Identification and correction of substitution errors in Moleculo long reads

Jared Price[1], Judson Ward[2], Joshua Udall[3], Quinn Snell[1] and Mark Clement[1]

[1]Department of Computer Science, Brigham Young University, Provo, UT 84602
[2]Department of Molecular Biology, Driscoll's, Watsonville, CA 95076
[3]Department of Plant and Wildlife Sciences, Brigham Young University, Provo, UT 84602

Abstract— **Moleculo DNA sequencing technology provides extremely accurate, phased, reads having an average length of over 4,000 bp. Very little is yet known about the precise characteristics of these reads. We estimate a lower bound for the single nucleotide substitution error rate of these reads, and provide probabilities for each type of substitution. We also present preliminary work on the development of an error correction algorithm for these reads which in its current implementation corrects 74,030 single nucleotide errors in a Moleculo data set obtained from *Rubus idaeus* 'Heritage'. We also demonstrate that the pattern of substitution errors shows no significant bias with respect to the position of an error along the body of a read.**

I. Introduction

Approximately 7 years ago the field of DNA sequencing began a major revolution in which Sanger sequencing became largely displaced by *next-generation* sequencing technologies capable of rapidly producing orders of magnitude more sequencing *reads* at dramatically reduced cost. This increase in throughput, however, was accompanied by a significant reduction in individual read length, thus greatly complicating analyses such as *de novo* assembly and haplotype phasing.

Today, the field is in the throes of another major revolution in which reads several times longer than traditional Sanger reads are becoming commonplace. As of June 2013, the *PacBio RS II* sequencing system produces reads with an average length of over 4,000 bp and a maximum length of over 20,000 bp [5]. This technology can exhibit error rates higher than 12% [4] but has been shown to be useful for de novo assembly when coupled with a disciplined error correction protocol [4].

Another technology exhibiting similar length characteristics but markedly different error characteristics has been developed by Moleculo [8] and recently acquired by Illumina. This technology works by first breaking the target DNA into large fragments of approximately 10,000 bp. Each of these large fragments is then used to produce a barcoded short-read Illumina library and the resulting reads sharing a particular barcode are assembled together to produce *virtual* long reads.

The characteristics of the reads produced by this technology are not yet well understood and no error correction algorithm has yet been designed specifically for the properties of these reads. This paper presents an initial exploration into these issues. It provides a reference-free assessment of the rate and types of substitution errors in Moleculo reads and describes the initial stages of the development of a k-mer based algorithm for their correction.

II. Related Work

A number of important k-mer based error correction algorithms currently exist for short-read high-coverage data such as that produced by the Illumina HiSeq. To our knowledge, the first description of a k-mer based error correction algorithm was given in Pevzner's landmark paper introducing the Eulerian path approach to fragment assembly [6]. This work has been built upon by a number of important algorithms including the error correction module of ALLPATHS-LG [2] and the Quake algorithm [3] which uses maximum likelihood in a framework that explicitly incorporates base qualities.

All of these methods rely on high coverage in order to make inference on which k-mers are likely to be in the target genome (*trusted* k-mers) and those which are likely to have errors. Although Moleculo long reads are built from barcoded *HiSeq* reads, the underlying HiSeq reads are not provided to the client. For this reason, error correction algorithms which rely on high coverage are not directly applicable to Moleculo reads. Instead, any k-mer based approach must use a separate, high-coverage, data set to build a model of trusted and untrusted k-mers. The k-mer model built from the high-coverage data set is then mapped onto the Moleculo reads in order to correct them. The practice of using high-quality short reads to error correct long reads is quickly becoming standard practice in the case of PacBio reads [1], [4].

Because the Moleculo technology is so new there is no body of research applying any of the standard error correction algorithms to Moleculo reads in the 2-data-set manner just described. The Quake algorithm separates its k-mer counting step from its read correction step. This may enable direct correction of Moleculo reads with the Quake algorithm by using a high-coverage data set in the k-mer counting step and the Moleculo data set in the correction step. Of course, this is an atypical usage of the algorithm and the performance of the algorithm under these conditions has yet to be tested.

