29661	GET3A	ARATH	ATPase GET3A
jg22575	GIGAN	ARATH	Protein GIGANTEA
jg37996	GIL1	ARATH	Protein GRAVITROPIC IN THE LIGHT 1
jg13893	GLUT1	ARATH	Glutamate synthase 1 [NADH], chloroplastic
jg26876	GPAT4	ARATH	Glycerol-3-phosphate 2-O-acyltransferase 4
jg39913	GRF1	ARATH	Growth-regulating factor 1
jg9737	GSHB	SOLLC	Glutathione synthetase, chloroplastic
jg39681	HDA19	ARATH	Histone deacetylase 19
jg37995	HEXO1	ARATH	Beta-hexosaminidase 1
jg3613	HNRPQ	ARATH	Heterogeneous nuclear ribonucleoprotein Q
jg6515	HPR	THEMA	Hydroxypyruvate reductase
jg15293	HST	TOBAC	Shikimate O-hydroxycinnamoyltransferase
jg5736	IDM1	ARATH	Increased DNA methylation 1
jg28048	IF2G	BOVIN	Eukaryotic translation initiation factor 2 subunit 3
jg9430	ILL1	ORYSJ	IAA-amino acid hydrolase ILR1-like 1
jg31945	INV2	DAUCA	Beta-fructofuranosidase, insoluble isoenzyme 2
jg37745	IRKI	ARATH	IRK-interacting protein
jg13779	ITPK1	MAIZE	Inositol-tetrakisphosphate 1-kinase 1
jg31626	LAC14	ARATH	Laccase-14
jg9669	LEP	ARATH	Ethylene-responsive transcription factor LEP
jg5907	LEU1A	SOLPN	2-isopropylmalate synthase A
jg28043	LIN1	LOTJA	Putative E3 ubiquitin-protein ligase LIN-1
jg23202	M3K1	ARATH	Mitogen-activated protein kinase kinase kinase 1
jg24959	MDIS2	ARATH	Protein MALE DISCOVERER 2
jg17175	NAGK	ARATH	Acetylglutamate kinase, chloroplastic
jg21226	NCED1	SOLLC	9-cis-epoxycarotenoid dioxygenase NCED1, chloroplastic
jg23934	NHL6	ARATH	NDR1/HIN1-like protein 6
jg5721	NPK1	TOBAC	Mitogen-activated protein kinase kinase kinase NPK1
jg15290	OCT7	ARATH	Organic cation/carnitine transporter 7
jg8502	PBL3	ARATH	Probable serine/threonine-protein kinase PBL3
jg38541	PDS5B	CHICK	Sister chromatid cohesion protein PDS5 homolog B
jg36561	PKL	ARATH	CHD3-type chromatin-remodeling factor PICKLE
jg14277	PLCD6	ARATH	Phosphoinositide phospholipase C 6
jg19370	PLSC	COCNU	1-acyl-sn-glycerol-3-phosphate acyltransferase
jg14279	POM1	SCHPO	DYRK-family kinase pom1
jg3616	PP223	ARATH	Putative pentatricopeptide repeat-containing protein g11460, mitochondrial
jg19871	PP351	ARATH	Pentatricopeptide repeat-containing protein g35850, mitochondrial
jg15506	PRRP1	ARATH	Proteinaceous RNase P 1, chloroplastic/mitochondrial
jg25640	PTA10	ARATH	Protein PLASTID TRANSCRIPTIONALLY ACTIVE 10
jg36560	PUB12	ORYSJ	U-box domain-containing protein 12
jg3839	Q944R4	ARATH	Glycosyl transferase family 21

jg1999	Q9M369	ARATH	Uncharacterized protein
jg38451	Q9SI11	ARATH	DUF868 family protein
jg38099	Q9SRR8	ARATH	2Fe-2S ferredoxin-like superfamily protein
jg27647	RAD50	ARATH	DNA repair protein RAD50
jg31552	RAVL3	ARATH	AP2/ERF and B3 domain-containing transcription factor g51120
jg6142	REN1	ARATH	Rho GTPase-activating protein REN1
jg12787	RH13	ORYSI	DEAD-box ATP-dependent RNA helicase 13
jg5897	SAD2	ARATH	Importin beta-like SAD2
jg32124	SCL9	ARATH	Scarecrow-like protein 9
jg38102	SCP20	ARATH	Serine carboxypeptidase-like 20
jg33999	SLK2	ARATH	Probable transcriptional regulator SLK2
jg5927	SMAL1	DANRE	SWI/SNF-related matrix-associated actin-dependent
jg28801	SQS2	PANGI	Squalene synthase 2
jg28752	SWI3D	ARATH	SWI/SNF complex subunit SWI3D
jg10543	TBB1	MAIZE	Tubulin beta-1 chain (Beta-1-tubulin)
jg11171	TIC	ARATH	Protein TIME FOR COFFEE
jg7191	TTM2	ARATH	Inorganic pyrophosphatase TTM2
jg5886	UBXN1	XENTR	UBX domain-containing protein 1
jg21820	UPL5	ARATH	E3 ubiquitin-protein ligase UPL5
jg9570	UREA	ARATH	Urease
jg6533	VP52A	ARATH	Vacuolar protein sorting-associated protein 52 A
jg5915	VTI12	ARATH	Vesicle transport v-SNARE 12
jg21273	Y2913	ARATH	G-type lectin S-receptor-like serine/threonine-protein kinase g19130
jg22611	Y4920	ARATH	Uncharacterized protein g37920
jg32715	Y5625	ARATH	B3 domain-containing protein g06250
jg23932	ZIP4	ARATH	Zinc transporter 4, chloroplastic
jg22221	ZNTB	YERE8	Zinc transport protein ZntB

137

138

Fig. S1. HiRise scaffolding workflow using a draft genome along with Chicago and HiC data for a final, high quality assembly.

