# Genome polarisation for detecting barriers to geneflow

Stuart J. E. Baird[1,*], Jan Petružela[1,2], Izar Jaroň[1], Pavel Škrabánek[3], and Natália Martínková[1,4,*]

[1]Institute of Vertebrate Biology, Czech Academy of Sciences, Květná 8, Brno, 60365, Czechia
[2]Department of Botany and Zoology, Masaryk University, Kotlářska 2, Brno, 62500, Czechia
[3]Institute of Automation and Computer Science, Brno University of Technology, Technická 2896/2, Brno, 61669, Czechia
[4]RECETOX, Masaryk University, Kotlarska 2, Brno, 62500, Czechia
[*]Correspondence: stuartj.e.baird@gmail.com, martinkova@ivb.cz

**Abstract**

Semi-permeable barriers to geneflow in principle allow distantly related organisms to capture and exchange pre-adapted genes and gene clusters through hybridisation. However, describing barriers to geneflow on a genomic scale is non-trivial. We extend classic diagnostic allele counting measures of geneflow across a barrier to the case of genome-scale data. Using a deterministic data labelling, small numbers of diagnostic markers can be replaced by large numbers of markers, each with a diagnostic index. Individuals' hybrid indices (genome admixture proportions) can then be calculated genome wide conditioned on marker diagnosticity; within diploid, haplodiploid and/or haploid genome compartments; or indeed over any subset of markers, allowing classical cline width/barrier strength comparisons along genomes. Further, blocks of genetic material that have introgressed across a barrier can be easily identified. Diagnostic index expectation maximisation (*diem*) polarises the labelling of bistate markers with respect to the sides of a barrier. An initial state of ignorance is enforced by starting with randomly generated marker polarisations. This means there is no prior on population or taxon membership of the genomes concerned. We suggest this simple property has wide reaching positive consequences for rigorous genome scale inference.

Divergence between taxa generates genome-wide correlations in SNP states (marker alleles). Geneflow across a barrier between taxa can be measured by counting alleles belonging on side A that have crossed to side B. The proportion of side A alleles an individual carries is its hybrid index; hybrid indices form clines across barriers from which we can estimate barrier strength (Barton 1979). Classically, reference sets of 'pure' individuals from each side of a barrier were used to arbitrate where alleles belong: a marker fixed for different states in each reference set is

diagnostic; which of its alleles belong on either side is considered proven. In modern terms, the use of diagnostic markers is reduced representation genomics favouring sites that enable geneflow to be detected. With genome scale sequencing we might hope to add many markers to a diagnostic set, but when adding a marker, a decision must be made regarding where its alleles belong.

While use of diagnostic markers has been described as introducing a bias against detection of geneflow (Gompert *et al.* 2017), this bias results only from declaring a priori that reference sets of individuals are 'pure'. The critical decision regarding which side of a barrier an allele belongs does not however require reference sets. For example, when Falush's linkage model is used within the STRUCTURE inference framework (Falush *et al.* 2003a,b), the SITEBYSITE option produces a separate output file that contains posterior population of origin assignments for each allele copy at each locus for each individual (Pritchard *et al.* 2010). This is not, however, a scalable solution, as convergence times under the linkage model increase rapidly with number of sites, even at $K = 2$. fineSTRUCTURE (Lawson *et al.* 2012) is more computationally efficient, but unlike STRUCTURE, requires declaration of 'pure' reference sets, and so re-introduces a bias against detection of geneflow.

Genome painting where alleles belong is difficult and computationally intensive using existing tools. Here we explore how reduced representation genomics for allele counting across a barrier can be adapted to high throughput sequencing datasets, and what measures characterising markers might be useful. We let correlation across the entire dataset arbitrate the allele/barrier-side decision. This is a data labelling task, and so open to expectation maximisation (EM) (Dempster *et al.* 1977).

## Basic concepts

Consider a marker with two common states in a set of genomes: 'A' and 'T'. If there is a barrier to geneflow within the genomes set, 'A' may be associated with genomes on one side, 'T' with those on the other side. Which way round these associations go is the *polarity* of the marker with respect to the barrier, and the evidential support for this decision can be calculated. We can represent an initial state of ignorance by using a random number generator to assign polarities to all markers. This is a *null polarisation*, and is equivalent to saying we work with no prior on population or taxon membership of the genomes. A null polarisation does not, however, remove information regarding associations between markers. If divergence has generated genome wide correlations in state, then a second marker with common states 'C' and 'G' will show 'A' genomes tend to have 'C' while 'T' genomes have 'C'. Of course, the converse can be true. A null polarisation by definition will get about half these cross-locus associations correct. This is sufficient signal that some of the remaining markers can be given the matching polarisation. The diagnostic index expectation maximisation (*diem*) algorithm iteratively co-polarises remaining markers. This process accelerates as further signal of genome-wide association is pulled out of

the data by an increasingly correct labelling of markers by barrier side. For a given initial null polarisation, *diem* will always halt at the same final marker labelling. These labelings can be used directly to visualise the nature of a barrier, and the labelled data is ideal for numerous downstream applications in population genomics.

While these basic concepts can be explained simply, questions remain: Can the algorithm be reliably implemented? What happens if there is no barrier within the set of genomes? How does computation time scale? Over what types of data will it be useful? We address these questions by applying the algorithm to two publicly available datasets.

# Results

We demonstrate *diem* on publicly available datasets: 31,074 SNPs from 115 bat genomes, and 38 million markers in 69 individuals for a genome barrier between species of mice. The bat data was chosen because current taxonomy, phylogeography and behavioural observation suggest there should be no barrier within this set of genomes, while the size of the dataset is representative of standard reduced representation genomics.

The mouse system on which we demonstrate *diem* was chosen for a number of reasons. First, Sanger sequencing was used to show that the worldwide human commensal *Mus musculus domesticus* had captured a warfarin (pesticide) resistance gene, *Vkorc1*, from a non-commensal mouse species with ∼5% divergence, *Mus spretus* (Orth *et al.* 2002; Song *et al.* 2011). Second, genomic and transcriptomic high throughput sequencing (HTS) data have been made publicly available for this system for a decade. third, the number of *domesticus/spretus* variable sites (38 million) exceeds the number of sites expected to vary between humans and Neanderthals (∼15 million). Demonstrating that *diem* scales to this mouse system means the method will scale to popular issues of hominid genome admixture. Fourth, the *domesticus/spretus* HTS data was recently analysed by co-authors of the original Sanger sequencing work (Banker *et al.* 2022).

