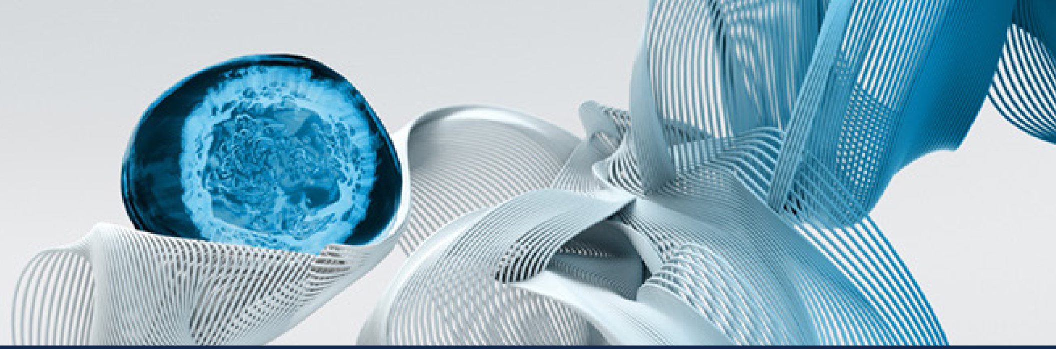

BioSkryb GENOMICS

Technical Note

Data Preprocessing in BaseJumper™ — Cells explored. Answers revealed.

Author

Tia Tate, PhD
Sr. Computational Biologist
Bioinformatics
BioSkryb Genomics

ResolveOME™ and ResolveDNA® Data Preprocessing with Fastp Using the Cloud-based BaseJumper™ Platform

Ensuring accurate, high quality, and high confidence sequence data is of utmost importance in the single-cell genomics and transcriptomics fields. A critical step in achieving this is effective removal of artifacts introduced during library preparation and sequencing. Removal of artifacts through data preprocessing has profound impact on downstream analyses. Therefore, it is essential to employ robust and dependable tools for preprocessing of single cell genomics and transcriptomics data. Here, we describe how the FASTQ preprocessing tool, fastp, is implemented within BioSkryb Genomic's cloud-based bioinformatics platform, BaseJumper™, and the impact on interpretation of ResolveDNA® and ResolveOME™ single-cell sequencing data.

