Evolutionary processes and cellular functions underlying divergence in Alexandrium minutum

Mickael Le Gac, Gabriel Metegnier, Nicolas Chomérat, Pascale Malestroit, Julien Quéré, Olivier Bouchez, Raffaele Siano, Christophe Destombe, Laure Guillou, Annie Chapelle

Original Article

Evolutionary processes and cellular functions underlying divergence in *Alexandrium minutum*.

Mickael le Gac[1*], Gabriel Metegnier[1,4], Nicolas Chomérat[2], Pascale Malestroit[1], Julien Quéré[1], Olivier Bouchez[3], Raffaele Siano[1], Christophe Destombe[4], Laure Guillou[5], Annie Chapelle[1]

[1]IFREMER, DYNECO/Pelagos, 29280 Plouzané, France

[2]IFREMER, Station de Biologie Marine, 29900 Concarneau, France

[3]GeT PlaGe, Genotoul, INRA Auzeville, Castanet Tolosan, France

[4]Sorbonne Universités, Université Pierre et Marie Curie - Paris 6, CNRS, PUCCh, UACH, UMI 3614, Evolutionary Biology and Ecology of Algae, Station Biologique de Roscoff, CS 90074, Place Georges Teissier, CS90074, 29688 Roscoff cedex, France

[5]Sorbonne Universités, Université Pierre et Marie Curie - Paris 6, CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, CS90074, 29688 Roscoff cedex, France

Keywords : Speciation, Pseudocryptic species, Dinoflagellates, Harmful Algal Blooms, Populations Genomics

*Corresponding author: Mickael Le Gac, Mickael.Le.Gac@ifremer.fr, Phone: (33)-298224358, Fax: (33)-298224548.

Running Title: A. minutum divergence

Abstract

Understanding divergence in the highly dispersive and seemingly homogeneous pelagic environment for organisms living as free drifters in the water column remains a challenge. Here, we analyzed the transcriptome wide mRNA sequences, as well as the morphology of 18 strains of *Alexandrium minutum*, a dinoflagellate responsible for Harmful Algal Blooms worldwide, to investigate the functional bases of a divergence event. Analysis of the joint site frequency spectrum (JSFS) pointed toward an ancestral divergence in complete isolations followed by a secondary contact resulting in gene flow between the two diverging groups, but heterogeneous across sites. The sites displaying fixed SNPs were associated with a highly restricted gene flow and a strong over-representation of non-synonymous polymorphism, suggesting the importance of selective pressures as drivers of the divergence. The most divergent transcripts were homologs to genes involved in calcium/potassium fluxes across the membrane, calcium transduction signal and saxitoxin production. The implication of these results in terms of ecological divergence and build-up of reproductive isolation are discussed. Dinoflagellates are especially difficult to study in the field at the ecological level due to their small size and the dynamic nature of their natural environment, but also at the genomic level due to their huge and complex genome and the absence of closely related model organism. This study illustrates the possibility to identify traits of primary importance in ecology and evolution starting from high throughput sequencing data, even for such organisms.

Introduction

The high number of unicellular eukaryote species coexisting in an apparently homogeneous pelagic environment has long puzzled ecologists (the paradox of the plankton, Hutchinson 1961). At the ecological scale the paradox may be resolved, at least partly, by invoking out of equilibrium dynamics (Roy and Chattopadhyay 2007). However, at the evolutionary scale the paradox remains extremely puzzling. In the marine environment numerous species have a pelagic stage often associated with long range dispersal creating high gene flow, opposing local adaptation and the speciation process (Palumbi 1992). For plants and animals with a benthic phase or animals able to swim against currents to remain in specific habitats, adaptive divergence for specific environmental conditions seems nevertheless possible (Bierne et al. 2003). For organisms remaining as free drifters in the water column, such as phytoplankton, the forces that may drive such divergence are virtually unknown. Theoretical works taking into account the huge population sizes, specific life history traits such as the ability to form resting cysts embedded in the sediment, and the dependency on hydrodynamics not only as a dispersive force of propagules but also as a force potentially impeding the organisms to remain in favorable environmental conditions during active growth are extremely scarce (but see Shoresh et al. 2008). Empirically speaking, a growing number of population genetic studies have highlighted that phytoplankton species may be spatially and temporally structured (Rynearson et al. 2004; Iglesias-Rodríguez et al. 2006; Masseret et al. 2009; Castelyn et al. 2010; Casabianca et al. 2012; Dia et al. 2014). Moreover, some works investigating ecological divergence between closely related species have highlighted vertical niche partitioning in foraminifer (Weiner et al. 2012), specialization for different light intensities and utilization of different parts of the light spectrum in cyanobacteria (Rocap et al. 2003; Stomp et al. 2007), as well as divergence in term of metal usage in chlorophytes (Palenik et al. 2007) and diatoms (Peers et al. 2006).

Dinoflagellates constitute an enigmatic group of mainly marine unicellular eukaryotes with lifestyles ranging from mixotrophic (autotrophic and predator) to fully heterotrophic for half of the species, sometimes producing toxins that have ecological, economic and sanitary impacts (Anderson et al. 2012a), and displaying many original genomic characteristics, including genome sizes among the largest of any organisms (up to 60 times the size of the human genome, Wisecaver and Hackett 2011).

The species belonging to the genus *Alexandrium* (Anderson et al. 2012b) are responsible for paralytic shellfish poisoning caused by the production of several toxins including saxitoxin (Cusick and Sayler 2013), a molecule classified as schedule 1 substance, in the sense of the Chemical Weapons Convention due to its very low lethal dose.

Thanks to recent development in sequencing technologies and bioinformatics tools, it is now becoming possible to investigate the genome-wide patterns of divergence (Seehausen et al. 2014). These developments are not only transforming our understanding of divergence from an individual gene to a whole genome perspective (Feder et al. 2013), but also enabling the investigation of genomic divergence in a wide variety of organisms spanning the entire tree of life, including organisms that are not closely related to any model organism (Ellegren 2014). So called reverse ecology approaches where genomic data is the starting point to identify traits of ecological and evolutionary interest (Li et al. 2008) are especially appealing for organisms that are difficult to study in the field, such as plankton species, to gain insight on the evolutionary processes at play during divergence and the affected cellular functions.

Here by sequencing and analyzing the mRNA sequences, as well as characterizing the morphology of 18 strains of *A. minutum* isolated from natural populations we highlight a divergence event. We investigated: 1. The model of divergence most likely to explain the observed joint site frequency spectrum among seven models of divergence, 2: Whether this event is driven by selective pressures, and 3: What are the underlying divergent cellular functions.

Material and Methods

RNA extraction, library preparation and sequencing

Starting from environmental samples, each *A. minutum* strain was founded by micropipetting a single cell into fresh medium under inverted microscope. Following isolation, the strains are maintained in the lab by biweekly dilution into fresh media. Under culture conditions, cells are haploid and divide by mitosis, each strain is thus composed of clonal individuals. A total of 18 strains isolated from various localities and time (Fig. 1) were grown to mid exponential phase in 100 ml of K medium at 18°C, 12/12 photoperiod, and 80 $\mu E.s^{-1}.m^{-2}$ of irradiance. Cell densities ranged from 5.10^{6} to $2.5.10^{7}$ $cell.l^{-1}$.

Cultures were centrifuged at 4,500 g for 8 min, sonicated on ice during 20 sec in RLT lysis buffer (Qiagen) containing β-mercaptoethanol. Extraction was performed using RNeasy plus mini kit (Qiagen) following the manufacturer protocol. Extracted RNA was quantified using a Biotek Epoch spectrophotometer and the quality estimated on RNA 6000 nano chips using a Bioanalyzer (Agilent). Reverse transcription of 4 µg of total RNA into cDNA and library preparation were performed at the GeT-PlaGe France Genomics sequencing platform (Toulouse, France) using the Illumina truseq RNA V2 kits. One library was generated per *A. minutum* strain. Library quality was assessed on a Bioanalyzer using high sensitivity DNA analysis chips and quantified using Kappa Library Quantification Kit. Paired-end sequencing was performed using 2 x 100bp cyles. The 18 libraries were sequenced on two Illumina Hiseq lanes.

Reads quality assessment and filtering

Galaxy interface (Giardine et al. 2005) was used to visualize sequencing outputs and filter out low quality reads. Visualization was performed using FastQC. Reads were truncated until the last nucleotide displayed a Phred score of at least 25. Reads shorter than 70 bp or with an average Phred score lower than 25 were removed. Cutadapt was used to remove sequences corresponding to the TruSeq indexed adapter, TruSeq Universal Adapter, dinoflagellate Spliced Leader (Zhang et al. 2007), as well as poly-A tails. For the 18 *A. minutum* strains sequenced, more than 68.10^9 bases were generated of which about 4.10^9 (~ 6%) were discarded after quality filtering.

Obtaining *A. minutum* reference transcriptome

After initial quality filtering, overlapping paired-end reads were merged using Flash (Magoc and Salzberg 2011). Sequences shorter than 70bp were removed. Merged paired-end reads, as well as non-overlapping paired-end and orphan reads from the 18 strains were used to perform a de novo assembly of *A. minutum* transcriptome using Trinity (Haas et al. 2013) after pooling the reads of the 18 strains. Only transcripts longer than 200bp were retained. A total of 216,203 transcripts were generated representing more than 178.10^6 bases of sequence and an average sequence length of 824 bases. When several isoforms were detected, only the longest one was retained for the analyses, representing 153,222 transcripts for a total of 117,601,765 bp with an average transcript size of 767 bp. Sequence similarity of the transcripts with genes of identified function in the UniProt databank was investigated using the bank to bank sequence similarity search tool ngKLAST v4.3 using the KLASTx algorithm (Nguyen and Lavenier 2009) with E-Value < 10^{-3} (32,948 transcripts with homologs). The transcripts were classified in various Gene Ontology categories (GO; http://geneontology.org/) based on this result. Independently from this annotation, Transdecoder (Haas et al. 2013) was used to determine Coding Sequences (CDS) from the transcripts (76,698 transcripts with CDS). When more than one possible frame was detected (17,492 CDS, ie ~ 23%), the CDS was not considered unless it contained mutations (see below), in which case the frame minimizing the number of non-synonymous mutations was retained (9,032 CDS). The effect of this choice on the ratio of non-synonymous mutations per transcript is illustrated in supplementary fig. S1 (Supplementary Material online). The analyses were also performed after excluding all the transcripts with more than one possible frame, without any impact on the conclusions (data not shown). As *A. minutum* is not closely related to any model organism, transcript annotation has to be taken with great caution. As a mean to both evaluate at what point annotations maybe meaningful, and decrease the amount of wrongly annotated transcripts the frames assigned by the ngKLAST annotation and the ones inferred from Transdecoder were compared. A total of 26,487 transcripts had frames assigned by both Transdecoder and ngKLAST of which 17,235 did match (65%). This is about 4 times more than expected if the annotation was biologically irrelevant (as there are 6 possible frames random matches are expected for 1/6 of the transcripts). When the two frames did not match, the frame inferred using Transdecoder was conserved, but the annotation was discarded.

Alignment to the reference transcriptome

The 18 strains were then individually aligned to the reference consisting of 153,222 transcripts with Bowtie2 (Langmead and Salzberg 2012) using paired-end reads. Only reads with a mapping score > 10 were retained. Alignments were sorted and duplicates removed using Samtools (Li et al. 2009). Taking into account all strains together, sites had an average sequencing depth of 462. Individually, the strains had an average sequencing depth ranging from 11 to 49.

