







SOFTWARE TOOL ARTICLE

REVISED Librarian: A quality control tool to analyse sequencing library compositions [version 2; peer review: 2 approved, 1 approved with reservations]Kartavya Vashishtha ¹, Caroline Gaud ², Simon Andrews ²,
Christel Krueger ^{2,3}¹Independent Researcher, New Delhi, India²Bioinformatics, Babraham Institute, Cambridge, CB22 3AT, UK³Bioinformatics, Altos Labs Cambridge Institute of Science, Cambridge, CB21 6GP, UK**v2** First published: 29 Sep 2022, 11:1122
<https://doi.org/10.12688/f1000research.125325.1>
Latest published: 24 Jan 2024, 11:1122
<https://doi.org/10.12688/f1000research.125325.2>**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.

Open Peer Review**Approval Status** ✓ ? ✓

	1	2	3
version 2 (revision) 24 Jan 2024	✓ view		✓ view
	↑		↑
version 1 29 Sep 2022	✓ view	? view	? view

1. **Andrew Keniry** , Walter and Eliza Hall Institute of Medical Research, Parkville, Australia
2. **Konstantin Okonechnikov** , German Cancer Research Center, Heidelberg, Germany
3. **Karim Gharbi** , The Earlham Institute, Norwich, UK

Any reports and responses or comments on the article can be found at the end of the article.

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/B000X0000) 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: © 2024 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 2; peer review: 2 approved, 1 approved with reservations]** F1000Research 2024, 11:1122 <https://doi.org/10.12688/f1000research.125325.2>

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

REVISED Amendments from Version 1

The revised version of this article contains clarification regarding the compositions of various library types (including a revised Figure 1) and includes new discussion points around subgrouping of library types, species effects and tool limitations. The revised manuscript goes alongside improvements of the Librarian tool itself (now including a web version, local installation with web query, and fully offline version). There is now also comprehensive online documentation and FAQs.

Any further responses from the reviewers can be found at the end of the article

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.¹ The resulting raw data should pass a number of quality control (QC) steps before analysis is performed.²⁻⁴ 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 bias³⁻⁵ to gauge the quality of the library preparation, or using multi-species alignments to confirm the expected species.⁵⁻⁷ 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 species⁸ 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-seq⁹ which is characterised by a strikingly low C content (Figure 1A). Other library preparation methods like ATAC-seq¹⁰ or ChIA-PET¹¹ produce nucleobase bias in specific regions of the read (Figure 1B, C), while ChIP-seq libraries largely reflect the genome composition (Figure 1D). 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 FastQC⁴ 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,¹² 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) database¹³ for high throughput sequencing datasets from mouse and human samples for the years 2018, 2019 and 2020.¹⁴ 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.