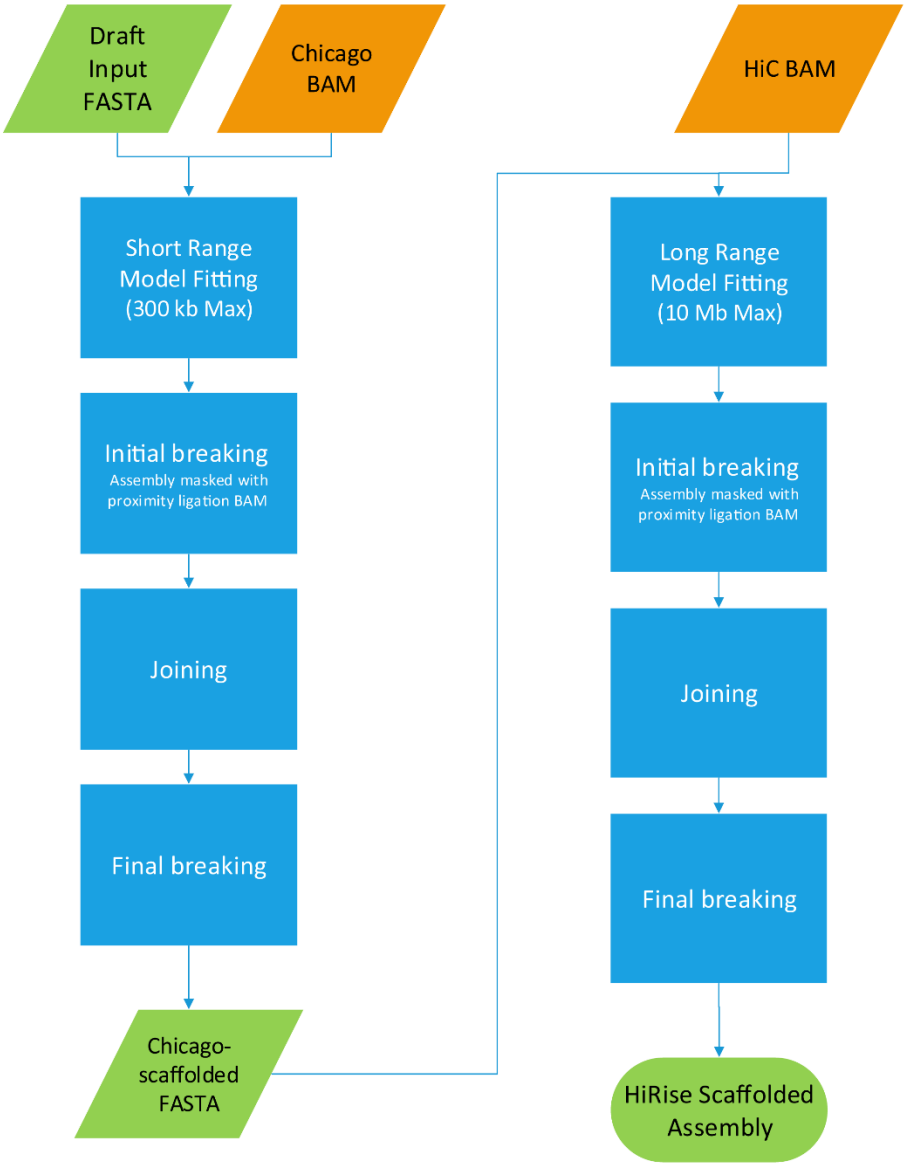


Fig. S2. Synteny plots between the 32 chromosomes of the final assembly against (A) the 196 scaffolds accounting for more than half of the draft genome; and (B) the 115 scaffolds accounting for more than 90% of the assembly based only on Chicago reads.

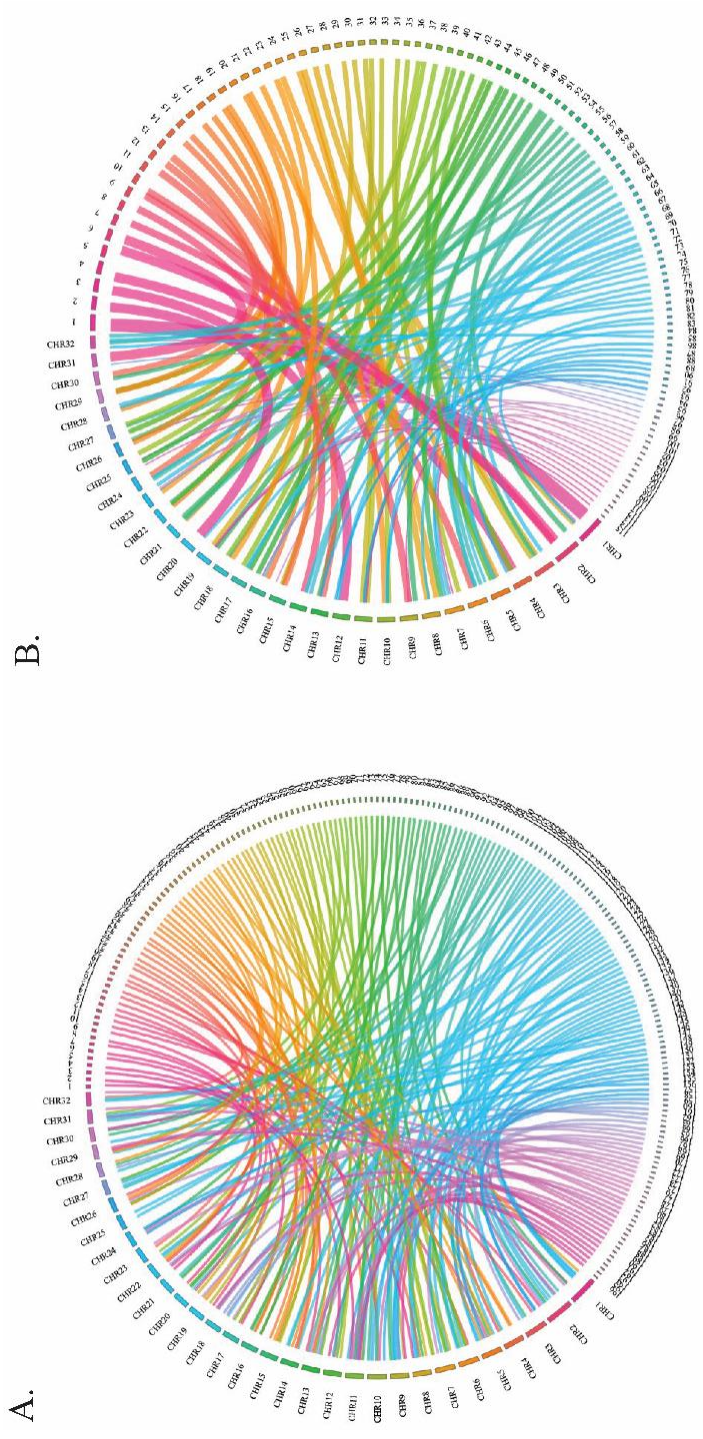


Fig. S3. Link density histogram for the assembly based on proximity ligation libraries. The horizontal and vertical axes give the mapping positions of the first and second read in the read pair respectively, grouped into bins. The color of each square gives the number of read pairs within that bin. Scaffolds shorter than 1 Mb are excluded.