Mutation analyses

For variant analyses, only transcripts with more than 100 sites covered more than ten times in each of the 18 strains were considered. Single Nucleotide Polymorphisms (SNPs) were detected using

FreeBayes (Garrison and Marth 2012). In culture conditions, *A. minutum* cells are in a vegetative, haploid stage. We took advantage of this to remove spurious SNPs and more specifically SNPs that may be identified because of genetic polymorphism within a single genome (in case of paralogy) and not between genomes. To do so, FreeBayes was run with three sets of parameters: 1. haploidy enforced, 2. diploidy enforced and 3. diploidy enforced with a minimal allele count supported by at least 5 reads to call a genotype. Mutations identified by Freebayes were then filtered using VCFtools (Danecek et al. 2011), only keeping positions involved in SNPs, with two alleles, a quality criterion > 40, and covered more than 10 times in each of the 18 sequenced strains. Because cultures are composed of haploid clones, diploid enforced genotypes must be homozygote. After filtering, the results of the three Freebayes runs were compared and only positions identified in the haploid enforced run and identified as homozygous in the two diploid enforced runs were considered. Genotypes identified as heterozygotes in the diploidy enforced runs were discarded. Genetic distance among any two strains was calculated as the proportion of variant sites. Hierarchical clustering analysis with complete linkage was performed in R using hclust.

To investigate the divergence between group A and B (see Results), the demographic history was analyzed from their joint site frequency spectrum (JSFS) using δaδi v1.7.0 (Gutenkunst et al. 2009). As proposed by Tine et al. (2014), we tested seven alternative models of historical divergence: Strict Isolation (SI), Isolation with Migration (IM), Ancient Migration (AM), Secondary Contact (SC), as well as a version of IM, AM, and SC including a restricted migration rate for a subset of SNPs (IM2m, AM2m, and SC2m). As the ancestral states of the SNPs could not be determined with confidence, we used folded joint frequency spectrum, i.e. the frequency spectrum based on minor allele count. The demographic history was inferred using all polymorphic sites, as well as using five subsets composed of a single randomly chosen synonymous polymorphic site per transcript. For each demographic model and each dataset, more than 30 runs were performed to identify the maximum likelihood and the corresponding parameter estimates. Using this modeling approach, the SNPs observed as fixed (one allele in all members of group A and the alternative allele in all members of group B) were identified as displaying a highly restricted gene flow between the two groups (see Results). We note that when we refer to fixed polymorphism, we considered the observed pattern in the 18 strains studied and do not extrapolate the fixation at the entire group level.

Fisher exact tests were used to investigate the deviation from random accumulation of fixed SNPs in the transcripts. False Discovery Rate (FDR) correction for multiple testing with a significance threshold set at q-value = 0.05 was used.

Following a McDonald and Kreitman (1991) approach, we used Fisher Exact tests to investigate whether NS mutations are over-represented in the fixed differences.

Over-representation of 1. SNPs, 2. Non-Synonymous (NS) SNPs, 3. Fixed SNPs, and 4. Fixed NS SNPs in GO categories was tested for GO categories represented by at least 5 transcripts, using Fisher Exact tests followed by FDR correction for multiple testing with a significance threshold set at q-value = 0.0001 (a very stringent FDR was set to balance the uncertainty of the GO annotations due to the absence of a closely related model organism). Only GO categories containing > 5 mutated transcripts were considered. Over-representation analyses were based on SNPs rather than on mutated transcripts to add more weight to the transcripts carrying multiple SNPs.

Saxitoxin, COI and rRNA genes

Two forms homolog to the cyanobacteria *sxtA* gene, named long and short forms, as well as one homologous of the *sxtG* cyanobacteria gene known to be involved in saxitoxin production were identified in *Alexandrium* (Stüken et al. 2011). We searched the *A. minutum* reference transcriptome generated above for the *A. fundyense sxtA* short (JF343238) and long (JF343239) forms well as *sxtG* (JX995121) using blastn 2.2 (Zhang et al. 2000). Similarly, we searched for published COI (AB374235) and rRNA (AY831408) sequences in the *A. minutum* reference transcriptome generated above using blastn 2.2.

Inter-group differential expression

Differential expression analyses were performed using the packages, DESeq2 (Love et al. 2014), edgeR (Robinson et al. 2010) and limma (Ritchie et al. 2015) in R. Only transcripts with a total read count higher than 200 were considered, representing 100,797 transcripts with a median coverage per transcript ranging from 42 to 188 reads for the different strains (mean range: 108-505). Hierarchical clustering was performed using hclust (R) based on the Euclidean distance calculated by the dist function (R) on the rlog transformed count matrix. Differential expression between the two groups of strains was tested with a significance FDR threshold set at q-value = 0.05, with rlog (Deseq2), TMM (edgeR), and voom (limma) normalization. The transcripts significant with the three methods (intersection) were considered as differentially expressed. We note that differentially expressed transcripts may be the result of differential regulation of gene expression in the two groups of strains, but also of deletion of the encoding genes in one of the two groups. Over-representation of GO categories was tested for GO categories represented by at least 5 transcripts, using Fisher Exact followed by a False Discovery Rate (FDR) correction for multiple testing with a significance threshold set at q-value = 0.01. Only GO categories containing > 5 differentially expressed transcripts were considered. We note that the presence of a conserved spliced leader in 5' of all dinoflagellate mRNA might indicate important post-transcriptional regulation of gene expression in these organisms (Zhang et al. 2007).

Morphological analyses of the strains

Thecal plate pattern and the presence of a ventral pore on the first apical plate (1′) of the different strains was analyzed after staining thecae with Fluorescent Brightener 28 (Sigma Aldrich) according to the method of Fritz and Triemer (1985). Strains were observed on a slide covered with a coverslip in epifluorescence microscopy after adding a drop of 1% (w/v) of the fluorophore and using a BX41 (Olympus, Tokyo) upright microscope fitted with a 100 W mercury lamp and epifluorescence (U-MWU2 filter cube).

Results

Genetic diversity

To investigate genetic diversity, we only considered transcripts with more than 100 sites covered more than ten times in each of the 18 sequenced strains, representing a total of 24,630,108 sites in 45,089 transcripts, and identified a total of 457,368 polymorphic sites (~1.9 % of the sites) in 41,698 transcripts (~92.5% of the transcripts, table 1). We performed a hierarchical clustering analysis based on the nucleotide divergence among any two strains (fig. 1a). Two groups of strains may be distinguished. The first group, hereafter named group A, of 15 strains composed of a slightly divergent strain isolated from Cork (Ireland), and two sub-clades grouping on the one hand all the strains isolated from the Penzé estuary (France) and on the other strains isolated from the Bay of Brest (France) and one strain isolated from the Rance estuary (France). In this group the median number of variable sites among any two strains is 99,224 (~22% of the variable sites), representing a nucleotide divergence of ~0.004, reflecting a high level of genetic diversity (fig. 1a, black). The second group, hereafter named group B, is composed of three strains, one isolated from the Bay of Brest and two from the Bay of Concarneau. Within this group B, the three strains are also very divergent genetically, with a median of 127,407 variable sites among strains (~28%), representing a nucleotide divergence of ~0.005. The intergroup median number of variable sites is 147,913 (~32%), representing a nucleotide divergence of ~0.006. A total of 193,325 variants are singletons, i.e. they were identified in a single strain, representing more than 42% of the identified variants. Looking at the repartition of these singletons in the 18 strains, the two groups of strains previously identified are again clearly visible. Within group A, the median number of singletons is 7,303 (fig. 1b, black). Within group B there is almost 4 times more singletons per strain (median=27,532) (fig. 1b, red).

Divergence between group A and B

To replace the divergence between group A and B in a classical phylogenetic context, we note that in the transcript corresponding to the ribosomal RNA, two SNPs observed as fixed (one allele in all members of group A and the alternative allele in all members of group B) are identified in the 5' external transcribed spacer but none in the region corresponding to the 18S, ITS1, 5.8S, ITS2, and LSU. Similarly, no SNP was identified in the transcript corresponding to the cytochrome c oxidase subunit I (COI), another gene often used to identify closely related species (table 2).

To better grasp the patterns of divergence between strains belonging to group A and B and gaining insights on the underlying evolutionary processes, we analyzed the demographic history of groups A and B using their joint site frequency spectrum (JSFS), exploring seven scenario of divergence (fig.2). The simplest model, involving divergence without any gene flow (SI, fig. 2, Supplementary table 1) did not explain the data as well as models involving some amount of gene flow after the split. Of these, the secondary contact (SC) model had the best likelihood, especially because it explained the low occurrence of minor allele only observed in group A (fig. 2). The only part of the JSFS not correctly explained by the SC model was an excess of observed fixed polymorphism compared to the model prediction (lowest residual values, fig. 2). The observed fixed polymorphism was correctly estimated when a heterogeneous migration rate across SNPs, with a fraction of the sites displaying a highly restricted gene flow between groups, was introduced in the divergence model (SC2m model, fig. 2, Supplementary table 1). Similarly, when considering subsets of the entire dataset composed of a single randomly chosen synonymous polymorphic site per transcript, the model with the highest likelihood was the SC2m model (Supplementary table 2). These analyses indicated an ancient divergence of the A and B groups in total isolation, followed by a secondary contact resulting in gene flow between the two groups that is heterogeneous across the genomes, with a fraction of the SNPs displaying highly restricted gene flow. As seen in Figure 2, the polymorphic sites that are observed as fixed between the two groups are the ones displaying restricted gene flow (part of the folded JSFS requiring a heterogeneous migration rate across genomes to be correctly explained, fig. 2).

In the dataset, 12,188 variant sites (5% of the variable sites, excluding singletons) display a fixed difference between group A and B (table 1). We focused on the fixed differences between groups A and B to determine if these SNPs are restricted to a few transcripts or randomly distributed in the transcripts. The 12,188 fixed differences occur in 6,215 transcripts but are over-represented, compared to the other differences, in 927 transcripts (Fisher exact test, q-value < 0.05, fig. 3a, red dots), representing 4,616 fixed mutations (38% of the fixed differences). This result clearly points toward a preferential accumulation of the fixed differences in some transcripts.

Next, we investigated whether mutations are synonymous (S) or non-synonymous (NS). Excluding singletons, a total of 44,880 NS and 176,609 S mutations were identified in 29,089 transcripts (table 1). Focusing on the fixed differences, 3,818 NS and 5,733 S mutations were detected in 4,916 transcripts, indicating that non-synonymous mutations are 2.77 times more frequent in the fixed differences compared to the other mutations (Fisher exact test, p-value < $2.2e^{-16}$). More precisely, the frequency of non-synonymous mutations in the transcripts is higher when considering fixed mutations (fig. 3b grey), not only in the transcripts where fixed mutations are over-represented (fig. 3b red), but also in the transcripts only displaying a few fixed mutations (fig 3b blue). This indicates that potential modification of protein functions associated with the divergence is not only linked to transcripts displaying numerous fixed mutations, but also to the transcripts only displaying a few fixed mutations.