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

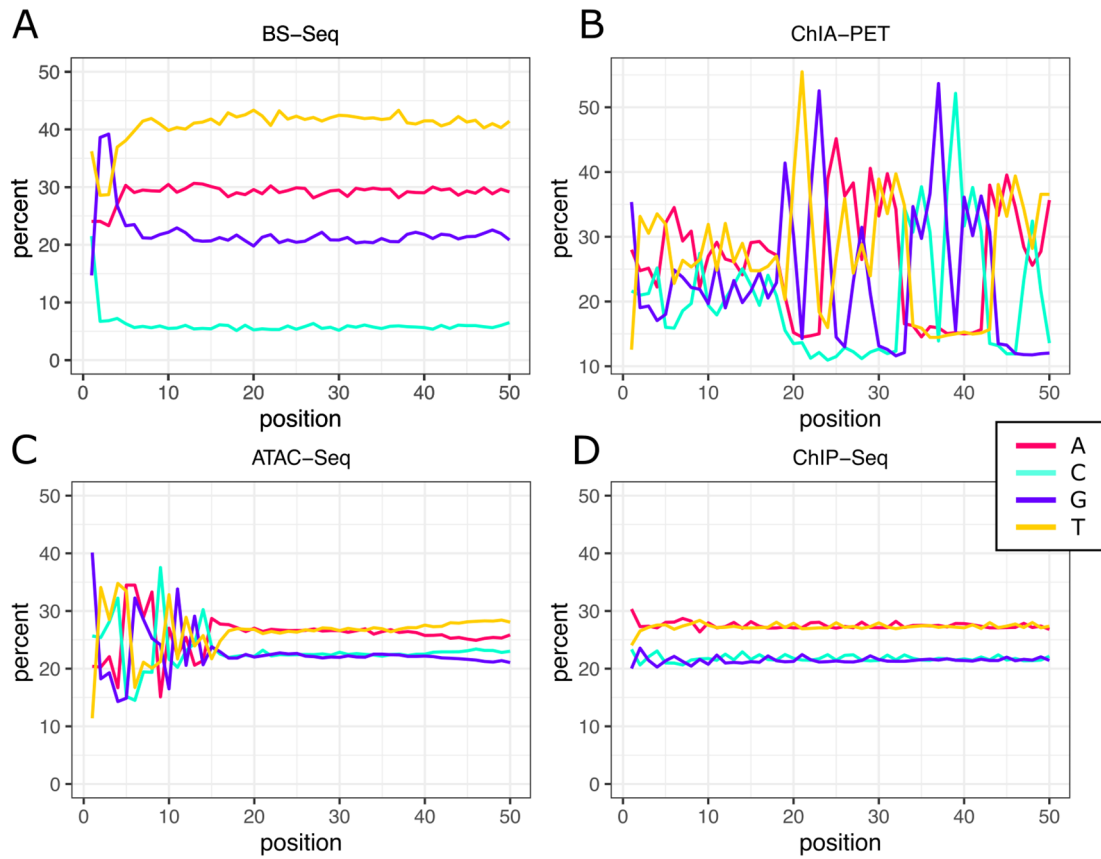


Figure 1. Per position base content for different library types. Base content across the first 50 positions of the sequencing reads was averaged for 436 Bisulfite-seq, 54 ChIA-PET, 416 ATAC-seq and 449 ChIP-seq libraries from mouse and human. Percentages are plotted for each of the four bases. Bisulfite-Seq is used to determine DNA methylation and includes conversion of unmethylated cytosines to thymines which results in the low prevalence of cytosine throughout the read. ChIA-PET is a technology that combines chromatin conformation (Hi-C) analysis with ChIP enrichment. During library preparation a linker is ligated to form paired-end tags which causes the characteristic nucleobase bias in the middle of the read. ATAC-seq is a widely used technique to assess chromatin accessibility and relies on the activity of the Tn5. Target sequence preference of this transposase leads to the observed bias at the start of the read. ChIP-seq allows for enrichment of genomic regions bound by a specific protein which can be pulled down using antibodies. However, as the majority of reads from these experiments originate in non-enriched regions, their base composition mainly reflects the GC content of the organism's genome.

Approximation and Projection (UMAP) dimensionality reduction¹⁵ (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). No systematic difference between mouse and human samples is observed (Supplemental Information¹⁴).

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.

