Lariat: Linked Read Aligner and reference bias https://github.com/10XGenomics/lariat

William H Heaton

July 9, 2016

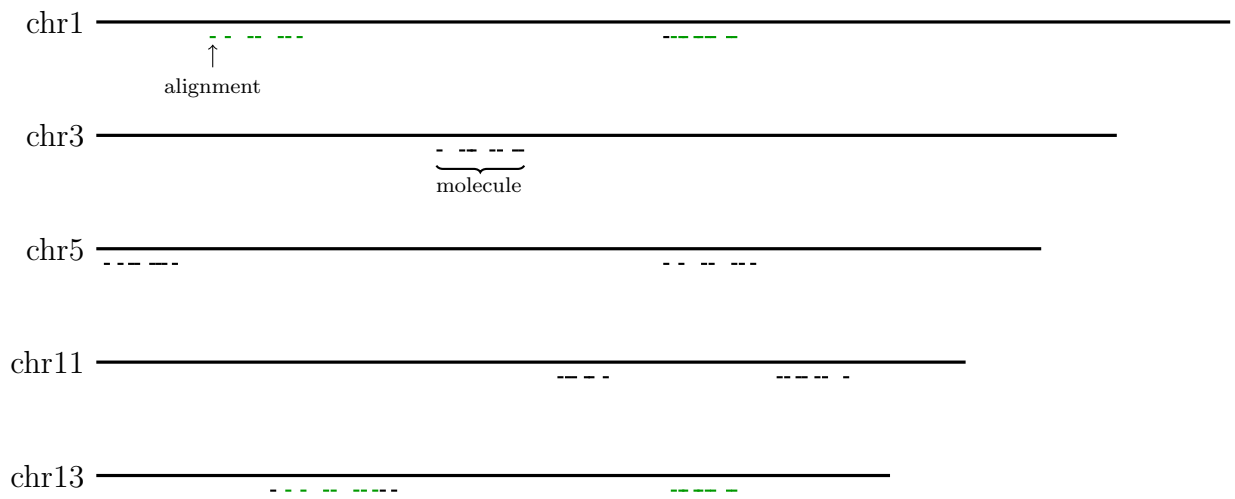
1 Introduction

Typical short read nextgen sequencing consists of aligning roughly 100 base reads to a genome of length 3.2 billion characters. Methods for search, mapping, and alignment to candidate positions are well known to the literature¡cite stuff¿. 10X technologies adds DNA barcodes to these short reads in the following manner. Long DNA molecules of roughly 100kb are put into a microfluidic system in which micro gel beads, each with many copies of a particular barcode DNA sequence, are mixed with long molecule DNA template solution and then separated from one another via oil to create an emulsion. Thus we have a gel bead with many copies of a particular barcode DNA sequence surrounded by solution containing several long DNA molecules surrounded by oil. There are perhaps a million different such droplets in reaction. From this emulsion, short reads are enzymatically amplified from the long molecules and the barcode DNA is attached to them. In this way, every read from the same source long molecule has the same barcode. And every molecule is associated with a different barcode with high probability. Because of this, within a set of reads with the same barcode, they are likely to have originated from a few long DNA molecules[8]. And thus when we map them to the reference genome, they should be high clustered into a few isolated locations. We will use this information to decide between multiple ambiguous mapping sites for reads in repeat regions.

2 Linked read aligning

Take all of the reads for a single barcode and find all alignments for each read. This is just for illustration. Scale is obviously way off. Typical molecules are 100kb. And for a typical whole genome 30x coverage run, each barcode would have on the order of a thousand reads coming from on the order of 10 long molecules. Notice that for each molecule we do not have coverage > 1. Some technologies such as molecule attempt to get multiple coverage of every physical molecule and reassemble that molecule into a "synthetic long read"[6]. Whereas others run in a mode known as "linked reads" where coverage of the physical molecule is < 1[7][1][8]. It is advantageous to run in "linked read" mode because you want to sample both haplotypes with high confidence. You need at least 10-20 coverage average of sampled haplotypes to very confidently sample both haplotypes at almost all loci. So if you are running in "synthetic long read" mode, you will at least $10 * 10$ total coverage (10 coverage of physical molecules to sample the haplotypes and 10 read coverage of each physical molecule to reconstruct the whole molecule using de novo assembly)[3].

Figure 1: Find all alignments for every read in a single barcode.

