Published online DD MM YYYY

Nucleic Acids Research, 2024, Vol. xx, No. xx ***1–3***
doi:10.1093/nar/gkn000

Re-implementation of ColoRMap: A method for correcting long reads with short reads

William Murray [1*]

[1]Department of Computer Science, University of Victoria, 3800 Finnerty Road, Victoria BC, V8P 5C2, Canada

ABSTRACT

ColoRMap is a hybrid method for correcting long reads by aligning high-quality Illumina short reads onto long reads. It then uses a graph-based model to construct sequences of overlapping short reads with minimal edit distance to the corresponding subregions of the long reads. In this study, we re-implement ColoRMap to replicate its effectiveness in improving alignments between long reads and a reference genome. Additionally, we investigate the impact of ColoRMap's parameters on the quality of read correction.

INTRODUCTION

Improving the quality of error prone long reads would allow them to be used in downstream analysis pipeline developed for Illumina technology. Hybrid long read correction methods utilize accurate short reads to correct long. Current hybrid methods such as (PacBioToCA,LSC,proovread) (3, 4, 5) correct long reads by using the consensus of the short read alignment onto the long reads. ColoRMap distinguishes itself from these methods as it does not rely on the consensus of short read mapping but instead constructs an sequence of overlapping short reads with a minimal edit distance to the long reads

METHODOLOGY

This version of ColoRMap implements only the shortest path method for correcting long reads and does not incorporate correction via One-End-Anchors. The source code for this version of ColoRMap is available at: https://github.com/will-murray/Colormap-Project.

Preliminaries

For a graph G, $V(G)$ and $E(G)$ are the set of vertices and edges of the graph respectively. If $(a,b) \in E(G)$ we write $a \sim b$. A short read s which has been aligned to l has the attributes s.start,s.end, which are 0-based positions on l where the alignment of s starts and ends, respectively. Additionally, s.seq is s's sequence.

Overview

The ColoRMap algorithm has 3 inputs: short reads S, long reads L, reference genome R and chunk size N. The outline of this implementation of ColoRMap goes as follows:

1. Partition S into chunks of size at most N and align each partition to L using BWA-MEM with options `-t 8 -aY -A 5 -B 11 -O 2,1 -E 4,3 -k 8 -W 10 -w 40 -r1 -D 0 -y 20 -L 30,30 -T 2.5`.

2. Construct a weighted alignment graph (WAG) G_l for each $l \in L$, where $V(G_l)$ is the set of short reads aligned with l. $E(V_l)$ is defined in the section below.

3. For each $l \in L$ and for each connected component $C \subseteq G_l$, use Dijkstra's single source shortest path algorithm to compute minimum weight $s-v$ path, where s and v are the leftmost and rightmost vertices in $V(C)$, respectively. Construct the string which is defined the reads in the $s-v$ path and replace $L[s.\text{start}:v.\text{end}]$ with this string (Using Python string slicing notation here).

Adjacency and Edge Weights in WAG

If $(a,b) \in E(G_l)$ if and only if (1) a.start $\leq$ b.start and a.end $\leq$ b.end, (2) b.start $\leq$ a.end - k + 1 and (3) the overlapping section between a and b are identical. Note that G_l is a DAG, so $(a,b) \in E(G_l)$ implies a directed edge **from** a **to** b. Moreover, the the weight of edge (a,b) is defined by the edit distance between the subsection of b.seq which is not a part of overlap between a and b.

RESULTS

Computing Resources

All tests were run using an Ubuntu VM with 11GB of base memory(RAM) and 126 GB of virtual disk space. This VM was hosted using Virtual Box 7.0.10. The host machine is a MacOS Big Sur with 16GB LPDDR3 memory and an intel 1.7 GHz Quad-Core Intel Core i7 processor and a 256 GB disk.

Comparison by Alignment to Reference Genome

The performance of the error correction was evaluated by comparing how the corrected and uncorrected reads align to the reference genome. In (1) this experiment was performed on data from 3 different organisms (E. coli, Yeast, Fruit Fly), however I did not perform any tests on the Fruit Fly Data due its size.

Data: The links to data in Table 1 of the Supplementary Appendix for (1) are broken, so I was not able to obtain exactly the same short and long read datasets which they used for evaluation. My solution was to find short and long read data that were generated from the same biosample. In the repository, the files `ecoli/init_ecoli.sh` and `yeast/init_yeast.sh` contain code to download the short reads, long reads and reference genome used for the following experiment respectively. Moreover when mapping the short reads onto the long reads with `bwa mem`, I consistently ran out of memory, so I used a subset of long reads.

Experiments : For the ecoli data set I ran 3 tests, each of which used 5000 long reads, with average read length 14617bp and the 1 million paired end short reads with length 150bp. I aligned short reads to long reads using `bwa mem` with options `-t 8 -aY -A 5 -B 11 -O 2,1 -E 4,3 -k 8 -W 10 -w 40 -r1 -D 0 -y 20 -L 30,30 -T 2.5`. I varied the number of shorts reads I aligned in a single call to `bwa mem` across these three experiments. More specifically, I aligned the short chunks of size 25000, 100000, 200000 reads for these 3 experiments respectively. Both the corrected and uncorrected long read were then aligned to the reference genome using `blasr` with option `--bestn 1` and the comparison of these two alignment is shown in **Table 1**. The tests listed below used the aforementioned parameters (chunk size, `bwa mem` option, `blasr` options) unless otherwise specified. In addition the last 2 rows of Table 1 contain the results from a test using 50000 long reads and 10 million short reads.

For the Yeast data set, only one small test was performed on 500 long reads with an average read length of 20139 bp using 10000 short reads with length 151bp.