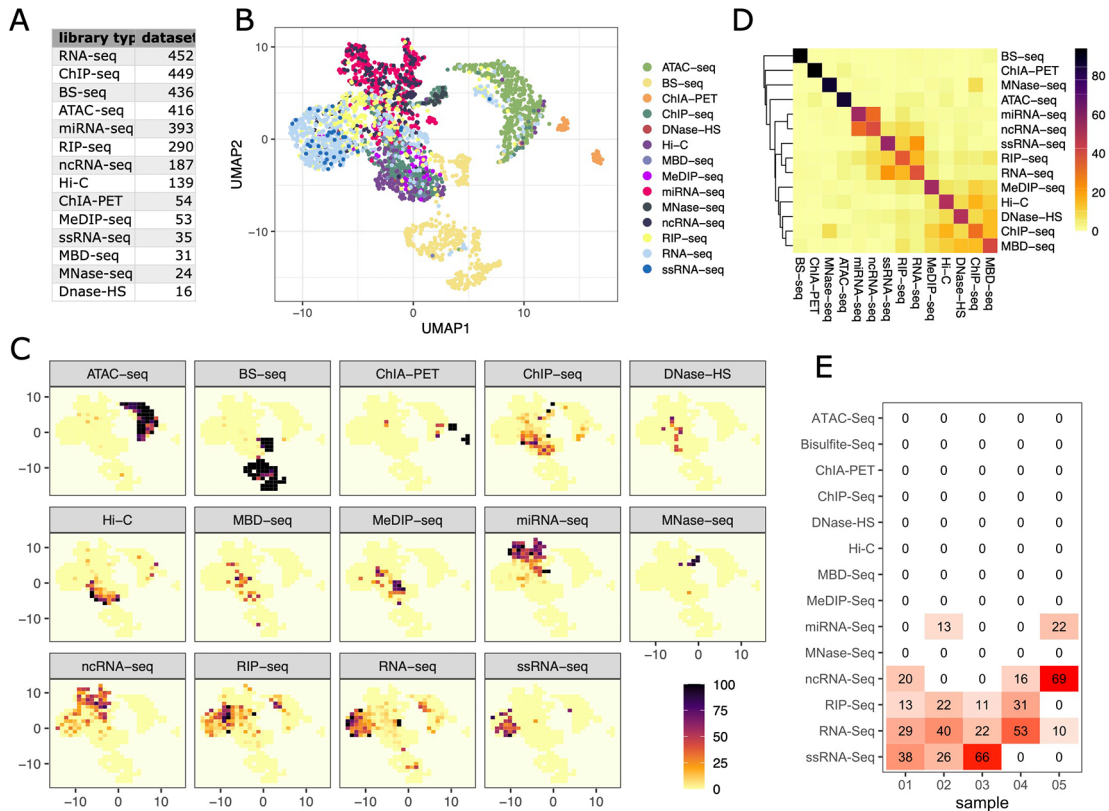


Figure 2. Library types can be distinguished by their base compositions. A) Number of samples per library type included in the analysis. B) UMAP representation of library compositions (reference map). C) Tile based probability map for each library type. Colour represents the percentage of a particular 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) Example of Librarian output: Librarian places each test sample in the two-dimensional space of the reference map. This heatmap shows the percent of each library type found in tile associated with the test library.

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 Librarian¹⁶ 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 reference map.

Results are presented graphically: The location of the test sample is indicated both on the reference 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 (Figure 2E). 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. Two modes of cli operation are available: Online query: 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. Offline: A standalone application is provided which includes the database and allows for analysis in closed environments.

Irrespective of platform, Librarian is only suitable to assess datasets which match the types that the reference map is built on. More specifically, test samples need to have been sequenced with Illumina technology, match any of the included library types and be of mammalian origin (ideally mouse or human).

Documentation for installation and usage, together with FAQs and best practices can be found at <https://desmondwillowbrook.github.io/Librarian/>.

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 input¹⁴). 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 output¹⁴). Any discrepancy to the expected library type should be considered a red flag and investigated further.

Two other use cases could be for a sequencing facility to run Librarian together with other QC packages and provide results to users together with FastQ files as standard. Also, when selecting publicly available datasets for meta-analysis, Librarian can be useful to identify subtypes or biases within the collection.

Discussion

Our analyses demonstrate that the base composition of sequencing libraries is heavily influenced by the method through which the library was prepared. While this may apply to any sequencing technology, we focussed our efforts on Illumina sequencing as it is by far the most commonly used technology and offers the most diverse applications. Checking the per position base composition 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,¹⁷ 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.

We also note that some library types produce distinct subclusters on the reference map. This is particularly obvious for BS-seq libraries. The reason behind this is that BS-seq encompasses a number of distinct library preparation and data processing protocols (e.g. whole genome bisulfite sequencing (WGBS) vs reduced representation bisulfite sequencing (RRBS), or post bisulfite adapter tagging (PBAT) vs non-directional libraries) which produce distinctive base composition profiles. However, only a limited number of metadata tags for library types are available when submitting sequencing data to the GEO database. As sequencing methods diversify, library types grouped by these tags may become more heterogeneous. This illustrates the 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 when additional information is shared by the public repository.

Data availability

Underlying data

Zenodo: Librarian manuscript data v1, <https://doi.org/10.5281/zenodo.10535987>.¹⁴

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 (subsamped for smaller file size))
- Use case output (Librarian plots generated from the use case input files)
- Supplementary Information (species differences)

GEO database query parameters: Organism: Mus musculus OR Organism: Homo sapiens AND Platform Technology Type: “high throughput sequencing” AND Publication Date: 2018/01/01 to 2020/12/31.

