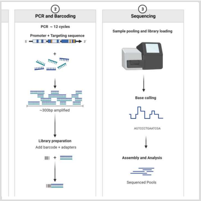

1

Mar 03, 2022

Validating Diversity in DNA Libraries through NGS V.1

Steffi Davison[1]

[1]B11 Biosciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico, USA

1

protocol .

LANL ABF

Steffi Davison

DISCLAIMER – FOR INFORMATIONAL PURPOSES ONLY; USE AT YOUR OWN RISK

The protocol content here is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to protocols.io is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with protocols.io, can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.

DNA libraries are important resources to derive targets to be used for a wide range of applications, from structural and functional studies to intracellular protein interference studies to developing new diagnostics and therapeutics. Whatever the goal, the key parameter for a **DNA library is its complexity** (also known as diversity), i.e. the number of distinct elements in the collection.

Quantitative evaluation of a DNA library complexity and quality has been for a long time inadequately addressed, due to the **high similarity and length of the sequences of the library**.

Complexity was usually inferred by the **transformation efficiency** and tested by **sequencing of a few random library elements**. Inferring complexity from such a small sampling is, however, very rudimental and gives limited information about the real diversity, because complexity does not scale linearly with sample size.

Next-generation sequencing (NGS) has opened new ways to tackle the DNA library complexity quality assessment. However, much remains to be done to fully exploit the potential of NGS for the quantitative analysis of DNA repertoires and to overcome current limitations. Even with the recent advances in NGS, it remains difficult to directly measure the representation of variant libraries, as the number of reads is insufficient to cover the size of a large library. As an example, a 1 kbp combinatorial DNA library with a billion variants has equivalent base pair content to that of 300 human genomes. Thus, brute force measurements of individual library members is impractical even with field-leading sequencing capabilities; i.e. >300 million reads at ~150 base lengths.

To obtain a more reliable DNA library complexity estimate, here we show a NGS approach to sequence DNA libraries on Illumina platform, coupled to with a bioinformatic analysis and software that allows to reliably estimate the complexity, taking in consideration the sequencing error

Jha RK, Narayanan N, Pandey N, Bingen JM, Kern TL, Johnson CW, Strauss CEM, Beckham GT, Hennelly SP, Dale T (2019). Sensor-Enabled Alleviation of Product Inhibition in Chorismate Pyruvate-Lyase.. ACS synthetic biology.
https://doi.org/10.1021/acssynbio.8b00465

Guido NJ, Handerson S, Joseph EM, Leake D, Kung LA (2016). Determination of a Screening Metric for High Diversity DNA Libraries.. PloS one.
https://doi.org/10.1371/journal.pone.0167088

Steffi Davison 2022. Validating Diversity in DNA Libraries through NGS . **protocols.io**
https://protocols.io/view/validating-diversity-in-dna-libraries-through-ngs-b5qtq5wn

protocol ,

Feb 28, 2022

Mar 03, 2022

58867

:

DISCLAIMER – FOR INFORMATIONAL PURPOSES ONLY; USE AT YOUR OWN RISK

The protocol content here is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to protocols.io is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with protocols.io, can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.

1 Library complexity inferred by the (1) transformation efficiency (**Transformation**) and/or (2) picking subset of clones/mutations for Sanger sequencing (**Sample set**) (example seen in Jha et al. 2019) or (3) sequencing the entire library (**NGS library**) 2d

Jha RK, Narayanan N, Pandey N, Bingen JM, Kern TL, Johnson CW, Strauss CEM, Beckham GT, Hennelly SP, Dale T (2019). Sensor-Enabled Alleviation of Product Inhibition in Chorismate Pyruvate-Lyase.. ACS synthetic biology. https://doi.org/10.1021/acssynbio.8b00465

Step 1 includes a Step case.