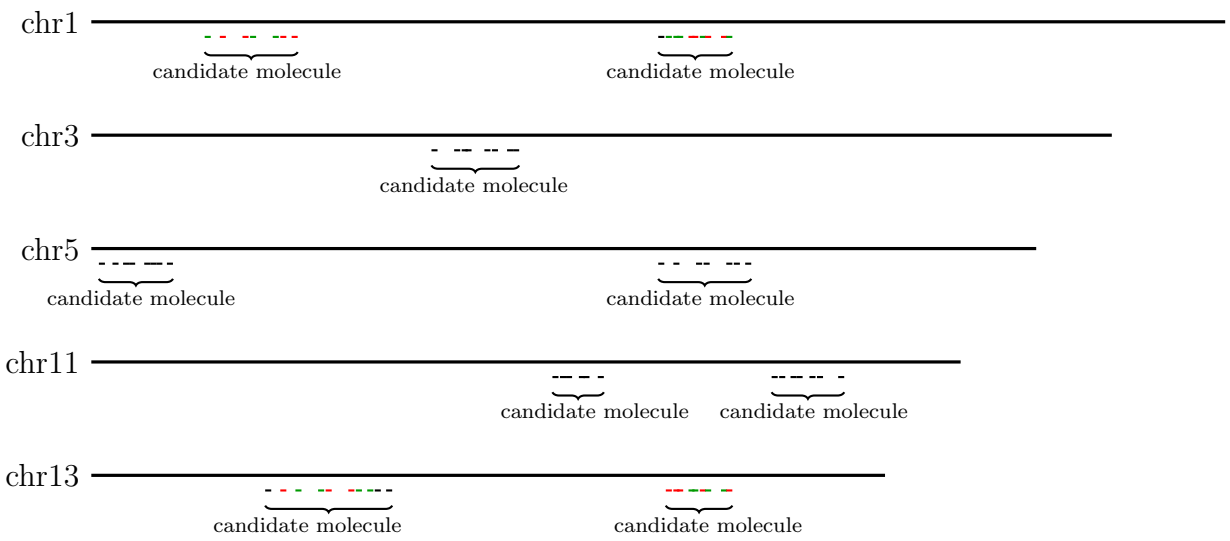


Green alignments are alignments of reads with multiple good alignment options. Black alignments are unique.

3 Algorithm

The basic algorithm follows that of the read fragment aligner (RFA) developed by Alex Bishara et al. in Serafim Batzoglou's lab at Stanford[citation]. Initially the best alignment for each read is marked as the current "active" alignment for that read choosing randomly among equally good alignments. Candidate molecules are then determined via clusters of alignments. [2]

Figure 2: Mark best alignment for each read. Calculate candidate molecules.

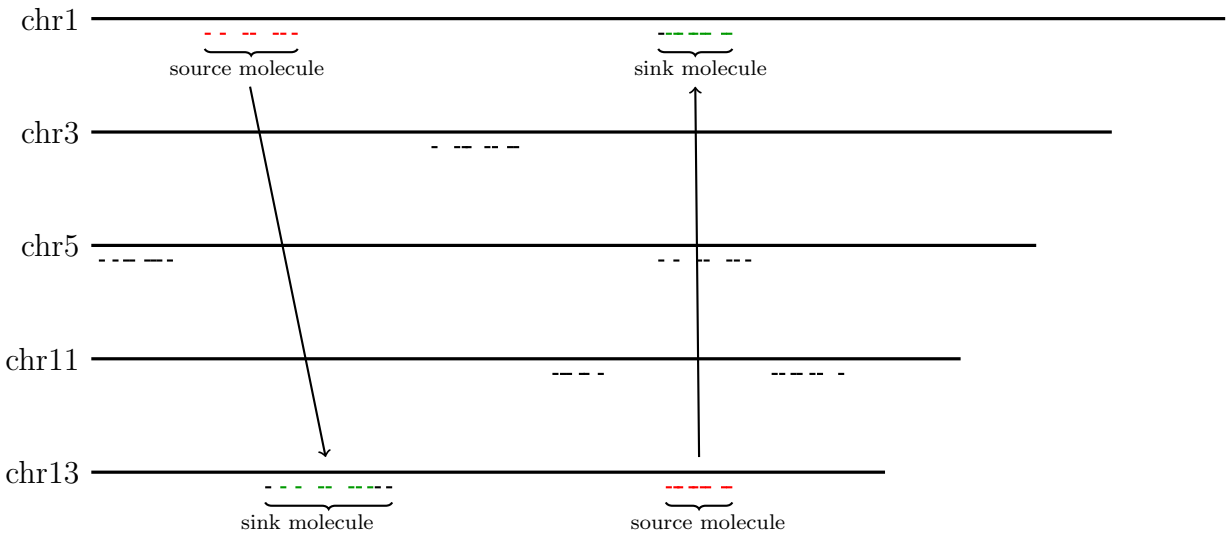


Green alignments are "active" alignments of reads with multiple good alignment options. Red alignments are "inactive" alignments of reads with multiple alignment options. Black alignments are unique.