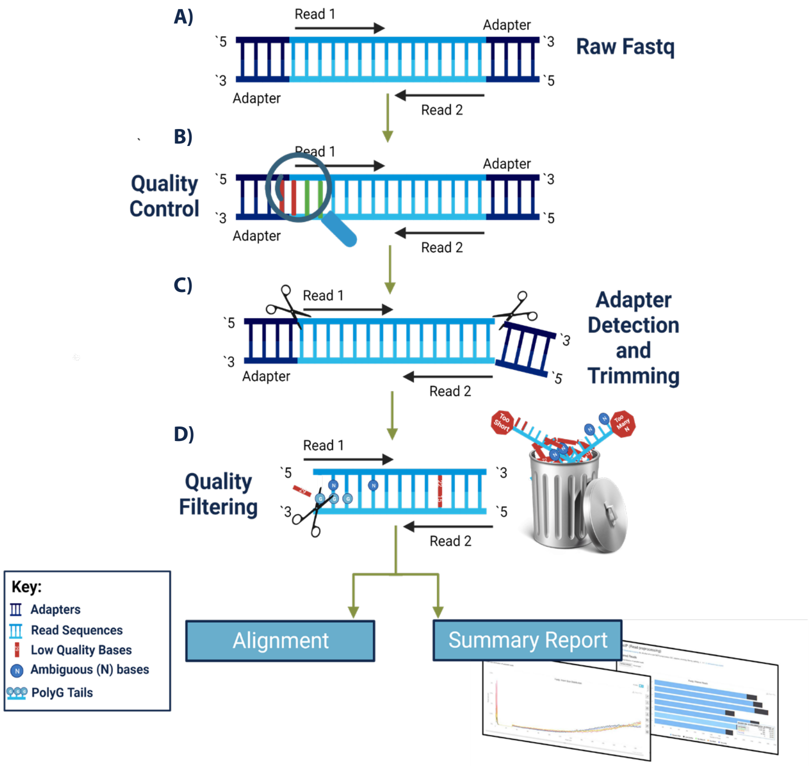

Figure 1. Common process of read trimming, clipping, and removal based on quality by fastp in paired end sequencing approaches. A) Original FASTQ file, direct from the sequencer capped by sequencing adapters. B) Fastp first investigates the call and quality of each base that comprises the read. C) If parts of the known adapter sequence are found at the 3' end of either read, those bases are removed. This indicates a library shorter than the intended read length and a potential for called bases to be technical artifacts. D) Following adapter removal, contiguous stretches of monomers or low quality bases are trimmed from the 3' end of the read. If after removal of all of the above, the read is less than the threshold, the entire read pair is removed prior to secondary analysis. Figure created with BioRender.

Introduction

BioSkryb Genomic's ResolveDNA workflow enables accurate, complete-genome assessment of single-nucleotide variation, copy number variation, and other genomic aberrations from single cells and the ResolveOME workflow offers the same with the addition of full-transcript RNA-seq from the same cell[1]. However, like all DNA and RNA sequencing workflows, a number of factors can negatively impact data quality. These include DNA contaminants or other impurities, low-quality sequencing reads, adapter sequences, over- or under-represented sequences, and others. Data preprocessing minimizes these biases, leading to more reliable and reproducible research findings. These indispensable preprocessing steps play an important role in facilitating meaningful and interpretable downstream analyses of data.

Fastp is a preprocessing tool that aims to enhance the management of genomic data for researchers[2]. With a strong emphasis on speed, accuracy, and versatility, fastp offers a comprehensive and user-friendly solution for processing high-throughput sequencing data. As a result, it has gained significant traction among scientists looking to streamline their sequencing data preprocessing workflows.

BioSkryb's BaseJumper is a cloud-based bioinformatics platform designed to perform analyses on the sequencing data produced from ResolveDNA, ResolveOME, and other sources to yield interpretable results[3]. It has pre-built workflows for primary and secondary analyses that automate preprocessing, including preprocessing by fastp. Here we describe how fastp is implemented within BaseJumper and the impact data preprocessing has on interpretation of single-cell genomics and transcriptomics experiments.

Native Functionality of Fastp

The FASTQ preprocessor, fastp, includes a suite of features that makes it an attractive option for DNA and RNA sequence data filtering and quality control. Following is a description of key features of fastp.

- **Adapter trimming:** Fastp automatically detects and removes adapter sequencing from single-end and paired-end reads.
- **Quality filtering:** The tool filters out low-quality, too short, or too many ambiguous (N) reads based on user-defined quality thresholds. This improves data reliability and reduces noise.
- **Base correction for homopolymer regions:** It corrects potential base errors in homopolymer regions that may arise during sequencing.
- **Length filtering:** Users can set minimum and maximum read length thresholds, allowing customization according to specific project requirements.
- **PolyG tail trimming:** Fastp trims polyG tails ≥10bp that may result from certain sequencing protocols (NextSeq/NovaSeq). This helps prevent bias in downstream analyses.
- **Duplication removal:** The tool identifies and eliminates duplicate reads. This is important for reducing redundancy and optimizing downstream analysis efficiency.
- **Error rate estimation:** Fastp provides error rate estimation based on quality scores, aiding in the assessment of data reliability.
- **Output unmatched reads:** Fastp saves unmatched reads (e.g. untrimmed reads) in separate files for further analysis.

BaseJumper Implementation of Fastp

Read trimming and sequencing quality evaluation using fastp is integrated into BaseJumper's RNA expression and DNA-QC pipelines. In both the RNA expression and DNA-QC pipelines, preprocessing with fastp takes place subsequent to read subsampling. Preprocessing is completed with a series of steps (Figure 1):

1. **Quality control (QC):** Fastp analyzes the quality of raw sequencing fastq inputs by calculating quality Phred scores for each base and then identifying and tagging low quality regions.
2. **Adapter detection and trimming:** Fastp scans the input reads for the presence of adapter sequences utilizing user specified adapter sequences as references.
3. **Quality filtering:** Based on the quality scores obtained during the QC analysis, fastp performs quality filtering. It removes reads or bases with low-quality scores that do not meet defined quality thresholds. This includes bases or reads that do not meet Phred quality score of 30, reads deemed too short (BaseJumper default is 15 base pairs), and reads with too many "N" or ambiguous bases (BaseJumper default is 5 bases).