## Bat SNPs dataset

Polarisation made little visual labelling difference in comparison to null in the bats dataset (Fig. 1A vs B). Pre-polarisation, hybrid indices were uniform $\approx 0.5$ (Fig. 1C) as expected for a random null. Post-polarisation, hybrid indices run from 0 to $\approx 0.2$ (Fig. 1C). Having found optimal marker polarities using *diem*, we conclude there is little or no evidence for a barrier to geneflow; These results are consistent with the current taxonomy of the group, and despite geographically widespread sampling. This is an important feature of *diem*: unlike Structure (Pritchard *et al.* 2010), the $K = 1$ (no barrier) and $K = 2$ (barrier) cases can be distinguished. The *diem* labelling of the bat dataset consists of a homogeneous set of genomes with five outliers. The outliers have large numbers of unknown marker states (Fig. 1B). These states do not affect

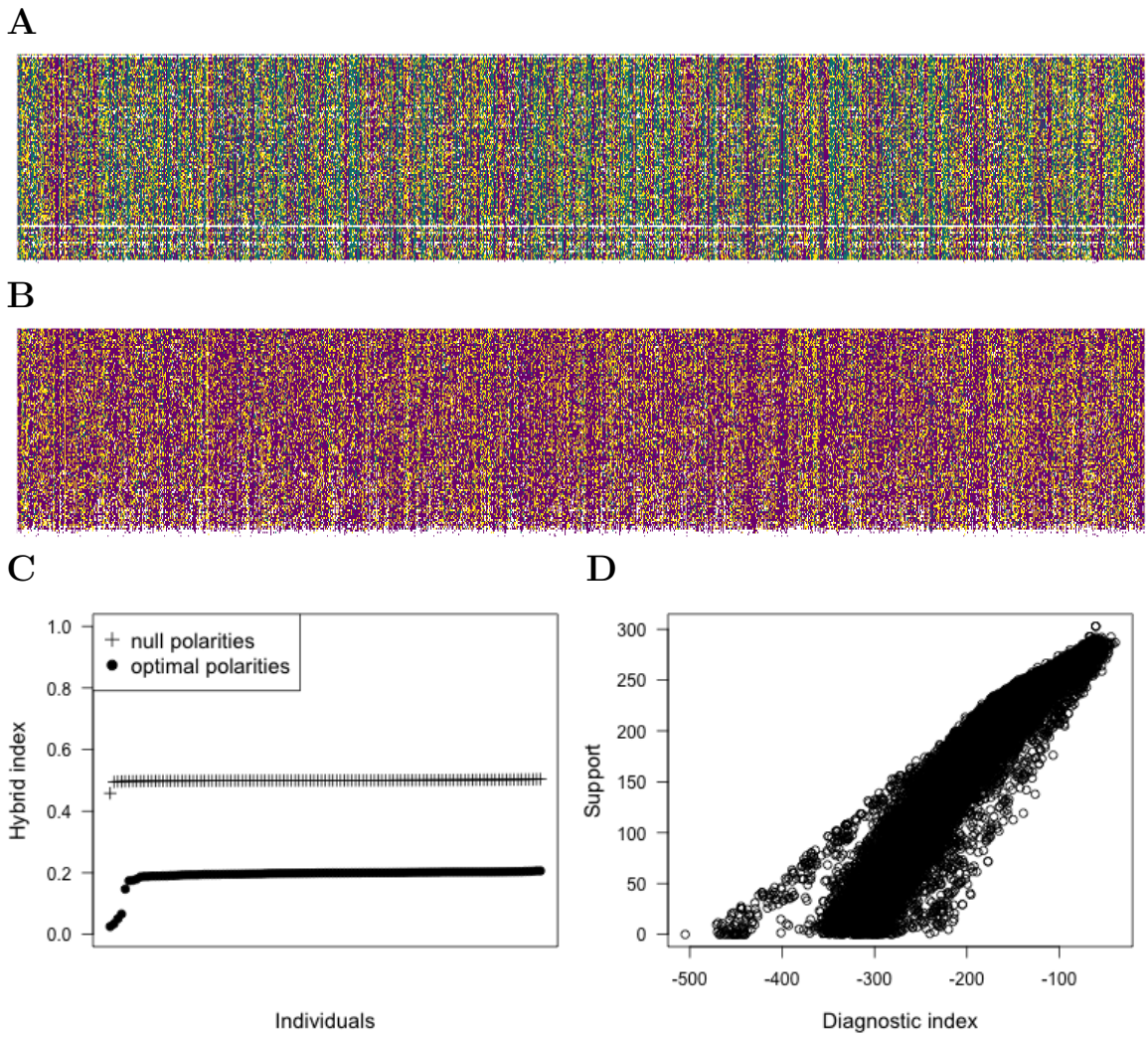


Figure 1: Genotypes of 31,074 SNP markers (columns) in 115 *Myotis myotis* bats (rows) (Harazim *et al.* 2021) with random null polarities (A), and with optimal polarities (B). Individuals in panel B are ordered according to their hybrid index. Purple – homozygotes for the allele attributed to barrier side A, green – homozygotes for the allele attributed to barrier side B, yellow – heterozygotes, white – unknown genotype state. C) Ordered individual hybrid indices for genotypes displayed in panels A and B. The optimal marker polarisation (circles) shows that the dataset consists of genetically similar individuals with the exception of five individuals with large proportion of unknown genotypes. D) Markers with optimal polarities will result in high diagnostic indices, and markers close to diagnostic markers will have high support.

*diem* polarisation, rather, they co-occur with a shared homozygous marker pattern. A plot of diagnostic indices and their polarisation supports shows a (typical) shark-fin shape, indicating that markers with high diagnostic index (likely given the data) also have high support for their polarisation (Fig. 1D).

## Mouse whole-genome dataset

The visualization of the mouse genomes highlights the difference between individuals sequenced for whole genomes (horizontal blocks) and for transcriptomes (sparse data in vertical blocks above

and below the whole-genome sequences) (Fig. 2A). The polarized genomic data (Fig. 2B) differentiate individuals with respect to the barrier to geneflow resulting in hybrid indices between 0.02 and 0.93 (Fig. 2C). The Linnean binomial taxon labels separate at the steepest change in the hybrid index. In Fig. 2B, we show the introgressed region, where two *M. musculus domesticus* individuals carry the introgressed block previously identified by Sanger sequencing (Song *et al.* 2011). The introgression present in two individuals cannot be distinguished in a genomic scan correlating marker state to the individual hybrid index (Fig. 2D), but is obvious in the visualization of the polarised genomes (Fig. 2B).

## Implementation validation

Using identical initial null polarities, we found 0 differences in converged *diem* marker polarities in the bats and mice datasets as implemented in R, Mathematica, and Python. As the *diem* algorithm is deterministic, alternate implementations should give the same output on the same input. The diemr vs diemmca, and diempy vs diemmca, implementations were cross-validated in this way on multiple datasets, including large. Finding no output differences, the chance of coding error in any implementation is then the chance that all have coding errors with perfectly matched effect despite idiomatic translation into different languages. While the algorithm is concise, it is of sufficient complexity that accidental perfect error matching of this sort seems far fetched.

