SUPPLEMENTARY METHODS

Detailed ddRAD lab protocol

We used a slightly modified version of the ddRADseq protocol as described by Peterson et al. [1]. In brief, we digested 500ng of genomic DNA using 10 U of PstI and 10 U of EcoRI restriction enzymes for 3 hours in 37° C. The reaction was completed without an enzyme deactivation step. The samples were then cleaned using 1.8 X SeraMag SPRI beads and eluted in 30ul TET buffer. We ligated adapters with inline barcodes flanked by the corresponding restriction site using T4 DNA Ligase. The ligation reaction was incubated in 23°C for 30 minutes, heat deactivated at 65° C for 10 minutes, and cooled at a rate of 2°C per 90 seconds until it reaches 23°C. Ten individuals with different inline barcodes were pooled and cleaned using 1.5X SeraMag SPRI beads and eluted in 20ul of PCR-grade water. We used the LabChip XT (PerkinElmer, LOC) to size-select 345 - 407bp. We used 3ul of the size-selected sample for two replicate PCR reactions using IS4 and P7 barcoded primers from Meyer and Kircher [2]. The samples were amplified using Phusion Hot Start High-Fidelity Taq. Enriched libraries were cleaned using 1.2X of SeraMag SPRI beads and parallel PCR reactions were pooled together. Final samples were quantified using the Qubit Flourimeter prior to pooling equimolarly. Approximately 10 x 10-individual pools were sequenced on a NextSeq500 for 150bp, single-read, mid-output and the rest were sequenced using a NextSeq500, 150bp, SR, high-output.

Data processing and analyses

Reference RAD locus set and mapping

In order to generate a reference set of RAD loci to map back to, we carried out the entire *pyrad* pipeline[3]. We clustered on 88% similarity and filtered for loci that did not have a minimum coverage of 2x for n - 5 individuals for each species. Lastly, we included two individuals of a different species to serve as an outgroup for polarizing the site frequency spectrum. Only loci with an outgroup are retained in the reference. We used a python script to extract a separate ingroup and outgroup reference set from the pyrad .loci file output.

We used a customized script and Trimmomatic (v.0.32) to filter our single-end, raw reads for mapping [4]. First we got rid of low complexity reads defined by being having at least 50% of the read be of a single base. We then carried out the following Trimmomatic call to trim and filter the reads even further. We cropped the first few bases of each read to get rid of adapter sequences and restriction cut sites. We used Bowtie2 (v2.2.2) to align the reads back onto the ingroup

reference RAD locus set [5]. The alignments were sorted and indexed using samtools and the resulting bam files were used for further ANGSD and ngsTools analyses [6-8]. We plotted the distribution of per locus average coverage to decide an upper limit for coverage. Approximately 95% of the reads fell under 40x average coverage.

Command:

```
java -jar trimmomatic SE -phred33 [SAMPLE].A.fq [SAMPLE].final.fq
SLIDINGWINDOW:4:15 MINLEN:36 HEADCROP:10
```

ANGSD and ngsTools data processing and analyses

SNP filtering

We filtered the reference data set further using the following command to get rid of other spurious loci. The resulting locus numbers can be found in Table S1.

Command:

```
angsd -P 2 -b bamlist.txt -ref ref.fa -out outdir/ -uniqueOnly 1 -remove_bads 1 -trim 0 -C
50 -baq 1 -minMapQ 20 -minQ 20 -minInd [Ntotal - 5] -setMinDepthInd 2 -
setMaxDepthInd 40 -doCounts 1 -GL 1 -doMajorMinor 1 -doMaf 1 -skipTriallelic 1 -
SNP_pval 1e-3 -deGeno 8 -doPost 1
```

Population structure

From the SNPs we recovered from the filtering we randomly selected a single SNP per locus to generate the distance matrix used for the principal coordinates analysis and population network graph. We reran ANGSD to get the genotype likelihoods only for those SNPs. We used ngsDist to generate a distance matrix to be used for the plots [9]. SplitsTree was used to generate the population network and the function cmdscale from base R (v 3.2.2) was used to generate the PCoA (Figure S1) [10].