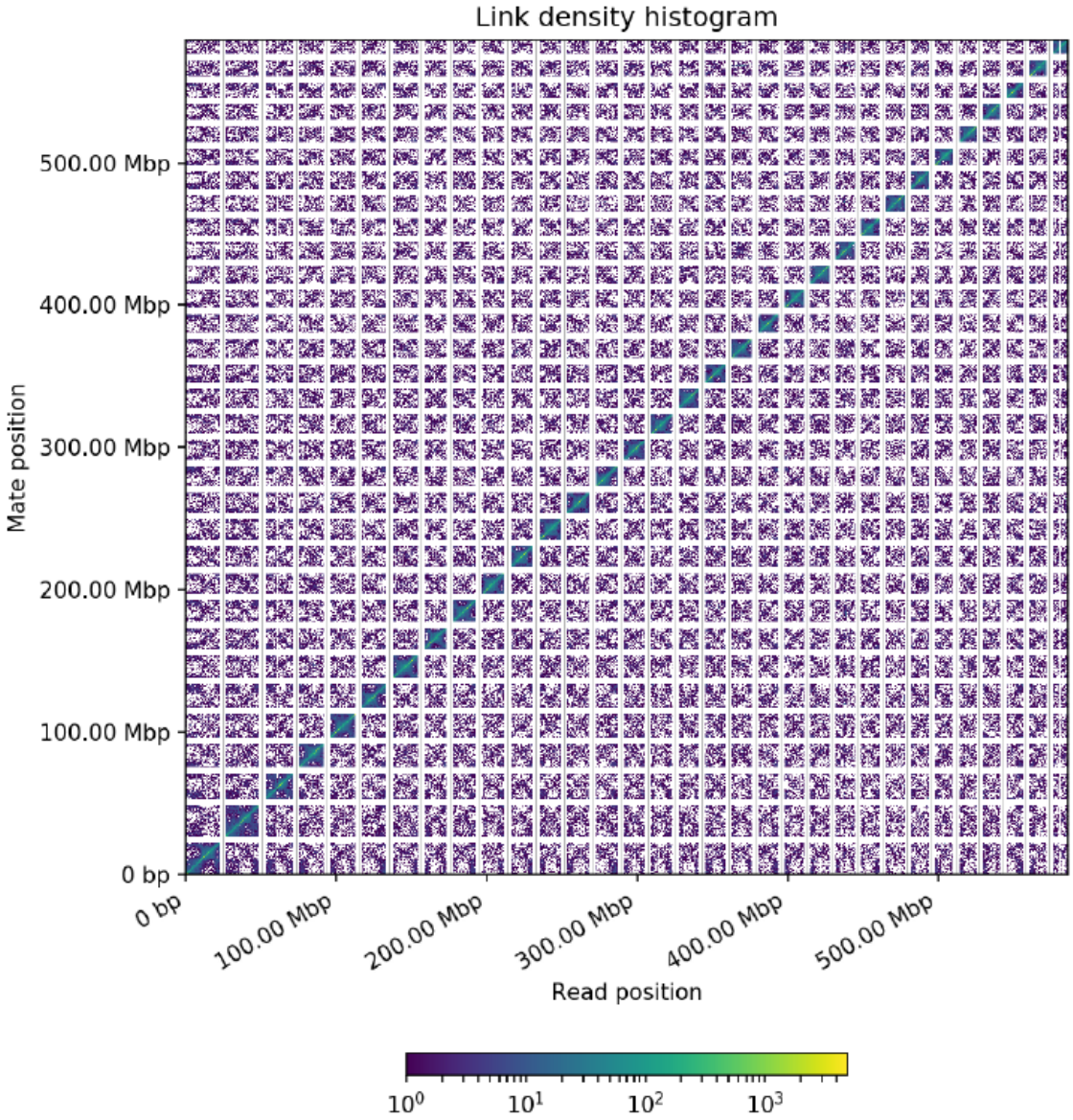


Fig. S4. Annotation workflow. Input/output files are showed in orange boxes, while intermediate steps are showed in blue. The steps corresponding to the predictive and functional annotation are showed in light yellow. File formats are reported in parentheses when applicable. Main output files (those reported in the annotation) are showed in bold.

