diemr: Genome polarisation for detecting barriers to geneflow in R

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, Brno, 62500, Czechia
[*]Correspondence: stuartj.e.baird@gmail.com, martinkova@ivb.cz

Abstract

Motivation: We extend classic allele counting measures of geneflow 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 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 standard cline width/barrier strength comparisons.

Results: Diagnostic index expectation maximisation (*diem*) estimates bistate marker labelling polarities with respect to the sides of a barrier, also returning polarity support and a likelihood-based diagnostic index. Polarised markers allow detection of geneflow at the scale of whole genomes.

Availability and implementation: The *diem* method is implemented in Mathematica and R. The Mathematica code is available at github through https://github.com/StuartJEBaird, and the R package diemr is available at CRAN through https://CRAN.R-project.org/package=diemr

1 Introduction

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).

2 Methods

The diagnostic index expectation maximisation (*diem*) algorithm is described in detail in the Supplementary Information. 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, ϕ. 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 ϕ, and weight ϵ, specifying how close to 'ideal' the purest individuals should be. For each marker, $\ln L$ is given as

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

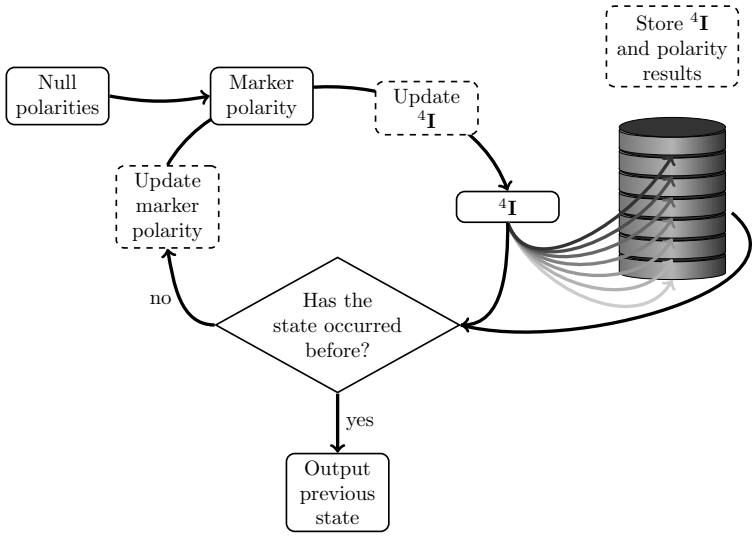

Figure 1: Diagnostic index expectation maximisation algorithm for detecting barriers to geneflow implemented in diemr.

where ${}_{\bullet}\vec{M}$ is the vector of one marker imported with polarity ϕ, ${}^4\mathbf{M}$ 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 ζ correction for small datasets (Suppmementary Information Eq. (9)), and ${}^4\mathbf{V}$ is ${}^4\mathbf{I}$ weighted by ϵ (Supplementary Information Eq. (20)). Given $\ln L$, we can calculate the diagnostic index d of the marker and its support δ as

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

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

diem is an iterative algorithm, where marker polarity is updated in a two step process. For each marker and the jth iteration,

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

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

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

2. Based on the comparison of the suggested marker polarity to the marker polarity used in the previous iteration, determine the new marker polarity for the jth iteration according to

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

diem then updates ${}^4\mathbf{I}$ with any new polarities and iterates, halting when the algorithm encounters a previous state (Fig. 1).

3 Implementation

The *diem* algorithm was developed in Mathematica and idiomatically translated to R, forming package diemr. The package checks input data format, polarises markers, calculates hybrid indices, heterozygosity, genotyping error rate, model of diagnostic (${}^4\mathbf{V}$), and runs the *diem* iterations to determine optimal marker polarities. Polarized genomes can be imported and visualized.

3.1 Parallelisation

For UNIX-based operating systems, the diemr package delegates 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.

3.2 Optimisation of the likelihood calculation

The likelihood calculation is required $(2 \times n)$ times for each iteration (cf. (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 (1) reduces to a non-zero position lookup of that vector, which is then summed.

3.3 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.

3.4 Optimisation of halting

In each iteration, *diem* checks first only the number of marker polarity changes (Fig. 1). 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.