## Benchmarking

We benchmarked the *diem* algorithm implemented in diemmca on a MacBook Pro (Retina, 15-inch, Mid 2015), processor 2.8 GHz Quad-Core Intel Core i7, memory 16 GB 1600 MHz DDR3, 1Tb flash storage and diempy on a computational cluster (MetaCentrum). We used the full mouse dataset (all reference scaffolds), thinned at random to 10 reduced levels and ran *diem* for up to ten null initial polarisations for each level. The results show that the algorithm scales linearly with $n$ markers (Fig. 3, Supplemental Table S1), with further speed-up over multiple cores, and is thus suitable for genome-scale analyses.

## Comparison with existing methods

We compare *diem* to two methods inspired by phylogenetics (Banker *et al.* 2022) and population genetics (Falush *et al.* 2003a,b). We used partial mouse chromosome 7 region that is known to include a several-Mbp-long *spretus* introgression into a *domesticus* genome (Song *et al.* 2011).

Banker *et al.* (2022) used a sliding window divergence-metric analysis to detect *domesticus/spretus* introgression for a subset of the ENA accessions we examine here. While we align genomes under a single reference (GRCm39, accession: GCA_000001635.9, assembled using long

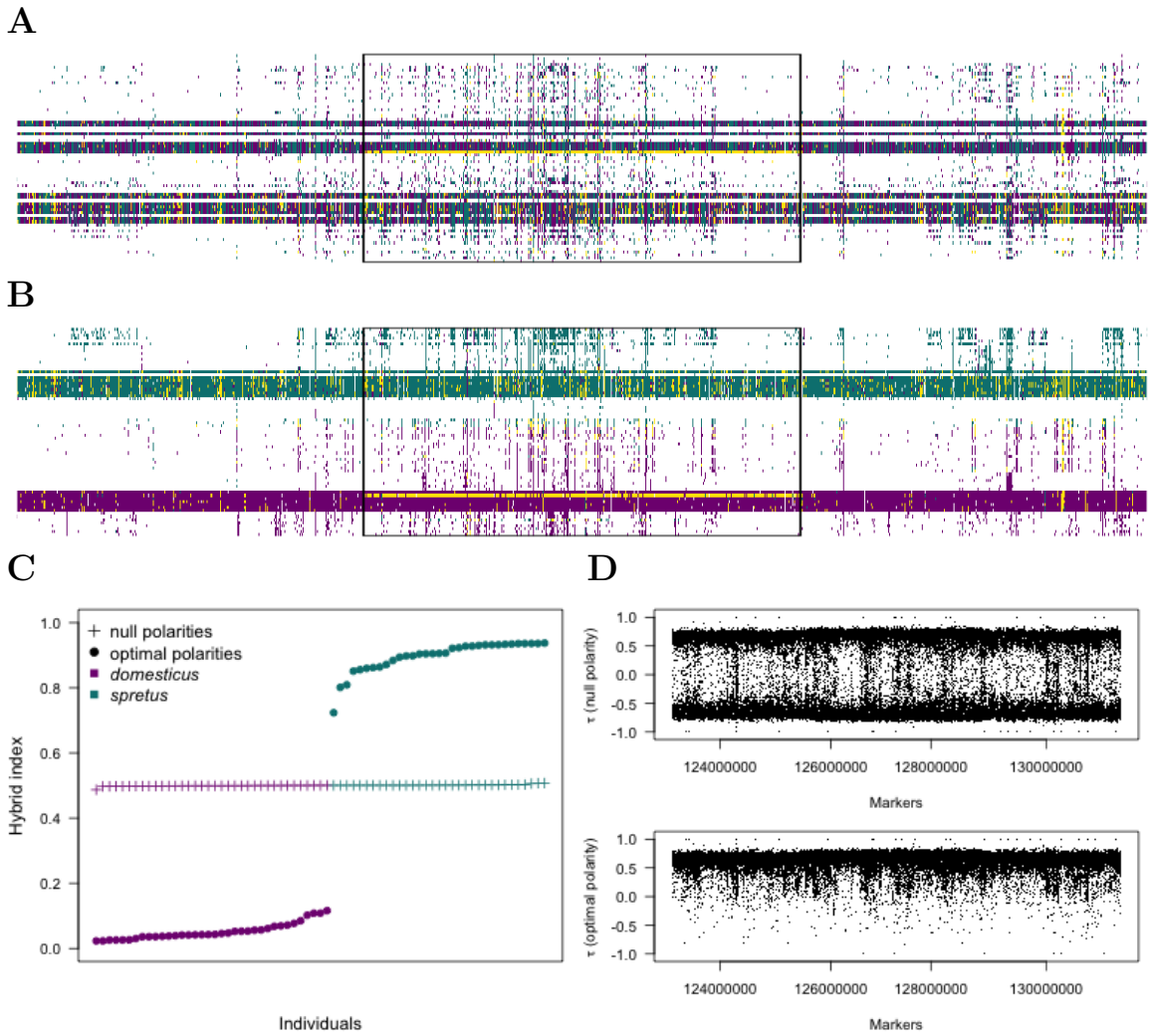


Figure 2: Genotypes of 117,403 markers (columns) on chromosome 7, positions 123,189,278–131,234,739 in 69 mice (rows) with random null polarities (A), and with optimal polarities (B). Individuals in panel B are ordered according to their genome-wide hybrid index. The rectangle delimitates the genomic block carrying warfarin resistance that includes *Vkorc1* introgressed from *M. spretus* to *M. musculus domesticus* (Song *et al.* 2011). Two individuals have the introgressed block in our dataset. One is predominantly heterozygous, and the other is homozygous. Purple – homozygotes for the allele attributed to barrier side A (*domesticus*), green – homozygotes for the allele attributed to barrier side B (*spretus*), yellow – heterozygotes, white – unknown genotype state. C) Ordered individual hybrid indices for whole-genome markers in mice. The optimal marker polarisation (circles) shows that the dataset consists of two main groups of genetically similar individuals. D) Kendall's $\tau$ correlation between hybrid indices and marker states in the genomic region displayed in panels A and B. The correlation coefficient is bimodal for markers with null polarity (upper panel) and unimodal for markers with optimal polarity with respect to barrier to geneflow (lower panel). Introgression in two individuals is not apparent in this genome region scan.