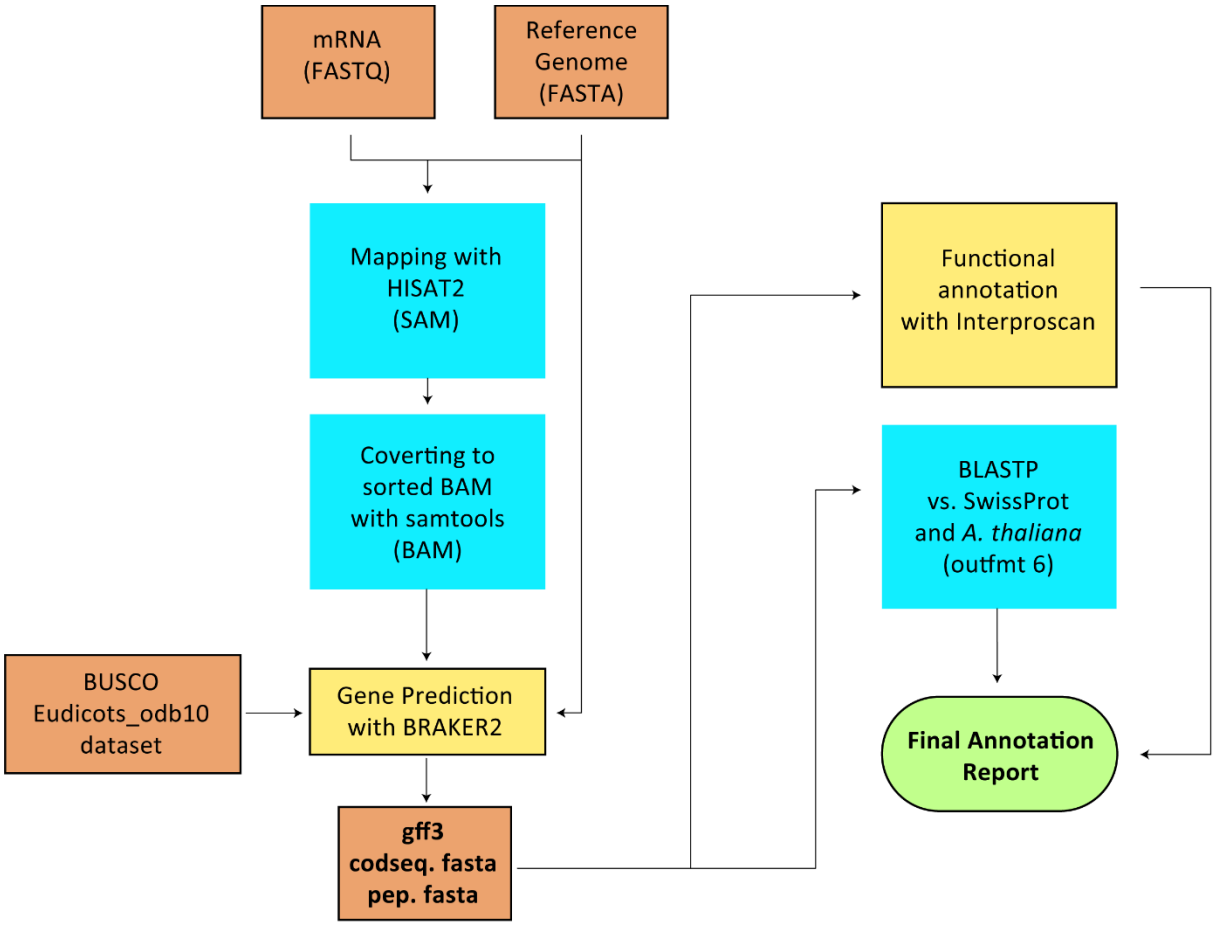
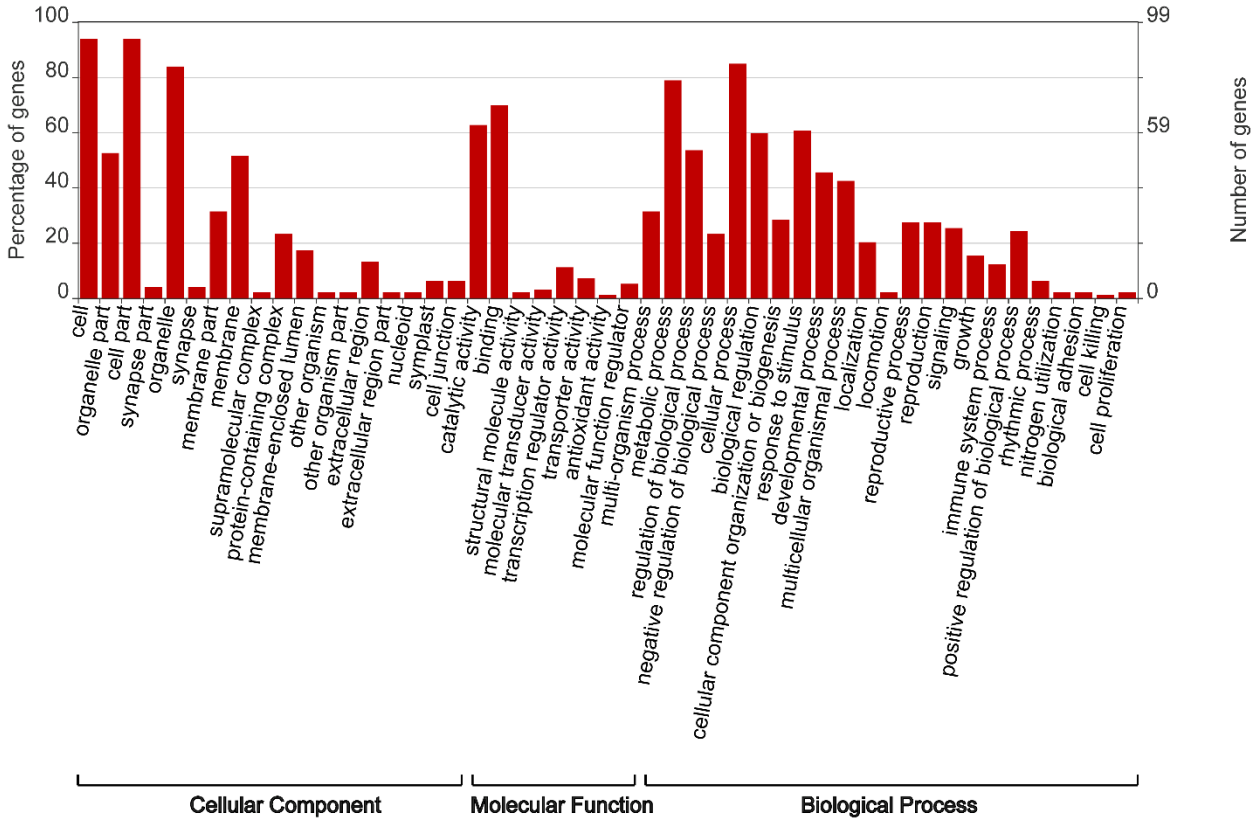


Fig. S5. WEGO plot of the GO terms associated with genes found to be divergent among *A. marina* populations of the Arabian Peninsula.