III. Materials and Methods

A. Moleculo data set

The Moleculo reads used in this study come from a *de novo* whole genome project targeting the highly heterozygous raspberry cultivar *Rubus idaeus* 'Heritage'. We obtained

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

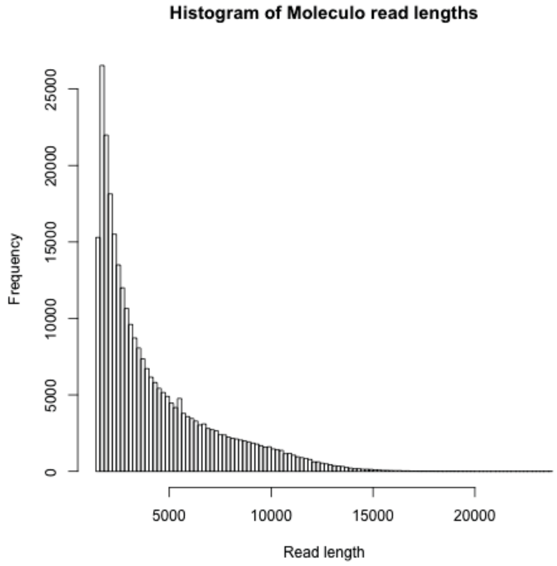

Fig. 1: Histogram of Moleculo read lengths (min = 1,500 bp; max = 23,760 bp; mean = 4,252 bp; median = 3,193 bp)

285,801 long Moleculo reads, comprising 1,215,243,110 bases. The read length distribution of these reads is summarized in Figure 1.

B. Definitions

Let R denote an arbitrary Moleculo read. For a read having length L, an individual nucleotide can be referred to using the symbol R_i where $i \in \{0, 1, 2, \ldots, L-1\}$. R_0 refers to the $5'$ terminus and R_{L-1} refers to the $3'$ terminus of the sequence.

In general usage, the term k-mer refers to a sequence of nucleotides, having length k, irrespective of the particular location at which the k-mer is found. For example, one might say that the 13-mer ACTGACTGATGCA is found 1000 times in a set of sequencing reads. We distinguish this usage of k-mer from the term *localized* k-mer. A *localized* k-mer is a sequence of nucleotides having length k that is also associated with a particular position in a sequencing read or genome. In this usage the *localized* 13-mer ACTGACTGATGCA beginning at a particular location is distinct from all other *localized* k-mers, including those that share the same sequence. When there can be no confusion a localized k-mer will sometimes be referred to simply as a k-mer.

Every position R_i has an associated *localized k-mer spectrum* which is the set of all localized k-mers which contain R_i as a base. The position R_i of such a spectrum is sometimes referred to as the anchor of the spectrum. When there can be no confusion a *localized k-mer spectrum* will be referred to simply as a k-mer spectrum. The maximum cardinality of any localized k-mer spectrum is equal to k. Spectra having cardinality k are referred to as *unconstrained* while spectra having cardinality less than k are said to be *constrained* (see Figure 2). When a localized k-mer spectrum is anchored on a single nucleotide substitution error, the spectrum is referred to as a *substitution spectrum*.

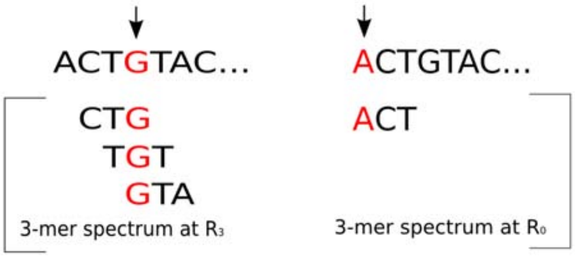

Fig. 2: The 3-mer spectrum at R_3 is *unconstrained* because it contains 3 3-mers. The 3-mer spectrum at R_0 is *constrained* because it only contains a single 3-mer.