NGS library

Sample set

Transformation

Complexity inferred by NGS of the library (library sequencing preparation) 2h 30m

step case

NGS library

2 The sample of the library, with length 0.5 µL , was prepared for “shotgun” sequencing on a high-throughput Illumina MiSeq sequencer (LBL) 1h

MiSeq

Sequencer

illumina **SY-410-1003**

3 For each sample, **3.75 µL** of library DNA was prepared in a **50 µL** reaction with **0.5 µL** of Illumina Nextera and **25 µL** TD buffer at 55°C for 5 minutes. 20m

4 This reaction was then purified using a 0.6 x SPRI magnetic bead-based purification. 20m

Magnetic Stand

Magnetic Stand

Thermo Scientific **MR02**

Any magnetic rack that fits your tubes will suffice.

5 The resulting material appears on a gel as a smear of DNA with varying lengths between ~100–717bp. The material was also quantified on Qubit and Bioanalyzer. 10m

Owl™ EasyCast™ B2 Mini Gel Electrophoresis Systems
electrophoresis system
Thermo Scientific **09-528-110B**

6 The prepared DNA was used as template in a **50 µL** index PCR with NEB Next mastermix and Illumina indexing primers; and re-purified using magnetic bead-based purification. 30m

7 In final preparation, each sample was diluted to , before running on an Illumina MiSeq sequencer using reagents from an Illumina MiSeq Reagent KitV2, to produce 150 base paired-end reads. 30m

Sequence alignment to reference genome 1h 30m

8 Read mapping can be performed by two methods: (1) an implementation of Smith-Waterman alignment and (2) Bowtie2, an aligner based on the Burrows-Wheeler transform. 1h

Barbitoff YA, Abasov R, Tvorogova VE, Glotov AS, Predeus AV (2022). Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery.. BMC genomics.
https://doi.org/10.1186/s12864-022-08365-3

Xia Z, Cui Y, Zhang A, Tang T, Peng L, Huang C, Yang C, Liao X (2021). A Review of Parallel Implementations for the Smith-Waterman Algorithm.. Interdisciplinary sciences, computational life sciences.
https://doi.org/10.1007/s12539-021-00473-0

9 Sequence alignment with the reference sequence of the library used for this study appears in ICE and DIVA platforms. Namely, reference genome **CJ019** (sequenced by Bill Alexander ANL 30m

2022 using Nanopore).

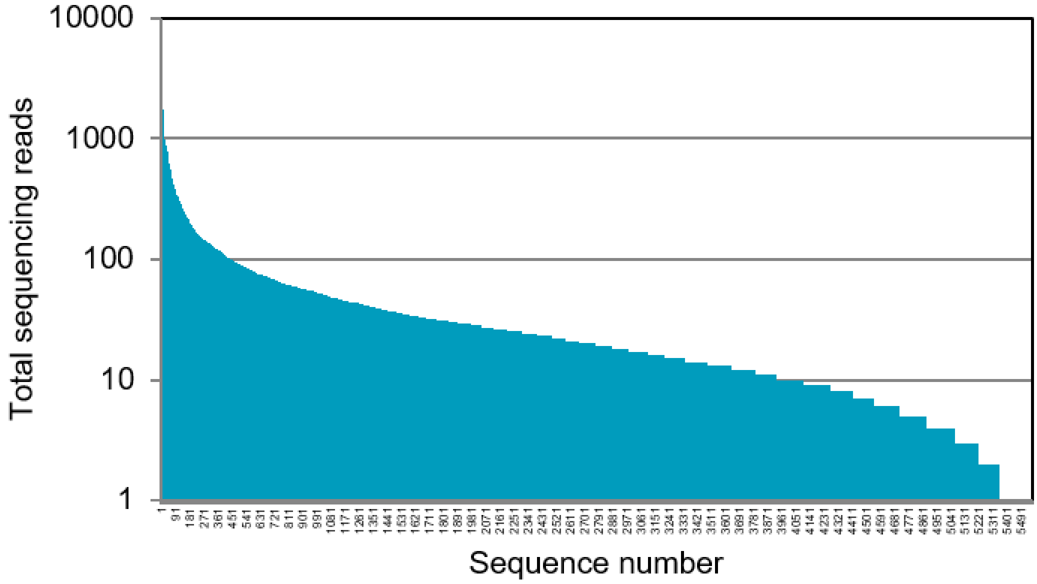

Figure 1. Library diversity screen counts.