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.10490625>.¹⁶

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. We also thank Phil Ewels for incorporating Librarian output into MultiQC.

References

1. **Sequencing | Key methods and uses.**
[Reference Source](#)
2. Wang L, Wang S, Li W: **RSeQC: quality control of RNA-seq experiments.** *Bioinformatics.* 2012; **28**: 2184–2185.
[PubMed Abstract](#) | [Publisher Full Text](#)
3. Okonechnikov K, Conesa A, García-Alcalde F: **Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data.** *Bioinformatics.* 2016; **32**: 292–294.
[PubMed Abstract](#) | [Publisher Full Text](#)
4. **Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data.**
[Reference Source](#)
5. Hadfield J, Eldridge MD: **Multi-genome alignment for quality control and contamination screening of next-generation sequencing data.** *Front. Genet.* 2014; **5**: 31.
6. Wingett SW, Andrews S: **FastQ Screen: A tool for multi-genome mapping and quality control.** 2018.
[Publisher Full Text](#) | [Reference Source](#)
7. Wood DE, Salzberg SL: **Kraken: ultrafast metagenomic sequence classification using exact alignments.** *Genome Biol.* 2014; **15**: R46.
[PubMed Abstract](#) | [Publisher Full Text](#)
8. Li X-Q, Du D: **Variation, Evolution, and Correlation Analysis of C+G Content and Genome or Chromosome Size in Different Kingdoms and Phyla.** *PLoS One.* 2014; **9**: e88339.
[PubMed Abstract](#) | [Publisher Full Text](#)
9. Bernstein AI, Jin P: **Chapter 3 - High-Throughput Sequencing-Based Mapping of Cytosine Modifications.** *Epigenetic Technological Applications.* Zheng YG, editor. Academic Press; 2015; 39–53.
[Publisher Full Text](#)

10. Buenrostro JD, Giresi PG, Zaba LC, *et al.*: **Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics.** *Nat. Methods.* 2013; **10**: 1213–1218.
[PubMed Abstract](#) | [Publisher Full Text](#)
11. Li G, Fullwood MJ, Xu H, *et al.*: **ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing.** *Genome Biol.* 2010; **11**(2): R22.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
12. Ewels P, Magnusson M, Lundin S, *et al.*: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**: 3047–3048.
[PubMed Abstract](#) | [Publisher Full Text](#)
13. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res.* 2002; **30**: 207–210.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Vashishtha K, Gaud C, Andrews S, *et al.*: Librarian manuscript data v2. [Dataset]. *Zenodo.* 2024.
[Publisher Full Text](#)
15. McInnes L, Healy J, Melville J: **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.** *ArXiv180203426 Cs Stat.* 2020.
16. Vashishtha K, Gaud C, Andrews S, *et al.*: *Kartavya Vashishtha/ Librarian-1.0.4.* *Zenodo.* 2024.
[Publisher Full Text](#)
17. Adey A, *et al.*: **Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition.** *Genome Biol.* 2010; **11**: R119.
[PubMed Abstract](#) | [Publisher Full Text](#)

Open Peer Review

Current Peer Review Status:   

Version 2

Reviewer Report 24 February 2024

<https://doi.org/10.5256/f1000research.161650.r240420>

© 2024 Gharbi K. This is an open access peer review report 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.



Karim Gharbi 

The Earlham Institute, Norwich, UK

The revised manuscript is significantly improved and addresses the majority of my original comments. The (current) limitations of Librarian have been clarified and accessibility has also been improved.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: genomics, next-generation sequencing, bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 15 February 2024

<https://doi.org/10.5256/f1000research.161650.r240418>

© 2024 Keniry A. This is an open access peer review report 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.



Andrew Keniry 

Molecular Medicine Division, Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, Australia

No further comments.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Epigenetics, development, cell biology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 18 October 2022

<https://doi.org/10.5256/f1000research.137618.r151968>

© 2022 Gharbi K. This is an open access peer review report 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.