For a particular set of trusted k-mers S, there is an associated *relapse rate* r. The *relapse rate* is the probability (weighted by k-mer count) that a single-nucleotide substitution in a set of trusted k-mers will produce another k-mer in the set. The procedure for calculating this probability has been incorporated into the *Bioluminescence* software suite [7]. The algorithm consists of a user-defined number of Bernoulli trials. Each Bernoulli trial consists of making a single nucleotide substitution in a particular trusted k-mer (each substitution is guaranteed to introduce a new base, in other words the algorithm never substitutes a nucleotide with itself). A trial is successful when the newly created nucleotide sequence is also in the trusted k-mer set.

The trials are performed in *trial sets*. A *trial set* is performed by first selecting a random k-mer from the trusted set and performing n Bernoulli trials using that sequence where n denotes the k-mer count. The use of *trial sets* ensures that a k-mer found at count 1000 carries twice as much weight as one found at count 500 and so on, however, this procedure can bias the results when the number of Bernoulli trials is small. The results reported in this paper perform 2,000,000 Bernoulli trials. All random operations are performed using cryptographically strong random number generators.

C. Estimating a lower bound on the error rate

In addition to the Moleculo sequence data, a set of Illumina short reads comprising approximately 45x coverage of the *Rubus idaeus* genome was generated. This data set was subjected to k-mer counting for k=31. K-mers observed at frequency 2 or greater are considered *trusted* while all other k-mers are considered *untrusted*. Each Moleculo read was examined to determine the number of its localized 31-mers that are trusted. Of the 285,801 Moleculo reads tested, 1,422 had 1,000 or more untrusted k-mers. These reads were excluded from the error rate analysis because they are likely contaminants.

A lower bound on the probability that a Moleculo base is an error was approximated using:

$$\frac{U/(31*(1-r))}{N} \tag{1}$$

where U is the total number of observed untrusted 31-mers, r is the relapse rate, and N is the total number of bases in the Moleculo reads.

This equation would yield a direct estimate of the error rate if no substitution spectra overlapped. Because of overlapping substitution spectra this number provides only a lower bound on the error rate, that is to say, the error rate must be somewhat worse than this number.

D. Error correction

Our development of an error correction algorithm specifically designed to correct long Moleculo reads using a k-mer model built from a separate high-coverage data set is in development. The completed implementation reported in this work is used to make inference about the properties of Moleculo reads. This implementation finds and reports single-nucleotide substitution errors which cause unconstrained *non-relapsing* substitution spectra. A substitution error causes a *non-relapsing* substitution spectrum when none of the k-mers in that position's localized k-mer spectrum are trusted. When such a spectrum is surrounded on each side by a trusted k-mer the position of the single nucleotide error is known. The current implementation corrects such positions when there is exactly one trusted k-mer which can be produced by mutating the erroneous base.

This implementation is too restrictive to be used as a general-purpose error correction utility but is useful for examining the types of substitution errors that are present in Moleculo reads, and this is the purpose for which it is used in this work. The algorithm is available as part of the *Bioluminescence* software package [7].

IV. Results

A. A lower bound on the error rate

The relapse rate for 31-mers in the *Rubus idaeus* data set was calculated using the method described in III-C to be 26.79%. Most readers will likely be surprised how high this number actually is. To understand why the rate is so high it is important to keep in mind that the relapse rate is calculated using weighted k-mer counts. A repetitive k-mer occurring in the set 3,000 times will have 100 times more effect on the rate than a k-mer occurring only 30 times. Furthermore, because we set the trusted threshold low (a count of 2 or greater) erroneous variants of true k-mers may be creeping into the trusted set causing the relapse rate to be higher. Finally, the unusually high heterozygosity of the target organism likely inflates this value above what would be seen in less heterozygous organisms.

A lower bound on the probability that a particular base is an error in a Moleculo read was estimated using the protocol described in Section III-C. To conceptualize why this number only provides a lower bound consider two single nucleotide errors which occur directly next to each other. Each one produces a substitution spectrum having 31 potentially untrusted k-mers. If the spectra did not overlap the 2 errors would introduce 62 potentially untrusted k-mers. Because they overlap, however, only 32 potentially untrusted k-mers are produced by the 2 errors (see Figure 3). Adjusting for overlapping substitution spectra can only cause the estimate of the error rate to increase so the number calculated using the technique described in Section III-C provides a lower bound.