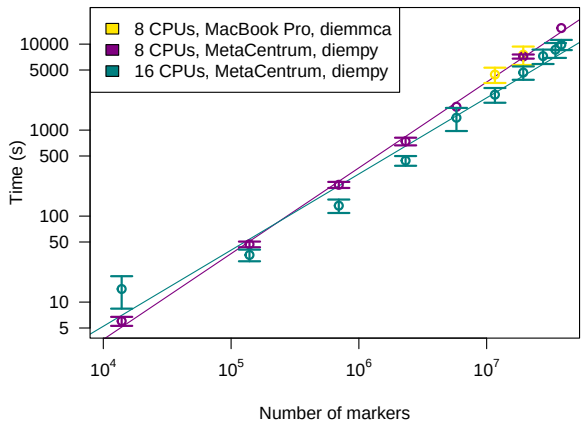


Figure 3: The *diem* computation time scales linearly with number of markers, and is reduced by making more CPUs available. Time trials were run on nested thinnings of the full mouse marker set, with replicates at each thinning level. Points and whiskers show mean ± SD in the thinning replicates. Yellow – absolute time of diemmca run on an 8 CPU Macbook Pro, purple (8 CPUs) and green (16 CPUs) – PBS monitor 'job time' for diempy jobs run on MetaCentrum computational cluster nodes (see Supplemental Table S1 for details), purple line – linear fit to all 8 CPU datapoints, irrespective of hardware, green line – linear fit to all 16 CPU datapoints, irrespective of hardware. Gradients $1.0728 \times 10^{-7}$ and $7.0277 \times 10^{-8}$ hrs/marker, respectively. Doubling CPU number gives $\sim$ 1.5 fold speed-up. diemmca and diempy results are identical on identical input. The Python version (diempy) currently requires more RAM resources due to differences in memory management between Mathematica kernels and Python parallel processes.

reads), the Banker *et al.* (2022) study involves distances measured relative to two different references: GRCm38 (aka mm10, short reads, proxy for *domesticus*) and GCA_001624865.1 (short reads, proxy for *spretus*). Direct comparison of the approaches' output is made possible by NCBI Genome Remapping Service (https://www.ncbi.nlm.nih.gov/genome/tools/remap), which allows us to exchange Banker *et al.* (2022) introgression locations in their Supplementary Table 3 very precisely between reference accessions. Figure 4 overlays these findings onto the *diem* results in the *Vkorc1* region.

The major axis of difference between the studies is censorship of sites inspected. We suggest this is the effect mentioned in our Introduction: a bias against detection of geneflow results from declaring a priori that reference sets of individuals are 'pure' (Gompert *et al.* 2017). Banker *et al.* (2022) only inspect introgression into *spretus* at sites where selected individuals a priori labelled *domesticus* have invariant state; conversely, they only inspect introgression into *domesticus* at sites where individuals a priori labelled *spretus* have an invariant state. Therefore, all sites with intra-taxon variation are censored; All sites with bi-directional introgression are censored. From Figure 4 it appears this censorship leaves very few sites with which to measure sequence divergence. In contrast, *diem* has zero prior on the membership of genomes in taxa. Sites invariant across all genomes are censored, but this leaves large absolute numbers of variable sites (genome size × divergence) that can be inspected for introgression with an approach that is scalable over $n$ markers (variable sites).

The STRUCTURE linkage model with the SITEBYSITE option (Falush *et al.* 2003a,b; Pritchard

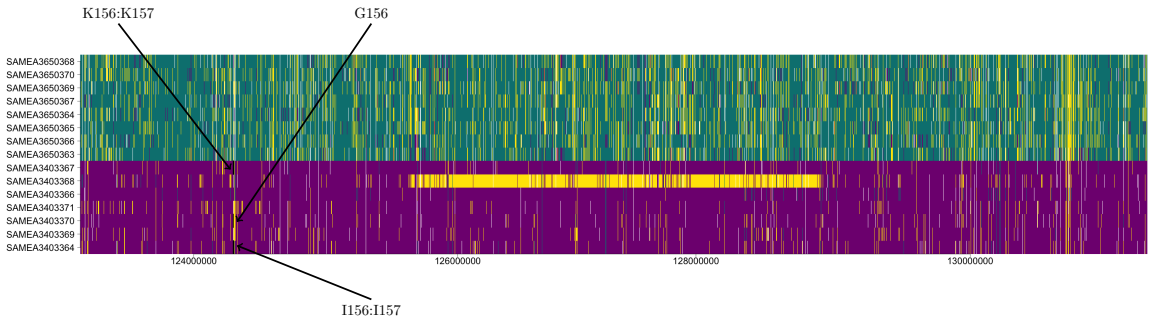


Figure 4: Comparison of *diem* genome polarization with a sequence-divergence based method to detect introgression (Banker *et al.* 2022). Partial mouse chromosome 7 genome polarisation corresponds to that in Fig. 2B, selecting individuals that were analysed both here and in Banker *et al.* (2022). Arrows point to black boxes delimiting introgressed regions identified in the respective cells of Suppementary Table 3 in Banker *et al.* (2022). The large, yellow block in sample SAMEA3403368 matches the introgression identified by Song *et al.* (2011) using Sanger sequencing.

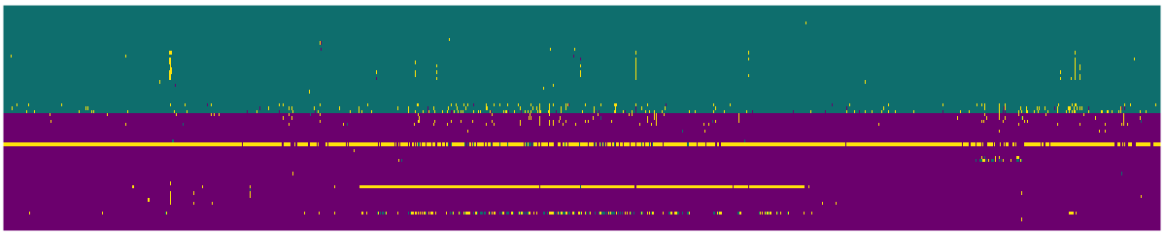

Figure 5: Genotype calls of mice from STRUCTURE SITEBYSITE linkage model. The individual ordering (rows) and colour encoding is identical to those in Fig. 2B. The horizontal, predominantly yellow line is an individual with accession number SAMEA3678857 that has been imputed as heterozygous by STRUCTURE (the transcriptomic data is very sparse in Fig. 2).

*et al.* 2010) clusters individuals while allowing for admixture or hybridization between linked markers. After 10,000 generations following a 5,000-generation burnin, the posterior allele assignment to *domesticus*/*spretus* statistically matches the *diem* result (G-test: $G = 2646211$, $\chi^2$ $df = 6$, $p < 0.001$; Fig. 5).

The major axis of difference between *diem* and STRUCTURE is that STRUCTURE imputes state whereas *diem* only polarises it. We compare STRUCTURE to *diem* using a contingency table summarising polarised vs imputed state for an interval of mouse chromosome 7 (Table 1). We see STRUCTURE imputation is consistent with *diem* homozygous states, but STRUCTURE overwrites unknown input states and tends to re-assign input heterozygous (1) state as homozygous (2). We can mimic this last aspect of the linkage model STRUCTURE imputation by re-calling all input states according only to the products of marginal genotype probabilities of individual genotypes and markers from genotypes polarised by *diem* (Table 1, *diem*-imputed). We see that imputation from *diem* marginals reproduces the tendency to re-assign heterozygous (1) input state as homozygous (2) in this dataset. We observe that the major difference between the *diem* marginals-only imputation and STRUCTURE linkage model imputation is a relative excess of STRUCTURE heterozygous calls. This is most noticeable for the individual with accession

Table 1: Comparison of *diem*-polarized genotypes of mice with STRUCTURE SITEBYSITE linkage model genotypes in the partial chromosome 7 (cf. Fig. 2).

| | STRUCTURE | | | *diem*-imputed | | |
|---|---|---|---|---|---|---|
| *diem* polarisation | *domesticus* (0) | heterozygous (1) | *spretus* (2) | *domesticus* (0) | heterozygous (1) | *spretus* (2) |
| Undetermined (U) | 3,056,099 | 132,580 | 2,728,672 | 3,228,406 | 35,881 | 2,653,064 |
| *domesticus* (0) | 950,357 | 5,231 | 33,551 | 954,643 | 1,607 | 32,889 |
| heterozygous (1) | 30,370 | 55,433 | 122,237 | 76,070 | 24,048 | 107,922 |
| *spretus* (2) | 8,816 | 1,727 | 975,734 | 12,721 | 1,386 | 972,170 |

SAMEA3678857, which has majority input state homozygous *domesticus* (Fig. 2B), but after STRUCTURE imputation is majority state heterozygous (Fig. 5).

