Fluidigm: An R-package for fluidigm data handling, sexing and clustering

Robert Ekblom, Helena Johansson, Mia Valtonen, Daniel Fischer

22 March 2024

Summary

The R-package `Fluidigm` is a comprehensive tool for analyzing genotypic data, as created by using nanofluidic dynamic arrays from Standard BioTools Inc., previously known as Fluidigm (Wang et al. 2009). Genotypic data refers here to the genetic makeup of an organism, encoded in its DNA, which determines its physical and physiological traits. This data can represent the entire genetic variation, with millions of single-nucleotide polymorphisms (SNP) and is crucial in understanding an organism's characteristics and can, among others, be used for species identification and sexing. However, the Fluidigm instrument generates very cost-effective up to 96 representative SNPs—key markers that provide insights into an individual's genetic diversity. These SNPs serve as valuable proxies for a targeted genetic analysis, enabling researchers to explore specific genetic variations within the genome.

The here presented R-package is written to process and analyze SNP data, which is sourced from a Fluidigm device. It offers valuable insights into the similarities and dissimilarities between different individuals or species, based on the called SNP variants. For instance, it can be utilized to distinguish between wolf and dog DNA based on collected and sequenced tissue samples. This capability makes it a practical tool in the field of genetic studies, wildlife ecology and biomonitoring.

Our software takes the raw genotypic data and transforms it into a format suitable for further analysis. It then estimates potential errors, e.g., in form of sequencing errors, from the data, ensuring its accuracy and reliability. The software also has the capability to perform sex assignment and species marker analysis, if required.

One of the key features of the software is its ability to calculate pairwise similarities between genotypes. This can help identify genomic markers with significant genetic variation, potentially associated with diverse species and even traits.

The software also provides visual representations of the data, generating histograms of the pairwise similarities. This allows users to better understand the overall structure and diversity of the genotypic data at hand.

In summary, our software serves as a robust tool for genetic analyses, offering a range of functionalities to ensure the accuracy, reliability, and comprehensibility of the analysed SNP data. It is designed to be user-friendly and customizable, catering to the needs of diverse users, from genetic researchers to non-specialists interested in genetic studies.

Statement of need

Genetic monitoring of wildlife population often relies on non-invasive samples such as scats and hairs. These samples tend to have DNA in low concentrations, or degraded DNA. As a standard in the field, such samples are genotyped more than once, and consensus genotypes created from the repeats. Commonly, low quality samples means that genotyping is substandard, such that a researcher may need to filter samples out based on thresholds for consensus genotype error rates or number of successful loci. This is the first R-package that

implements these functions for SNP data to our knowledge. Similar functions for microsatellite markers are implemented in ConGenR (Lonsinger and Waits 2015).

SNP panels are becoming increasingly sophisticated by incorporating more and more functions such as diagnostic species markers (Harmoinen et al. 2024) or Y and/or X chromosomal markers (Ekblom et al. 2021) for diagnosing sex. This requires flexible analytical tools to handle these data in a way that is informative, for example by determining sex according to rules set by the user, or sum the number of successful markers to indicate whether a species assignment is reliable. Similarly, flexibility in the number of SNPs and number of samples per analysis batch is important for users who may seek these functions from PLINK (ped,map) filesets rather than data generated by Standard Biotools platforms.

The Fludigm software described here is currently in use in annual wolf monitoring in Finland, on invasive and non-invasive samples, using two SNP panels; 1) annual monitoring SNP panel for determining sex, putative species, and individual ID (Heikkinen et al. 2023). A publication is in process describing this panel and use of this software (Harmoinen et al. 2024) 2) a wolf dog hybridisation SNP panel (Harmoinen et al. 2021). The original scripts were used for wolverine (Ekblom et al. 2021).

Functionality

The package is constructed using five main building bricks, each producing outputs that serve as inputs for the subsequent component. These components are designed to be executed in sequential order:

1. `fluidigm2PLINK(...)`
2. `estimateErrors(...)`
3. `calculatePairwiseSimilarities(...)`
4. `getPairwiseSimilarityLoci(...)`
5. `similarityMatrix(...)`

The initial step in the package's functionality involves creating a `ped/map`-file pair from the `csv`-output typically generated by the Fluidigm machine. The basic usage of the function is as follows:

```
fluidigm2PLINK(file = "example_data.csv",
               map = "example_data.map",
               out = "new_data")
```

The `estimateErrors` function is designed to process the generated PLINK ped files and estimate errors. It offers a comprehensive analysis of genotyping data, ensuring the accuracy and reliability of the results. This function is particularly useful in large-scale genetic studies where error estimation is vital for maintaining data integrity.

One of the standout features of `estimateErrors` is its capability to perform sex assignment and species marker analyses, if required. This is accomplished by providing Y and X markers (using dedicated options in the function call) for sexing and species-identification markers for species analysis.

The function is highly customizable, allowing users to specify various parameters such as the path to the ped input file, the database name, and whether new samples should be added to the database. It also allows users to control the number of replicates to keep, the markers for sexing and species identification, and the thresholds for various error checks. Further, the function can create a database to store already sequences samples and use them in subsequent analyses.

In addition, `estimateErrors` can generate plots for visual inspection of the data and provides verbose output for a detailed function feedback. It returns a list containing a matrix indicating if genotypes are called correctly for replicates and/or if genotypes are missing, and a matrix with summary statistics.

