Benchmark

Quality control and statistical evaluation of combinatorial DNA libraries using nanopore sequencing

Cédric Lood[1,2], Hans Gerstmans[1,3,4], Yves Briers[3], Vera van Noort[2,5] & Rob Lavigne*[,1]

*[1]Department of Biosystems, Laboratory of Gene Technology, KU Leuven, Leuven, Belgium; [2]Department of Microbial & Molecular Systems, Centre of Microbial & Plant Genetics, Laboratory of Computational Systems Biology, KU Leuven, Leuven, Belgium; [3]Department of Biotechnology, Laboratory of Applied Biotechnology, Ghent University, Ghent, Belgium; [4]Department of Biosystems, MeBioS-Biosensors Group, KU Leuven, Leuven, Belgium; [5]Institute of Biology, University of Leiden, Leiden, The Netherlands; *Author for correspondence: rob.lavigne@kuleuven.be*

BioTechniques 69: 00–00 (November 2020) 10.2144/btn-2020-0060
First draft submitted: 28 April 2020; Accepted for publication: 28 July 2020; Published online: 1 October 2020

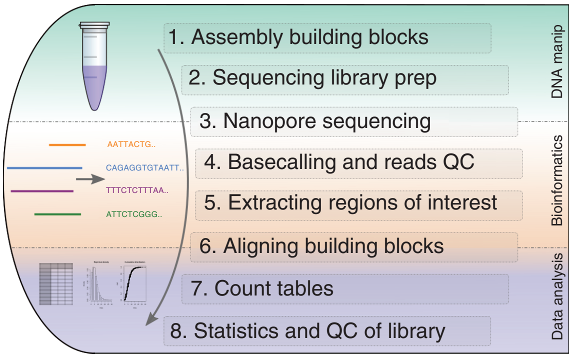
ABSTRACT

Protein engineering and synthetic biology applications increasingly rely on the assembly of modular libraries composed of thousands of different combinations of DNA building blocks. At present, the validation of such libraries is performed by Sanger sequencing analysis on a small subset of clones on an *ad hoc* basis. Here, we implement a systematic procedure for the comprehensive evaluation of combinatorial libraries, immediately after their creation *in vitro*, using long reads sequencing technology. After an initial step of nanopore sequencing, we use straightforward bioinformatics tools to tabulate the composition and synteny of the building blocks in each read. We subsequently use exploratory statistics to assess the library and validate its diversity before carrying downstream cloning and screening assays.

METHOD SUMMARY

Nanopore sequencing enables a comprehensive assessment of combinatorial libraries of DNA sequences assembled in single tube reactions. The inspection of such DNA libraries after their creation *in vitro* allows investigating potential biases in the DNA assembly method, possible issues in the upstream library of DNA building blocks that feed the *in vitro* assembly reaction and the correct synteny of the constructs.

GRAPHICAL ABSTRACT



1. Assembly building blocks
2. Sequencing library prep
3. Nanopore sequencing
4. Basecalling and reads QC
5. Extracting regions of interest
6. Aligning building blocks
7. Count tables
8. Statistics and QC of library

KEYWORDS:

combinatorial DNA library • constructs validation • DNA assembly • GoldenBraid assembly • GoldenGate assembly • nanopore sequencing • quality control • synthetic biology • VersaTile shuffling

Advances in nucleic acids assembly have enabled researchers to construct fragments from multiple DNA sequences orderly pieced together via single-tube reactions, such as GoldenGate cloning [1] and GoldenBraid assembly [2]. These developments have also opened the door to the creation of combinatorial libraries with techniques like VersaTile that use custom repositories of DNA building blocks and assemble them in random or rational manners, creating vast libraries with thousands of different combinations [3].