The mouse introgressions highlighted across the genome by *diem* are bidirectional (Fig. 6, Table 2). While *M. musculus domesticus* gained a large introgressed block on chromosome 7 carrying resistance to warfarin, introgressions to *M. spretus* genome include genomic blocks on multiple chromosomes with genes involved in chemical sensing.

# Discussion

Genome polarisation dissolves the concept of populations: they exist nowhere in the *diem* algorithm. This might seem surprising for a method inspired by population genetics, but allowing for continuous change in genotypes in fact dates back to the neo-Darwinian synthesis (Haldane 1948), while inferring barriers to geneflow from abrupt change in correlated genotypes (Barton 1979) lies at the base of (arguably) the most replicable and productive treatment of biological field data: hybrid zone analysis. We have shown that the *diem* algorithm is robust in detecting a single homogeneous set of individuals (Fig. 1) and a barrier between two taxa (Fig. 2), including detection of introgressed blocks (Figs. 2B, 6), *diem* allows analyses of samples with mixed levels of reduced genome representation, the implementation scales to genome-size data (Fig. 3), and diemr, diempy and diemmca provide matching solutions that make the approach available to a wide potential userbase. The *diem* approach lies upstream of well established analyses in population genetics, empowering them with thousands or millions of informatively labelled markers. It also enables phylogenetic analyses: from a *diem* polarisation, genome regions homozygous for no admixture can be identified, greatly aiding short read assembly. *diem* naturally extends to long read data. As with short read data, within-read phasing is ignored, and we look forward with interest to *diem* analyses over mixed short/long read sequencing. The *diem* approach extends to higher ploidy, though above ploidy 2 the heterozygous state is approximate.

Polarising genomes allows us to study a continuous genomic space, investigating barriers to geneflow and the capture of genetic material at a regional scale. An emphasis on sequences (along-genome summaries) in previous phylogenetic-inspired approaches is replaced by an emphasis on markers (across-genome summaries) for population genomics. Perhaps most promisingly, *diem* informs hypothesis formation at the genome scale. For example, we can formulate a new hypoth-