Sources of Poor Data Quality

The fastp quality outputs generated using BaseJumper provide valuable information about the quality of the raw sequencing data obtained from ResolveDNA and ResolveOME experiments. This information can in turn be used to evaluate performance of experimental steps. For example, if there is a high output value for number of reads deemed too short (e.g. ≤15 base pairs) this could indicate that there are issues with library preparation, fragmentation, or sequencing chemistry. Additionally, low quality outputs mean there are reads with low Phred scores caused by poor base call qualities. Technical issues with sequencing or poor reagent quality could be sources for low quality outputs. Finally, a high output of ambiguous bases (too many N-calls) may be caused by sequencing sample degradation.

Impact of Data Preprocessing

Employing fastp for data preprocessing in BaseJumper has a profound impact on data processing efficiency. Fastp is designed to handle large scale datasets efficiently. This enables faster data processing and reduces computational burden on BaseJumper servers. Additionally, because fastp removes low-quality reads and adapter sequences, the tool reduces the overall data size, which is beneficial for storage and subsequent computational analyses in BaseJumper.

Data preprocessing using fastp also ensures downstream analyses of ResolveDNA and ResolveOME data are reliable and reproducible. Specifically, preprocessing ensures only high-quality single-cell profiles are retained for analysis by identifying and filtering out low-quality cells from ResolveDNA and ResolveOME experiments. Secondly, fastp preprocessing ensures that adapters and other artifacts that can interfere with alignment and lead to biased results are removed from ResolveDNA and ResolveOME data. Additionally, data preprocessing facilitates batch correction which is important for integrating single cell RNAseq data from different experiments, platforms, or batches. This reduces technical variation and enables more accurate cell type identification from RNAseq data. Finally, accurate trimming permits high confidence variant calling and structural variant identification.

Because data preprocessing enhances reliability and reproducibility of sequencing data analyses it ultimately enables confident conclusions to be drawn from ResolveDNA and ResolveOME experiments. Data preprocessing allows for reliable identification of differentially expressed genes and cell type specific markers as well as better identification of genetic variants associated with diseases. This is especially impactful for biomarker discovery, pinpointing genetic disease drivers and therapeutic targets, identifying rare tumor subclones and shedding light on tumor heterogeneity, and robustly characterizing off-target effects of gene editing.

Conclusion

Data preprocessing is a critical step in achieving accurate, high quality, and high confidence sequence data for single-cell genomics and transcriptomics analysis pipelines. BaseJumper harnesses the power of fastp to efficiently assess the quality of sequencing reads, detect and trim adapters, and filter out reads that fail to meet pre-defined quality thresholds. Data preprocessing using BaseJumper ensures downstream analyses of sequencing data are reliable and reproducible and ultimately enables confident conclusions to be drawn from ResolveDNA and ResolveOME experiments.

References

1. Marks, J.R. *et al.* Unifying comprehensive genomics and transcriptomics in individual cells to illuminate oncogenic and drug resistance mechanisms. bioRxiv. [preprint] July 19, 2023. Available from: https://doi.org/10.1101/2022.04.29.489440
2. Chen, S. *et al.* fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics **34**, i884-i890 (2018). Available from: https://doi.org/10.1093/bioinformatics/bty560
3. Bioskryb Genomics. *BaseJumper Docs*, https://docs.basejumper.bioskryb.com/. Accessed 25 August 2023

For more information or technical assistance:
techsupport@bioskryb.com

Published by:

BioSkryb
GENOMICS

2810 Meridian Pkwy, Suite 110
Durham, NC 27713
bioskryb.com

All data on file.
BIOSKRYB, RESOLVEDNA, RESOLVEOME, and BASEJUMPER are trademarks of BioSkryb, Inc.
All other product names and trademarks are the property of their respective owners.

TAS_056, 09/2023

For Research Use Only. Not for use in diagnostic procedures.