The basic usage in the running example is as follows:

```
estErr.out <- estimateErrors(file="new_data.ped")
```

The `calculatePairwiseSimilarities` function is designed to calculate pairwise similarities between genotypes. This function serves as a wrapper for the PLINK software.

The function requires a file path to the filtered ped/map file pair, without the ped/map file extension. This file contains the genotype data that the function will process and which was provided by `estimateErrors` earlier. These files have then a `.GOOD.map/ped`-file extension.

Optionally, the function can also accept a path to an existing sample database. Or, in case of its absense, it can also create such a database. If provided, the function will merge the output with this existing database. If a database is not provided, the function will proceed with the existing data only.

The basic call is as follows:

```
calculatePairwiseSimilarities(file="new_data.GOOD")
```

The `getPairwiseSimilarityLoci` function, which is a wrapper for a Perl script, performs pairwise comparisons of genotypes. Specifically, it counts the number of complete pairwise comparisons, with no missing alleles, between each genotype.

In the context of this script, a pairwise comparison involves comparing two genotypes locus by locus. A locus is defined here as a specific SNP location. When the script compares two genotypes, it checks each locus to see if the alleles, which are essentially versions of a gene, are the same or different.

If a locus has no missing alleles in both genotypes, it is considered a complete pairwise comparison. The script counts the number of these complete pairwise comparisons for each pair of genotypes. This count is then written to an output file.

This analysis can be instrumental in genetic studies to discern the similarities or dissimilarities between different individuals or species. It can aid in identifying regions of the genome with significant genetic variation, potentially associated with diverse traits or susceptibility to certain diseases.

The function does not return a value in the R environment. Instead, it generates an output file with a `.pairs`-extension in the same directory as the input file. This output file encapsulates the results of the pairwise similarity loci analysis.

The basic usage of the function is:

```
getPairwiseSimilarityLoci(file="new_data.GOOD")
```

The `similarityMatrix` function serves as the final component in this genetic analysis pipeline. It conducts a pairwise similarity analysis on the genotypic data, comparing each pair of genotypes in the dataset to ascertain their similarity.

The function accepts a main file and, optionally, separate `MIBS`, `PAIRS`, and `PED` files. If these separate files are not provided, the function presumes they share the same base name as the main file, with their respective extensions.

The function reads the genotype data from these files and computes the pairwise similarities. These similarities are then exported to a `CSV` file. All pairwise similarities exceeding a specified threshold (default is 0.85) are included in this output.

If the plots parameter is set to `TRUE`, the function also generates histograms of the pairwise similarities and saves them as a `PNG` file. These plots offer a visual representation of the distribution of the pairwise similarities, facilitating a better understanding of the overall structure and diversity of the genotypic data.

If a group is specified, the function conducts additional analyses for each sample in the group. This includes generating individual output files for each sample, which can be beneficial for performing sample-wise statistics.

The basic usage is:

```
similarityMatrix(file="new_data.GOOD")
```

Availability

The package is available in a stable release version on Cran, see https://cran.r-project.org/web/packages/Fluidigm. Latest development versions can be found in the `dev` branch of the corresponding GitHub repository (https://github.com/fischuu/Fluidigm), whereas the `main` branch is always aligned to the latest Cran release.

Acknowledgement

The original perl code from the `getPairwiseSimilarityLoci` function is based on the code behind the following URL: https://github.com/douglasgscofield/bioinfo/blob/main/scripts/plink-pairwise-loci.pl

Conflict of Interest

The authors declare no conflict of interest.

Bibliography

Ekblom, R., M. Aronsson, F. Elsner-Gearing, M. Johansson, T. Fountain, and J. Persson. 2021. "Sample Identification and Pedigree Reconstruction in Wolverine (*Gulo Gulo*) Using SNP Genotyping of Non-Invasive Samples." *Conservation Genetics Resources* 13: 261–74. https://doi.org/10.1007/s12686-021-01208-5.

Harmoinen, J., A.v. Thaden, J. Aspi, L. Kvist, B. Cocchiararo, A. Jarausch, G. Andrea, et al. 2021. "Reliable Wolf-Dog Hybrid Detection in Europe Using a Reduced SNP Panel Developed for Non-Invasively Collected Samples." *BMC Genomics* 22: 473. https://doi.org/10.1186/s12864-021-07761-5.

Harmoinen, J., M. Valtonen, D. Fischer, and et al. 2024. "Development of a Cost-Effective, Multifunctional SNP Panel and Analysis Workflow for Genetic Species Monitoring."

Heikkinen, S., M. Valtonen, H. Johansson, I. Helle, A. Herrero, S. Mäntyniemi, and I. Kojola. 2023. "Susikanta Suomessa Maaliskuussa 2023." *Luonnonvara- Ja Biotalouden Tutkimus* 70: 120. https://doi.org/10.10024/552010.

Lonsinger, R., and L. Waits. 2015. "ConGenR: Rapid Determination of Consensus Genotypes and Estimates of Genotyping Errors from Replicated Genetic Samples." *Conservation Genetics Resources* 7: 10.1007/s12686-015-0506-7.

Wang, J., M. Lin, A. Crenshaw, A. Hutchinson, B. Hicks, M. Yeager, S. Berndt, et al. 2009. "High-Throughput Single Nucleotide Polymorphism Genotyping Using Nanofluidic Dynamic Arrays." *BMC Genomics* 10 (1): 561. https://doi.org/10.1186/1471-2164-10-561.