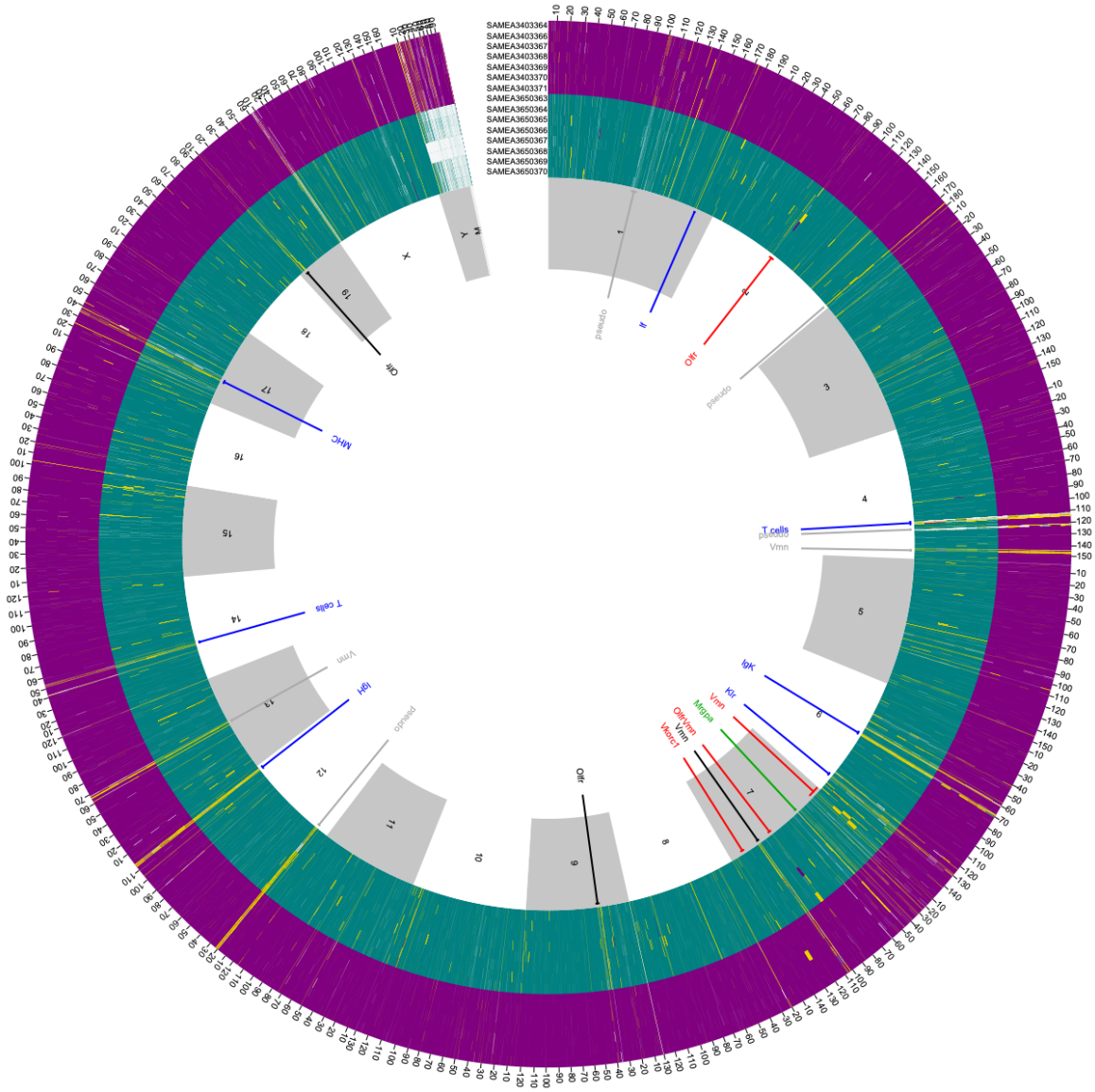


Figure 6: Visualising a mouse genome barrier: 38 million variable sites aligned to the GRCm39 reference and coloured using *diem* output. Purple – *M. musculus domesticus*, green – *M. spretus*, yellow – heterozygous or overmerged onto the reference, white – no call. Of the full *diem* analysis, only the fifteen ENA whole-genome submissions (accession numbers top, centre) are shown, ordered by their *diem*-estimated hybrid index. The transition from purple to green genomes is sharp, suggesting the genome barrier is strong. Inner grey/white ring – reference chromosome, outer ticks – reference chromosome position in Mb. Central annotation, major features: Red – regions of introgression including *Vkorc1*; the *M. spretus* version of this gene provides pesticide resistance to *M. musculus domesticus* carriers: SAMEA3403368 is a heterozygous carrier, *Vkorc1* being embedded in a ∼3.4Mb long yellow block on purple background, demonstrating that the genome barrier is permeable. All other major introgressions are from *M. musculus domesticus* to *M. spretus* and occur in chemosensory gene clusters (Vmn: Vomeronasal, Olfr: Olfactory). Remaining annotated features: yellow – false heterozygous state due to overmerging, blue – wild mouse variation in immune gene clusters, and chemosensory gene clusters overmerges onto the GRCm39 reference, assembled with long reads from the lab mouse strain C57BL/6J, inbred for many hundreds of generations. Overmerged regions are associated with high pseudogenic C57BL/6J annotation density, sometimes to the extent that entire gene clusters are pseudogenic (grey). Table 2 details the annotated regions.

| Reference Chr | Annotation feature | Annotation colour | Reference region description | Region start | Region end | Region length | Region state | N features | protein coding gene | poly. pseudogene | pseudogene | gene segment | % interval with feature |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chr1 | pseudo | Grey | hotspot of pseudogenisation | 99,566,121 | 103,351,383 | 3,785,263 | overmerged | 19 | 0 | 0 | 19 | 0 | 11.01 |
| Chr1 | If | Blue | Interferon activated genes (innate immunity) (Ifi) | 173,296,051 | 173,859,376 | 563,326 | overmerged | 12 | 12 | 0 | 0 | 0 | 52.91 |
| Chr2 | Olfr | Red | Olfactory (Olfr) gene cluster with introgression | 86,001,215 | 90,232,018 | 4,230,804 | introgressed | 228 | 191 | 0 | 37 | 0 | 7.05 |
| Chr2 | Olfr | | SAMEA3650366 HET DOM–>SPRET introgression (39561 markers) | 87,399,047 | 90,101,056 | 2,702,010 | introgressed | | | | | | |
| Chr2 | Olfr | | SAMEA3650367 HOM DOM–>SPRET introgression (14902 markers) | 87,727,609 | 88,717,904 | 990,296 | introgressed | | | | | | |
| Chr2 | pseudo | Grey | hotspot of pseudogenisation | 176,745,857 | 176,920,756 | 174,900 | overmerged | 7 | 0 | 0 | 7 | 0 | 12.01 |
| Chr4 | T cells | Blue | Selection and upkeep of intraepithelial T cells (Skint) | 111,776,643 | 114,102,225 | 2,325,583 | overmerged | 11 | 11 | 0 | 0 | 0 | 63.2 |
| Chr4 | pseudo | Grey | hotspot of pseudogenisation | 120,971,986 | 121,336,647 | 364,662 | overmerged | 9 | 0 | 0 | 9 | 0 | 2.63 |
| Chr4 | Vmn | Grey | Vomeronasal gene cluster (All C57BL/6J pseudogenised) (Vmn) | 145,502,182 | 147,606,271 | 2,104,090 | overmerged | 6 | 0 | 0 | 6 | 0 | 0.09 |
| Chr6 | IgK | Blue | Immunoglobulin kappa chain variable (Igkv) | 67,532,620 | 70,703,738 | 3,171,119 | overmerged | 168 | 0 | 0 | 0 | 108 | 2.62 |
| Chr6 | Klr | Blue | Klr killer cell lectin–like receptor | 128,586,190 | 131,224,325 | 2,638,136 | overmerged | 30 | 26 | 0 | 4 | 0 | 17.06 |
| Chr7 | Vmn | Red | Vomeronasal (Vmn) gene cluster with introgression (internal Olfr cluster has NOT introgressed) | 4,775,372 | 19,967,753 | 15,192,382 | introgressed | 141 | 91 | 0 | 50 | 0 | 6.7 |
| Chr7 | Vmn | | SAMEA3650367 HET DOM–>SPRET introgression (35745 markers) | 3,354,799 | 6,330,489 | 2,975,691 | introgressed | | | | | | |
| Chr7 | Vmn | | SAMEA3650368 HET DOM–>SPRET introgression (37467 markers) | 6,990,142 | 10,775,605 | 3,785,464 | introgressed | | | | | | |
| Chr7 | Vmn | | SAMEA3650366 HET DOM–>SPRET introgression (34916 markers) | 7,206,868 | 10,782,768 | 3,575,901 | introgressed | | | | | | |
| Chr7 | Mrgpa | Green | MAS–related GPR (Mrgp) The only non–chemosensory gene cluster with overmerging | 46,835,465 | 47,962,036 | 1,126,572 | overmerged | 22 | 13 | 0 | 9 | 0 | 12.56 |
| Chr7 | OlfrVmn | Red | Olfactory (Olfr) & Vomeronasal (Vmn) gene cluster with introgression | 84,459,635 | 86,687,176 | 2,227,542 | introgressed | 61 | 32 | 2 | 27 | 0 | 20.41 |
| Chr7 | OlfrVmn | | SAMEA3650363 HET DOM–>SPRET introgression (45004 markers) | 84,447,662 | 86,880,049 | 2,432,388 | introgressed | | | | | | |
| Chr7 | OlfrVmn | | SAMEA3650366 HOM DOM–>SPRET introgression (34766 markers) | 84,570,624 | 86,443,012 | 1,872,389 | introgressed | | | | | | |
| Chr7 | Vmn | Black | Vomeronasal (Vmn) gene cluster | 102,125,980 | 105,018,524 | 2,892,545 | overmerged | 158 | 130 | 7 | 21 | 0 | 7.58 |
| Chr7 | Vkorc1 | Red | SAMEA3403368 HET SPRET–>DOM introgression (44977 markers) | 125,567,611 | 128,954,684 | 3,387,074 | introgressed | | | | | | |
| Chr9 | Olfr | Black | Olfactory (Olfr) gene cluster | 37,622,644 | 39,986,558 | 2,363,915 | overmerged | 123 | 91 | 10 | 22 | 0 | 8.16 |
| Chr12 | pseudo | Grey | hotspot of pseudogenisation | 17,950,735 | 21,826,708 | 3,875,974 | overmerged | 78 | 0 | 0 | 78 | 0 | 9.68 |
| Chr12 | IgH | Blue | immunoglobulin heavy variable (Ighv) | 113,413,208 | 115,821,921 | 2,408,714 | overmerged | 73 | 0 | 0 | 0 | 0 | 0.83 |
| Chr13 | Vmn | Grey | Vomeronasal gene cluster (All C57BL/6J pseudogenised) (Vmn) | 65,501,070 | 67,021,082 | 1,520,013 | overmerged | 13 | 0 | 0 | 13 | 0 | 6.48 |
| Chr14 | T cells | Blue | T cell receptor alpha variable (Trav) | 52,665,424 | 54,215,135 | 1,549,712 | overmerged | 130 | 0 | 0 | 0 | 112 | 4.32 |
| Chr17 | MHC | Blue | Histocompatibility 2 (H2) | 34,193,633 | 37,794,443 | 3,600,811 | overmerged | 48 | 36 | 4 | 8 | 0 | 8.5 |
| Chr19 | Olfr | Black | Olfactory (Olfr) gene cluster | 11,805,441 | 13,897,336 | 2,091,896 | overmerged | 92 | 67 | 3 | 22 | 0 | 8.29 |