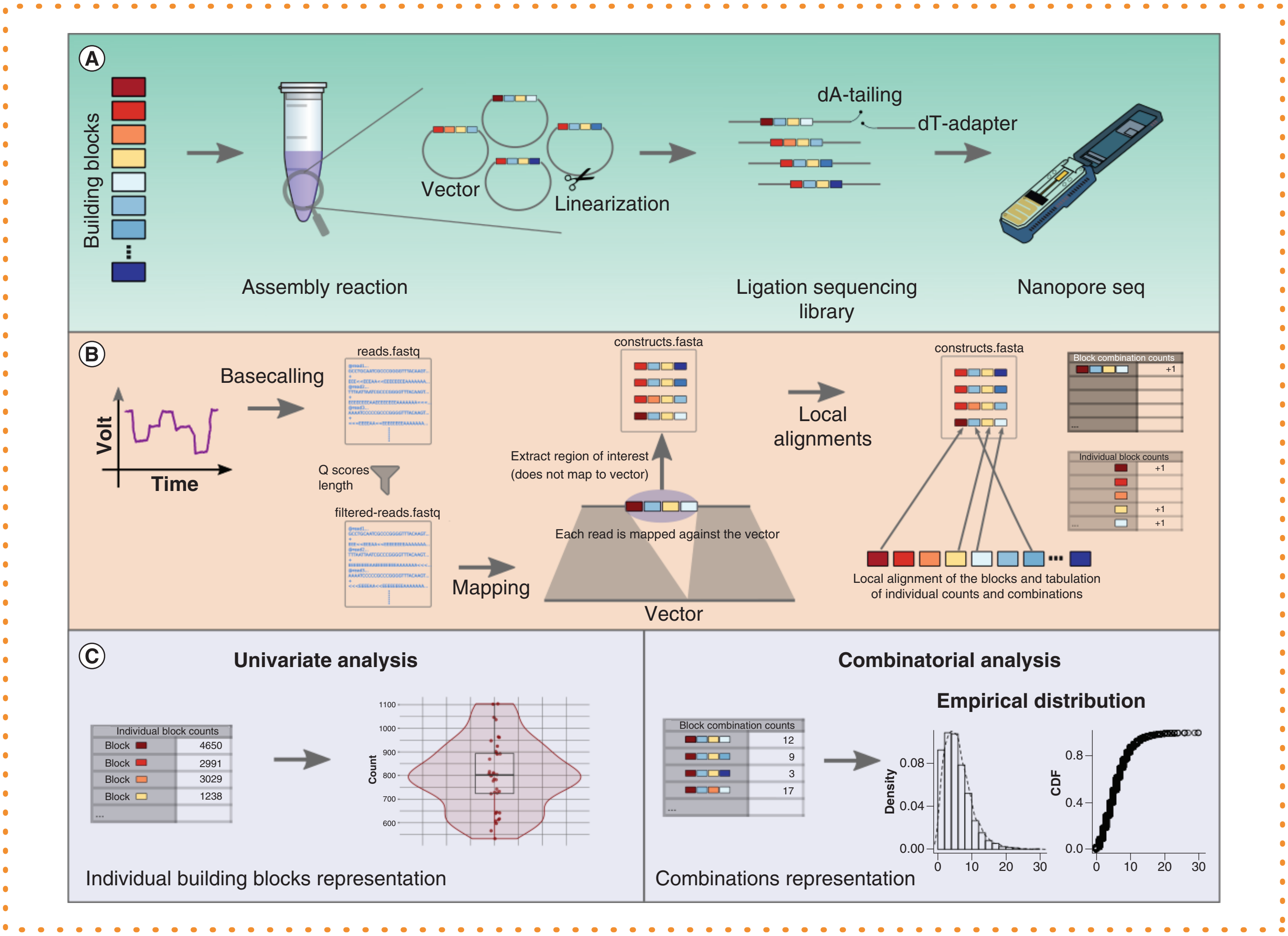


Figure 1. Assessment of a combinatorial library with nanopore sequencing. (A) Creation of the library and sequencing. After the creation of the library *in vitro*, we prepare the constructs for sequencing using a first step of restriction to linearize the vectors, and a library preparation by ligation (ONT, 1D sequencing kit) that then goes onto a nanopore sequencer. **(B)** Bioinformatics pipeline. The raw sequencing data is first basecalled and then filtered for quality based on length and Phred scores. Next, after mapping-based filtering of vector sequences, the regions of interest can be extracted using SE-MEI. We iteratively determine their make-up using local alignment and tabulate both the individual building blocks, as well as their combinations. **(C)** Data exploration and statistical analysis. The count tables can be readily used for exploratory analysis. On a block-by-block basis, we can reveal specific missingness that could point to problems with the upstream repository of building blocks, whereas by looking at combinations, we can determine the empirical distribution of the constructs and further fit models to assess the library completeness.
ONT: Oxford Nanopore Technology.