Appendix I

Scripts

Genome annotation

```
#!/usr/bin/env bash

## RNA-seq mapping using HISAT2

#TASK tags=Sample_Flower
hisat2 -x avicennia_marina_03Aug2019_ioBoG_CR_1kb --dta -p 24
-1 Sample_Flower_read1_trimmomatic_1PE.gz \
-2 Sample_Flower_read2_trimmomatic_2PE.gz \
-S Sample_Flower_hisat2.sam

#TASK tags=Sample_Leaves
hisat2 -x avicennia_marina_03Aug2019_ioBoG_CR_1kb --dta -p 24
-1 Sample_Leaves_read1_trimmomatic_1PE.gz \
-2 Sample_Leaves_read2_trimmomatic_2PE.gz \
-S Sample_Leaves_hisat2.sam

#TASK tags=Sample_Pooled
hisat2 -x avicennia_marina_03Aug2019_ioBoG_CR_1kb --dta -p 24
-1 Sample_Pooled_read1_trimmomatic_1PE.gz \
-2 Sample_Pooled_read2_trimmomatic_2PE.gz \
-S Sample_Pooled_hisat2.sam

#TASK tags=Sample_Roots
hisat2 -x avicennia_marina_03Aug2019_ioBoG_CR_1kb --dta -p 24
-1 Sample_Roots_read1_trimmomatic_1PE.gz \
-2 Sample_Roots_read2_trimmomatic_2PE.gz \
-S Sample_Roots_hisat2.sam


#TASK tags=Sample_Seeds
hisat2 -x avicennia_marina_03Aug2019_ioBoG_CR_1kb --dta -p 24
-1 Sample_Seeds_read1_trimmomatic_1PE.gz \
-2 Sample_Seeds_read2_trimmomatic_2PE.gz \
-S Sample_Seeds_hisat2.sam


#TASK tags=Sample_Stems
hisat2 -x avicennia_marina_03Aug2019_ioBoG_CR_1kb --dta -p 24
-1 Sample_Stems_read1_trimmomatic_1PE.gz \
-2 Sample_Stems_read2_trimmomatic_2PE.gz \
-S Sample_Stems_hisat2.sam


## Samtools

samtools view -Su Sample_Flower_hisat2.sam > Sample_Flower_hisat2.bam
samtools view -Su Sample_Leaves_hisat2.sam > Sample_Leaves_hisat2.bam
samtools view -Su Sample_Pooled_hisat2.sam > Sample_Pooled_hisat2.bam
samtools view -Su Sample_Roots_hisat2.sam > Sample_Roots_hisat2.bam
```

```
samtools view -Su Sample_Seeds_hisat2.sam > Sample_Seeds_hisat2.bam
samtools view -Su Sample_Stems_hisat2.sam > Sample_Stems_hisat2.bam

samtools sort -@ 24 -o Sample_Flower_hisat2.sorted.bam
Sample_Flower_hisat2.bam
samtools sort -@ 24 -o Sample_Leaves_hisat2.sorted.bam
Sample_Leaves_hisat2.bam
samtools sort -@ 24 -o Sample_Pooled_hisat2.sorted.bam
Sample_Pooled_hisat2.bam
samtools sort -@ 24 -o Sample_Roots_hisat2.sorted.bam Sample_Roots_hisat2.bam
samtools sort -@ 24 -o Sample_Seeds_hisat2.sorted.bam Sample_Seeds_hisat2.bam
samtools sort -@ 24 -o Sample_Stems_hisat2.sorted.bam Sample_Stems_hisat2.bam


## Gene annotation using BRAKER2

GENEMARK_PATH=~/programs/GeneMark/gmes_linux_64/

braker.pl --cores 24 --genome=Amar_genome_softmasked.fasta \
  --prot_seq=busco_eudicots_proteins_datasetodb10_amar.faa \
  --bam=Sample_Flower_hisat2.sorted.bam,Sample_Leaves_hisat2.sorted.bam,
Sample_Pooled_hisat2.sorted.bam,Sample_Roots_hisat2.sorted.bam,Sample_Seeds_h
isat2.sorted.bam,Sample_Stems_hisat2.sorted.bam  \
   --softmasking  --etpmode  \
   --ALIGNMENT_TOOL_PATH=~/programs/ProtHint/bin/   --gff3


## Functional annotation

# InterPro
/soft/interproscan-5.31-70.0/interproscan.sh -i augustus.hints.aa -t p -d
functional_annot -goterms -iprlookup

# Blastp Vs SwissProt and Arabidopsis thaliana annotation
blastp -query augustus.hints.aa -db
blast_db/uniprot_filtered_reviewed_yes.fasta -out Amar_Vs_sprot.outfmt6 -
outfmt '6 std qlen slen' -evalue 1e-3 -num_threads 12 -max_target_seqs 5
blastp -query augustus.hints.aa -db
blast_db/uniprot_arabidopsis_thaliana.fasta -out Amar_Vs_atha.outfmt6 -outfmt
'6 std qlen slen' -evalue 1e-3 -num_threads 12 -max_target_seqs 5

# Create annotation table using an in-house Perl script
perl functional_annot.pl augustus.hints.aa Amar_Vs_sprot.outfmt6
Amar_Vs_atha.outfmt6 augustus.hints.aa.tsv
blast_db/uniprot_filtered_reviewed_yes.tab
blast_db/uniprot_arabidopsis_thaliana.tab gene_ontology_ext_Sep2020.obo
```

SNP calling

```
#!/bin/bash

#-------------------------------------------------------------------
## TrimGalore
#-------------------------------------------------------------------

fq_files=`find . -name "*_1.fq.gz" | sort`
adapter_list=SampleAdpter_list.txt
out_trim=A_ReadsTrim

for files in $fq_files
do

    reads1=$(basename "$files")
    reads2=${reads1%_1.fq.gz}_2.fq.gz

    IFS='_'
    read -a strarr <<< "$reads1"
    IFS=' '

    name=${strarr[0]}
    adapt1=$(sed -n "s/${name}\t//p" $adapter_list)


    trim_galore --paired $reads1 $reads2 --output_dir $out_trim --stringency
1 --clip_R1 12 --clip_R2 12 --length 90 --no_report_file --adapter $adapt1 --
adapter2 AGATCGGAAGAGCGTCGTGTAGGGAAAGA

done


#-------------------------------------------------------------------
## MAPPING
#-------------------------------------------------------------------

# Input:
genome=Amar_genome.fasta

# Getting the complete path of all .fq files:
read_files=`find $out_trim -name "*_val_1.fq.gz"`

# Output:
out_sams=B1_MappedReads_SAM
out_bams=B2_MappedReads_BAM

# Build bwa index:
bwa index $genome

# Align reads against ref (Info for AddOrReplaceReadGroups id obtained from
samples names):
cd $out_trim

for reads in $read_files
do
```

```
325
326         val1=$(basename "$reads")
327         val2=${val1%_1_val_1.fq.gz}_2_val_2.fq.gz
328
329         name=${val1%_val_1.fq.gz}
330         IFS='_'
331         read -a strarr <<< "$name"
332         IFS=' '
333
334         rg_id=${strarr[2]}.${strarr[3]#L}
335         rg_pl=ILLUMINA
336         rg_pu=${strarr[2]}.${strarr[0]}.${strarr[3]#L}
337         rg_lb=${strarr[1]}
338         rg_sm=${strarr[0]}
339         sample=${rg_sm}_${strarr[2]}_${strarr[3]}
340
341         echo -e "\n1.STARTING BWA FOR $val1 and $val2\n"
342         time bwa mem -t 28 -M $genome $val1 $val2 > $out_sams/${sample}.sam
343
344         echo -e "\n2.STARTING BAM SORTING AND ADDING RGID INFO FOR: $sample\n"
345
346         time picard AddOrReplaceReadGroups \
347              VALIDATION_STRINGENCY=LENIENT \
348              INPUT=$out_sams/${sample}.sam \
349              OUTPUT=$out_bams/${sample}.bam \
350              RGID=$rg_id RGLB=$rg_lb RGPL=$rg_pl RGPU=$rg_pu RGSM=$rg_sm \
351              SORT_ORDER=coordinate \
352              CREATE_INDEX=true \
353              QUIET=false
354
355         echo -e "\n3.COMPUTING MAPPING STATS WITH SAMTOOLS FOR $sample\n"
356         samtools flagstat $out_bams/${sample}.bam
357
358         rm $out_sams/${sample}.sam
359
360
361     done
362
363     ## Mark duplicates:
364     bam_files=`find $out_bams -name "*.bam"`
365     out_bams2=B3_NoDups_BAM
366
367     for bam in $bam_files
368     do
369
370         sample=`basename ${bam%.bam}`
371         time picard MarkDuplicates \
372         VALIDATION_STRINGENCY=LENIENT \
373         INPUT=$bam \
374         OUTPUT=$out_bams2/${sample}_dedup.bam \
375         METRICS_FILE=$out_bams2/${sample}_metrics \
376         MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=900 \
377         ASSUME_SORTED=true
378
379         time picard BuildBamIndex \
380         VALIDATION_STRINGENCY=LENIENT \
381         INPUT=$out_bams2/${sample}_dedup.bam \
```

```
    CREATE_INDEX=true

done


#-----------------------------------------------------------------------
## Variant Calling
#-----------------------------------------------------------------------
gatk=gatk-package-4.1.8.1-local.jar

## Generate g.vcf files with GATK:
out_gvcfs=C_GVCFs
bam_files2=`find $out_bams2 -name "*.bam"`

for bam in $bam_files2
do

    filename=$(basename "$bam" _dedup.bam)
    java -Xmx32G -jar $gatk HaplotypeCaller -I $bams -R $genome --emit-ref-
confidence GVCF --native-pair-hmm-threads 8 -O $out_gvcfs/${filename}.g.vcf


done

## Combine g.vcf files (Two iterations, here shown the second one, after
merging g.vcf files by population):
cd $out_gvcfs

java -Xmx100G -jar $gatk CombineGVCFs -R $genome \
    --variant QRM.g.vcf \
    --variant RGB.g.vcf \
    --variant SND.g.vcf \
    --variant ESL.g.vcf \
    --variant FLM.g.vcf \
    --variant NRS.g.vcf \
    -O Mangroves.g.vcf

## Joint Call:
java -Xmx100G -jar $gatk GenotypeGVCFs \
-R $genome \
-V Mangroves.g.vcf \
-O Mangroves.vcf

## Retaining biallelic SNPs with vcftools from samples with less than 25% of
missing (txt file list):
vcftools --vcf Mangroves.vcf --remove-indels --min-alleles 2 --max-alleles 2
--keep sample_list025.txt --minDP 4 --maxDP 50 --minQ 40 --maf 0.0001 --hwe
0.00001 --recode --out Mangroves025_biall_dp450_q40_hwe000001

## Hard filtering, GATK Best Practices recommendations:
java -Xmx100G -jar $gatk VariantFiltration -R $genome -V
Mangroves025_biall_dp450_q40_hwe000001.recode.vcf \
--filter-name "Filter1_QD" \
--filter-expression "QD < 2.0" \
--filter-name "Filter2_FS" \
--filter-expression "FS > 60.0" \
--filter-name "Filter3_MQ" \
```

```
--filter-expression "MQ < 40.0" \
--filter-name "Filter4_MQRK" \
--filter-expression "MQRankSum < -12.5" \
--filter-name "Filter5_RPRK" \
--filter-expression "ReadPosRankSum < -8.0" \
--filter-name "Filter6_SOR" \
--filter-expression "SOR > 3.0" \
-O Mangroves025_biall_dp450_q40_hwe000001_HF.vcf
```