Table 2: Genomic regions introgressed and overmerged between *M. musculus domesticus* and *M. spretus* shown in Fig. 6. Positions refer to mouse genome reference GRCm39, NCBI Assembly Accession: GCA_000001635.9.

esis for the across-mouse-species admixture event analysed here. An event seen as unidirectional capture of pesticide genes with some evidence of bidirectional geneflow can be reconsidered as an exchange of genes suited to the anthropocene. While long-time synanthropic *M. musculus domesticus* received pesticide resistance from *M. spretus*, which previously lived further from human influence, no species can now live far from humans, and *M. spretus* may benefit from sets of chemosensory genes evolved by *M. musculus domesticus* in human proximity.

# Methods

The diagnostic index expectation maximisation (*diem*) algorithm is described in detail in the Supplemental Methods. It inputs genomic marker state over individuals encoded as homozygotes for the two most frequent alleles, their heterozygotes, and 'Undetermined' for individuals with genetic state that cannot be expressed in terms of the two most common alleles. Which allele is correlated with which barrier side is called marker polarity, $\phi$. Given a polarisation of all markers, for each individual, the genome can be summarised as the count of these four states. The entire sample can then be summarised as a matrix ${}^4\mathbf{I}$ with four columns and number of rows equal to the number of individuals. We can calculate the log-likelihood $\ln L$ of each marker, given the ${}^4\mathbf{I}$ matrix, polarity of the marker $\phi$, and weight $\epsilon$, specifying how close to 'ideal' the purest individuals should be. For the $i$th marker, $\ln L$ is given as

$$\ln L\left({}_{\bullet}\vec{M}_i|\phi_i, {}^4_{\zeta}\mathbf{I}, \epsilon\right) = \sum {}^4\mathbf{M}_i \odot \ln\left(\frac{1}{n+4\zeta}\cdot {}^4\mathbf{V}\right), \tag{1}$$

where ${}_{\bullet}\vec{M}_i$ is the marker $i$ vector imported with polarity $\phi_i$, ${}^4\mathbf{M}_i$ is the respective 4-genomic state count matrix of the single marker, $n$ is the number of markers, ${}^4_{\zeta}\mathbf{I}$ is ${}^4\mathbf{I}$ with a $\zeta$ correction for small datasets (Suppmementary Methods Eq. 9), ${}^4\mathbf{V}$ is ${}^4\mathbf{I}$ weighted by $\epsilon$ (Supplemental Methods Eq. 20), and $\odot$ is Hadamart product. Given $\ln L$, we can calculate the diagnostic index $d$ of the marker and its support $\delta$ as

$$d_i = \max\left\{\ln L\left({}_{\bullet}\vec{M}_i|\phi_i, {}^4_{\zeta}\mathbf{I}, \epsilon\right)\middle|\phi_i \in \{0,1\}\right\}, \tag{2}$$

$$\delta_i = d_i - \min\left\{\ln L\left({}_{\bullet}\vec{M}_i|\phi_i, {}^4_{\zeta}\mathbf{I}, \epsilon\right)\middle|\phi_i \in \{0,1\}\right\}. \tag{3}$$

*diem* is an iterative algorithm, where marker polarity is updated in a two step process. For the $i$th marker and the $j$th iteration,

1. determine suggested polarity $\hat{\phi}_i$ at $d_i$ according to

$$\hat{\phi}_i^{(j)} = \begin{cases} 0, & \text{if } \ln L\left({}_{\bullet}\vec{M}_i|\phi_i^{(j-1)}, {}^4_{\zeta}\mathbf{I}^{(j)}, \epsilon\right) > \ln L\left({}_{\bullet}\vec{M}_i|\neg\phi_i^{(j-1)}, {}^4_{\zeta}\mathbf{I}^{(j)}, \epsilon\right) \\ 1, & \text{otherwise,} \end{cases} \tag{4}$$

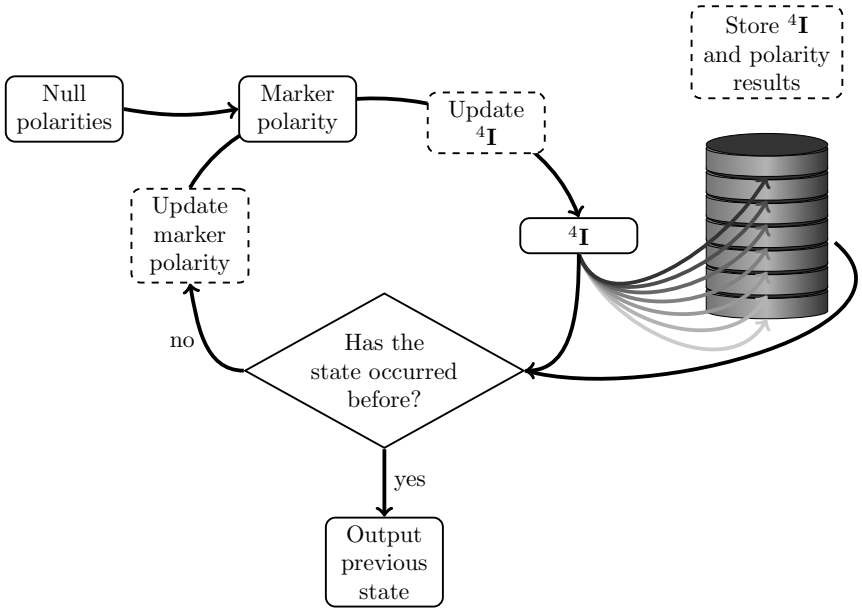


Figure 7: Diagnostic index expectation maximisation algorithm for detecting barriers to geneflow implemented in diemmca, diempy, and diemr.

where $\phi_i^{(j-1)}$ is the polarity of the $i$th marker in the $(j-1)$th iteration, and ${}^4_\zeta\mathbf{I}^{(j)}$ is the $\zeta$ corrected ${}^4\mathbf{I}$ in the $j$th iteration.