Functional genetic divergence

We used two approaches to investigate the functions of the genes displaying fixed differences. In the first one, we analyzed the repartition of SNPs associated to the different gene product properties, as defined by Gene Ontology (GO). A total of 9,508 transcripts representing 82,805 mutations could be associated to GO categories (table 1). We tested whether mutations are over or under represented in the different GO categories. Considering the entire dataset, 24 GO categories display an excess of mutations and 147 display less mutations than expected (fig. 4, 171 GO categories Overall). The non-synonymous mutations were over-represented in 6 categories and underrepresented in 6 (fig. 4, 12 GO categories Overall Non-synonymous). Focusing on the fixed differences, mutations were over-represented in 33 categories (Non-synonymous mutations, 4 categories) and not under-represented in any GO category (fig. 4 Fixed and Fixed Non-synonymous). These fixed differences are found in a total of 328 transcripts. Of special interest are 130 transcripts involved in 5 GO categories related to calcium binding and fluxes across membranes (fig. 4 red) and 44 in 4 GO categories related to potassium fluxes across membranes (fig. 4 blue).

In a second approach to grasp the functional bases of the divergence, we focused on the 25 transcripts displaying most fixed genetic divergence between the divergent groups (lowest q-value, fig. 3a, i.e. ~0.5‰ (25/45,089) transcripts displaying the highest level of genetic divergence (table 2). Out of these 25 transcripts, 14 were identified as homologs to genes encoding for proteins with known functions. Of extreme interest was the presence of four transcripts homologs to genes involved in calcium mediated transduction signals: two involved in calcium transport (Polycystin-2 and Sodium/calcium exchanger 3), one intermediate messenger transducing calcium signals by binding calcium ions (Calmodulin-like protein 6), and one calcium-dependent protein kinase thought to function in signal transduction pathways that utilize changes in cellular Ca^{2+} concentration to couple cellular responses to extracellular stimuli (Calcium-dependent protein kinase 13). Even more interesting, was the genetic divergence of a transcript corresponding to the short form of the *sxtA* gene, a gene known to be involved in saxitoxin production in cyanobacteria. It contains domains 1 to 3 homologous to the *sxtA* genes found in cyanobacteria and a last translated region that has no homolog in databases except the end of the short *sxtA* form from *A. fundyense* (Stüken et al. 2011). Moreover, these fixed differences include numerous NS mutations (9), two of them being in the first domain (sxtA1, corresponding to the amino acids 28-531), one in the second domain (sxtA2, amino acids 535-729), none in the third (sxtA3, amino acids 750-822) and six in the last translated part of the transcript (amino acids 822-976) (Fig. 2*C*).

Differential gene expression

We analyzed the mRNA sequences to investigate differential gene expression *in vitro*. First, a clustering analysis based on the expression levels clearly indicates that the two groups of strains identified above using genetic information are also identified using global expression data (fig. 5a). Differential expression was analyzed between group A and group B strains, and a total of 1,518 transcripts were identified as differentially expressed (q-value<0.05; fig. 5b; Supplementary Table 3), but no gene ontology category was identified as over or under-represented in the differentially expressed transcripts at a FDR level < 0.1.

Morphology

The 18 strains were stained with Fluorescent Brightener 28 and observed blindly, i.e. without knowing which strains belonged to group A and B in epifluorescence microscopy to analyze the thecal plate pattern. No difference of the thecal organization was found among strains which all possessed the typical plate pattern of *A. minutum*. However, the presence of a ventral pore on the right side of the 1′ plate was found on the three strains belonging to group B while the 15 strains belonging to group A lacked this feature (fig. 6).

Discussion

Analyzing mRNA sequences in 18 *A. minutum* strains, we identified the divergence of two groups, represented by 15 and 3 strains, respectively. The identification of these two groups was incidental, explaining the unbalanced sampling and illustrating the possibility for reverse ecology approaches to uncover cryptic diversity. This divergence was not detectable using the classical barcoding loci ITS and COI, but the analysis of a transcriptome wide SNPs dataset pointed toward the presence of two distinct evolutionary units. A genetic distance analysis clearly indicated the presence of the two groups. A consistent observation is the difference in the number of singletons identified in the strains belonging to the groups A and B, clearly pointed toward the sampling of two independent genetic entities. Differential expression, although less dramatic than genetic divergence, also goes in the same direction, with the strains belonging to the two groups displaying the most difference in terms of global expression profile. One of the strains was isolated from the natural environment in 1989 (Am89) and maintained ever since (i.e. during 24 years, corresponding to ~3,000 generations of cellular division) in batch culture involving bi-weekly transfer in the culture media used in the present work. The other strains were isolated from 2010 to 2013 and maintained in the same culture regime (6 months to 3 years of lab culture, 60-350 generations). Despite the difference in the time spent in the laboratory environment and thus experiencing the associated strong selective pressures, the strain Am89 is genetically indistinguishable from the other strains belonging to group A. It illustrated that, compared to the standing genetic variation encountered in natural populations, there are very few mutations that occurred during the long term maintenance in culture. In term of gene expression, Am89 clusters at the base of group A, pointing toward a more extensive evolution of gene expression profile, but still insufficient to overcome the difference in global expression profile occurring between groups A and B.

Following the analysis of the mRNA sequences and the identification of the two diverging groups, a morphological difference, the presence/absence of a ventral pore was identified. This morphological character seems diagnostic of the two groups, suggesting the occurrence of two pseudo-cryptic (or pseudo-sibling) species (Knowlton 1993), but caution must be taken due to the very limited sampling of one of the two group. Nonetheless, this morphological character is especially interesting to replace our study in a biogeographical context. Indeed, this morphological feature has been reported in *A.*

minutum studies with some indication that the morphotype with ventral pore may be more frequent in Southern Europe and the one lacking the ventral pore more frequent in Northern Europe (Hansen et al. 2003). Interestingly the two types have also been reported in mixed communities (Western Ireland, Hansen et al. 2003; in the present study Am1072 (group B) and Am1080 (group A) were isolated from the same day and locality) which rules out complete allopatry.

Using the SNPs dataset, we investigated the process of divergence between groups A and B. We compared the joint site frequency spectrum of these two groups to the patterns expected following seven models of divergence. The most likely scenario involves an ancient divergence in complete isolation followed by a secondary contact involving gene flow between the two groups. Quite interestingly, the introduction of a heterogeneous migration rate across the genome, with a fraction of the genome displaying a highly restricted gene flow, considerably improved the likelihood of the models. So far only a handful of studies have considered heterogeneous gene flow across the genomes when investigating the divergence of population/species. We note that models of divergence in isolation followed by a secondary contact allowing gene flow between diverging populations/species, but at different rates across the genome, are, so far, almost always the best at explaining the observed allelic frequencies in ascidian (Roux et al. 2013), mussels (Roux et al. 2014), fishes (Tine et al. 2014, Le Moan et al. 2016, Rougemont et al. 2016), and Ascomycota (Gladieux et al. 2015). Here, an extremely low migration rate at a fraction of the genome is required to explain the observed pattern of fixed polymorphism, i.e. of polymorphism with all members of group A displaying one allele, and all members of group B displaying the alternative allele. This fixed polymorphism corresponds to about 5% of the SNPs displaying a heterogeneous distribution in the various transcripts. This is similar to the pattern reported in studies investigating recent or ongoing speciation events at the genome scale (Seehausen et al. 2014). However, one of the caveat of using transcriptome and not genome wide data is that the information regarding the physical linkage between the genes encoding for the transcripts is lacking. As a result, we do not know whether the transcripts displaying high levels of genetic divergence are physically linked in a few genomic islands of divergence (Turner et al. 2005) or if they are spread out in the genome.

We compared the proportion of non-synonymous polymorphism segregating and fixed between the two groups and identified a strong excess of non-synonymous polymorphism in the fixed mutations. SNPs fixation within each group, but divergence between groups associated with overrepresentation of NS SNPs is difficult to explain with demographic fluctuations or relaxed selection and points toward the importance of selection as a driving force of the divergence between the two groups. This pattern could reflect classic selective sweeps, i.e. the fixation of adaptive mutations in either group, and the associated hitchhiking of physically linked neutral mutations (Nielsen 2005). Interestingly, an excess of fixed non-synonymous mutations was also identified in transcripts only displaying a few fixed polymorphic sites, often associated with segregating polymorphism. This excess of NS mutations suggests that mutations associated with the functional divergence of the two divergent groups are not systematically associated with a selective sweep, i.e. may get to fixation without a drastic reduction of diversity at neighboring sites. Indeed the pattern of linkage disequilibrium associated to an adaptive mutation is influenced by numerous factors including, the strength of selection, local levels of recombination, and whether adaptive mutations are *de novo* mutations or were segregating in ancestral populations before becoming adaptive (Fay and Wu 2000; Przeworski et al. 2005; Lee et al. 2014).

The selective pressures responsible for restricted and heterogeneous gene flow may be directly linked to the ecological divergence of the two groups. It could for example be the case, if the two groups occupy geographically and ecologically distinct habitats and only encounter each other and exchange genes at localized hybrid zones. In this case, introgression of neutral SNPs from one group to the other would occur more or less freely, while the introgression of the SNPs responsible for local adaptation of each group would be counter selected. An alternative scenario could involve the build-up of reproductive isolation between the two groups. For example, gene flow could be restricted overall if members of the two groups are not likely to recognize each other as proper mates, and negative epistasis between sets of SNPs could lead to reduced hybrid fitness depending (hybrid maladaptation) or not (genetic incompatibilities) on the environmental conditions. Distinguishing between these different scenarios (none of them being mutually exclusive) would require extensive sampling from the natural environment, crossing experiments and fitness assays that are beyond the scope of the present work. However, investigating the cellular functions of the transcripts displaying restricted gene flow between the two groups could help pointing in one direction.

Transcripts related to potassium and calcium fluxes across membranes were identified as carrying more fixed polymorphism than expected. Moreover, among the transcripts displaying the highest levels of divergence, four could be related to calcium mediated transduction signals, and one was homologous to *sxtA*, a gene involved in saxitoxin production (Stüken et al. 2011). Two genetically divergent forms of *sxtA* have been identified in *Alexandrium* transcriptomes (Stüken et al. 2011). Here, the *sxtA* identified as highly divergent between the two groups corresponds to the short form, which is probably not involved in saxitoxin production (Murray et al. 2015; the long form was also identified in all strains, but without displaying a pattern of divergence, data not shown). As a result, we hypothesize that the molecule of interest associated with the divergence of the two groups might not be the saxitoxin itself but another compound synthesized via the saxitoxin biosynthesis pathway. There may be a direct link between *sxtA*, genes related to calcium and potassium fluxes, and calcium mediated signal transduction. Indeed, although the saxitoxin toxicity occurs through the blocking of mammal sodium channels, it is also known to bind to mammal calcium and potassium channels, modifying calcium and potassium fluxes without entirely blocking them (Cusick and Sayler 2013). This analysis points toward a molecular mechanism that may be at play during the divergence of the two groups, but does not indicate whether it is related to ecological divergence or the build-up of reproductive isolation. In favor of the build-up of reproductive isolation, saxitoxin has been proposed to act as a sex pheromone in natural environment (Wyatt and Jenkinson 1997; Cusick and Sayler 2013) and another guanidine alkaloids marine toxin, the tetrodotoxin, has been shown to act as sex pheromone (Matsumura 1995). However, some *Alexandrium* strains do not produce the toxin and would thus be unable to attract proper mates, but as discussed above, the molecule at play here is probably not the saxitoxin itself but a related molecule. In favor of an ecological divergence, we may cite the proposed role of saxitoxin as a grazer deterrent (Cusick and Sayler 2013), but it would require a specialized relationship to exert a selective pressure responsible for the observed divergence. Finally, unicellular motility is often linked to calcium fluxes across the membrane (Verret et al. 2010), with potential implications in both ecological divergence and reproductive isolation.

