DNABarcodeCompatibility: an R-package for optimizing DNA-barcode combinations in multiplex sequencing experiments

Céline Trébeau, Jacques Boutet de Monvel, Fabienne Wong Jun Tai, Christine Petit, Raphael Etournay

HAL Id: pasteur-02175109

https://pasteur.hal.science/pasteur-02175109

Submitted on 5 Jul 2019

OXFORD

Data and text mining

DNABarcodeCompatibility: an R-package for optimizing DNA-barcode combinations in multiplex sequencing experiments

Céline Trébeau[1,2,3], Jacques Boutet de Monvel[1,2,3],
Fabienne Wong Jun Tai[1,2,3], Christine Petit[1,2,3,4,5]
and Raphaël Etournay [ID] [1,2,3,*]

[1]Unité de Génétique et Physiologie de l'Audition, Département Neuroscience, Institut Pasteur, 75015 Paris, France, [2]UMRS 1120, Institut National de la Santé et de la Recherche Médicale, 75015 Paris, France, [3]Sorbonne Université, 75006 Paris, France, [4]Collège de France, 75005 Paris, France. [5]Institut de la Vision, Paris, France and [5]Institut de la Vision, 75012 Paris, France

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Abstract

Summary: Using adequate DNA barcodes is essential to unambiguously identify each DNA library within a multiplexed set of libraries sequenced using next-generation sequencers. We introduce DNABarcodeCompatibility, an R-package that allows one to design single or dual-barcoding multiplex experiments by imposing desired constraints on the barcodes (including sequencer chemistry, barcode pairwise minimal distance and nucleotide content), while optimizing barcode frequency usage, thereby allowing one to both facilitate the demultiplexing step and spare expensive library-preparation kits. The package comes with a user-friendly interface and a web app developed in Java and Shiny (https://dnabarcodecompatibility.pasteur.fr), respectively, with the aim to help bridge the expertise of core facilities with the experimental needs of non-experienced users.

Availability and implementation: DNABarcodeCompatibility can be easily extended to fulfil specific project needs. The source codes of the R-package and its user interfaces are publicly available along with documentation at [https://github.com/comoto-pasteur-fr] under the GPL-2 licence.

Contact: raphael.etournay@pasteur.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Inherent to next-generation sequencing (NGS) techniques is the requirement to provide high enough sequence complexity of barcodes for the design of multiplex experiments, in order to allow for proper demultiplexing, i.e. identifying back each individual sample from a pooled library. In general, finding optimal combinations of compatible barcodes under given experimental constraints (single or dual barcoding, sequencer chemistry, DNA composition, barcode pairwise distance and error correction properties, availability in the laboratory) can become challenging for users. Existing software such as DNABarcodes, checkMyIndex, Illumina Experiment Manager, Barcosel (Buschmann, 2017; Somervuo *et al.*, 2018; Varet and Coppée, 2018) do not provide any integrated tool to automatically refine sets of compatible barcodes given all the above-mentioned constraints. Yet, a rational and general approach allowing to optimize the design of multiplexed NGS experiments under various constraints is highly desirable. To this end, we developed DNABarcodeCompatibility, a generic R-package allowing one to generate potentially large libraries of barcode combinations compatible with most sets of multiplexing constraints and optimized to achieve least-heterogeneity in terms of barcode usage. It is accompanied by a user-friendly interface designed with Illumina barcode-combination rules in mind, while allowing the user to apply constraints

specific to other NGS platforms on the distance between barcodes and on the nucleotide composition.

2 Results

2.1 Algorithm (DNABarcodeCompatibility R-package)

We describe here only the main steps and ideas. For details see the Supplementary Material. The inputs of the algorithm are a list of n distinct barcodes $\{i_1, \ldots, i_n\}$, the number N of required libraries, and the multiplex level k to be used for the experiment (so that $N = ak$, where a is the number of lanes of the flow cells).

Step 1. This step consists of identifying a set of barcode combinations that are compatible with the constraints set for the experiment (including pairwise barcode distance, nucleotide content, and sequencer chemistry). Given the number n of barcodes and the multiplex level k, the total number of barcode combinations (compatible or not) reads $\binom{n}{k}$. If this number is not too large (Supplementary File S1), the algorithm will perform an exhaustive search and return all compatible combinations of k barcodes. Otherwise, it will proceed sequentially by picking up barcode combinations at random, and stopping when a set of N_{comp} distinct compatible combinations has been generated, that is large enough for the next step, entropy maximization, to be effective.