Karim Gharbi 

The Earlham Institute, Norwich, UK

In this manuscript, Vashishtha et al describes the implementation of a novel quality control (QC) tool for next-generation sequencing (NGS) datasets, which uses nucleotide composition profiles along sequence reads to infer the likely library preparation method used to generate the data. The authors first demonstrate that nucleotide composition is strongly influenced by library method as recorded in the GEO database for a selection of human and mouse NGS datasets. Having established this result, they implemented a program tool (Librarian) to compare library nucleotide composition profiles to a collection of reference datasets and identify libraries with unexpected profiles, which may be indicative of potential failure during library preparation and/or sample/data mix-ups. The tool, which is available as a web application and command line tool, extracts nucleotide composition from user supplied FASTQ files and returns similarity scores against existing profiles stored in the Librarian database.

The manuscript is well-written, and the authors provide strong evidence of library method influencing nucleotide composition in the read output files. This will be a familiar observation for those experienced with generating and/or analysing diverse NGS datasets, but the manuscript is a welcome documentation and quantification of these patterns. As a tool, Librarian has the potential to become an important step in the QC of NGS data, alongside other, more generic QC tools, such as FastQC, and help detect quality issues early in data processing. However, I have some concerns about the limitations of the software as currently implemented, which I feel are not sufficiently discussed in the manuscript and could cause significant confusion in the hands of less experienced users. The comments below are intended to help the authors improve the current manuscript and indicate areas for future improvement to increase the usability of the tool.

Major comments

- Please comment on the applicability of Librarian to data generated with other NGS technologies than Illumina. If not tested or not applicable, this should be highlighted in the discussion.
- Please provide a rationale for trimming reads to 50 bases and only considering read 1 to build

the database of nucleotide composition profiles, i.e., why is this sufficient to accurately capture the nucleotide composition of each library type. Some methods result in asymmetric library fragments (e.g., 10X Genomics), with different nucleotide compositions in read 1 and read 2, which in itself can be diagnostic of the library type.

- The selection of GEO datasets to build reference profiles seems restrictive and potentially biased. Please can you provide evidence that Librarian is applicable to other species than human/mouse, especially species with divergent GC content.

- The date range filter (01/01/2018 - 31/12/2020) is also likely to have resulted in more recent library types to be excluded from the analysis and therefore the reference dataset. 10X Genomics library types are highly popular, but surprisingly absent. Other library types may have been missed too.

- Transposon-based library preparation is increasingly popular and applied across a wide range of library types, including single-cell RNA and DNA sequencing, whole-genome sequencing, ATAC-seq, enrichment capture etc. The authors briefly acknowledge this in the discussion, but it appears to be a major limitation of the tool, i.e., transposon-insertion signature at the start of the reads will likely obscure the underlying library type, causing most transposon-based libraries to cluster together. This should be explicitly documented and investigated further, if possible.

- More generally speaking, I would strongly encourage the authors to explicitly identify library types and species "supported" by Librarian, indicating that submission of other library types and/or species may result in inconclusive or potentially misleading results (I acknowledge that the software will accept any FASTQ file).

Minor comments

- Please briefly comment on the observed pattern for ChIA-PET and ChIP-seq libraries, i.e., why are these expected and consistent with the library method. ChIA-PET is not a widely used library method. A short description should be included in the text for context.

- Please add legend to Figure 1 with key matching coloured lines to individual bases.

- I would suggest meta-analysis of public datasets as another important use case for Librarian, e.g., as a clean-up tool prior to meta-analysis or identifying patterns/biases in library type, or subtypes.

- Please clarify whether Librarian can be set up with a local, custom server in addition to query against an online database via the web app or command line tool.

- The tabular data in figure 2A shows library types with fewer than 25 samples despite these being classified as under-represented libraries and excluded from the analysis in the text.

Overall, I believe that the premise of Librarian is a very good idea and thank the authors for their efforts in releasing the program as a publicly available tool. I look forward to reading their responses, and future iterations of the software addressing current limitations.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: genomics, next-generation sequencing, bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 14 October 2022

<https://doi.org/10.5256/f1000research.137618.r151966>

© 2022 Okonechnikov K. This is an open access peer review report 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.



Konstantin Okonechnikov 

German Cancer Research Center, Heidelberg, Germany

The manuscript describes the quality control (QC) tool Librarian that predicts the similarity of sequencing library correctness in comparison to the control cohort based on the analysis of reads. Initially for this purpose, the composition of nucleotide variance in the first 50 bp of read was used as input to create a large reference control cohort from publicly available data. Visualisation of these merged nucleotide profiles via UMAP allows to observe clear groups formation based on the data type of a dataset. Novel sample check is a projection into this reference UMAP. Testing the online tool confirmed its usefulness: from inspection of own data, the majority of cases were distinguished correctly. Such projection of a novel dataset into reference would be a useful QC step for any sequencing experiment. However, the manuscript and the software description could be improved in order to provide more details about the tool as well as explain certain missing

blocks.