To conclude, using a reverse ecology approach based on the mRNA sequencing and morphology analysis of several strains of the dinoflagellate *A. minutum*, two diverging groups, co-occurring in nature, were identified. The most likely scenario of divergence involved ancient divergence in complete isolation followed by a secondary contact resulting in gene flow, heterogeneous across the genome, between the diverging groups. The SNPs subjected to restricted gene flow also display an overrepresentation of fixed non-synonymous polymorphism. This highlights the importance of the functional aspect of the divergence, and identifies selection as a potential major evolutionary force driving this event. At the molecular level the functions associated with the divergence are especially related to toxin production and calcium/potassium fluxes with potential implications in terms of ecological divergence and build-up of reproductive isolation that remain to be tested.

Acknowledgments: This work was supported by a financial support from the Région Bretagne and the ANR Hapar (ANR-14-CE02-0007-01). We acknowledge RIC and Abims facilities for bioinformatics support, the laboratories LSEM and PFOM for providing access to some of their equipment, Elisabeth Nézan, and Pierre-Alexandre Gagnaire for discussion, four anonymous reviewers, Tatiana Giraud, and Myriam Valero for comments on the manuscript.

References

Anderson DM, Alpermann TJ, Cembella AD*, et al.* (2012a) The globally distributed genus Alexandrium: Multifaceted roles in marine ecosystems and impacts on human health. *Harmful Algae* **14**, 10-35.

Anderson DM, Cembella AD, Hallegraeff GM (2012b) Progress in understanding harmful algal blooms: Paradigm shifts and new technologies for research, monitoring, and management. In: *Annual Review of Marine Science* **4**, 143-176.

Bierne N, Bonhomme F, David P (2003) Habitat preference and the marine-speciation paradox. *Proceedings of the Royal Society B-Biological Sciences* **270**, 1399-1406.

Casabianca S, Penna A, Pecchioli E, Jordi A, Basterretxea G and Vernesi C (2011) Population genetic structure and connectivity of the harmful dinoflagellate *Alexandrium minutum* in the Mediterranean Sea. *Proceedings of the Royal Society B-Biological Sciences* **279**, 129-138.

Casteleyn G, Leliaert F, Backeljau T, *et al.* (2010) Limits to gene flow in a cosmopolitan marine planktonic diatom. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 12952-12957.

Cusick KD, Sayler GS (2013) An overview on the marine neurotoxin, saxitoxin: genetics, molecular targets, methods of detection and ecological functions. *Marine Drugs* **11**, 991-1018.

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

Dia A, Guillou L, Mauger S, *et al.* (2014) Spatiotemporal changes in the genetic diversity of harmful algal blooms caused by the toxic dinoflagellate *Alexandrium minutum*. *Molecular Ecology* **23**, 549-560.

Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution* **29**, 51-63.

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405-1413.

Feder JL, Flaxman SM, Egan SP, Comeault AA, Nosil P (2013) Geographic mode of speciation and genomic divergence. *Annual Review of Ecology, Evolution, and Systematics, Vol 44* **44**, 73-97.

Fritz L, Triemer RE (1985) A rapid simple technique utilizing Calcofluor White M2R for the visualization of dinoflagellate thecal plates. *Journal of Phycology* **21,** 662-664.

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. Preprint arXiv:1207.3907

Giardine B, Riemer C, Hardison RC, *et al.* (2005) Galaxy: A platform for interactive large-scale genome analysis. *Genome Research* **15**, 1451-1455.

Gladieux P, Wilson BA, Perraudeau F, *et al.* (2015) Genomic sequencing reveals historical, demographic and selective factors associated with the diversification of the fire-associated fungus Neurospora discreta. *Molecular Ecology* **24**, 5657–5675.

Gutenkunst RN, Hernandez RD, Williamson SH, and Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP data. *PLoS Genetics* **5**, e1000695.

Haas BJ, Papanicolaou A, Yassour M, *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* **8**, 1494-1512.

Hansen G, Daugbjerg N, Franco JM (2003) Morphology, toxin composition and LSU rDNA phylogeny of *Alexandrium minutum* (Dinophyceae) from Denmark, with some morphological observations on other European strains. *Harmful Algae* **2**, 317-335.

Hutchinson G (1961) The paradox of the plankton. *The American Naturalist* **95**, 137-145.

Iglesias-Rodríguez MD, Schofield OM, Batley J, Medlin LK, and Hayes PK (2006) Intraspecific genetic diversity in the marine coccolithophore *Emilianta huxleyi* (Primnesiophyceae) : The use of microsatellite analysis in marine phytoplankton population studies. *Journal of Phycology* **42,** 526-536.

Knowlton N (1993) Sibling Species in the Sea. *Annual Review of Ecology and Systematics* **24**, 189-216.

Le Moan A, Gagnaire P-A, Bonhomme F (2016) Parallel genetic divergence among coastal–marine ecotype pairs of European anchovy explained by differential introgression after secondary contact. *Molecular Ecology*. doi: 10.1111/mec.13627.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-359.

Lee YCG, Langley CH, Begun DJ (2014) Differential strengths of positive selection revealed by hitchhiking effects at small physical scales in *Drosophila melanogaster*. *Molecular Biology and Evolution* **31**, 804-816.

Li H, Handsaker B, Wysoker A, *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079.

Li YF, Costello JC, Holloway AK, Hahn MW (2008) "Reverse ecology" and the power of population genomics. *Evolution* **62**, 2984-2994.

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**.

Magoc T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-2963.

Masseret E, Grzebyk D, Nagai S *et al.* (2009) Unexpected genetic diversity among and within populations of the toxic dinoflagellate *Alexandrium catenella* as revealed by nuclear microsatellite markers. *Applied and Environmental Microbiology* **75,** 2037-2045.

Matsumura K (1995) Tetrodotoxin as a pheromone. *Nature* **378**, 563-564.

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the adh locus in *Drosophila*. *Nature* **351**, 652-654.

Murray SA, Diwan R, Orr RJS, Kohli GS, John U (2015) Gene duplication, loss and selection in the evolution of saxitoxin biosynthesis in alveolates. *Molecular Phylogenetics and Evolution* **92**, 165-180.

Nguyen VH, Lavenier D (2009) PLAST: parallel local alignment search tool for database comparison. *Bmc Bioinformatics* **10**.

Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics* **39**, 197-218.

Palenik B, Grimwood J, Aerts A, *et al.* (2007) The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 7705-7710.

Palumbi SR (1992) Marine speciation on a small planet. *Trends in Ecology & Evolution* **7**, 114-118.

Peers G, Price NM (2006) Copper-containing plastocyanin used for electron transport by an oceanic diatom. *Nature* **441**, 341-344.

Przeworski M, Coop G, Wall JD (2005) The signature of positive selection on standing genetic variation. *Evolution* **59**, 2312-2323.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shy W, and Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47.

Robinson MD, McCarthy DJ and Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140.

Rocap G, Larimer FW, Lamerdin J, *et al.* (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424**, 1042-1047.

Rougemont Q, Gaigher A, Lasne E, Côte J, Coke M, Besnard A-L, Launey S and Evanno G (2015) Low reproductive isolation and highly variable levels of gene flow reveal limited progress towards speciation between European river and brook lampreys. *Molecular Ecology* **28**, 2249-2263.

Roux C, Tsagkogeorga G, Bierne N, and Galtier N (2013) Crossing the species barrier: genomic hotspots of introgression between two highly divergent *Ciona intestinalis* species. *Molecular Biology and Evolution* **30**, 1574–1587.

Roux C, Fraïsse C, Castric V, Vekemans X, Pogson GH and Bierne N (2014) Can we continue to neglect genomic variation in introgression rates when inferring the history of speciation? A case study in a *Mytilus* hybrid zone. *Journal of Evolutionary Biology* **27**, 1662–1675.

Roy S, Chattopadhyay J (2007) Towards a resolution of 'the paradox of the plankton': A brief overview of the proposed mechanisms. *Ecological Complexity* **4**, 26-33.

Rynearson TA, and Armbrust VE (2004) Genetic differentiation among populations of the planktonic marine diatom *Ditylum brightwellii* (Bacillariophycae). *Journal of Phycology* **40**, 34-43.

Seehausen O, Butlin RK, Keller I, *et al.* (2014) Genomics and the origin of species. *Nature Reviews Genetics* **15**, 176-192.

Shoresh N, Hegreness M, Kishony R (2008) Evolution exacerbates the paradox of the plankton. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 12365-12369.

Stomp M, Huisman J, de Jongh F, *et al.* (2004) Adaptive divergence in pigment composition promotes phytoplankton biodiversity. *Nature* **432**, 104-107.

Stüken A, Orr RJS, Kellmann R, *et al.* (2011) Discovery of nuclear-encoded genes for the neurotoxin saxitoxin in dinoflagellates. *Plos One* **6**, 12.

Tine M, Kuhl H, Gagnaire P-A, et al. (2014) European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nature Communications* **5**, 5770.

Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic islands of speciation in *Anopheles gambiae*. *Plos Biology* **3**, 1572-1578.

Verret F, Wheeler G, Taylor AR, Farnham G, Brownlee C (2010) Calcium channels in photosynthetic eukaryotes: implications for evolution of calcium-based signalling. *New Phytologist* **187**, 23-43.

Weiner A, Aurahs R, Kurasawa A, Kitazato H, Kucera M (2012) Vertical niche partitioning between cryptic sibling species of a cosmopolitan marine planktonic protist. *Molecular Ecology* **21**, 4063-4073.

Wisecaver JH, Hackett JD (2011) Dinoflagellate genome evolution. *Annual Review of Microbiology* **65**, 369-387.

Wyatt T, Jenkinson IR (1997) Notes on *Alexandrium* population dynamics. *Journal of Plankton Research* **19**, 551-575.

Zhang H, Hou Y, Miranda L, *et al.* (2007) Spliced leader RNA trans-splicing in dinoflagellates. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 4618-4623.

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology* **7**, 203-214.

Data Accessibility:

Raw reads and Reference transcriptome: European Nucleotide Archive
http://www.ebi.ac.uk/ena/data/view/PRJEB15046

SNP and differential expression information: SEANOE database http://doi.org/10.17882/45445

Author Contributions:

MLG, CD and LG designed research, MLG, GM, NC, PM, JQ and OB performed research, MLG and NC analyzed the data, MLG, GM, NC, RS, CD, LG and AC wrote the paper.

Figure legends

Fig. 1. Genetic divergence. (a) Hierarchical clustering analysis displaying the genetic distance among *A. minutum* strains based on nucleotide divergence, names of the strains and year of isolation are indicated; (b) number of singletons par strain; (c) origin of the strains. The strains from group A are in black and the ones from group B in red.

Fig. 2. Results of model fitting for seven alternative models of divergence. The observed folded Allele Frequency Spectrum (AFS), as well as for each model, the residuals of the modeled AFS are presented. SI is the strict isolation model. IM is the Isolation-with-Migration model, AM the Ancient Migration model, and SC is the Secondary Contact model. All three models of divergence-with-gene-flow were implemented using one, shared migration rate in each direction (m1>2, m2>1) across the genome (homogeneous migration), or with two categories of migration rates in each direction across the genome (heterogeneous migration). The data are best explained by the SC2m model.

