Librarian: A quality control tool to analyse sequencing library compositions [version 1; peer review: awaiting peer review]

Kartavya Vashishta¹, Caroline Gaud², Simon Andrews², Christel Krueger²,³

¹Independent Researcher, New Delhi, India
²Bioinformatics, Babraham Institute, Cambridge, CB22 3AT, UK
³Bioinformatics, Altos Labs Cambridge Institute of Science, Cambridge, CB21 6GP, UK

Abstract
Background: Robust analysis of DNA sequencing data needs to include a set of quality control steps to ensure that technical bias is kept to a minimum. A metric easily obtained is the frequency of each of the nucleobases for each position across all sequencing reads. Here, we explore the differences in nucleobase compositions of various library types produced by standard experimental methodologies.

Methods: We obtained the compositions of nearly 3000 publicly available datasets and subjected them to Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction for a two-dimensional representation of their composition characteristics.

Results: We find that most library types result in a specific composition profile. We use this to give an estimate of how strongly the composition of a test library resembles the profiles of previously published libraries, and how likely the test sample is to be of a particular type. We introduce Librarian, a user-friendly web application and command line tool which enables checking base compositions of test libraries against known library types.

Conclusions: Library preparation methods strongly influence the per position nucleobase content. By comparing test libraries to a database of previously published library types we can make predictions regarding the library preparation method. Librarian is a user-friendly tool to access this information for quality assurance purposes as discrepancies can flag potential irregularities very early on.

Keywords
high throughput sequencing, quality control, sequencing libraries, FastQ, base composition
This article is included in the Bioinformatics gateway.

**Corresponding author:** Christel Krueger (ckrueger@altoslabs.com)

**Author roles:** Vashishtha K: Software, Writing – Review & Editing; Gaud C: Software, Writing – Review & Editing; Andrews S: Conceptualization, Funding Acquisition, Software, Writing – Review & Editing; Krueger C: Conceptualization, Formal Analysis, Software, Visualization, Writing – Original Draft Preparation

**Competing interests:** No competing interests were disclosed.

**Grant information:** This research was supported by the Babraham Institute's United Kingdom Research and Innovation - Biotechnology and Biological Sciences Research Council (UKRI- BBSRC) core capability grant (reference number BBS/E/B/000X0000) and the Epigenetics Institute Strategic Programme (reference number BBS/E/B/000C0425).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Copyright:** © 2022 Vashishtha K et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Vashishtha K, Gaud C, Andrews S and Krueger C. **Librarian: A quality control tool to analyse sequencing library compositions [version 1; peer review: awaiting peer review]** F1000Research 2022, 11:1122

https://doi.org/10.12688/f1000research.125325.1

**First published:** 29 Sep 2022, 11:1122 https://doi.org/10.12688/f1000research.125325.1
Introduction

High-throughput sequencing is now a routine technology for the analysis of biological phenomena. A multitude of methods have been developed that obtain genome-wide information on the transcriptome, protein-DNA binding, chromatin compaction, chromosomal conformation and DNA modifications to name but a few. While these approaches address different biological questions and employ various sample preparation techniques, the workflow mostly converges at a stage where adapter flanked short DNA sequences, so called libraries, are subjected to Illumina sequencing. The resulting raw data should pass a number of quality control (QC) steps before analysis is performed. These can be roughly split into two categories, pre-mapping QC, for example monitoring of base call quality scores, and post-mapping QC, for example overall enrichment scores in ChIP-seq data. For example, raw sequencing data can be queried for adapter contamination and GC bias to gauge the quality of the library preparation, or using multi-species alignments to confirm the expected species. Early detection of technical biases or problems during sample preparation is important for rigorous data analysis and conservation of resources.