Step 2. We then use a Shannon entropy maximization approach to select the N/k compatible combinations to be used in the experiment out of the N_{comp} combinations found in Step 1. Namely, these combinations are chosen in such a way that the resulting distribution of barcodes has maximum entropy $S = -\sum_i f_i \log f_i$, where $f_1, \ldots, f_n$ denote the frequencies of the various barcodes occurring in the selection. This ensures that this distribution is as uniform as possible given the constraints imposed by the particular set of compatible combinations at hand, thereby optimally balancing the consumable usage to spare library-preparation kits.

It can be shown (Supplementary File S2) that the maximum value of the entropy that can be attained for a selection of a library of N barcodes among n, with possible repetitions, is given by equation 1:

$$S_{max} = -(n - r)\frac{\lfloor\frac{N}{n}\rfloor}{N}\log\left(\frac{\lfloor\frac{N}{n}\rfloor}{N}\right) - r\frac{\lceil\frac{N}{n}\rceil}{N}\log\left(\frac{\lceil\frac{N}{n}\rceil}{N}\right) \quad (1)$$

in which r denotes the rest of the division of N by n, and $\lceil N/n \rceil$ ($\lfloor N/n \rfloor$) stands for the upper (lower) integer part of N/n. This expression reduces to $\log(\min(N,))$, that is the entropy of the uniform distribution, if either $N < n$ or n divides N ($r = 0$). In our algorithm, we use the S_{max} value as a stopping criterion to perform a randomized greedy search for a selection whose index distribution has maximum entropy. Simulations suggest that this algorithm is near optimal, in the sense that with high probability it finds a solution whose entropy is maximum or very close to maximum. In most cases, the solution found is much improved in terms of reducing the heterogeneity of barcode frequencies, compared to a randomly chosen (non-optimized) selection (see Supplementary File S1).

2.2 Graphical user interface

The DNABarcodeCompatibility user interface harbours two panels for input and output, respectively (Fig. 1) Once barcodes are loaded into the interface, the user can easily manipulate the list of barcodes by clicking on them. The interface provides tools to visualize and filter barcodes according to their chemical properties (GC content, i.e. the number of G or C nucleotides divided by the barcode length, and homopolymer length at most 2) and pairwise separation
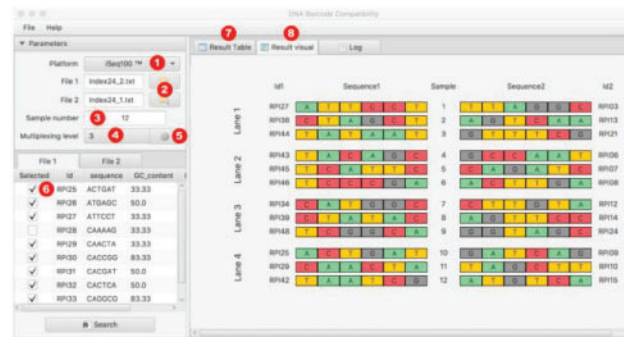
Fig. 1. Graphical user interface: dual-barcoding example. Left panel: input (1–6). Right panel: graphical output of the results (7–8). The user selects the platform (1), loads the barcode datasets (2), selects the number of libraries (3) and the multiplex level (4). Barcodes can be filtered according to their chemical properties and desired pairwise distance (5). They can also be manually unchecked (6). Results are shown in text mode (7), and graphically (8)

properties according to either the Hamming, SeqLev or phaseshift distance (Bystrykh, 2012; Buschmann, 2017). Next, the application seeks an optimized set of compatible barcodes given the number of libraries and multiplex level. Finally, the interface also allows one to visualize how barcoded libraries are distributed among lanes of flow cells.

3 Conclusion

DNABarcodeCompatibility provides a rational optimized design of multiplex sequencing experiments integrating barcode constraints from various NGS platforms including (i) barcode chemical properties, (ii) barcode pairwise distance, (iii) Illumina barcode-combination rules and (iv) barcode frequency usage. Controlling these barcode constraints is not only critical to ensure proper demultiplexing, but also to optimize consumable usage and reduce costs. The major strength of the package lies on its generic features that are applicable to wide range of multiplex pooling design on most NGS platforms. This should be useful notably for spatially and temporally resolved transcriptomics (see Supplementary File S3 for a comparison of DNABarcodeCompatibility with other existing software).

References

Buschmann,T. (2017) DNABarcodes: an R package for the systematic construction of DNA sample tags. *Bioinformatics*, **33**, 920–922.

Bystrykh,L.V. (2012) Generalized DNA barcode design based on Hamming codes. *PLoS One*, **7**, e36852.

Somervuo,P. *et al.* (2018) BARCOSEL: a tool for selecting an optimal barcode set for high-throughput sequencing. *BMC Bioinformatics*, **19**, 257.

Varet,H. and Coppée,J.-Y. (2018) checkMyIndex: a web-based R/Shiny interface for choosing compatible sequencing indexes. *Bioinformatics*.