References cited in this Supplementary Information

Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics, 11.10. 11-11.10. 33.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., 2011. The variant call format and VCFtools. Bioinformatics 27, 2156-2158.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Anglais 43, 491-498.

Krijthe, J., van der Maaten, L., Krijthe, M.J., 2018. Package 'Rtsne'. GitHub.

Krueger, F., 2015. Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files.

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

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. science 326, 289-293.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 20, 1297-1303.

476 Murray, K.D., Borevitz, J.O., 2017. Axe: rapid, competitive sequence read demultiplexing using
477 a trie. bioRxiv, 160606.
478 Putnam, N.H., O'Connell, B.L., Stites, J.C., Rice, B.J., Blanchette, M., Calef, R., Troll, C.J.,
479 Fields, A., Hartley, P.D., Sugnet, C.W., 2016. Chromosome-scale shotgun assembly using an in
480 vitro method for long-range linkage. Genome research 26, 342-350.
481 Turner, S.D., 2016. qqman: an R package for visualizing GWAS results using QQ and manhattan
482 plots. bioRxiv. 2014. DOI 10, 005165.
483 Weir, B.S., Cockerham, C.C., 1984. Estimating F-statistics for the analysis of population
484 structure. Evolution. 38, 1358-1370.

485 Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan,
486 T., Shakir, K., Roazen, D., Thibault, J., 2013. From FastQ data to high-confidence variant calls:
487 the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics, 11.10.
488 11-11.10. 33.
489 Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E.,
490 Lunter, G., Marth, G.T., Sherry, S.T., 2011. The variant call format and VCFtools.
491 Bioinformatics 27, 2156-2158.
492 DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis,
493 A.A., Del Angel, G., Rivas, M.A., Hanna, M., 2011. A framework for variation discovery and
494 genotyping using next-generation DNA sequencing data. Anglais 43, 491-498.
495 Krijthe, J., van der Maaten, L., Krijthe, M.J., 2018. Package 'Rtsne'. GitHub.
496 Krueger, F., 2015. Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently
497 apply quality and adapter trimming to FastQ files.

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

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. science 326, 289-293.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 20, 1297-1303.

Murray, K.D., Borevitz, J.O., 2017. Axe: rapid, competitive sequence read demultiplexing using a trie. bioRxiv, 160606.

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., 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome research 26, 342-350.

Turner, S.D., 2016. qqman: an R package for visualizing GWAS results using QQ and manhattan plots. bioRxiv. 2014. DOI 10, 005165.

Weir, B.S., Cockerham, C.C., 1984. Estimating F-statistics for the analysis of population structure. Evolution. 38, 1358-1370.

Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics, 11.10. 11-11.10. 33.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., 2011. The variant call format and VCFtools. Bioinformatics 27, 2156-2158.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Anglais 43, 491-498.

Krueger, F., 2015. Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files.

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

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. science 326, 289-293.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 20, 1297-1303.

Murray, K.D., Borevitz, J.O., 2017. Axe: rapid, competitive sequence read demultiplexing using a trie. bioRxiv, 160606.

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., 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome research 26, 342-350.

Turner, S.D., 2016. qqman: an R package for visualizing GWAS results using QQ and manhattan plots. bioRxiv. 2014. DOI 10, 005165.

Weir, B.S., Cockerham, C.C., 1984. Estimating F-statistics for the analysis of population structure. Evolution. 38, 1358-1370.

Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics, 11.10. 11-11.10. 33.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., 2011. The variant call format and VCFtools. Bioinformatics 27, 2156-2158.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Anglais 43, 491-498.

Krueger, F., 2015. Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files.

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

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. science 326, 289-293.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., 2010. The Genome Analysis Toolkit: a MapReduce

framework for analyzing next-generation DNA sequencing data. Genome research 20, 1297-1303.

Murray, K.D., Borevitz, J.O., 2017. Axe: rapid, competitive sequence read demultiplexing using a trie. bioRxiv, 160606.

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., 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome research 26, 342-350.

Weir, B.S., Cockerham, C.C., 1984. Estimating F-statistics for the analysis of population structure. Evolution. 38, 1358-1370.

Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics, 11.10. 11-11.10. 33.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., 2011. The variant call format and VCFtools. Bioinformatics 27, 2156-2158.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Anglais 43, 491-498.

Krueger, F., 2015. Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files.

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

588 Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A.,
589 Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., 2009. Comprehensive mapping of long-range
590 interactions reveals folding principles of the human genome. science 326, 289-293.
591 McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella,
592 K., Altshuler, D., Gabriel, S., Daly, M., 2010. The Genome Analysis Toolkit: a MapReduce
593 framework for analyzing next-generation DNA sequencing data. Genome research 20, 1297-
594 1303.
595 Murray, K.D., Borevitz, J.O., 2017. Axe: rapid, competitive sequence read demultiplexing using
596 a trie. bioRxiv, 160606.
597 Putnam, N.H., O'Connell, B.L., Stites, J.C., Rice, B.J., Blanchette, M., Calef, R., Troll, C.J.,
598 Fields, A., Hartley, P.D., Sugnet, C.W., 2016. Chromosome-scale shotgun assembly using an in
599 vitro method for long-range linkage. Genome research 26, 342-350.

600 Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan,
601 T., Shakir, K., Roazen, D., Thibault, J., 2013. From FastQ data to high-confidence variant calls:
602 the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics, 11.10.
603 11-11.10. 33.
604 Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E.,
605 Lunter, G., Marth, G.T., Sherry, S.T., 2011. The variant call format and VCFtools.
606 Bioinformatics 27, 2156-2158.
607 DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis,
608 A.A., Del Angel, G., Rivas, M.A., Hanna, M., 2011. A framework for variation discovery and
609 genotyping using next-generation DNA sequencing data. Anglais 43, 491-498.

Krueger, F., 2015. Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files.

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

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. science 326, 289-293.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 20, 1297-1303.

Murray, K.D., Borevitz, J.O., 2017. Axe: rapid, competitive sequence read demultiplexing using a trie. bioRxiv, 160606.

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., 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome research 26, 342-350.

Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics, 11.10. 11-11.10. 33.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., 2011. The variant call format and VCFtools. Bioinformatics 27, 2156-2158.

633 DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis,
634 A.A., Del Angel, G., Rivas, M.A., Hanna, M., 2011. A framework for variation discovery and
635 genotyping using next-generation DNA sequencing data. Anglais 43, 491-498.
636 Kim, D., Langmead, B., Salzberg, S., 2015. hisat2. Nature methods.
637 Krueger, F., 2015. Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently
638 apply quality and adapter trimming to FastQ files.
639 Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler
640 transform. Bioinformatics 25, 1754-1760.
641 Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A.,
642 Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., 2009. Comprehensive mapping of long-range
643 interactions reveals folding principles of the human genome. science 326, 289-293.
644 McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella,
645 K., Altshuler, D., Gabriel, S., Daly, M., 2010. The Genome Analysis Toolkit: a MapReduce
646 framework for analyzing next-generation DNA sequencing data. Genome research 20, 1297-
647 1303.
648 Murray, K.D., Borevitz, J.O., 2017. Axe: rapid, competitive sequence read demultiplexing using
649 a trie. bioRxiv, 160606.
650 Putnam, N.H., O'Connell, B.L., Stites, J.C., Rice, B.J., Blanchette, M., Calef, R., Troll, C.J.,
651 Fields, A., Hartley, P.D., Sugnet, C.W., 2016. Chromosome-scale shotgun assembly using an in
652 vitro method for long-range linkage. Genome research 26, 342-350.

653 Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan,
654 T., Shakir, K., Roazen, D., Thibault, J., 2013. From FastQ data to high-confidence variant calls:

655 the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics, 11.10.
656 11-11.10. 33.
657 Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E.,
658 Lunter, G., Marth, G.T., Sherry, S.T., 2011. The variant call format and VCFtools.
659 Bioinformatics 27, 2156-2158.
660 DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis,
661 A.A., Del Angel, G., Rivas, M.A., Hanna, M., 2011. A framework for variation discovery and
662 genotyping using next-generation DNA sequencing data. Anglais 43, 491-498.
663 Krueger, F., 2015. Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently
664 apply quality and adapter trimming to FastQ files.
665 Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler
666 transform. Bioinformatics 25, 1754-1760.
667 Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A.,
668 Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., 2009. Comprehensive mapping of long-range
669 interactions reveals folding principles of the human genome. science 326, 289-293.
670 McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella,
671 K., Altshuler, D., Gabriel, S., Daly, M., 2010. The Genome Analysis Toolkit: a MapReduce
672 framework for analyzing next-generation DNA sequencing data. Genome research 20, 1297-
673 1303.
674 Murray, K.D., Borevitz, J.O., 2017. Axe: rapid, competitive sequence read demultiplexing using
675 a trie. bioRxiv, 160606.

676 Putnam, N.H., O'Connell, B.L., Stites, J.C., Rice, B.J., Blanchette, M., Calef, R., Troll, C.J.,
677 Fields, A., Hartley, P.D., Sugnet, C.W., 2016. Chromosome-scale shotgun assembly using an in
678 vitro method for long-range linkage. Genome research 26, 342-350.

679 Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan,
680 T., Shakir, K., Roazen, D., Thibault, J., 2013. From FastQ data to high-confidence variant calls:
681 the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics, 11.10.
682 11-11.10. 33.

683 Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E.,
684 Lunter, G., Marth, G.T., Sherry, S.T., 2011. The variant call format and VCFtools.
685 Bioinformatics 27, 2156-2158.

686 DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis,
687 A.A., Del Angel, G., Rivas, M.A., Hanna, M., 2011. A framework for variation discovery and
688 genotyping using next-generation DNA sequencing data. Anglais 43, 491-498.

689 Krueger, F., 2015. Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently
690 apply quality and adapter trimming to FastQ files.

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

693 Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A.,
694 Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., 2009. Comprehensive mapping of long-range
695 interactions reveals folding principles of the human genome. science 326, 289-293.

696 Luu, K., Bazin, E., Blum, M.G., 2017. pcadapt: an R package to perform genome scans for
697 selection based on principal component analysis. Molecular ecology resources 17, 67-77.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 20, 1297-1303.

Murray, K.D., Borevitz, J.O., 2017. Axe: rapid, competitive sequence read demultiplexing using a trie. bioRxiv, 160606.

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., 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome research 26, 342-350.

Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics, 11.10. 11-11.10. 33.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., 2011. The variant call format and VCFtools. Bioinformatics 27, 2156-2158.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Anglais 43, 491-498.

Krueger, F., 2015. Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files.