FastQ is a file format commonly used for storing unmapped sequencing data. One of the metrics that can be obtained from such files is the summarised base composition across the sequencing reads. For each position in the read the respective content of the bases adenine (A), thymine (T), guanosine (G), and cytosine (C) can be determined. For a theoretic random genomic library the expectation would be four horizontal lines reflecting the overall base composition of the genome. Since the GC content of DNA varies according to species sequencing libraries will show different composition profiles depending on which organism was sequenced. Less intuitively, libraries produced by different experimental protocols may show vastly different sequence compositions (Figure 1). A prominent example is Bisulfite-seq used for DNA methylation analysis: during sample preparation unmethylated genomic Cs are converted to Ts resulting in libraries with a strikingly low C content. Another instructive example is ATAC-seq. Here, the fragments to be sequenced are produced by a transposase which shows target sequence preference; ATAC-seq libraries are therefore compositionally

![Figure 1. Per position base content for different library types. Base content across the first 50 positions of the sequencing reads was averaged for 54 ChIA-PET, 436 Bisulfite-seq, 416 ATAC-seq and 449 ChIP-seq libraries from mouse and human. Percentages are plotted for each of the four bases.](image-url)
biased at the start of the read. Expanding on these observations, we asked if base compositions could be used to distinguish different library types more generally.

The ‘Per base sequence content’ module of the widely used QC tool FastQC provides composition information for individual samples, but makes no comparison. Any judgement of whether a particular composition profile is expected for the analysed sample type would require highly specialised niche knowledge which cannot generally be expected of individual researchers. Using the tool MultiQC, researchers can collate composition information from multiple individual FastQC reports and visualise them together. This is useful to compare the base compositions of different samples in an experiment and can flag up outliers, but it does not allow for placing samples in the general base composition landscape.

Here, we describe how sample preparation protocols for sequencing libraries result in characteristic composition signatures, and introduce a new quality control tool to check any sequence library against the expected composition of its preparation method.

Methods
To get an overview of expected library compositions we queried the open Gene Expression Omnibus (GEO) database for high throughput sequencing datasets from mouse and human samples for the years 2018, 2019 and 2020. Mice and humans are among the most studied species and are similar in overall GC content (42% and 41%, respectively) making them a good choice to look for compositional differences of different library types. Search results were filtered to exclude library type ‘OTHER’ as well as under-represented types (fewer than 25 samples), and over-represented library types (e.g. ribonucleic acid (RNA)-seq) were capped at 500 samples. Figure 2A shows the number of samples per library type for which per position base compositions could be retrieved.

Figure 2. Library types can be distinguished by their base compositions. A) Number of samples per library type included in the analysis. B) UMAP representation of library compositions (reference map). C) Tile based probability map for each library type. Colour represents the percentage of a particular library type per tile. D) Heatmap illustrating the specificity of each library type for tiles of the reference map. All samples were assigned to a reference map tile and colour represents the average percentage of each library type for these tiles. E) Librarian tile probability output: Percent of each library found in the reference map tile associated with the test library.
We then determined how frequently the bases A, T, G and C were found at the first 50 positions in the read (read1 for paired-end data). To visualise sample groupings, the resulting composition data was subjected to Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction14 (using the umap R package with parameters n_neighbors = 15, min_dist = 8) and a two-dimensional representation is shown in Figure 2B ('reference map'). Interestingly, visually distinct clusters are formed largely along library types, with some library types having very specific base compositions (e.g. Bisulfite-seq, ChIA-PET, ATAC-seq) while others are largely overlapping (e.g. RNA-seq and ssRNA-seq).

To explore how well represented each library type was in each region of the reference map, we split the map into tiles and calculated the percentage of each library type per tile normalised to the total number of samples. Figure 2C shows that, indeed, some tiles are exclusively occupied by a certain library type while others are less specific. To get an overall measure for how well library types could be distinguished, we first annotated each of the samples included in the analysis with its reference map tile. We then averaged the percentages of the library types represented by the tile across all samples of a particular library type to produce the confusion matrix visualised in Figure 2D. While most tiles are very indicative of a certain library type, we also find tiles which are co-occupied by more than one type, for example ncRNA-seq and miRNA-seq. Base composition similarity of certain library types comes as no surprise as the probed material and involved preparation methods can be largely overlapping.

Having concluded that different library types result in largely distinct base compositions along the sequencing reads, we propose to include checking library compositions as a pre-mapping quality control step in the analysis of high throughput sequencing data. This will help flag technical irregularities during sample preparation or potential sample swaps early on and avoid bias during downstream analyses. To make this generally accessible, we developed Librarian15 which allows the user to relate the base composition of any newly sequenced library to other samples in the database.

**Implementation**

Librarian will first extract base composition of the first 50 positions of randomly selected 100,000 reads from a supplied FastQ file. It will then project the compositions of the test library onto the manifold created by all libraries in the database as described above, thereby assigning it to a tile on the compositions map.