For example, recent research focused on the establishment of a combinatorial library of protein-based, engineered lysins [4]. These designer lysins are modular and contain different functional domains. Specifically, four modules were combined in a defined order, with an outer membrane permeabilizing peptide (38 different peptides), a linker oligo (one of two), a cell wall-binding domain (six in total), and finally an enzymatic activity domain (EAD; 21 active enzymes). From this 9576-variants library (38 × 2 × 6 × 21 for all combinations of the four modules, and assembled length between 712 and 1186 bp), a screen was developed to identify the most active enzymes against pathogenic *Acinetobacter baumannii* strains in human serum. Validating such combinatorial libraries is usually achieved by picking a subset of *Escherichia coli* clones for Sanger sequencing. However, in this traditional verification step, only a minimal fraction (typically less than 1%) is actually verified and therefore remains oblivious towards cloning and combination biases that may occur.

Here, we present a method to validate libraries of constructs after their creation *in vitro* to confirm the actual search space of downstream screening applications, and to assess potential biases in the reaction or problems with the upstream repository of building block vectors. This quality control relies on PCR-free, long-read sequencing of the library, preserving syntenic information by spanning the entire assembled constructs [5]. We have made an example sequencing dataset available via the NCBI database (BioProject PRJNA646613) and the related analysis via GitHub: https://github.com/LoGT-KULeuven/DNALibNanoQC.

The analysis pipeline we implemented can be broken down into three main categories (Figure 1). The first category includes the DNA manipulations where we prepare the constructs and sequence them (Figure 1A). The second category includes the bioinformatics that takes us from the sequencing results to the count tables, through mapping, local alignments and tabulation of the results (Figure 1B). And finally, the third category consists of the data analysis with the visual exploration of the count tables and statistical modeling that allows a quantitative assessment of the library (Figure 1C).

After the initial creation of the library via the assembly reaction, destination vectors carrying the constructs need to be linearized using a restriction enzyme (blunt end or sticky end). Here, care should be taken to select an enzyme whose restriction site is unique and found only in the vector, outside of the region of interest containing the assembly product. Once the vectors have been linearized, the restriction reaction can be cleaned using AMPureXP beads (Beckman Coulter, CA, USA).

In the next step, we prepare the library for nanopore sequencing. The approach we recommend is that of ligation library preparation that preserves the synteny of the DNA molecules (no fragmentation). For this purpose, we use the Ligation Sequencing Kit SQK-LSK109 available from Oxford Nanopore Technology (ONT) with an initial step of end-repair/dA-tailing with the NEBNext Ultra End Repair/dA-Tailing Module (New England Biolabs, MA, USA). Depending on the complexity of the library being assessed (i.e., the number of combinations possible), an optional step of barcoding can be included at this stage if multiple samples are to be evaluated via a single run of the nanopore sequencer. This can be done using the Native Barcoding Extension kit EXP-NBD104 (ONT).

The sequencing can take place on a MinION device equipped with a flowcell of type FLO-MIN106D (ONT), which can produce up to 50 Gbp of sequencing data, or on a flongle flowcell FLO-FLG001, which produces up to 2 Gbp of data (ONT), depending on the depth of sequencing required to evaluate the library. A typical sequencing run lasts for 48 h but can be interrupted earlier once the desired depth has been achieved, a figure that can be estimated from the real-time reporting on the nanopore control software MinKNOW. Given that the flongle is single use, it is advised to let the run proceed to completion.

Once the sequencing data have been acquired, if it has not been done during the run, we basecall it with the software Guppy (ONT), which generates the fastq files for downstream mapping analysis, as well as deconvolutes the samples in the case they were barcoded and multiple samples were pooled together in a single sequencing library. The fastq files are checked for quality using the NanoPack software suite [6], specifically NanoPlot for a visual assessment of the quality and NanoFilt to exclude reads that are shorter than the minimum size of your construct or those of poor quality (e.g., Q $<$10). This results in a new, quality-controlled fastq file that is taken further in the analysis.

To analyze the makeup of each assembled construct, we first consolidate all of the sequences of the building blocks used for the assembly into a single, multi fasta file, where each header is unique and contains a descriptive label for the sequence of the oligo/domain/gene. We then extract the loci in the vector that contain the regions of interest (Figure 1B). This last step is achieved by mapping the sequencing reads against the native destination vector with bwa-mem or minimap2 [7,8], then extracting the regions that do not map using the SE-MEI software. For each of these extracted regions, we then iterate over their building blocks composition by using the multi fasta file and the *pairwise2.align.localms* function from the Biopython library [9]. For each extracted region of interest, the start and end positions of all building blocks are determined, and these values are exported into a tabular form for further processing. One can also optionally add localization constraints to detect potential clashes between building blocks or missing positions.