DISCUSSION

Table 1 provides evidence that this implementation of ColoRMap increases the number of long reads which align to a reference genome. This result is consistent with the results in (1). It was noted in (1) that `bwa mem` does not report all mappings, and using a smaller chunk size would provide higher accuracy at the cost of runtime. This was completely validated by the first 4 rows in Table 1 since chunk size = 25000 performed worse that chunk size = 100,000 on all metrics except for identity. With that in mind, more than 3 samples is necessary to establish a relationship between the chunk size and the performance metrics in Table 1.

Negative Results

During testing of this method many negative results were obtained. Here I will discuss those results and some potential causes.

Low Coverage Alignments: In preliminary testing, I corrected 100 ecoli long reads with 10000 paired-end short reads. For most long reads, the WAG produced nearly as many components as vertices, with many components containing just one vertex. Correcting l with the shortest path in each component often led to replacing parts of the long read with highly dissimilar short reads. Using `bwa mem` with a low seed length (`-k 8`) and minimum score (`-T 2.5`) allowed for alignment of mismatched reads with short regions of strong matches (2). This resulted in 90% of corrected reads aligning to the reference genome, compared to 98% for uncorrected reads.

Some Reads Get Worse: Table 2 shows ColoRMap will alter some long reads so that they no longer can align to the reference genome. This was the case for 178 of the 50000 corrected long reads in this sample. It would be of interest to examine how the coverage of the alignment from short reads to long reads may affect the number of long reads that fall into this category. In addition the average identity for the 5000 and 50000 long reads tests is highest on uncorrected reads which

Table 1. Comparing the Alignment of Corrected vs Uncorrected E-coli long reads

Type	Chunk size	Total reads	Aligned reads	Aligned bp	% Aligned bp	% Matched bp	Avg Identity	Alignment size
Uncorrected	–	5000	4969	70641286	96.65%	96.34%	**99.67**%	–
ColoRMap	25000	5000	4985	71225052	97.45%	96.48%	99.00%	2050316
ColoRMap	100000	5000	**4991**	**71430050**	**97.73**%	**96.71**%	98.95%	2200335
ColoRMap	200000	5000	4987	71421252	97.72%	96.65%	98.90%	**2400362**
Uncorrected	–	50000	49627	703208136	96.65%	96.34%	**99.67**%	–
ColoRMap	100000	50000	**49729**	706634883	**97.12**%	**96.45**%	98.30%	9706985

(1) **Chunk size** = number of short reads aligned to the long reads in a single call to `bwa mem`. (2) **Aligned reads** = number of long reads that aligned to the reference genome. (3) **Aligned bp** = number of base pairs that aligned to the reference genome, **% Aligned bp** = Aligned bp/total number of base pairs in long reads.(4) **Matched bp** =total number of long reads base pairs with match in thier aligned position on the reference genome. (5) **Avg Identity** = number of aligned bp/ number of matched base pairs

.

could indicate that many reads align to the reference genome before and after correction, but the correction process makes decreases the pecertange of base pairs which align.

Runtime

My implementation of this was done in C++, and was written with a single thread.

Graph Construction: This algorithm constructs a graph for every long read. The space complexity for constructing a single graph G_l for the long read l, given S short reads aligned to L is $O(S \cdot C \cdot k^2)$ where k is the short read length and C is the average coverage of across short reads in S. In particular, For each short read i, we must query if i is adjacent to all of the reads which it overlaps. The code I used for edit distance is $O(k^2)$. The space complexity of G_l is $O(S \cdot C)$,

Shortest Path Correcting a long read l with its graph G_l costs $O(VlogV + E)$, as this step is constrained by the call to Dijkstra's.

Runtime in practice Running the entire pipeline on 5000 long reads and 1 million short reads (150 bp each) took 90 minutes. Moreover, with 50,000 long reads and 10 million short reads the runtime took 4 hours. This larger test scaled well compared to the initial test as alignment produced on this data resulted in graphs who's average vertex count was far lower.

CONCLUSION

In this study, we successfully re-implemented ColoRMap, a hybrid method for correcting long reads using short reads, and evaluated its effectiveness in improving long read alignments to a reference genome. Our results show that, consistent with previous work, ColoRMap increases the number of long reads that align with the reference genome.

While the implementation yielded positive results for E. coli and yeast datasets, challenges arose with certain test conditions, particularly in cases of low coverage alignments, which resulted in dissimilar short reads replacing sections of long reads. These issues highlight that ColoRMap is highly contingent on alignment produced `bwa mem`.

Further study of ColoRMap should include an analysis of the relationship between alignment coverage and the aforementioned performance metrics. Moreover, this method may be improved by implementing threshold for determining when an alignment might introduce errors into a long read.

Conflict of interest statement. None declared.

Table 2. Alignment Before and After Correction.

–	Alinged,U	Not aligned,U
Aligned,C	49449	178
Not aligned,C	280	93

U denotes Uncorrected, **C** denotes corrected. These statistics were derived from alignments produced from 50000 corrected and uncorrected Ecoli long reads.

REFERENCES

1. Haghshenas, E., Hach, F., Sahinalp, S. C., & Chauve, C. (2016). CoLoRMap: Correcting Long Reads by Mapping Short Reads. *Bioinformatics*, **32**(17), i545–i551. doi:10.1093/bioinformatics/btw463.
2. Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324
3. Au, K., Underwood, J., Lee, L., Wong, K., et al. (2012). Improving PacBio long read accuracy by short read alignment. *PLoS One*, *7*, e46679. doi:10.1371/journal.pone.0046679
4. Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D., and Phillippy, A. M. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, *30*, 693–700. doi:10.1038/nbt.2280
5. Hackl, T., Hedrich, R., Schultz, J., and Förster, F. (2014). proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics*, *30*, 3004–3011. doi:10.1093/bioinformatics/btu392