Results are presented graphically: The location of the test sample is indicated both on the compositions map and on the plots displaying the probability of each library type per tile. Moreover, the percentages for each library type for the tile assigned to the test library are plotted as a heatmap. This lets the user easily gauge how similar the test library is to a collection of published library types.

**Operation**

Librarian is available as a web app and a command line tool. In the web app, one or more FastQ files are selected and processed locally to produce the library compositions. Client-side processing avoids upload of large FastQ files and potentially sensitive data. The resulting library composition is compared to the database on the server, and the graphical output can be viewed and downloaded in svg format from the web page.

Librarian can also be run as a command line client application on Linux. Download and install instructions are provided via GitHub (see Software availability). Multiple FastQ files can be processed in the same query and summarised output plots are produced. Just as for the web app, library compositions are compared to the online database to ensure integration of future database expansions with additional library types.

**Use case**

As a use case we assume that a researcher has submitted three samples for sequencing and has now received FastQ files from the provider (use case input13). They want to check if the data conforms to the expectation of the respective library preparation (i.e. RNA-seq, BS-seq and ATAC-seq). Using the Librarian web app, they choose the FastQ files from a directory on their computer and are presented with a graphical representation of how similar their libraries are to published ones regarding their base composition, and a prediction of how likely these samples are to be of a particular library type (use case output13). Any discrepancy to the expected library type should be considered a red flag and investigated further.

Another use case would be for a sequencing facility to run Librarian together with other QC packages and provide results to users together with FastQ files as standard.
Discussion
Our analyses demonstrate that the base composition of sequencing libraries is heavily influenced by the method through which the library was prepared. This finding can be used as an early quality assurance step for newly sequenced or publicly available data. A sample not matching its expected composition should raise a red flag and the underlying cause should be investigated before moving on with the analysis. While this could point to a sample swap or problem during library preparation, it is also possible that it is caused by a non-standard preparation method.

Of note, within our database of published sequencing libraries we find a small subset of samples which cluster with a different library type. This is nicely illustrated by a group of RNA-seq samples which fall into a region of the map which is otherwise very specific for ATAC-seq. Closer inspection of these examples reveals that their libraries were produced by tagmentation, a process that generates short DNA fragments using the same transposase as ATAC-seq. This clearly demonstrates that sequence bias at the start of the read introduced thereby has more of an impact on base composition than the difference between RNA producing genomic regions and generally open chromatin. The limited number of available tags for library types on public sequencing data repositories means that there is inherent heterogeneity within the groups. The example also illustrates that there is a need to update the library database as new methods are developed and certain commercial library preparation kits change popularity over time. We have therefore built Librarian in a way that can easily incorporate future developments.

Data availability
Underlying data

This project contains the following underlying data:

- Composition data (output from the original GEO database queries, and datasets included in the Librarian database (filtered list))
- Use case input (example FastQ files (subsampled for smaller file size))
- Use case output (Librarian plots generated from the use case input files)


Data are available under the terms of the GNU General Public License v3.0.

Software availability
Software available from: https://www.bioinformatics.babraham.ac.uk/librarian/ [Librarian web app]

Source code available from: https://github.com/DesmondWillowbrook/Librarian [Librarian command line download and install instructions]

Archived source code at time of publication: https://doi.org/10.5281/zenodo.7003739.

Licence: GNU General Public License 3.0

Author contributions
Kartavya Vashishtha: Software, Writing – Review & Editing
Caroline Gaud: Software, Writing – Review & Editing
Simon R. Andrews: Conceptualization, Funding Acquisition, Software, Writing – Review & Editing
Christel Krueger: Conceptualization, Formal Analysis, Software, Visualization, Writing – Original Draft Preparation
Acknowledgements

Librarian was originally started as an open project at the Cambridge Bioinformatics Hackathon (www.cambiohack.uk, 21st-23rd Sep 2020) with initial ideas from many people including Stephen Kanyerezi and Lordstrong Akano. We would like to thank Felix Krueger for useful discussions and critical reading of the manuscript.

References

1. Sequencing | Key methods and uses. Reference Source
The benefits of publishing with F1000Research:

• Your article is published within days, with no editorial bias
• You can publish traditional articles, null/negative results, case reports, data notes and more
• The peer review process is transparent and collaborative
• Your article is indexed in PubMed after passing peer review
• Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com