Using the information gathered in the previous step, we tabulate the different building blocks present in each of the reads and form count tables where each of the items has a certain 'count' value for its absolute representation in the total sequenced DNA. This is either done in a 'block-by-block' manner library (i.e., block 1 appears *x* times, block 2 appears *y* times), giving a total count of representation for each individual building block or, in a combinatorial manner, where each assembly of a specific composition is counted (e.g., 67 different blocks with 9576 combinations in the work of Gerstmans *et al.* [3]).

Finally, these count tables are analyzed by importing them into the software suite RStudio with the R programming language [10]. Suggested initial qualitative analysis include univariate data visualization such as plotting the counts of building blocks to assess potential problems with the initial library (Figure 1C, left), and plotting of combinations of building blocks to observe if specific patterns of missingness or enrichment are present. In addition, these count tables are analyzed quantitatively by using exploratory statistics (Figure 2). Indeed, one can plot the empirical distribution of the frequencies or explore different distribution models for the combinations using the R package fitdistrplus [11]. Once a model has been selected for the distribution (e.g., Poisson distribution), it is possible to test the empirical distribution against the model fitted with the dataset via the *chisq.test* function in R.

The techniques described here can be adapted to multiple situations and library types, including combinatorial libraries encoding modular proteins, mono- and multivalent antibodies, yeast two-hybrid libraries, and synbio circuits. At this moment, there are some limitations inherent to the nanopore sequencing technology. As we are essentially analyzing single reads, we come close to the (ONT company estimated) 5% error rate and homopolymer base calling issues currently present in that platform [12]. In practice, this means that if for a given position in an assembled module you have two different candidates that differ by a few SNPs, you may not be able to distinguish them due to the signal-to-noise ratio. As a rule of thumb, the pairwise distance between blocks should be greater than the calculated sequencing error rate of the reads dataset. This value can be calculated by spiking the library with the positive control DNA CS (ONT) and using a tool like QualiMap [13] to estimate the general error rate. If the distance is lower, then we would suggest to group and count 'highly similar' building blocks as a unique block and divide the counts proportionally while we look forward to continuous improvements in the nanopore sequencing technology in the future that will render such considerations moot [14]. Another limitation we have observed is that it can be hard to predict the sequencing depth that is desirable to assess a library with a given complexity. Indeed, post sequencing analysis may reveal that a significant number of reads do not carry assembled constructs (e.g., empty vectors) and are