Fig. 3. Fixed polymorphism. (a) Repartition of the transcripts based on the number sites displaying fixed and segregating polymorphism. Red dots indicate over-representation of fixed polymorphism (q-value < 0.05). For the 25 most divergent transcripts, homology with genes involved in calcium transduction signal (red) and saxitoxin production (violet) are indicated. (b) Frequency of NS polymorphism considering segregating polymorphism (grey), fixed polymorphism in transcripts where fixed polymorphism is over-represented (red), and fixed polymorphism in transcripts without over-representation of fixed polymorphism (blue). (c) Fixed amino acid substitutions in SxtA.

Fig. 4. Venn diagram indicating the number of Gene Ontology (GO) categories displaying deviation from random accumulation of mutations (q-value < 0.0001), considering all mutations (Overall), the NS mutations (Overall Non-Synonymous), the fixed mutations (Fixed), and the NS fixed mutations

(Non-Synonymous Fixed). For the analyses focusing on the fixed mutations, the name of the GO categories is given, as well as the number of transcripts mutated, number of mutations, and q-values. Black arrows indicate over-representation of fixed mutations and white arrows indicate under-representation of mutations overall.

Fig. 5. Gene expression. (a) Hierarchical clustering based on the expression Euclidean distance (rlog). The strains from group A are in black and the ones from group B in red. (b) MA plot showing for each transcript the fold change (groupB/groupA) as a function of the average expression. Transcripts identified as differentially expressed are in red (q-value < 0.05).

Fig 6. Epifluorescence micrographs of the 18 strains showing the presence (red arrow) or the absence (blue arrow) of a ventral pore on the first apical plate of the theca. Scale bars: 20 µm.

Table 1: Summary of transcripts and mutations analyzed, considering the entire dataset (Total), the transcripts displaying mutations (Mutated), the transcripts displaying mutations excluding singletons (Mutated no singleton), and the transcript displaying fixed mutations (Fixed).

	Number of Transcripts	Length	Transcripts with CDS	Length CDS (NS)	Transcripts with homolog	Length Annotated (NS)
Total	45,089	24,630,108[a]	32,797	20,396,618[a]	10,454	7,703,971[a]
Mutated	41,698	457,368[b]	31,111	376,242[b] (85,923[b])	10,029	139,286[b] (26,725[b])
Mutated no singleton	38,116	264,573[b]	29,089	221,489[b] (44,880[b])	9,508	82,805[b] (14,007[b])
Fixed	6,215	12,188[b]	4,916	9,551[b] (3,818[b])	1,670	3,408[b] (1,183[b])

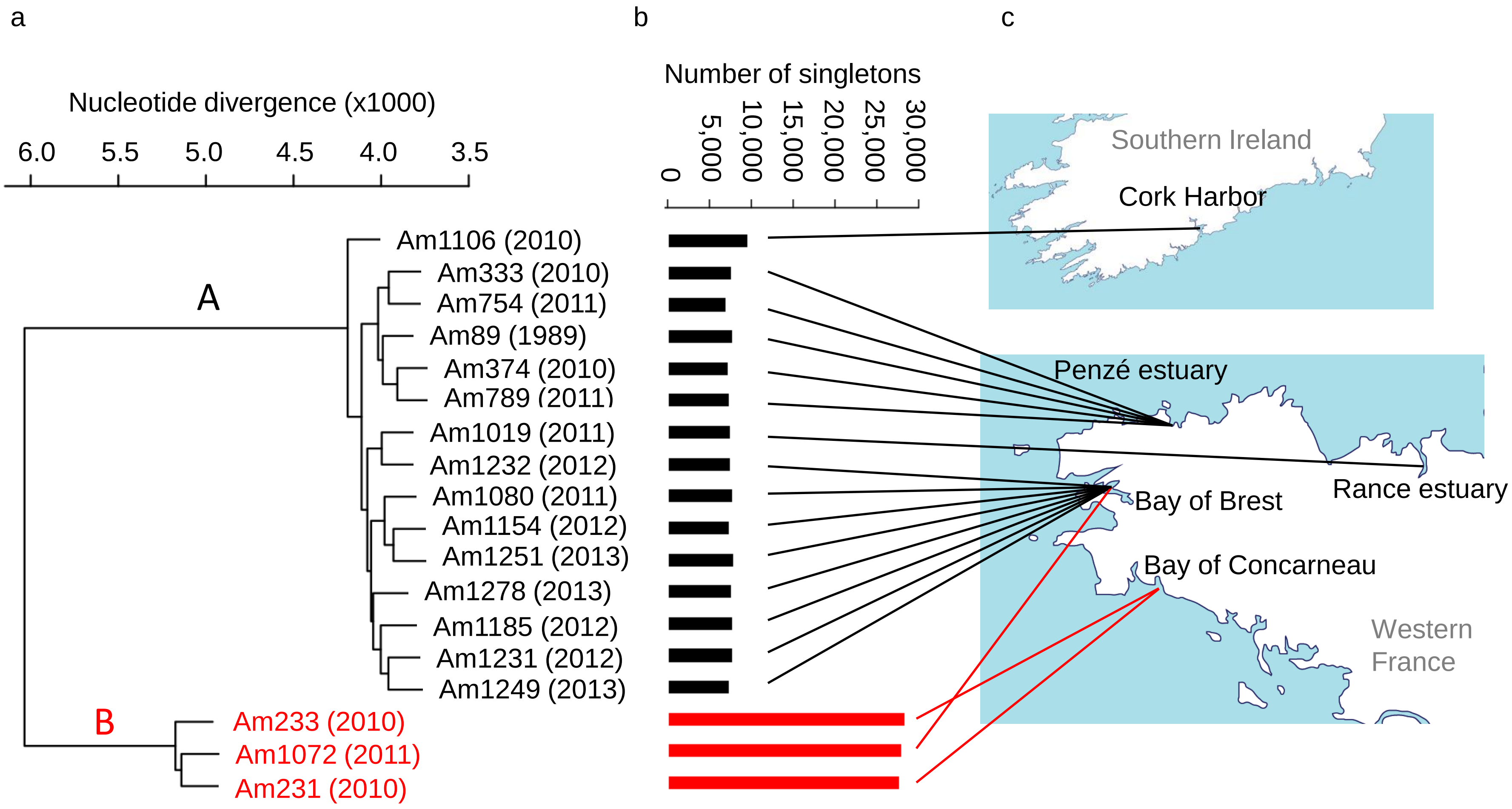
[a] Length of the transcripts

[b] Number of mutations

Table 2: Most divergent transcripts between A and B, and loci classically used in phylogenetic studies.

	Name	Fixed Mutations	Not Fixed Mutations	Homologs		E-value	Identity
25 most divergent transcripts between A and B	**comp60373_c0_seq1**	**22**	**2**	**CALL6_HUMAN**	**Calmodulin-like protein 6**	**1.10^{-10}**	**38.1%**
	comp98959_c0_seq1	18	0	TGS1_HUMAN	Trimethylguanosine synthase	3.10^{-41}	39.1%
	comp102434_c0_seq1	**18**	**3**	**PKD2_MOUSE**	**Polycystin-2**	**2.10^{-19}**	**35.4%**
	comp124736_c0_seq1	15	0	NA			
	comp95518_c0_seq1	**16**	**2**	**NAC3_HUMAN**	**Sodium/calcium exchanger 3**	**2.10^{-120}**	**33.3%**
	comp86525_c0_seq1	13	0	NA			
	comp101280_c0_seq1	15	4	NA			
	comp101305_c0_seq1	13	1	NA			
	comp75832_c0_seq1	13	2	CMBL_RAT	Carboxymethylenebutenolidase homolog	6.10^{-12}	22.8%
	comp96757_c0_seq1	13	2	NA			
	comp96807_c0_seq1	12	1	NEK5_HUMAN	Serine/threonine-protein kinase Nek5	5.10^{-05}	26.5%
	comp124661_c0_seq5	14	5	PGMC2_ARATH	Glucose phosphomutase 2	1.10^{-164}	48.9%
	comp78930_c0_seq1	11	0	NA			
	comp94714_c0_seq1	15	8	NA			
	comp104352_c0_seq1	12	2	NAAA_MOUSE	N-acylethanolamine-hydrolyzing acid amidase	8.10^{-33}	29.9%
	comp82584_c0_seq1	11	1	NA			
	comp115853_c0_seq1	13	5	PAMO_THEFY	Phenylacetone monooxygenase	5.10^{-05}	34%
	comp95265_c0_seq1	12	3	NA			
	comp105111_c2_seq1	10	0	EF1A_CRYPV	Elongation factor 1-alpha	3.10^{-96}	46.2%
	comp106635_c0_seq1	**10**	**0**	**CDPKD_ARATH**	**Calcium-dependent protein kinase 13**	**1.10^{-31}**	**24.9%**
	comp119140_c0_seq2	10	0	WIPF1_MOUSE	WAS/WASL-interacting protein family member 1	2.10^{-05}	34.3%
	comp86654_c0_seq1	11	2	NA			
	comp121041_c0_seq1	**15**	**13**	**F5BWX9_ALEFU**	**SxtA short isoform precursor**	**0.0**	**63%**
	comp117520_c0_seq1	14	14	MSL7_MYCMM	Beta-ketoacyl-acyl-carrier-protein synthase I	7.10^{-22}	30.4%
	comp66739_c0_seq1	10	1	ATAD3_BOVIN	ATPase family AAA domain-containing protein 3	7.10^{-91}	40%
COI	comp126209_c0_seq1	0	0	AB374235	A. catenella cox1	0.0	99%
rRNA	comp93300_c0_seq1	2 (ETS)	0	AY831408	A. minutum CCMP113 ETS-18S-ITS1-5.8S-ITS2-LSU	0.0	99%

Upper part, transcripts displaying the highest level of divergence between group A and B (klastx against UniProt/Swissprot). Transcripts with homologs involved in saxitoxin production and calcium signal transduction are indicated in violet, and red, respectively. Lower part, loci classically used in phylogenetic studies (blastn).


a
b
c
Nucleotide divergence (x1000)
6.0
5.5
5.0
4.5
4.0
3.5
Number of singletons
0
5,000
10,000
15,000
20,000
25,000
30,000
A
B
Am1106 (2010)
Am333 (2010)
Am754 (2011)
Am89 (1989)
Am374 (2010)
Am789 (2011)
Am1019 (2011)
Am1232 (2012)
Am1080 (2011)
Am1154 (2012)
Am1251 (2013)
Am1278 (2013)
Am1185 (2012)
Am1231 (2012)
Am1249 (2013)
Am233 (2010)
Am1072 (2011)
Am231 (2010)
Southern Ireland
Cork Harbor
Penzé estuary
Rance estuary
Bay of Brest
Bay of Concarneau
Western
France