4 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

Table 1: Benchmark results of *diem* implementation in diemr on bats and mice datasets.

Dataset	Individuals	Markers	Compartments	CPU time	System time	Elapsed
Bats	115	31,074	3	2m41s	8s	1m7s
Mice	69	1,974,410	8	3h31m51s	2m13s	1h23m19s

On iMac 2.7 GHz Quad-Core Intel Core i5, 8 GB 1600 MHz DDR3 RAM, using 3 processors.

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*, Supplementary Information Table S1) mapped onto mouse reference GRCM39. Including only variable sites, the mouse dataset consisted of 1,974,410 markers on chromosome 7.

4.1 Bat SNPs dataset

The *diem* converged after 67 seconds (Table 1). Using random null polarities, the bats dataset showed vertical structure indicative of genomic blocks (Fig. 2A). The individual hybrid indices were ≈ 0.5 (Fig. 2C). Having found optimal marker polarities using *diem* in diemr, we conclude that the dataset consists of a homogeneous set of individuals with five outliers, likely artefacts of large number of unknown marker states (Fig. 2B). The hybrid indices of the polarised data are ≈ 0.2 (Fig. 2C), and the plot of diagnostic indices and their supports shows a typical shark-fin shape, indicating that markers with high diagnostic index have also high support (Fig. 2D).

4.2 Mouse whole-genome dataset

The house mouse chromosome 7 analysis converged after 4,999 seconds (Table 1). The visualization of the mouse genomes with null polarities shows distinct difference in individuals sequenced for whole genomes (horizontal blocks) and for transcriptomes (sparse data in vertical blocks above and below the whole-genome sequences) (Fig. 3A). The polarized genomic data (Fig. 3B) differentiate individuals with respect to the barrier to geneflow resulting in the hybrid indices in $[0.2, 0.8]$ (Fig. 3C). The Linnean binomial taxon labels separate at the steepest change in the hybrid index. This dataset was chosen for two reasons. First, high divergence between the taxa ($\sim 5\%$) allows *diem* to be benchmarked for large n. Second, previous studies suggested geneflow between these taxa (Orth et al. 2002; Banker et al. 2022), and in particular capture of *M. spretus* warfarin-resistance genes by *M. musculus domesticus* (Song et al. 2011). In Fig. 3B, we show the introgressed region, where two *M. musculus domesticus* individuals carry the introgressed block. The introgression present in two individuals cannot be distinguished in a genomic scan correlating marker state to the individual hybrid index (Fig. 3D), but is apparent in the visualization of the polarised genomes (Fig. 3B).