Commands:

```
angsd -P 2 -b bamlist.txt -ref ref.fa -out outdir/ -uniqueOnly 1 -remove_bads 1 -trim 0 -C
50 -baq 1 -minMapQ 20 -minQ 20 -minInd [Ntotal - 5] -setMinDepthInd 2 -
setMaxDepthInd 40 -doCounts 1 -GL 1 -doMajorMinor 1 -doMaf 1 -skipTriallalic 1 -
SNP_pval 1e-3 -doGeno 8 -doPost 1 -sites ALL.sites.txt -rf ALL.chrs
```

```
ngsDist -verbose 1 -geno ALL.geno.gz -probs -n_ind [20] -n_sites [20000] -labels pops.label -o ALL.dist -n_threads 4
```

Site frequency spectrum (SFS)

Two different SFS were generated, one to use as summary statistics for the demographic modeling and the other to estimate per locus F_{ST}. The demographic modeling assumes unlinked sites so only a single SNP per locus was used to generate the SFS. Additionally, intersection of the same SNPs were necessary between the 4 populations we were comparing. On the other hand, per locus F_{ST} needs the SFS containing information for all SNP variants. An initial run of ANGSD was run per population to get information regarding which RAD loci were recovered for all 4 populations. We then created a sites file containing all the SNPs and one SNP/locus from the shared loci. We reran ANGSD separately for these sites files. Finally, we used realSFS to extract 2DSFS for each population pair and obtain summary statistics. The SFS was used directory for the demographic modeling and an F_{ST} per locus file was generated to be further processed by a personal R script.

Commands:

Initial run

```
for POP in CYP QLD NT PNG; doangsd -P 2 -b $POP.bamlist -ref [ingroup_ref.fa] -fold 0 -anc [outgroup_ref.fa] -out out/$POP -uniqueOnly 1 -remove_bads 1 -trim 1 -C 50 -baq 1 -minMapQ 20 -minQ 20 -minInd 3 -setMinDepthInd 2 -setMaxDepthInd 40 -doCounts 1 -GL 1 -doSaf 1 -sites ALL.sites -rf ALL.chrs; done
```

Generate all SNP intersect

```
realSFS print CYP.saf.idx PNG.saf.idx NT.saf.idx QLD.saf.idx | cut -f 1-2 > intersect.txt
```

Second round - same command but for different SNP sets

```
cut -f1 intersect.txt | uniq > intersect.chrs
```

```
for POP in CYP QLD NT PNG; do angsd -P 2 -b $POP.bamlist -ref [ingroup_ref.fa] -fold 0 -anc [outgroup_ref.fa] -out out/$POP -uniqueOnly 1 -remove_bads 1 -trim 1 -C 50 -baq 1 -minMapQ 20 -minQ 20 -minInd 3 -setMinDepthInd 2 -setMaxDepthInd 40 -doCounts 1 -GL 1 -doSaf 1 -sites intersect.txt -rf intersect.chrs; done
```

2DSFS estimate - same command for both

```
realSFS -P 8 CYP.saf.idx PNG.saf.idx > CYP.PNG.sfs
```

Generate F_{ST} index

```
realSFS fst index CYP.saf.idx PNG.saf.idx -sfs CYP.PNG.sfs -fstout CYP.PNG
```

Generate F_{ST} to be processed by R script

```
realSFS fst print CYP.PNG.fst.idx > CYP.PNG.fst
```

Absolute DNA divergence - D_{XY}

The same SNP set that was used for the F_{ST} estimate was also used for the D_{XY} estimate. We used ANGSD to generate allele frequencies for all alleles in each SNP to calculate D_{XY}. We used the .mafs.gz output into an R script with the F_{ST} file to generate per locus and global estimates of divergence.

Commands:

```
for POP in CYP NT PNG QLD; do angsd -P 2 -b $POP.bamlist -ref [ingroup_ref.fa] -anc [outgroup_ref.fa] -out out/$POP -uniqueOnly 1 -remove_bads 1 -trim 0 -C 50 -baq 1 -minMapQ 20 -minQ 20 -minInd 3 -setMinDepthInd 2 -setMaxDepthInd 40 -doCounts 1 -GL 1 -doMajorMinor 5 -doMaf 1 -skipTriallelic 1 -sites intersect.txt -rf intersect.chrs -dumpCounts 2; done
```

Population divergence - D_A

To estimate Da we needed to estimate population nucleotide diversity first and subtract the average from D_{XY}. To do this we used ANGSD for estimation of θ and π. We need the SFS per population to estimate θ. We used the same files that obtained for the F_{ST} estimates to estimate an SFS per population and used that as input for the ANGSD run. We used an R script to estimate the per population diversity (Table S2).

