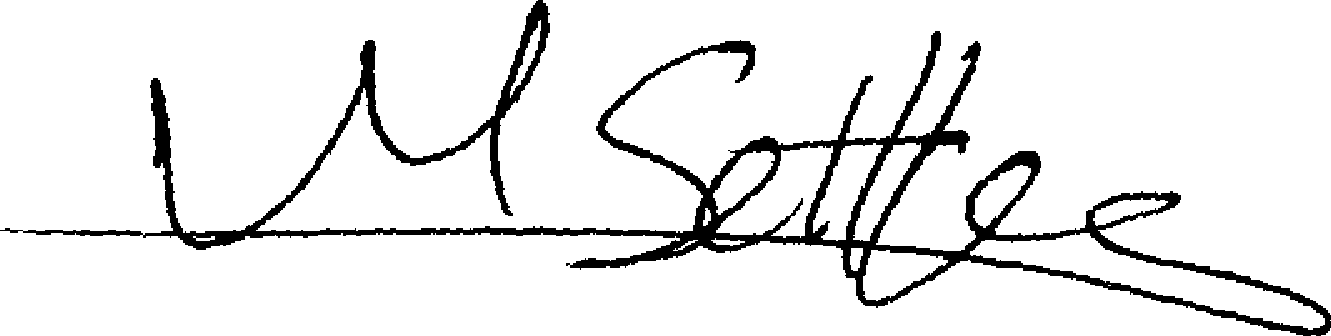
April 15, 2014

Dear Editor:

We would like to submit a presubmission enquiry regarding a suitability of our manuscript, “SeqyClean: a pipeline for high-throughput sequence data preprocessing” for Nature Methods. We developed, implemented and evaluated pre-processing pipeline, SeqyClean, to reduce technical and biological noise from High-Throughput Sequence (HTS) data that are introduced during sample preparation and instrument imprecision. SeqyClean currently is the most comprehensive HTS preprocessing tool and significantly outperforms other known tools in terms of quality of *de novo* assembly and mapping of sequence data.SeqyClean facilitates comprehensive sequence cleaning, making downstream analysis less error-prone and more accurate. In particular: (1) We show that the SeqyClean pipeline gives reasonably good results (for correctly chosen input parameters) in genome assembly and mapping, compared to the best available alternative pre-processing applications; (2) We show that SeqyClean has significantly higher values of N50 and mean contig lengths for assembly, compared to other HTS data pre-processing applications; (3) We show that SeqyClean outputs fewer short contigs (less than 1000 base pairs) than alternative software, producing assemblies close to reference genomes in publicly available sources such as GenBank; and (4) We show that SeqyClean significantly reduces false-positive SNPs, indels and number of misaligned sequences for downstream mapping analysis.

I have attached a summary of the manuscript. Thank you in advance for considering this pre-submission inquiry.

Sincerely,



Matt Settles, Ph.D.

Research Assistant Professor, Department of Biological Sciences

Director, UI IBEST Genomics Resources Core Facility

Initiative for Bioinformatics and Evolutionary Studies (IBEST)

University of Idaho, Moscow, ID

**TITLE:** SeqyClean: a pipeline for high-throughput sequence data preprocessing

**AUTHORS:** Ilya Y. Zhbannikov1, Samuel S. Hunter2, James A. Foster2,3, Matthew L. Settles2,3\*

1Bioinformatics and Computational Biology Program, University of Idaho, Moscow, ID 83844-3051

2Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID 83844-3051

3Department of Biological Sciences, University of Idaho, Moscow, ID 83844-3051

\* Corresponding Author: Matthew L. Settles; [msettles@uidaho.edu](mailto:msettles@uidaho.edu); Phone (208) 885-6051

**ABSTRACT**

An important problem in genomic studies is a very large amount of sequence data, and accompanying noise. Sequence noise has significant impact on downstream analysis. Although a plethora of preprocessing software applications developed to reduce sequence noise has already been proposed, many of them are not capable to handle data from multiple technologies and a few are able to provide reduction of more than one kind of noise, thereby limiting applicability of a majority of preprocessing applications. We propose SeqyClean, a comprehensive preprocessing software pipeline developed to alleviate these limitations.

