2016 HitSeq Abstract, Orellana. *et al.*

EDGAR: Full-length RNA transcript identification by hybrid sequencing and best edit-distance graph alignment of a single molecule read

Christian F. Orellana, Jacob E. Bogerd, Nathaniel Moorman, Paul Armistead, Corbin D. Jones, Jan F. Prins – UNC Chapel Hill

Background. Ideally, we would characterize the RNA transcriptome by sequencing full length RNA molecules harvested from a cell. Single molecule sequencing could identify novel transcripts produced in virus-infected cells, show novel splicing as a result of disease, and identify linked SNPs in transcripts isoforms. However, current single molecule sequencing technologies from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies operate at the limits of detection and are fundamentally susceptible to noise, resulting in missed or repeated nucleotides as well as errors in nucleotide identity, reaching a 10% error rate or more [1]. By circularizing the cDNA copied from the RNA, a PacBio sequencer has the opportunity to read a single molecule multiple times (subreads) and correct the separate observations with each other to create a circular consensus sequence (CCS). However this comes at the cost of limited length RNA transcripts as the overall number of nucleotides that can be sequenced before the observation fails is in the 1-10 kb regime. On the other hand, short read bulk sequencing is highly accurate but cannot reliably determine full length transcripts because we end up sequencing fragments of many different molecules and attempt to piece them together to infer the identity of the original full length transcripts, and this problem is fundamentally underdetermined [2].

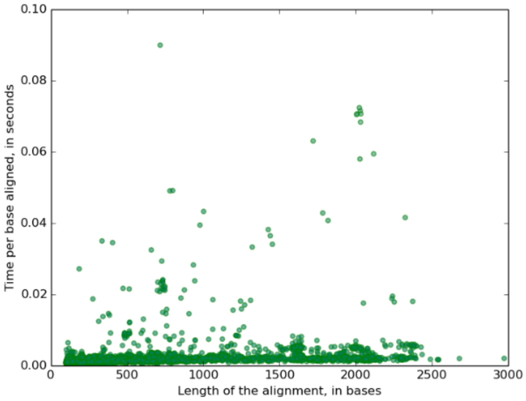
Hybrid long and short read sequencing of two aliquots of an RNA sample has been proposed as a way to combine the accuracy of short read sequencing with the full-length transcript identity of single molecule sequencing. A pioneering method introduced by Au et al. [3] corrects noisy long reads by replacing the local sequence by short reads matching closely to a given interval. In both CCS and hybrid sequencing methods, one shortcoming is correction without a broader context such as the reference genome. In the hybrid sequencing approach, a diploid cell with heterozygous SNPs may be difficult to “correct” because short reads may observe both SNPs. We propose an alternative approach to hybrid sequencing that addresses both these issues by finding the smallest edit-distance alignment of a noisy long read to a path in a directed graph constructed from short reads, either by assembly, or by (spliced) alignment to a reference genome, to form an accurate account of possible transcripts as paths in the graph. Only some of the paths will correspond to actual transcripts. We will describe our approach using the latter form of the graph, termed a splice graph.

Method. A splice graph G is a weighted, directed, multigraph in which nodes represent genomic coordinates in a reference genome and edges represent possible connections between those coordinates: exonic edges represent transcribed sequences, and splice edges join disparate exons. Additional edges are included to represent observed insertions, deletions, and SNPs. Given a single molecule observed as a noisy sequence S, identification of the transcript corresponding to S is reduced to finding that path P in G with smallest edit distance to S. The problem appears to have high complexity, since the total number of paths through the splice graph can be exponential in the number of edges in the graph, and infinite in the presence of cycles. In addition, the traditional minimum edit distance algorithm has cost proportional to the product of the lengths of the sequences being compared, compounding the cost. We developed the EDGAR algorithm to solve this problem with expected cost *linear* in the length of S, using three main strategies. (1) We perform a rapid search of approximate matches between S and the exonic edges of the graph. This search yields “seeds” that serve as starting points for alignments. The algorithm explores paths through the graph in both directions starting from a seed. (2) We define a local bound (r, n) on the number of errors permitted – in any window of length n, at most r edits are permitted. This captures our view that on a sufficiently large scale (ten to hundreds of nts) the errors are distributed uniformly. Thus, if the wrong path is being explored, the local error bound will be exceeded in distance n with high probability, and the path will be discarded, limiting the exponential growth of paths explored. The local error bound also enables a linear time alignment algorithm since the number of cells within a fixed distance r in a traditional dynamic programming tableau is linear in the length of S. (3) When multiple paths starting from one seed reach a given node in G, having aligned an identical subsequence of S, then all paths other than the minimum cost among this set can be deleted, thereby choosing early among paths that differ in a small feature like a SNP.