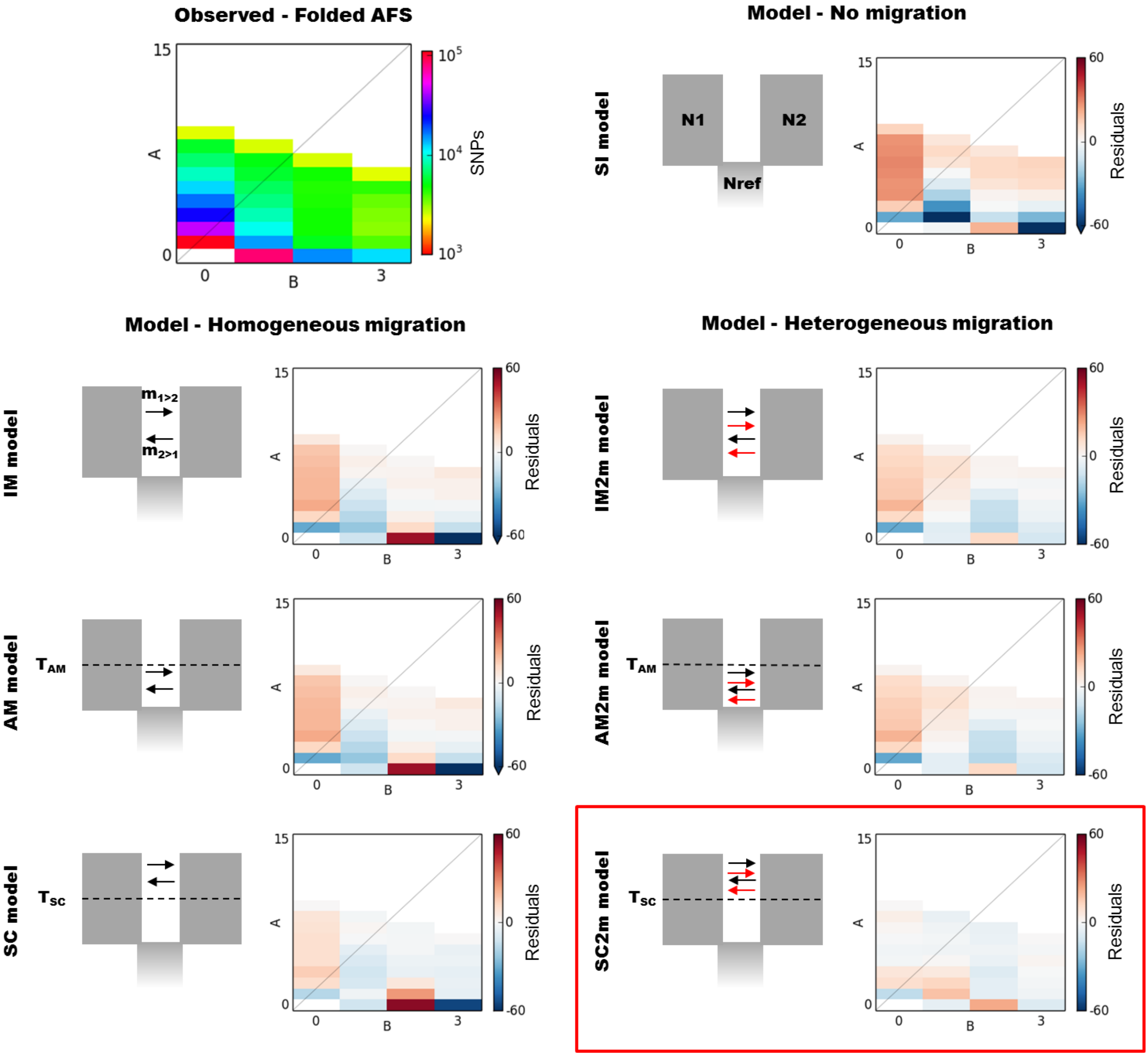
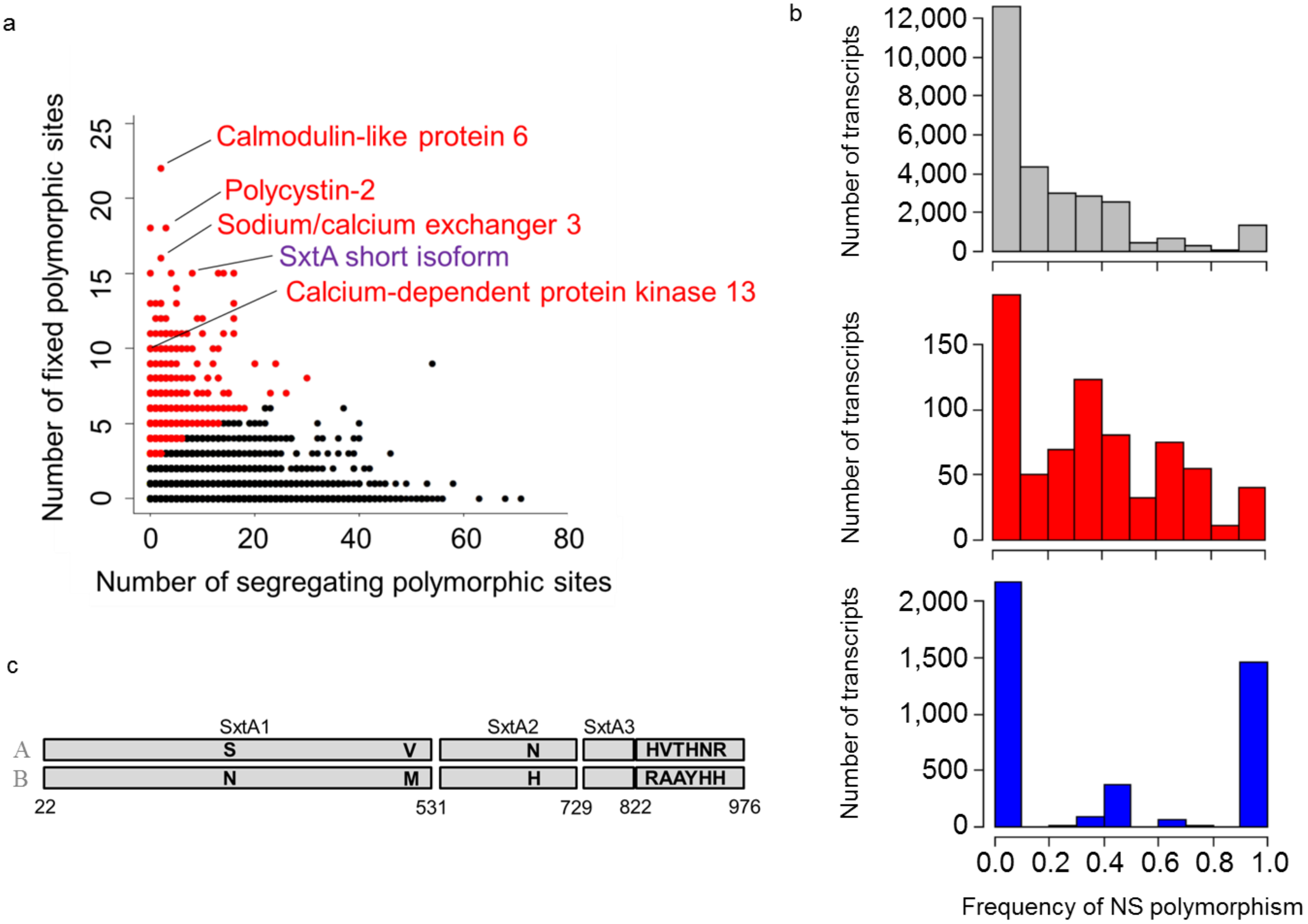
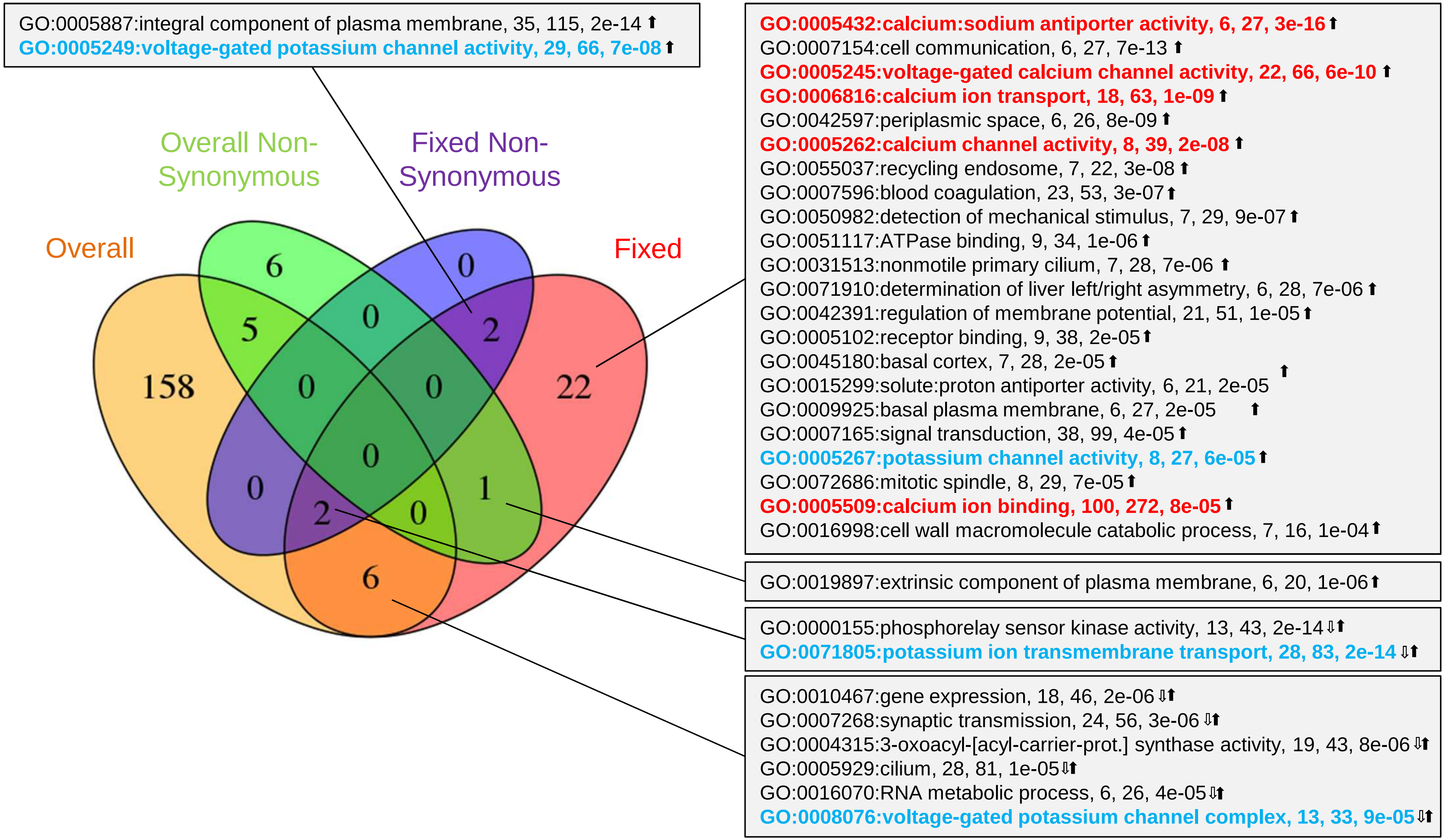