- In general, the manuscript clearly describes the technique, however, only one possible limitation of the method is stated (effect of a cut in RNA-seq fragments leading to similarity to ATAC-seq in Discussion). More variance factors could be inspected to avoid misleading conclusions about the analysis results. For example, the tissue materials can be obtained from frozen tissue (FFPE), is there any impact of this preparation procedure? In my inspection, the standard RNA-seq datasets were distinguished quite well, but FFPE RNA-seq demonstrated the closest similarity to MBPD and MeDIP-seq. The scRNA-seq protocols are also included, however, they vary since they could be either full gene body covering or only 5'/3' segment of a gene. Could this have an impact on read nucleotide variance?
- The reads selection is performed with 100K subsampling - how was this amount selected? What is the effect of the total number of reads? Is it sufficient to provide only a subset of them? In this case, what is the suggested limit?
- Also, 50 bp read segment is used as the reference, but how was this selection made? Currently, the main standard for sequencing is 100-150 bp. Would it be more beneficial to use a larger segment of the read for reference generation? Or do quality issues in longer reads have a negative impact?
- How strong is the species effect? Are there variances observed between mice and human materials in full UMAP, e.g. clusters formation? Does it make sense to create own reference for such a procedure, especially when working on other species, e.g. Drosophila?

Further additional comments could help to improve the manuscript for easier reading:

- In Figure 1, the nucleotide type color legend is missing, also segments are not cited in the text directly by suffix (a,b,c,d). Figure 1a demonstrates ChIA-PET, but not clear why it is included since it's not stated in the manuscript text.
- Figure 2a: Are the amounts of mice and humans mixed? What is the variance?
- Figure 2c: Several enrichment UMAP locations for certain data types are far from each other, e.g. RNA-seq. How to interpret this? Could it be certain subclusters, further splitting the dataset types?
- When downloading example datasets (sample FASTQ files), the archive cannot be opened. Also, there is no documentation available regarding input format preparation, e.g. fastq are not allowed to be gzipped, it's not clear without testing.
- Github documentation on the establishment/launch lacks some details. Would be useful to extend it especially to state what are the system environment requirements before installation.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

No

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics data analysis in pediatric neurooncology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 06 October 2022

<https://doi.org/10.5256/f1000research.137618.r151965>

© 2022 Keniry A. This is an open access peer review report 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.



Andrew Keniry 

Molecular Medicine Division, Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, Australia

Vashishtha and colleagues perform an analysis of the base composition from ~3000 publicly available sequencing data sets and show that these segregate by library type in a UMAP analysis of fastq files. The authors suggest that this analysis could be a useful pre-mapping QC step to identify incorrect libraries early in analysis pipelines and provide a tool, Librarian, for performing this test. Such an analysis could certainly be useful and has the potential to be widely adopted. I have a few suggestions that could improve the manuscript:

1. The authors show that Librarian identifies libraries prepared by different techniques, but is it able to identify a failed sample within a specific technique? Can the method be tested on failed samples such as ChIP-seq without enrichment, BS-seq with a low conversion efficiency, or samples with high duplication?
2. BS-seq seems to segregate into multiple clusters. Is there an easily identifiable reason for

this – perhaps enrichment techniques or developmental stage?

3. I'm not sure of the logistics of this, but Librarian may be more widely used if it was available as an option within the already widely used fastqc.
4. An example of the Librarian output would be beneficial.
5. The terms 'reference map' and 'compositions map' seem to be used interchangeably. For simplicity, one term should be used throughout.
6. Fig 1A shows the base composition of ChIA-pet data. As this is a less well known technique, it would be beneficial to have an explanation of the base composition results.
7. Fig 1 is missing a legend explaining which base each colour represents.
8. I'm not certain what Fig 2E is showing. Could the authors provide more explanation in the legend? There is also no reference to this figure in the text.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Epigenetics, development, cell biology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

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

F1000Research