719 Li, H., 2011. A statistical framework for SNP calling, mutation discovery, association mapping
720 and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987-
721 2993.
722 Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler
723 transform. Bioinformatics 25, 1754-1760.
724 Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A.,
725 Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., 2009. Comprehensive mapping of long-range
726 interactions reveals folding principles of the human genome. science 326, 289-293.
727 Luu, K., Bazin, E., Blum, M.G., 2017. pcadapt: an R package to perform genome scans for
728 selection based on principal component analysis. Molecular ecology resources 17, 67-77.
729 McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella,
730 K., Altshuler, D., Gabriel, S., Daly, M., 2010. The Genome Analysis Toolkit: a MapReduce
731 framework for analyzing next-generation DNA sequencing data. Genome research 20, 1297-
732 1303.
733 Murray, K.D., Borevitz, J.O., 2017. Axe: rapid, competitive sequence read demultiplexing using
734 a trie. bioRxiv, 160606.
735 Putnam, N.H., O'Connell, B.L., Stites, J.C., Rice, B.J., Blanchette, M., Calef, R., Troll, C.J.,
736 Fields, A., Hartley, P.D., Sugnet, C.W., 2016. Chromosome-scale shotgun assembly using an in
737 vitro method for long-range linkage. Genome research 26, 342-350.

738 Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan,
739 T., Shakir, K., Roazen, D., Thibault, J., 2013. From FastQ data to high-confidence variant calls:
740 the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics, 11.10.
741 11-11.10. 33.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., 2011. The variant call format and VCFtools. Bioinformatics 27, 2156-2158.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Anglais 43, 491-498.

Krueger, F., 2015. Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files.

Li, H., 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987-2993.

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

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. science 326, 289-293.

Luu, K., Bazin, E., Blum, M.G., 2017. pcadapt: an R package to perform genome scans for selection based on principal component analysis. Molecular ecology resources 17, 67-77.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 20, 1297-1303.

764 Murray, K.D., Borevitz, J.O., 2017. Axe: rapid, competitive sequence read demultiplexing using
765 a trie. bioRxiv, 160606.
766 Putnam, N.H., O'Connell, B.L., Stites, J.C., Rice, B.J., Blanchette, M., Calef, R., Troll, C.J.,
767 Fields, A., Hartley, P.D., Sugnet, C.W., 2016. Chromosome-scale shotgun assembly using an in
768 vitro method for long-range linkage. Genome research 26, 342-350.

769 Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E.,
770 Lunter, G., Marth, G.T., Sherry, S.T., 2011. The variant call format and VCFtools.
771 Bioinformatics 27, 2156-2158.
772 Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler
773 transform. Bioinformatics 25, 1754-1760.
774 Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A.,
775 Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., 2009. Comprehensive mapping of long-range
776 interactions reveals folding principles of the human genome. science 326, 289-293.
777 McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella,
778 K., Altshuler, D., Gabriel, S., Daly, M., 2010. The Genome Analysis Toolkit: a MapReduce
779 framework for analyzing next-generation DNA sequencing data. Genome research 20, 1297-
780 1303.
781 Putnam, N.H., O'Connell, B.L., Stites, J.C., Rice, B.J., Blanchette, M., Calef, R., Troll, C.J.,
782 Fields, A., Hartley, P.D., Sugnet, C.W., 2016. Chromosome-scale shotgun assembly using an in
783 vitro method for long-range linkage. Genome research 26, 342-350.

784 Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E.,
785 Lunter, G., Marth, G.T., Sherry, S.T., 2011. The variant call format and VCFtools.
786 Bioinformatics 27, 2156-2158.

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

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 20, 1297-1303.

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., 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome research 26, 342-350.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., 2011. The variant call format and VCFtools. Bioinformatics 27, 2156-2158.

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

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 20, 1297-1303.

Andrews, S., 2010 FastQC: a quality control tool for high throughput sequence data.

Auwera, G.A., M.O. Carneiro, C. Hartl, R. Poplin, G. del Angel *et al.*, 2013 From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*:11.10. 11-11.10. 33.

Basher, Z., D.A. Bowden, and M.J. Costello, 2018 GMED: Global Marine Environment Datasets for environment visualisation and species distribution modelling. *Earth Syst. Sci. Data Discuss* Accessed at http://gmed.auckland.ac.nz on Oct. 3rd 2019.

Danecek, P., A. Auton, G. Abecasis, C.A. Albers, E. Banks *et al.*, 2011 The variant call format and VCFtools. *Bioinformatics* 27 (15):2156-2158.

DePristo, M.A., E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 43 (5):491-498.

Flynn, J.M., R. Hubley, C. Goubert, J. Rosen, A.G. Clark *et al.*, 2019 RepeatModeler2: automated genomic discovery of transposable element families. *bioRxiv*:856591.

Krijthe, J., L. van der Maaten, and M.J. Krijthe, 2018 Package 'Rtsne'. GitHub.

Krueger, F., 2015 Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files.

Krzywinski, M., J. Schein, I. Birol, J. Connors, R. Gascoyne *et al.*, 2009 Circos: an information aesthetic for comparative genomics. *Genome Research* 19 (9):1639-1645.

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25 (14):1754-1760.

Lieberman-Aiden, E., N.L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy *et al.*, 2009 Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326 (5950):289-293.

Lyu, H., Z. He, C.I. Wu, and S. Shi, 2018 Convergent adaptive evolution in marginal environments: unloading transposable elements as a common strategy among mangrove genomes. *New phytologist* 217 (1):428-438.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20 (9):1297-1303.

Murray, K.D., and J.O. Borevitz, 2017 Axe: rapid, competitive sequence read demultiplexing using a trie. *bioRxiv*:160606.

Putnam, N.H., B.L. O'Connell, J.C. Stites, B.J. Rice, M. Blanchette *et al.*, 2016 Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Research* 26 (3):342-350.

Smit, A., R. Hubley, and P. Green, 2015 RepeatMasker Open-4.0. 2013–2015.

Tatusova, T.A., and T.L. Madden, 1999 BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS microbiology letters* 174 (2):247-250.

Turner, S.D., 2016 qqman: an R package for visualizing GWAS results using QQ and manhattan plots. bioRxiv. 2014. *DOI* 10 (1101):005165.

Weir, B.S., and C.C. Cockerham, 1984 Estimating F-statistics for the analysis of population structure. *Evolution.* 38:1358-1370.

Xu, S., Z. He, Z. Guo, Z. Zhang, G.J. Wyckoff *et al.*, 2017 Genome-wide convergence during evolution of mangroves from woody plants. *Molecular Biology and Evolution* 34 (4):1008-1015.

Ye, J., L. Fang, H. Zheng, Y. Zhang, J. Chen *et al.*, 2006 WEGO: a web tool for plotting GO annotations. *Nucleic Acids Research* 34 (suppl_2):W293-W297.

854