Commands:

```
for POP in CYP PNG QLD NT; do realSFS -P 8 CYP.saf.idx > CYP.sfs
```

```
for POP in CYP PNG QLD NT; do angsd -P 2 -b $POP.bamlist -ref [ingroup_ref.fa] -anc [outgroup_ref.fa] -out out/$POP -uniqueOnly 1 -remove_bads 1 -trim 1 -C 50 -baq 1 -minMapQ 20 -minQ 20 -minInd 3 -setMinDepthInd 2 -setMaxDepthInd 40 -doCounts 1 -GL 1 -doSaf 1 -doThetas 1 -pest out/$POP.sfs -sites intersect.txt -rf intersect.chrs; done
```

Demographic modeling

We followed similar guidelines to Roux et al. to simulate our demographic models [11]. Unlike Roux et al., who used various population genetic statistics as their summary statistics, we used the ANGSD output for the SFS as the summary statistic [11]. Using Hudson's ms we simulated a single variant per locus to simulate the ddRAD output and calculated the 2DSFS [12]. We simulated six different scenarios: isolation-with-migration (IM), IM with heterogeneous effective population size per locus (hetN), IM with heterogeneous migration rate per locus (hetM), IM+hetNhetM, strict isolation (SI) and SI+hetN. The heterogeneity in effective population size and migration rate is to account for the variation in nucleotide diversity and migration across the genome. Similar to the Roux et al. simulations, we used a beta binomial distribution to sample heterogeneity in N and m across the loci [11]. Initially, we sample the two separate beta shape parameters from a uniform distribution. We then assign a random proportion of the loci a heterogeneous N or m and the rest of the loci with homogeneous N or m (sampled from a uniform distribution). The each locus assigned heterogeneous receives a different value calculated from the homogeneous N or m parameter multiplied by a value from a beta distribution with the shape parameters sampled previously. Heterogeneous parameters were written out in a tbs file to be taken by the ms calls. All other priors were sampled from uniform distributions. The following ms calls were run under the various scenarios and the output was converted to a 2DSFS. We ran 300k simulations under each model for each sample number combination (ie. pop1 = 5 samples x pop2 = 4 samples).

Strict isolation (SI)

```
ms [total_samples] [n_loci] -s 1 -I 2 [samples_pop1] [samples_pop2] 0 -m 1 2 0 -m 2 1 0 -n 1 [Ne_pop1_prior] -n 2 [Ne_pop2_prior] -ej [Tsplit_prior] 2 1 -eN [Tsplit_prior] [Ne_anc_prior] > ms_SI.txt
```

Strict isolation with heterogeneous N_e (SIhetN)

```
ms [total_samples] [n_loci] -s 1 -I 2 [samples_pop1] [samples_pop2] 0 -m 1 2 0 -m 2 1 0 -n 1 tbs -n 2 tbs -ej [Tsplit_prior] 2 1 -eN [Tsplit_prior] tbs > ms_SIhetN.txt
```

Isolation-with-migration (IM)

ms [total_samples] [n_loci] -s 1 -I 2 [samples_pop1] [samples_pop2] 0 -m 1 2 [m1_prior] -m 2 1 [m2_prior] -n 1 [Ne_pop1_prior] -n 2 [Ne_pop2_prior] -ej [Tsplit_prior] 2 1 -eN [Tsplit_prior] [Ne_anc_prior] > ms_IM.txt

Isolation-with-migration with heterogeneous N_e (IMhetN)

ms [total_samples] [n_loci] -s 1 -I 2 [samples_pop1] [samples_pop2] 0 -m 1 2 [m1_prior] -m 2 1 [m2_prior] -n 1 tbs -n 2 tbs -ej [Tsplit_prior] 2 1 -eN [Tsplit_prior] tbs > ms_IMhetN.txt

Isolation-with-migration with heterogeneous migration (IMhetM)

ms [total_samples] [n_loci] -s 1 -I 2 [samples_pop1] [samples_pop2] 0 -m 1 2 tbs -m 2 1 tbs -n 1 [Ne_pop1_prior] -n 2 [Ne_pop2_prior] -ej [Tsplit_prior] 2 1 -eN [Tsplit_prior] tbs > ms_IMhetN.txt