2. Based on the comparison of the suggested polarity to the polarity used in the previous iteration, determine the $i$th marker polarity for the $j$th iteration according to

$$\phi_i^{(j)} = \begin{cases} 0, & \text{if } \phi_i^{(j-1)} = \hat{\phi}_i^{(j)} \\ 1, & \text{otherwise.} \end{cases} \tag{5}$$

*diem* then updates ${}^4\mathbf{I}$ with any new polarities and iterates, halting when the algorithm encounters a previous state (Fig. 7).

## Implementation

The *diem* algorithm was developed in Mathematica as diemmca and idiomatically translated to Python, forming package diempy, and to R, forming package diemr. The diemmca and diempy calculate *diem* genome polarisation including calculating hybrid indices, heterozygosity, genotyping error rate, model of diagnostic (${}^4\mathbf{V}$), and running the *diem* iterations to determine optimal marker polarities. The diemr package also checks input data format, and polarized genomes can be imported and visualized.

### Parallelisation

For UNIX-based operating systems, the *diem* implementations delegate processing genomic compartments (such as marker sets on chromosomes or scaffolds) to multiple cores. When the last

core is finished, the core reports are summarized for genome-wide updates.

### Optimisation of the likelihood calculation

The likelihood calculation is required $2 \times n$ markers times for each iteration (cf. Eq. 4). However, the non-zero positions in ${}^{4}\mathbf{M}$ of a marker are invariant and need be recorded only once during *diem* initialization. Further, the ${}^{4}\mathbf{I}$ matrix need only be computed once per iteration. Flattening the values $\ln\left(\frac{1}{n+4\zeta} \cdot {}^{4}\mathbf{V}\right)$ to a vector, the likelihood calculation for a marker (Eq. 1) reduces to a non-zero position lookup of that vector, which is then summed.

### Optimisation of updating

The ${}^{4}\mathbf{I}$ counts need to be initialised, but EM updates involve changing the polarity of only a subset of markers. The change to ${}^{4}\mathbf{I}$ caused by these polarity switches is straightforward to calculate, and ${}^{4}\mathbf{I}$ is updated only by this change.

### Optimisation of halting

In each iteration, *diem* first checks the number of marker polarity changes (Fig. 7). Only if the same number of polarity changes has been encountered before, does the algorithm check whether those changes involve the same markers as previously encountered, indicating arrival at a previous model state.

## Validation

We cross validated *diem* implementation in diemr and diemmca on two datasets. The first dataset consisted of 115 *Myotis myotis* bats scored for 31,074 SNP markers (Harazim *et al.* 2021, raw reads available at NCBI BioProject Accession: PRJNA681157). The second dataset included transcriptomic and genomic reads of 69 mice (36 *Mus musculus domesticus* and 33 *Mus spretus*, Supplemental Table S2) mapped onto mouse reference GRCm39, NCBI Assembly Accession: GCA_000001635.9. The full mouse dataset was of 38 million variable sites, and we used 117,403 SNPs at chr7:123189278–131234739 as a focus for comparative analyses. The *diem* implementation in diempy was cross validated with diemmca for the mouse dataset.

# Acknowledgements

We thank Beate Nürnberger for discussions and constructive criticism. SJEB thanks the Lohse Lab, University of Edinburgh, Scotland, and the Barton Lab, IST, Austria, for discussions during development. In 1996 and 1999, SJEB worked in Montpellier side by side with Annie Orth, who deserves credit for the initial discovery of geneflow between *M. musculus domesticus* and *M. spretus*.

**Author contributions: Conceptualization: SJEB; Methodology: SJEB, NM; Software: SJEB, NM, JP, IJ; Validation: SJEB, NM, PŠ; Writing - original draft: SJEB, NM; Writing - review & editing: SJEB, NM, JP, IJ, PŠ.**

# Funding

This work has been supported by the Institute of Vertebrate Biology of the Czech Academy of Sciences (RVO: 68081766), and Czech Science Foundation "We Support Fundamental Research", RECETOX Research Infrastructure (No. LM2018121) financed by the Ministry of Education, Youth and Sports, and the Operational Programme Research, Development and Education (the CETOCOEN EXCELLENCE project No. CZ.02.1.01/0.0/0.0/17_043/0009632). MetaCentrum computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA CZ LM2018140) supported by the Ministry of Education, Youth and Sports of the Czech Republic.

# Code access

The *diem* method is implemented in Mathematica, Python and R. The Mathematica and Python codes are available at github through https://github.com/StuartJEBaird as diemmca and diempy, respectively, and the R package diemr is available at CRAN through https://CRAN.R-project.org/package=diemr.

# References

Banker, S. E., Bonhomme, F., and Nachman, M. W. (2022). Bidirectional Introgression between *Mus musculus domesticus* and *Mus spretus*. *Genome Biology and Evolution*, **14**, evab288.

Barton, N. (1979). The dynamics of hybrid zones. *Heredity*, **43**, 341–359.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1), 1–38.

Falush, D., Stephens, M., and Pritchard, J. (2003a). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Falush, D., Wirth, T., Linz, B., Pritchard, J. K., Stephens, M., Kidd, M., Blaser, M. J., Graham, D. Y., Vacher, S., Perez-Perez, G. I., Yamaoka, Y., Mégraud, F., Otto, K., Reichard, U., Katzowitsch, E., Wang, X., Achtman, M., and Suerbaum, S. (2003b). Traces of human migrations in *Helicobacter pylori* populations. *Science*, **299**, 1582–1585.

Gompert, Z., Mandeville, E. G., and Buerkle, C. A. (2017). Analysis of population genomic data from hybrid zones. *Annual Review of Ecology, Evolution, and Systematics*, **48**, 207–229.

Haldane, J. (1948). The theory of a cline. *Journal of Genetics*, **48**, 277–284.

Harazim, M., Piálek, L., Pikula, J., Seidlová, V., Zukal, J., Bachorec, E., Bartonička, T., Kokurewicz, T., and Martínková, N. (2021). Associating physiological functions with genomic variability in hibernating bats. *Evolutionary Ecology*, **35**, 291–308.

Lawson, D. J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genetics*, **8**, e1002453.

Orth, A., Belkhir, K., Britton-Davidian, J., Boursot, P., Benazzou, T., and Bonhomme, F. (2002). Hybridation naturelle entre deux espèces sympatriques de souris *Mus musculus domesticus* L. et *Mus spretus* Lataste. *Comptes Rendus Biologies*, **325**(2), 89–97.

Pritchard, J. K., Wen, X., and Falush, D. (2010). *Documentation for Structure software: Version 2.3*.

Song, Y., Endepols, S., Klemann, N., Richter, D., Matuschka, F.-R., Shih, C.-H., Nachman, M. W., and Kohn, M. H. (2011). Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. *Current Biology*, **21**, 1296–1301.

# Supplemental Material

**Supplemental Methods** – *diem*: Diagnostic index expectation maximisation.

Table S1: Comma-separated table with benchmarking results of *diem* implemented in diempy on the MetaCentrum computational cluster.

Table S2: Tab-separated table of BioSample Accession Numbers, geographic origin and hybrid indices (HI) of *Mus musculus domesticus* and *Mus spretus* analysed by *diem*.