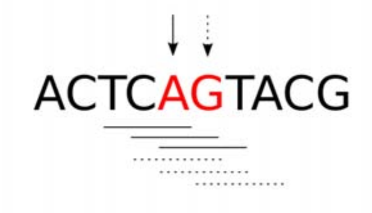

Fig. 3: Base A (pointed to with the solid arrow) and base G (pointed to with the dashed arrow) are single nucleotide substitution errors. Each produces a substitution 3-mer spectrum containing 3 potentially untrusted 3-mers (shown in solid and dashed lines respectively). If the errors had been far apart the spectra would not overlap and 6 potentially untrusted 3-mers would be found. When directly next to each other (as shown in the image) they only produce 4 potentially untrusted 3-mers.

The calculated lower bound on the probability that a Moleculo base is an error is 0.00014 (1.4 errors per 10,000 bases). This suggests that although the Moleculo reads are likely to be extremely high quality the Phred scores reported with the reads should be understood in relative, rather than absolute, terms. The average Phred score reported in our Moleculo data set was 49.935. This translates to an error rate of 1.015 errors per 100,000 bases. Our analysis suggests that the actual error rate is at least an order of magnitude higher.

With that said, even if the correction for overlapping spectra caused a full order of magnitude increase in the error rate (extremely unlikely) the estimated error rate would be 1.4 errors in 1,000 bases making this technology the most accurate sequencing technology available. Our analysis suggests that the accuracy of Moleculo reads is similar to that of high-quality consensus sequence.

B. Error identification and correction

A run of the restrictive error correction algorithm described in Section III-D found and corrected 74,030 single nucleotide errors in the *Rubus idaeus* data set. The relative frequency at which each possible type of correction occurred was also recorded. These results demonstrate that nearly 70 % of all the miscalls in the *Rubus idaeus* data set consist of incorrectly calling a C when T was the correct base, or incorrectly calling a G when A was the correct base. The complete results for the relative frequency at which each type of correction was made is given in Figure 4.

Standard Illumina reads have a well known error-profile bias in which nucleotides near the $5'$ end of the read are much more likely to be correct than nucleotides near the $3'$ end of the read. Because of the way in which Moleculo reads are produced we expect that the position of errors along the body of a read should be much more randomly distributed. To test

Correction	Count	Percentage
A → C	2150	2.9 %
A → G	6046	8.2 %
A → T	1269	1.7 %
C → A	1411	1.9 %
C → G	659	0.9 %
C → T	25609	34.6 %
G → A	25368	34.3 %
G → C	692	0.9 %
G → T	1431	1.9 %
T → A	1327	1.8 %
T → C	5913	8.0 %
T → G	2155	2.9 %

Fig. 4: Table showing the relative frequency of various types of substitution error corrections (original base → corrected base; Count = the number of occurrences of the specified correction; Percentage = the percentage of all corrections accounted for by the specified correction)

this hypothesis each position at which a single nucleotide error correction occurred was translated to a *scaled position* ranging from 0 at the $5'$ terminus to 1 at the $3'$ terminus.

Let y be the random variable whose values are the *scaled positions* at which single nucleotide errors were corrected. The probability density function for this random variable is reported in Figure 5 and shows no significant position bias. The randomness of the error distribution makes a number of downstream analyses easier.

V. Discussion

This study provides confirming evidence that the new Moleculo technology provides long reads at exceptionally high accuracy. These reads exhibit error rates much more similar to high-quality consensus sequence than to the raw reads of any widely-available technology. Of course, Moleculo reads are produced through a consensus process so this is not an unexpected result. However, to our knowledge, no independent analysis had yet confirmed the high accuracy of these reads.

The combination of long read length and consensus-quality accuracy will likely lead to the adoption of the Moleculo technology as the dominant sequencing technology for a number of challenging applications such as *de novo* assembly and haplotype phasing.