Results. We tested our method on a hESC hybrid dataset of Illumina and PacBio reads [4] using synthetic and experimental long read data. We generated splice graph G_1 from the short read alignments in chr1 without calling SNPs (thus exonic edges reflect the reference sequence) and generated 1000 random paths through the graph, adding errors to the sequences associated with each path with 9% probability at each nt. The error could be a replacement, insertion, or a deletion of a random nucleotide in the ratio learned from alignments of PacBio reads. This resulted in a synthetic dataset in which each read was at least 1000 nt long with an average length of 1486, for which we know the original paths sampled. Using $(r, n) = (20,100)$, EDGAR aligned 955 of these reads fully to the right path, 43 aligned partially (i.e. exceeded the error threshold at some point), and 2 aligned with one wrong exon at the end. Using $(r, n) = (15,100)$ EDGAR aligned 723 reads fully, 271 reads partially, and 6 reads with at least one incorrect exon at

	corrected CCS	uncorrected CCS	subreads
#reads/full/part align	994/994/0 (*)	994/926/52	909/225/559
avg read length	2078	2155	1920
snps incl/matched	16689/16689 = 100%	15837/16665 = 95%	7929/8909 = 89%

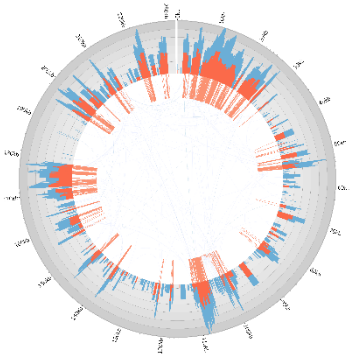
the end of the path, or in one case skipping a 3 nt early 5’-end splice site of an exon. This establishes the accuracy of the method. Next we used 994 long read CCS from [4] that were corrected using the method of Au, *et al*. [3] which had full length alignment to chr1 on graph G_1' (which includes edges for observed SNPs) using $(r, n) = (20{,}100)$ and consider these alignments as ground truth (we note these alignments deviated from the exact path by an average of 0.8% suggesting there is value in using the graph for context). We aligned uncorrected CCS and individual subreads to G_1' and report above the number of alignments that match the path of the corrected CCS reads, the average read length, and the number of SNPs included in the alignment and called the same way as in the corrected CCS. The rest of the subreads align to a path that is different from the corrected CCS in at least one exon. We see that SNP calls are reasonably accurate even in subreads. They can be further improved by considering multiple subreads from the same molecule, or by using additional information from the PacBio quality scores.

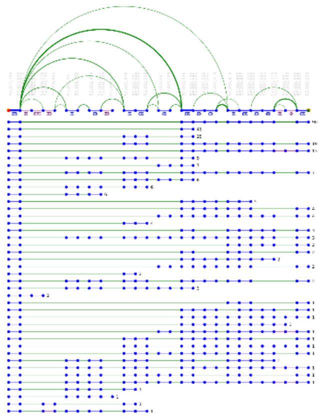


We measured the run time for aligning the long circular consensus sequences provided by Au, *et al*. The figure to the left shows the time in seconds per base aligned, which remains fairly constant for a wide range of read lengths, and varies by less than a factor of 10 over the experiments. The expected time is near the bottom of the range shown.

Applications. We tested our method on two datasets generated at UNC:

(1) Viral transcriptome. The short read alignments to the Human Herpes Virus (HHV5) genome during human cell infection present a high number of cryptic splices (likely due to repeated regions). However, when we aligned the long reads to the splice graph, many of the splices presented by the short reads were not used by any full-length transcript. The figure to the right shows the viral genome in a circular plot. The bars around the circle compare short read coverage (in blue) with long read coverage (in red). The lines in the middle of the circle represent splices confirmed by PacBio long reads (in red) and the ones present only in the short read alignments (in blue). We used 5’ and 3’ linkers in the protocol to identify full length transcripts which were detected as part of the alignment process.





(2) Novel transcripts in a human cancer cell line. We used our method to analyze the transcriptomes of 10 genes of interest in a human cell line, in order to find novel transcripts. The results of this investigation are under review for publication. The figure to the left shows the full length transcripts found in one of the genes of interest.

Conclusions. The fundamental difference of EDGAR compared with long read correction is the use of context of the short reads represented in the underlying directed graph, identifying a limited set of splices or variants available to the transcript being aligned. Additionally, when used with a splice graph generated from short read alignments to the genome, our method yields an alignment of the long read to the genome, as opposed to just a correction. In this case, long read correction and transcript identification are both achieved simultaneously.

[1] Chaisson, Mark J., and Glenn Tesler. "Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory." BMC bioinformatics 13.1 (2012): 238.

[2] V. Lacroix, M. Sammeth, R. Guigo, A. Bergeron, “Exact transcriptome reconstruction from short sequence reads”, WABI 2008 LNCS 5251:50-63, 2008.

[3] Au KF, Underwood JG, Lee L, Wong WH (2012) Improving PacBio Long Read Accuracy by Short Read Alignment. PLoS ONE 7(10): e46679. doi:10.1371/journal.pone.0046679

[4] Au KF, Sebastiano V, Afshar PT, et al. Characterization of the human ESC transcriptome by hybrid sequencing. Proceedings of the National Academy of Sciences of the United States of America. 2013;110(50):E4821-E4830. doi:10.1073/pnas.1320101110.