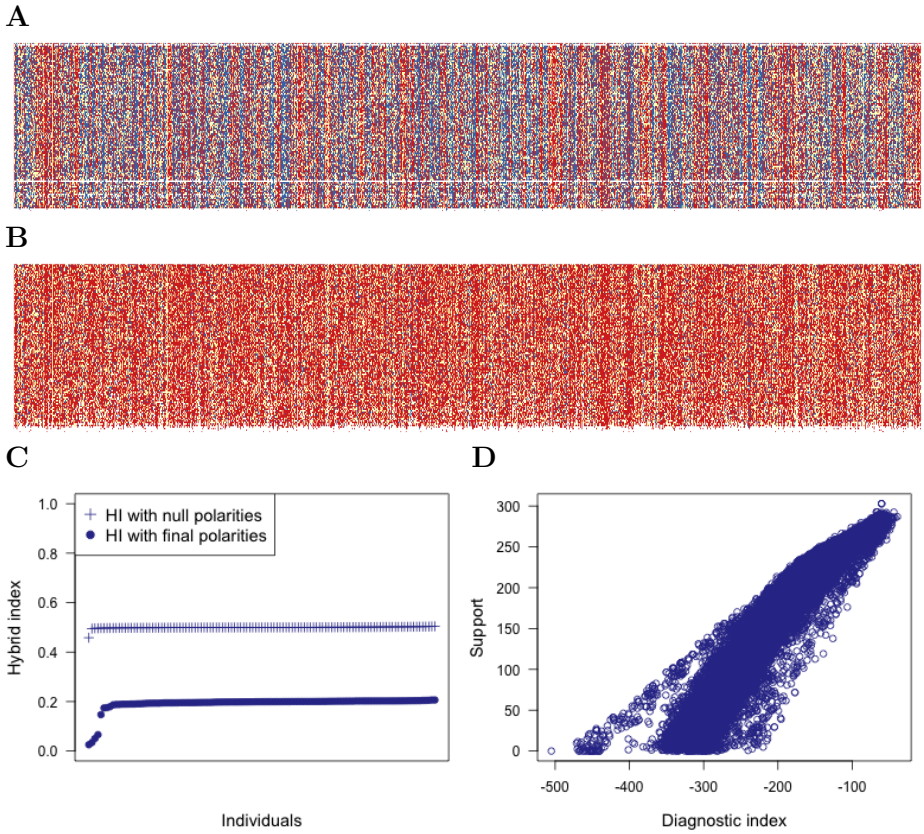

Figure 2: Genotypes of 31,074 SNP markers (columns) in 115 *Myotis myotis* bats (rows) (Harazim et al. 2021) with random null polarities (A), and with optimal polarities (B). Individuals in panel B are ordered according to their hybrid index. Red – homozygotes for the allele attributed to barrier side A, blue – 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.

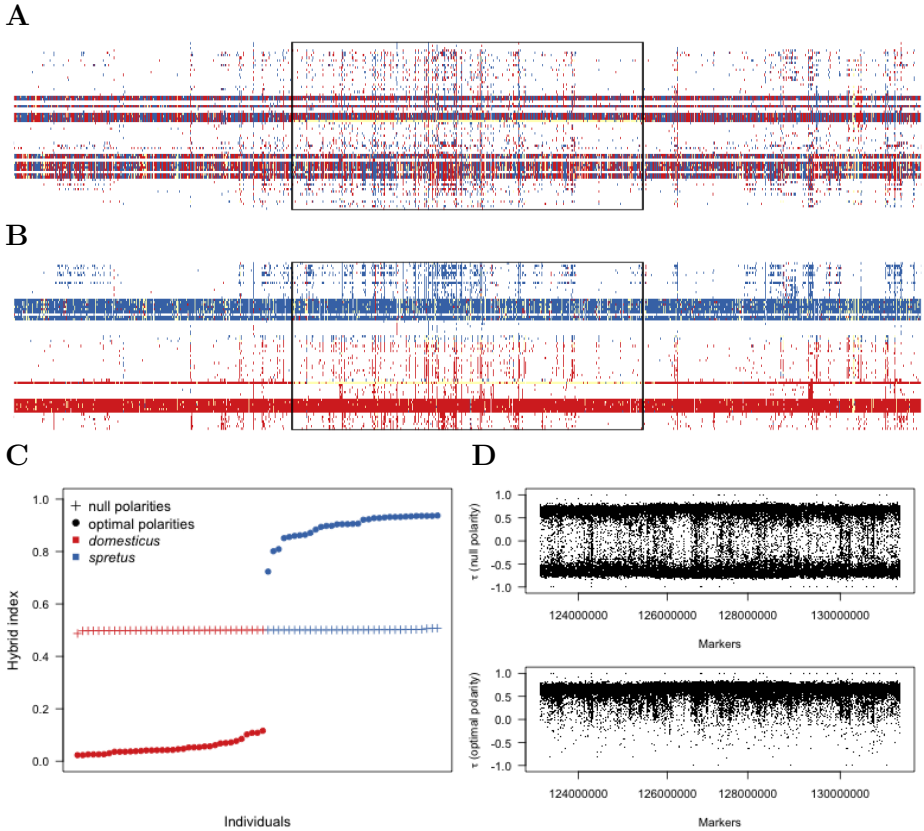

Figure 3: Genotypes of 117,403 markers (columns) on chromosome 7, positions 123,189,278:131,234,739 in 69 mice (rows) with random null polarities (A), and with optimal polarities (B). Individuals in panel B are ordered according to their hybrid index. The rectangle delimitates the genomic block carrying warfarin resistance 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. Red – homozygotes for the allele attributed to barrier side A (*domesticus*), blue – homozygots for the allele attributed to barrier side B (*spretus*), yellow – heterozygotes, white – unknown genotype state. C) Ordered individual hybrid indices for markers on the complete chromosome 7 in mice. The optimal marker polarisation (circles) shows that the dataset consists of two main groups of genetically similar individuals. D) Kendall's τ 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 the genomic scan.