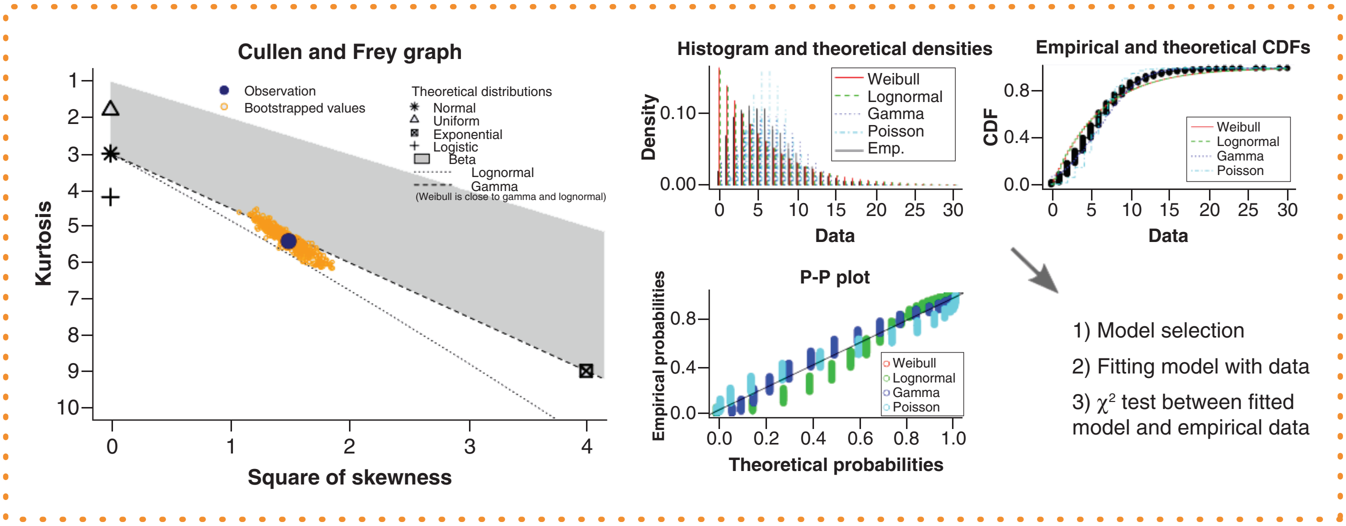


Figure 2. Statistical modeling of combinatorial library. We explore the statistic of the combinations present in the library. (Left) Using the *descdist* function available in the R package fitdistrplus, we can explore graphically how well the data fit in a set of defined distribution models (Cullen and Frey graph). (Right) We can fit our data specifically to candidate distributions via the *fitdist* function and visually explore which model fits our data best. The function also gives us the parameters of the different models, and we use these values to validate the selected model's goodness of fit using the *chisq.test* function in R.

not useful to the evaluation. In the example dataset and analysis provided with the article, about 4.3% of the reads initially present in the raw sequencing data were carried forward to the analysis of the block composition. A possible avenue to mitigate such loss could be to add an enrichment step after the assembly reaction by using capture techniques such as CRISPR-Cap [15]. If not possible, one could also use (multiple) nanopore flowcells that can yield higher throughput, such as those from PromethION.

In conclusion, this nanopore sequencing-based approach represents a necessary and important tool to assess the quality of combinatorial DNA libraries in a statistically sound manner. The approach will serve the experimental design in the field of synthetic biology and protein engineering.

Author contributions

C Lood and R Lavigne designed the study, and C Lood performed the experiments and wrote the manuscript. H Gerstmans, Y Briers and R Lavigne provided feedback on the VersaTile library analysis. All authors revised the manuscript and discussed the results.

Financial & competing interests disclosure

This work was supported by the Research Foundation–Flanders (FWO) through SB PhD fellowships for C Lood and H Gerstmans (grants 1S64720N and 1S32219N, respectively), and research project grants (grants FWOKAN2015002001 and 3G066919). H Gerstmans, R Lavigne and Y Briers are co-inventors on a patent application related to VersaTile (grant WO2018114980). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Open access

References

1. Engler C, Kandzia R, Marillonnet S. A one pot, one step, precision cloning method with high throughput capability. *PLoS ONE* 3(11), e3647 (2008).
2. Sarrion-Perdigones A, Falconi EE, Zandalinas SI *et al.* GoldenBraid: an iterative cloning system for standardized assembly of reusable genetic modules. *PLoS ONE* 6(7), e21622 (2011).
3. Gerstmans H, Grimon D, Gutiérrez D *et al.* A VersaTile driven platform for rapid hit-to-lead development of engineered lysins. *Sci. Adv.* 6(23), eaaz1136 (2020).
4. Briers Y, Walmagh M, Van Puyenbroeck V *et al.* Engineered endolysin-based 'artilysins' to combat multidrug-resistant Gram-negative pathogens. *mBio* 5(4), e01379-14 (2014).
5. Van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The third revolution in sequencing technology. *Trends Genet.* 34(9), 666–681 (2018).
6. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 34(15), 2666–2669 (2018).
7. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv:13033997 (2013).
8. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34(18), 3094–3100 (2018).
9. Cock PJA, Antao T, Chang JT *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25(11), 1422–1423 (2009).

10. R Development Core Team. R: a language and environment for statistical computing. (2020).
11. Delignette-Muller ML, Dutang C. fitdistrplus: an R package for fitting distributions. *J. Stat. Softw.* 64(i04), 1–34 (2015).
12. Senol Cali D, Kim JS, Ghose S, Alkan C, Mutlu O. Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions. *Brief. Bioinform.* 20(4), 1542–1559 (2019).
13. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* 32(2), 292–294 (2016).
14. Noakes MT, Brinkerhoff H, Laszlo AH *et al.* Increasing the accuracy of nanopore DNA sequencing using a time-varying cross membrane voltage. *Nat. Biotechnol.* 37(6), 651–656 (2019).
15. Lee J, Lim H, Jang H *et al.* CRISPR-Cap: multiplexed double-stranded DNA enrichment based on the CRISPR system. *Nucleic Acids Res.* 47(1), e1 (2019).