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

Robert Ekblom, Helena Johansson, Mia Valtonen, Daniel Fischer

15 March 2024

Summary

The R-package `Fluidigm` is a comprehensive tool for analyzing genotypic data, as created by using nanofluidic dynamic arrays from a sequencer called 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 present 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, in form of sequencing errors, in 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

A Statement of need section that clearly illustrates the research purpose of the software and places it in the context of related work. A list of key references, including to other software addressing related needs. Note that the references should include full names of venues, e.g., journals and conferences, not abbreviations only understood in the context of a specific discipline. Mention (if applicable) a representative set of past or ongoing research projects using the software and recent scholarly publications enabled by it.

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 PLINK ped files and estimate errors. It offers a comprehensive analysis of genotyping data, ensuring the accuracy and reliability of your 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. (INTRODUCE HERE STILL THE IDEA OF THE DATABASE, WHAT IS IT AND WHATFOR IS IT NEEDED!)

In addition, `estimateErrors` can generate plots for visual inspection of the data and provides verbose output for a detailed analysis. 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 genotype database. Or, in case of its absense, it can also create such a database. If provided, the function will merge the genotype 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 the `.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 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 a available in a stable 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 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 FUNDING?!

Conflict of Interest

The authors declare no conflict of interest.

Bibliography

Wang, Jun, Mingyan Lin, Andrew Crenshaw, Andrew Hutchinson, Brian Hicks, Meredith Yeager, Sonja 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.