***Background***

With the ability to easily outsource genome scale DNA sequencing and advances in modern computation tools, small labs are able to handle research tasks that were previously available only to large genomic centers [1]. However, analysis of high-throughput DNA sequence data (HTS) is a non-trivial problem. Data from high-throughput sequencing machines contain various types of artificial inclusions (sequencing adapters, vectors, contaminants), and sequence errors caused by instrument imprecision, resulting in a poor quality of analysis. Therefore there is a need for data preprocessing before downstream analysis tasks such as mapping and assembly applications. Aggressive preprocessing of HTS data, before mapping and/or assembly, in order to remove a majority of inclusions (vectors, adapters and contaminants) and potentially erroneous nucleotides is important because it can improve the quality, speed and reliability of analysis [2].

Current sequence preprocessing tools are divided into tools that perform a single preprocessing stage and those that provide several preprocessing stages, packaged together. Many of them have only some of the stages described above, or lack the ability to process sequence data from different sequencing technologies, since different technologies use different output file formats.

SeqyClean is a comprehensive preprocessing software pipeline for high throughput sequence data. The purpose of SeqyClean is to incorporate all of the sequence preprocessing stages together into one bioinformatics pipeline that works with the two most common sequence data formats, SFF and FASTQ files (both single-end and paired-end).

***Methodology***

SeqyClean is an open-source software application. The workflow diagram comprises of the following stages: (1) Input data preprocessing; (2) PCR-Duplicates removal; (3) Overlapping and adapter removal for paired-end reads; (4) Trimming poly A/T tails; (5) Vector trimming; (6) Contaminant removal; (7) Adapter trimming for single-end reads (454, Illumina single-end); (8) Quality trimming; (9) Establish final trim points; and (10) Generating output files and summary statistics. An advantage of SeqyClean is its modular structure, in which the user can implement different preprocessing strategies, rather than a strictly determined workflow. It also incorporates all of the preprocessing stages described above into single bioinformatics pipeline, providing the most powerful sequence cleaning. SeqyClean also has a number of additional options that are useful in preprocessing, such as a minimum read length cutoff and conversion from the newer CASAVA 1.8 style FASTQ read IDs to the older pre-CASAVA 1.8 style read ids.

***Significance***

SeqyClean successfully recognizes and removes technological components (adapters, primers, barcodes), contaminants and vectors. Further, SeqyClean provides quality trimming, poly A/T trimming, overlapping of paired-end reads and PCR-duplicate detection and removal. Further, SeqyClean allows the user to choose which stages to perform and to adjust the default parameters within stages, as the experimental conditional may need. We show that SeqyClean greatly improves both genome *de-novo* assembly and genome mapping and outperforms other best available preprocessing applications, such as Lucy [3], AlienTrimmer [2], Btrim [4], Ea-utils [5], Skewer [6], Trimmomatic [7], AdapterRemoval [8] and Sickle [9] in terms of quality of downstream analysis.

***References***

1. Schatz MC, Delcher AL and Salzberg SL, Assembly of large genomes using second-generation sequencing, Genome Res, 2010 20: 1165-1173.
2. Criscuolo, A., & Brisse, S. (2013). Alientrimmer: A tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. Genomics.
3. Chou, H., & Holmes, M. H. (2001). DNA sequence quality trimming and vector removal. Bioinformatics/computer Applications in The Biosciences, 17, 1093–1104. doi: 10.1093/ bioinformatics/17.12.1093.
4. Kong, Y (2011) Btrim: A fast, lightweight adapter and quality trimming program for next-generation sequencing technologies, Genomics, 98, 152-153.
5. Erik Aronesty (2013). TOBioiJ : "Comparison of Sequencing Utility Programs", DOI:10.2174/1875036201307010001.
6. Hongshan Jiang, Rong Lei, Shou-Wei Ding and Shuifang Zhu, Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads, BMC Bioinformatics.2014, 15:182.
7. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170.
8. Lindgreen S, AdapterRemoval: easy cleaning of next-generation sequencing reads, BMC Res Notes. 2012 Jul 2;5:337.
9. Joshi NA, Fass JN. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.29) [Software]. Available at <https://github.com/najoshi/sickle>.