In particular, these reads promise to revolutionize the study of polymorphic genomes. Careful separation of haplotypes is among the most challenging of bioinformatics tasks. Often just a few bases in a thousand separate one haplotype from the other. The haplotypes can be kept separate across long distances only when the reads are highly likely to cover more than one polymorphic position and each polymorphic position can be reliably distinguished from any error. Moleculo reads are the first reads we are aware of that have the correct combination of length and accuracy to allow for *de novo* construction of long haplotypes across a wide array of organisms with varying levels of heterozygosity.

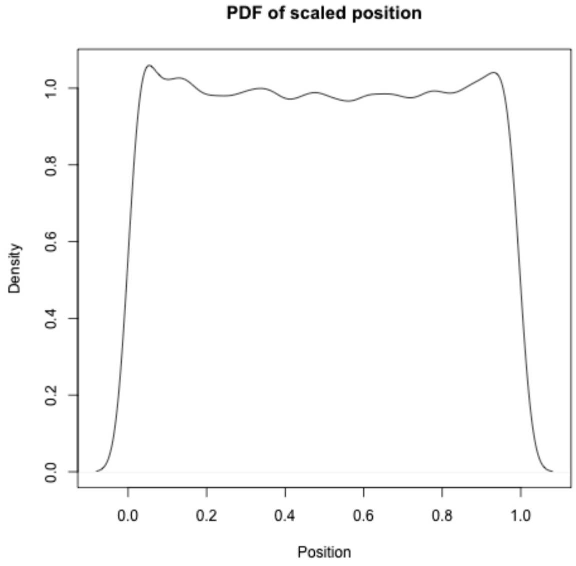

Fig. 5: The probability density function for the random variable y where each value of y is the *scaled position* of a corrected base (0 = $5'$ terminus; 1 = $3'$ terminus)

References

[1] K. F. Au, J. G. Underwood, L. Lee, and W. H. Wong, "Improving PacBio long read accuracy by short read alignment," *PLoS ONE*, vol. 7, no. 10, p. e46679, 10 2012. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0046679

[2] S. Gnerre, I. MacCallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea, S. Sykes, A. M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E. S. Lander, and D. B. Jaffe, "High-quality draft assemblies of mammalian genomes from massively parallel sequence data," *Proceedings of the National Academy of Sciences*, vol. 108, no. 4, pp. 1513–1518, 2011. [Online]. Available: http://www.pnas.org/content/108/4/1513.abstract

[3] D. R. Kelley, M. C. Schatz, S. L. Salzberg *et al.*, "Quake: quality-aware detection and correction of sequencing errors," *Genome Biol*, vol. 11, no. 11, p. R116, 2010.

[4] S. Koren, M. C. Schatz, B. P. Walenz, J. Martin, J. T. Howard, G. Ganapathy, Z. Wang, D. A. Rasko, W. R. McCombie, E. D. Jarvis *et al.*, "Hybrid error correction and de novo assembly of single-molecule sequencing reads," *Nature biotechnology*, vol. 30, no. 7, pp. 693–700, 2012.

[5] Pacific Biosciences. (2013) PacBio RS II Sequencing System. [Online]. Available: http://files.pacb.com/pdf/PacBio_RS_II_Brochure.pdf

[6] P. A. Pevzner, H. Tang, and M. S. Waterman, "An eulerian path approach to dna fragment assembly," *Proceedings of the National Academy of Sciences*, vol. 98, no. 17, pp. 9748–9753, 2001. [Online]. Available: http://www.pnas.org/content/98/17/9748.abstract

[7] J. Price. (2013, May) The Bioluminescence software suite. [Online]. Available: http://dna.cs.byu.edu/~jcp44/bioluminescence/

[8] G. Waldbieser, M. Kertesz, D. Pushkarev, T. Blauwkamp, and J. Liu, "Production of long (1.5kb 15.0kb), accurate, DNA sequencing reads using an Illumina HiSeq2000 to support de novo assembly of the Blue Catfish genome," in *The international conference on the status of plant and animal genome research*, 2013.