Figure 2: Results of model fitting for seven alternative models of divergence. The observed folded Allele Frequency Spectrum (AFS), as well as for each model, the residuals of the modeled AFS are presented. SI is the strict isolation model. IM is the Isolation-with-Migration model, AM the Ancient Migration model, and SC is the Secondary Contact model. All three models of divergence-with-gene-flow were implemented using one, shared migration rate in each direction (m1>2, m2>1) across the genome (homogeneous migration), or with two categories of migration rates in each direction across the genome (heterogeneous migration). The data are best explained by the SC2m model,


a
Calmodulin-like protein 6
Polycystin-2
Sodium/calcium exchanger 3
SxtA short isoform
Calcium-dependent protein kinase 13
Number of fixed polymorphic sites
0
5
10
15
20
25
0
20
40
60
80
Number of segregating polymorphic sites
b
Number of transcripts
12,000
10,000
8,000
6,000
4,000
2,000
0
Number of transcripts
150
100
50
0
Number of transcripts
2,000
1,500
1,000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
Frequency of NS polymorphism
c
SxtA1
SxtA2
SxtA3
A
S
V
N
HVTHNR
B
N
M
H
RAAYHH
22
531
729
822
976



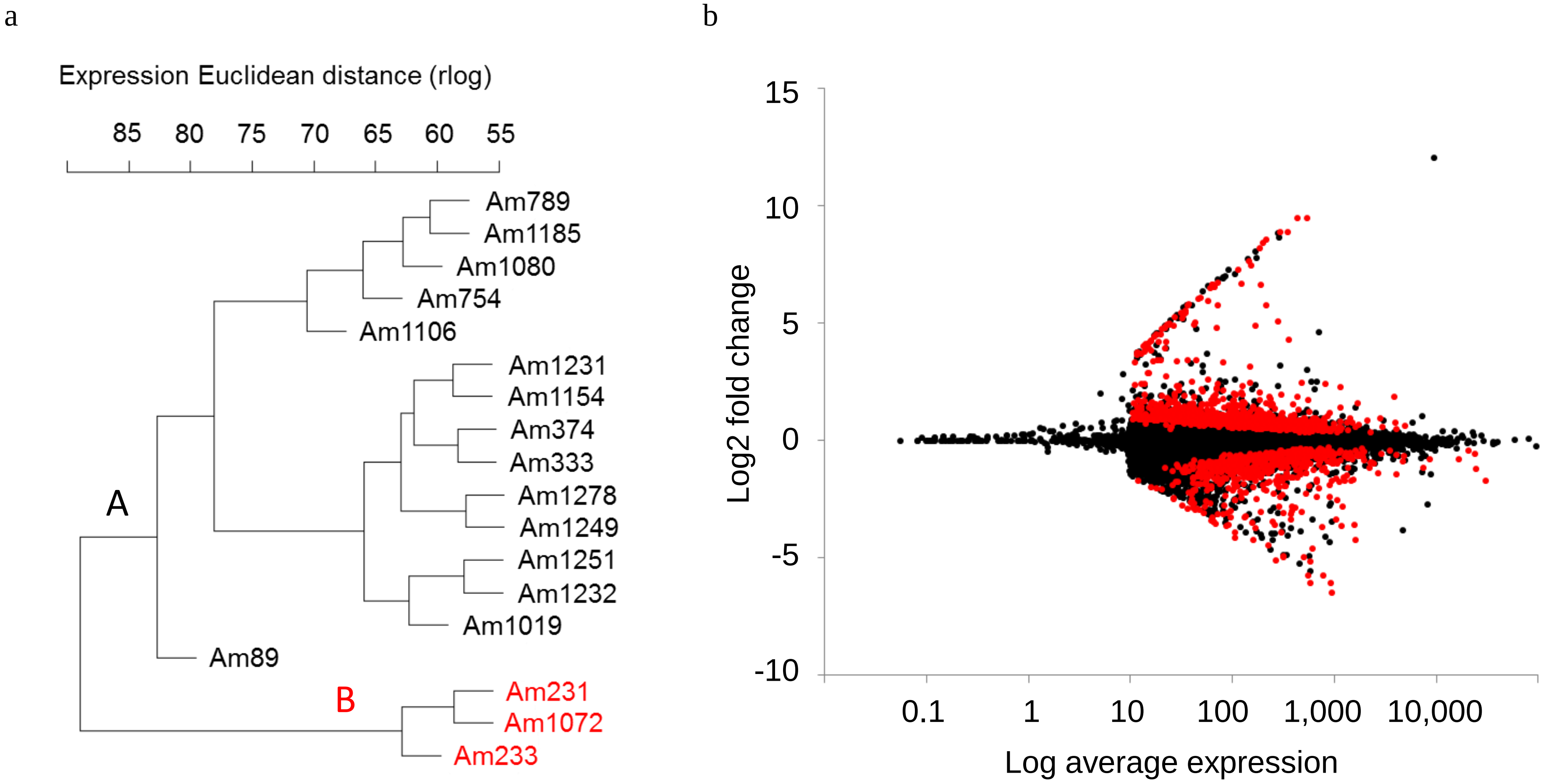
GO:0005887:integral component of plasma membrane, 35, 115, 2e-14 ⬆
GO:0005249:voltage-gated potassium channel activity, 29, 66, 7e-08 ⬆

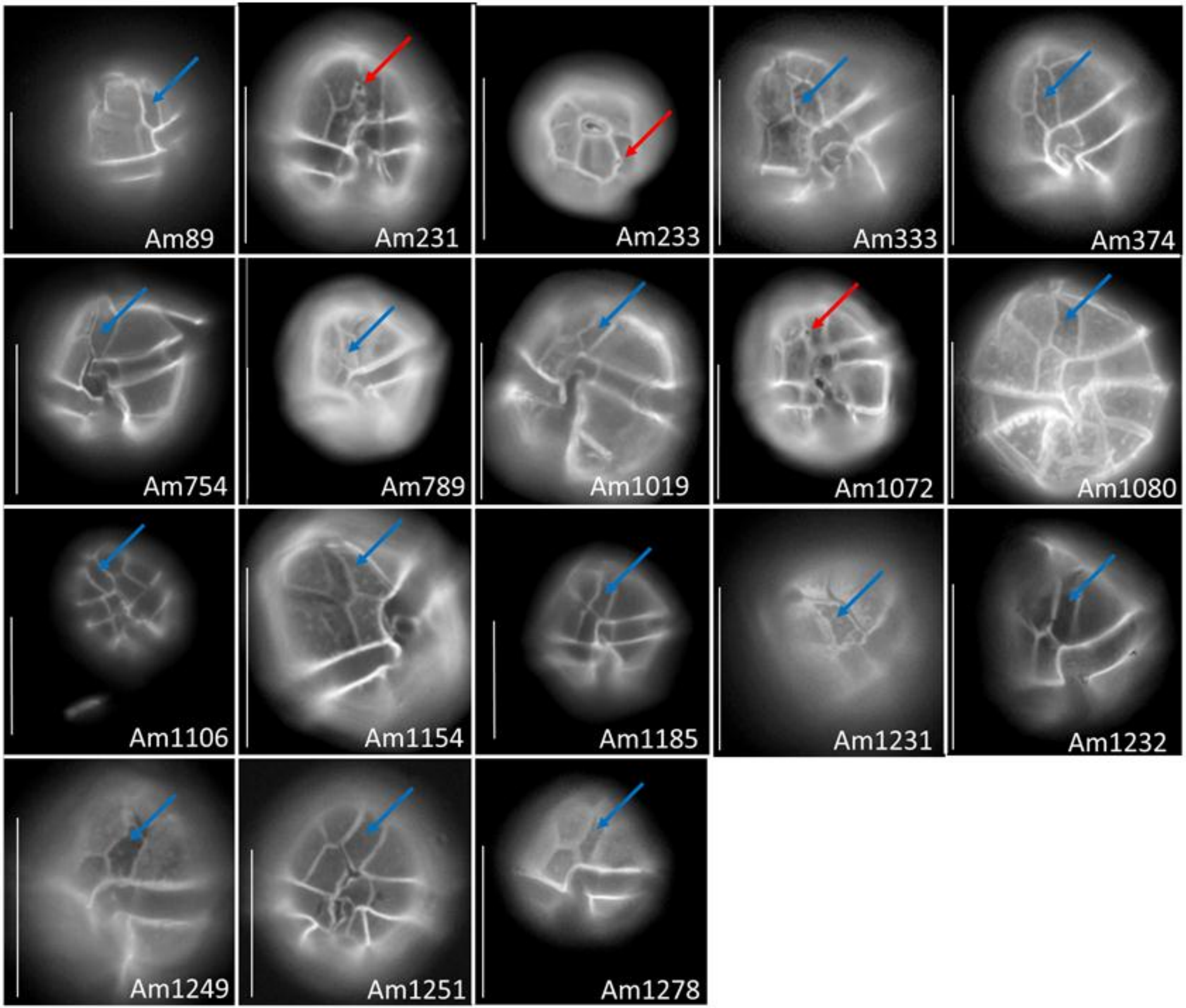
GO:0005432:calcium:sodium antiporter activity, 6, 27, 3e-16 ⬆
GO:0007154:cell communication, 6, 27, 7e-13 ⬆
GO:0005245:voltage-gated calcium channel activity, 22, 66, 6e-10 ⬆
GO:0006816:calcium ion transport, 18, 63, 1e-09 ⬆
GO:0042597:periplasmic space, 6, 26, 8e-09 ⬆
GO:0005262:calcium channel activity, 8, 39, 2e-08 ⬆
GO:0055037:recycling endosome, 7, 22, 3e-08 ⬆
GO:0007596:blood coagulation, 23, 53, 3e-07 ⬆
GO:0050982:detection of mechanical stimulus, 7, 29, 9e-07 ⬆
GO:0051117:ATPase binding, 9, 34, 1e-06 ⬆
GO:0031513:nonmotile primary cilium, 7, 28, 7e-06 ⬆
GO:0071910:determination of liver left/right asymmetry, 6, 28, 7e-06 ⬆
GO:0042391:regulation of membrane potential, 21, 51, 1e-05 ⬆
GO:0005102:receptor binding, 9, 38, 2e-05 ⬆
GO:0045180:basal cortex, 7, 28, 2e-05 ⬆
GO:0015299:solute:proton antiporter activity, 6, 21, 2e-05 ⬆
GO:0009925:basal plasma membrane, 6, 27, 2e-05 ⬆
GO:0007165:signal transduction, 38, 99, 4e-05 ⬆
GO:0005267:potassium channel activity, 8, 27, 6e-05 ⬆
GO:0072686:mitotic spindle, 8, 29, 7e-05 ⬆
GO:0005509:calcium ion binding, 100, 272, 8e-05 ⬆
GO:0016998:cell wall macromolecule catabolic process, 7, 16, 1e-04 ⬆

GO:0019897:extrinsic component of plasma membrane, 6, 20, 1e-06 ⬆

GO:0000155:phosphorelay sensor kinase activity, 13, 43, 2e-14 ⇩⬆
GO:0071805:potassium ion transmembrane transport, 28, 83, 2e-14 ⇩⬆

GO:0010467:gene expression, 18, 46, 2e-06 ⇩⬆
GO:0007268:synaptic transmission, 24, 56, 3e-06 ⇩⬆
GO:0004315:3-oxoacyl-[acyl-carrier-prot.] synthase activity, 19, 43, 8e-06 ⇩⬆
GO:0005929:cilium, 28, 81, 1e-05 ⇩⬆
GO:0016070:RNA metabolic process, 6, 26, 4e-05 ⇩⬆
GO:0008076:voltage-gated potassium channel complex, 13, 33, 9e-05 ⇩⬆


a
Expression Euclidean distance (rlog)
85
80
75
70
65
60
55
Am789
Am1185
Am1080
Am754
Am1106
Am1231
Am1154
Am374
Am333
Am1278
Am1249
Am1251
Am1232
Am1019
A
Am89
B
Am231
Am1072
Am233
b
15
10
5
0
-5
-10
Log2 fold change
0.1
1
10
100
1,000
10,000
Log average expression


Am89
Am231
Am233
Am333
Am374
Am754
Am789
Am1019
Am1072
Am1080
Am1106
Am1154
Am1185
Am1231
Am1232
Am1249
Am1251
Am1278

Supplementary Table 1: **Results of model fitting for seven alternative models of divergence**. *SI* is the strict isolation model. *IM* is the Isolation-with-Migration model, *AM* the Ancient Migration model, and *SC* is the Secondary Contact model. All three models of divergence-with-gene-flow were implemented using one, shared migration rate in each direction (*m*12, *m*21) across the genome (homogeneous migration), or with two categories of migration rates in each direction across the genome (heterogeneous migration).