4.3 Implementation validation

Using identical null polarities, we found 0 differences in marker polarities between converged *diem* states in the bats and mice datasets as implementated in R and Mathematica. As the *diem* algorithm is deterministic, alternate implementations should give the same output on the same input. The diemr and diemmca implementations were cross-validated in this way on multiple datasets, including large. Finding no differences, the chance of coding error in either is then the chance that both 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.

5 Conclusions

Genome polarisation deconstructs the concept of a population. We have shown that the *diem* algorithm is robust in detecting a single homogeneous set of individuals (Fig. 2), two taxa (Fig. 3), including detection of introgressed blocks (Fig. 3B), *diem* allows analyses of samples with genomic and transcriptomic data, the implementation scales well with genome-size data (Table 1), and diemr and diemmca provide matching solutions. Polarising genomes allows us to study a continuous genomic space, investigating barriers to geneflow and capture of genetic material at a regional scale.

Acknowledgements

We thank Beate Nürnberger for discussions and constructive criticism.

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).

References

Sarah E Banker, François Bonhomme, and Michael W Nachman. Bidirectional Introgression between *Mus musculus domesticus* and *Mus spretus*. *Genome Biology and Evolution*, 14: evab288, 2022. doi: 10.1093/gbe/evab288.

N Barton. The dynamics of hybrid zones. *Heredity*, 43:341–359, 1979. doi: 10.1038/hdy.1979.87.

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

D Falush, M Stephens, and JK Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164:1567–1587, 2003a. doi: 10.1093/genetics/164.4.1567.

Daniel Falush, Thierry Wirth, Bodo Linz, Jonathan K. Pritchard, Matthew Stephens, Mark Kidd, Martin J. Blaser, David Y. Graham, Sylvie Vacher, Guillermo I. Perez-Perez, Yoshio Yamaoka, Francis Mégraud, Kristina Otto, Ulrike Reichard, Elena Katzowitsch, Xiaoyan Wang, Mark Achtman, and Sebastian Suerbaum. Traces of human migrations in *Helicobacter pylori* populations. *Science*, 299:1582–1585, 2003b. doi: 10.1126/science.1080857.

Zachariah Gompert, Elizabeth G. Mandeville, and C. Alex Buerkle. Analysis of population genomic data from hybrid zones. *Annual Review of Ecology, Evolution, and Systematics*, 48: 207–229, 2017. doi: 10.1146/annurev-ecolsys-110316-022652.

Markéta Harazim, Lubomír Piálek, Jiri Pikula, Veronika Seidlová, Jan Zukal, Erik Bachorec, Tomáš Bartonička, Tomasz Kokurewicz, and Natália Martínková. Associating physiological functions with genomic variability in hibernating bats. *Evolutionary Ecology*, 35:291–308, 2021. doi: 10.1007/s10682-020-10096-4.

Daniel John Lawson, Garrett Hellenthal, Simon Myers, and Daniel Falush. Inference of population structure using dense haplotype data. *PLoS Genetics*, 8:e1002453, 2012. doi: 10.1371/journal.pgen.1002453.

Annie Orth, Khalid Belkhir, Janice Britton-Davidian, Pierre Boursot, Touria Benazzou, and François Bonhomme. Hybridation naturelle entre deux espèces sympatriques de souris *Mus musculus domesticus* L. et *Mus spretus* Lataste. *Comptes Rendus Biologies*, 325(2):89–97, 2002. doi: 10.1016/S1631-0691(02)01413-0.

Jonathan K. Pritchard, Xiaoquan Wen, and Daniel Falush. *Documentation for structure* software*: Version 2.3*, 2010. URL `https://web.stanford.edu/group/pritchardlab/structure_software/release_versions/v2.3.4/structure_doc.pd`

Ying Song, Stefan Endepols, Nicole Klemann, Dania Richter, Franz-Rainer Matuschka, Ching-Hua Shih, Michael W. Nachman, and Michael H. Kohn. Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. *Current Biology*, 21:1296–1301, 2011. doi: 10.1016/j.cub.2011.06.043.