Isolation-with-migration with heterogeneous N_e and migration (IMhetNhetM)

ms [total_samples] [n_loci] -s 1 -I 2 [samples_pop1] [samples_pop2] 0 -m 1 2 tbs -m 2 1 tbs -n 1 tbs -n 2 tbs -ej [Tsplit_prior] 2 1 -eN [Tsplit_prior] tbs > ms_IMhetNhetM.txt

Prior to the ABC run, the first and last bins were removed. These bins correspond to no variation either fixed for the ancestral or alternate allele and these bins are not output by ms. The rest of the bins were also converted to ratios to account for variation in locus numbers of the different systems. We calculated the support values for each model from the 1500 closest SFS simulations using a neural network. The ABC probabilities for each population pair for each model can be found in Table S3.

Speciation model support

We fit our data under various proposed trajectories of parapatric speciation. These trajectories all share a parameter, z_c, where there is no gene flow between populations. Under the threshold model, all populations below z_c have high migration and all above have no migration. Under the constant growth model, there is a steady increase in incompatibilities until z_c, when speciation has completed. Under the accelerated model incompatibilities start accumulating

slowly but rapidly increases up to z_c. In the decelerating model, incompatibilities start accumulating rapidly but slows down as it approaches z_c. Lastly, in the sigmoid model incompatibilities start to accumulate slowly, then rapidly transitions to a tipping point, then slows down again when nearing completion of speciation. Mathematical descriptions of each model can be found in Table S4.

We simulated our observed Fst values under the different speciation trajectories and assigned a uniform prior for the unknown parameters (a, z_c, z_o). The range of the prior for a varied depending on whether the model is constant ($a = 1$), decelerating ($0 < a < 1$), or accelerating ($1 < a < 10$). The point at which speciation is completed, z_c, was bound between $0.5 < z_c < 1$ as we know from our data that most population pairs still have some, albeit low, support for gene flow at Fst ~ 0.5 therefore z_c cannot be below 0.5.

Parameter estimation of z_o and b for the sigmoidal model yielded a mean of 0.33 and 4.91, respectively. Posterior distribution of z_c is identical to the prior distribution as we do not have enough data to define the divergence at which there is no gene flow in any populations. The estimate of z_c is, therefore, not informative.

Test of robustness of our ABC models yielded good support under each scenario. Of the simulated data sets per model 92.28% of the accelerated model, 99.78% of the constant model, 97.20% of the decelerate model, 81.84% of the sigmoid model, and 100% of the threshold model were assigned correctly. Of the remaining simulations under the sigmoid model that was not assigned correctly 13.39% were assigned to accelerate and 3.71% were assigned to decelerate. Mean model posterior probabilities can be found in Table S9.

References

1. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PloS one. 2012 May 31;7(5):e37135.

2. Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protocols. 2010 Jun 1;2010(6):pdb-rot5448.

3. Eaton DA. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics. 2014 Mar 5;30(13):1844-9.

4. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Apr 1;30(15):2114-20.

5. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012 Apr 1;9(4):357-9.

6. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9.

7. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. BMC bioinformatics. 2014 Nov 25;15(1):356.

8. Fumagalli M, Vieira FG, Linderoth T, Nielsen R. ngsTools: methods for population genetics analyses from next-generation sequencing data. Bioinformatics. 2014 Jan 23;30(10):1486-7.

9. Vieira FG, Lassalle F, Korneliussen TS, Fumagalli M. Improving the estimation of genetic distances from Next-Generation Sequencing data. Biological journal of the Linnean Society. 2015 Mar 30;117(1):139-49.

10. Huson DH. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics (Oxford, England). 1998 Jan 1;14(1):68-73.

11. Roux C, Fraisse C, Romiguier J, Anciaux Y, Galtier N, Bierne N. Shedding light on the grey zone of speciation along a continuum of genomic divergence. PLoS biology. 2016 Dec 27;14(12):e2000234.

12. Hudson RR. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics. 2002 Feb 1;18(2):337-8.

13. Hijmans RJ, Elith J. Species distribution modeling with R. R package version 0.8-11. 2013 Dec 13.