Model	**k**	**MLE**	**AIC**	**Δi**	***L*(Mi|y)**	**Theta**	**nu1**	**nu2**	**m12**	**m21**	**me12**	**me21**	**Ts**	**Tps**	**P**
SI	4	-12154	24316	22172	0	110008	0.98	1.67	-	-	-	-	0.34	-	-
IM	6	-5974	11960	9816	0	48985	1.97	2.04	0.16	0.64	-	-	3.34	-	-
AM	7	-5971	11956	9812	0	41411	2.33	2.40	0.14	0.55	-	-	4.22	0,00	-
SC	7	-4372	8758	6614	0	78799	1.31	1.36	0.38	2.46	-	-	1.07	0.19	-
IM2M	9	-2082	4182	2038	0	41054	1.98	4.14	0.61	1.34	0.06	0.00	3.86	-	0,30
AM2M	10	-2112	4244	2100	0	39529	2.02	4.31	0.64	1.26	0.06	0.00	4.00	0.00	0.30
SC2M	10	-1062	2144	0	1	82530	0.8	2.3	6.09	6.01	0.26	0.00	0.94	0.10	0,43

k The number of free parameters in the model
MLE maximum likelihood estimate
AIC Akaike Information Criterion
Δi Difference in AIC between model i and the best model (*SC2M*)
***L*(Mi|y)** Relative likelihood of model i compared to the best model (*SC2M*)
Theta Theta parameter for the ancestral population before split ($\theta = 2N\text{ref}\mu$), with *N*ref being the effective size of the ancestral population, and μ the per-site mutation rate per generation.
nu1 The effective size of the A species relative to *N*ref
nu2 The effective size of the B species relative to *N*ref
m12 The neutral movement of genes from the B to the A lineage in units of *N*ref*m*2>1 generations
m21 The neutral movement of genes from the A to the B lineage in units of *N*ref*m*1>2 generations
me12 The effective migration rate of “genomic-island” genes from the B to the A lineage
me21 The effective migration rate of “genomic-island” genes from the A to the B lineage
Ts The time of split in units of *N*ref generations
Tps The time of migration stop (*AM* model) or start (*SC* model) post-split in units of *N*ref generations
P The proportion of the SNPs experiencing reduced effective migration rate

Supplementary Table 2: Results of model fitting for seven alternative models of divergence, using a single randomly chosen SNP per transcript. A total of 5 different subsets were tested. Abbreviation as in Supplementary Table 1.

Subset	Model	k	MLE	Theta	nu1	nu2	m12	m21	me12	me21	Ts	Tps	P
1	SI	4	-1731	5054	0.11	0.31	-	-	-	-	0.03	-	-
	IM	6	-1222	6299	0.17	0.38	3.75	4.99	-	-	0.36	-	-
	IM2M	9	-1014	1265	0.73	3.62	2.37	0.97	0.05	0.02	6.28	-	0.90
	AM	7	-1339	1046	1.07	3.87	0.76	0.49	-	-	9.97	0.00	-
	AM2M	10	-1731	5050	0.12	0.32	0.06	0.00	0.00	0.00	0.00	0.03	0.00
	SC	7	-607	921	1.11	0.59	0.69	4.68	-	-	14.89	0.35	-
	SC2M	**10**	**-379**	**728**	**0.91**	**1.39**	**1.55**	**7.98**	**0.09**	**0.00**	**17.85**	**0.23**	**0.95**
2	SI	4	-1701	5058	0.12	0.38	-	-	-	-	0.03	-	-
	IM	6	-1181	14411	0.07	0.16	9.06	14.98	-	-	0.34	-	-
	IM2M	9	-989	1215	0.60	3.97	2.72	0.93	0.01	0.01	7.53	-	0.95
	AM	7	-1382	907	1.39	4.80	0.65	0.42	-	-	9.46	0.00	-
	AM2M	10	-1701	5061	0.12	0.38	0.00	0.00	0.00	0.00	0.00	0.03	0.00
	SC	7	-635	788	1.62	1.39	0.55	3.04	-	-	14.33	0.47	
	SC2M	**10**	**-288**	**431**	**1.01**	**1.11**	**1.20**	**7.25**	**0.46**	**0.00**	**36.84**	**0.28**	**0.97**
3	SI	4	-1752	5076	0.10	0.32	-	-	-	-	0.03	-	-
	IM	6	-1352	4262	0.26	1.02	3.27	2.40	-	-	1.18	-	-
	IM2M	9	-1017	1214	0.54	3.87	2.96	0.79	0.01	0.01	7.77	-	0.96
	AM	7	-1507	914	1.87	3.81	0.43	0.50	-	-	9.95	0.01	
	AM2M	10	-1752	5077	0.10	0.32	0.09	0.02	0.00	0.00	0.00	0.03	0.00
	SC	7	-643	570	1.20	0.69	0.70	4.53	-	-	29.91	0.46	
	SC2M	**10**	**-372**	**539**	**0.60**	**0.99**	**1.97**	**11.20**	**0.30**	**0.00**	**28.33**	**0.16**	**0.96**
4	SI	4	-1620	5064	0.11	0.32	-	-	-	-	0.03	-	-
	IM	6	-1125	11891	0.08	0.19	7.97	12.40	-	-	0.35	-	-
	IM2M	9	-1066	966	1.44	3.68	1.01	0.62	0.01	0.01	8.85		0.92
	AM	7	-1202	1663	0.53	2.50	1.69	0.86	-	-	9.76	0.00	-
	AM2M	10	-1620	5063	0.11	0.32	0.00	0.00	0.00	0.00	0.00	0.03	0.00
	SC	7	-673	553	1.69	1.07	0.54	2.70	-	-	25.48	0.68	-
	SC2M	**10**	**-498**	**903**	**0.86**	**2.90**	**2.51**	**9.32**	**0.10**	**0.00**	**11.28**	**0.17**	**0.97**
5	SI	4	-1749	5126	0.10	0.34	-	-	-	-	0.03	-	-
	IM	6	-1186	9711	0.09	0.25	9.14	9.98	-	-	0.34	-	-
	IM2M	9	-1031	1265	0.52	3.86	2.65	0.77	0.01	0.01	7.64	-	0.96
	AM	7	-1319	1179	0.91	3.35	0.88	0.48	-	-	9.98	0.00	-
	AM2M	10	-1749	5117	0.11	0.36	0.22	0.37	0.00	0.00	0.00	0.03	0.05
	SC	7	-556	659	1.15	0.66	0.67	4.81			24.53	0.38	-
	SC2M	**10**	**-329**	**551**	**0.60**	**0.99**	**1.96**	**11.37**	**0.30**	**0.00**	**28.82**	**0.17**	**0.96**

Supplementary Material

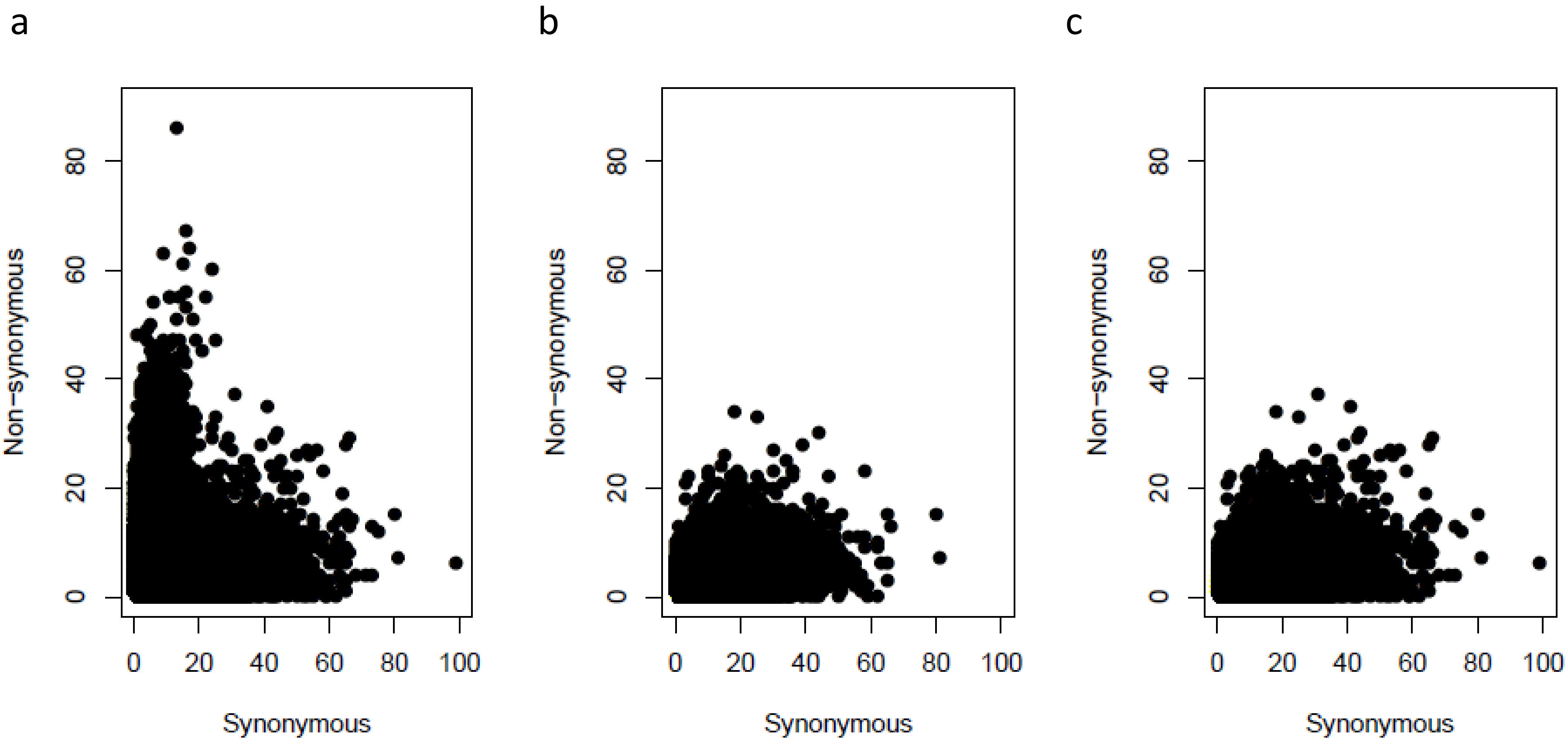


Fig. S1. Selecting the reading frame minimizing the proportion of non-synonymous mutations when several reading frames are possible. (A) considering all transcripts and all possible reading frames, (B) only considering transcripts with a single possible reading frame, (C) considering all transcript and the reading frame minimizing the number of non-synonymous mutations.