4 Inference: group moves

At this point, the optimal active alignment configuration is inferred. The algorithm tries to change which alignment for a given read is the active alignment for that read. It does this not read by read, but in specific groups. It takes two candidate molecules, one as the source molecule and one as the sink molecule. Then, all active alignments in the source molecule with alternative alignments in the sink molecule are changed from active to inactive and their corresponding alignments in the sink molecule are changed to active. We then score this new configuration via a probabilistic model based on the alignment quality, sequencing error, genetic variation as well as parsimony of the number of physical molecules generating these reads. Using this probabilistic score we can do hill climbing or metropolis hastings search for the optimal configuration.

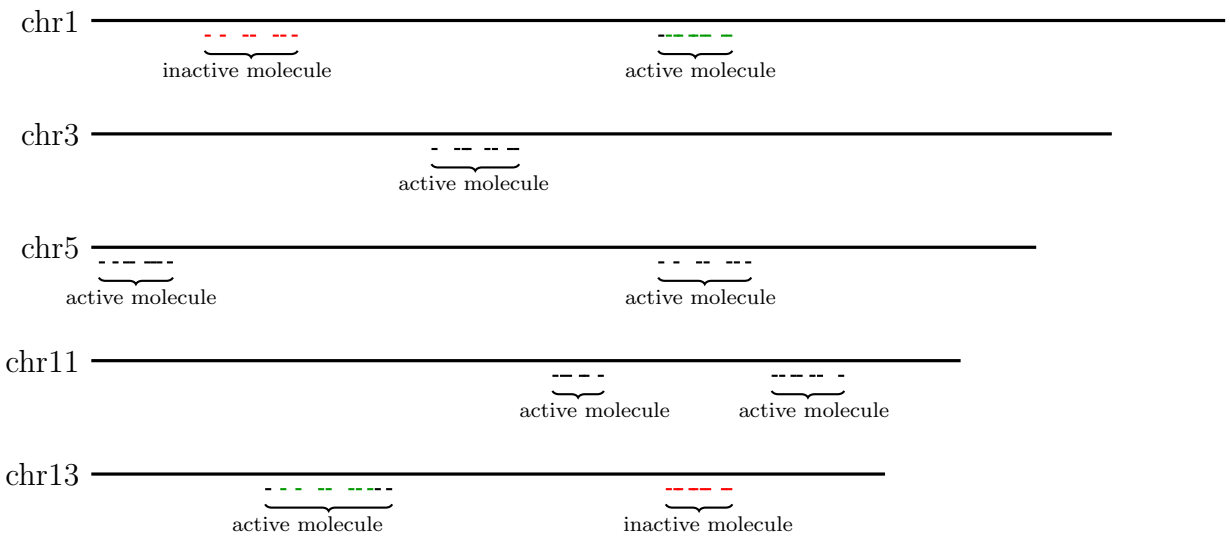
Figure 3: Try group moves between candidate molecules.



Two moves are found to produce a higher probabilistic score.

After alignment configuration inference and before mapq calculation, we take stock of which candidate molecules we believe to be real molecules, which we term "active molecules". We do this by calling candidates that have no active alignments or very few active alignment "inactive candidates".

Figure 4: Final configuration.



Green alignments are "active" alignments of reads with multiple good alignment options. Red alignments are "inactive" alignments of reads with multiple alignment options. Black alignments are unique.

5 Mapping quality calculation without molecule information

We wish to calculate the probability that our mappings are correct based on a model of how the reads were generated from the sample. It is useful to consider the assumptions we must make to do so.

- All mismatches and indels versus the reference genome are independent.
- All mismatches and indels are created from sequencing errors or biochemistry errors. This will become important later when we talk about reference bias.
- The locus from which this read was generated exists in the reference genome. This is important when there is not the case but a distance paralogue of that sequence does exist in the reference genome.

From our knowledge of sequencing and biochemistry error rates, we can come up with a probability this data was generated from the underlying sequence at this reference location.

I will denote A_i as the i'th alignment of this read[5].

$$p(A_i|data) = \frac{p(data|A_i)p(A_i)}{p(data)} \tag{1}$$

$$p(A_i|data) = \frac{p(data|A_i)p(A_i)}{\sum_k p(data|A_k)p(A_k)} \tag{2}$$

If we assume a uniform prior on different alignments, the above equation simplifies to the following.

$$p(A_i|data) = \frac{p(data|A_i)}{\sum_k p(data|A_k)} \tag{3}$$

Simply normalizing all alignment options to sum to 1. From the posterior probability of the best alignment, we can create a fred scaled quality score with $-10\log_{10}(1-p)$.

6 Mapping quality with molecule information

With the molecule information, we can better estimate $p(data|A_i)$. There is some probability that a read originated from a long molecule and some probability that it originated from a short molecule. If it originated from a short molecule, the chance that a long molecule (aka active molecule) from the same barcode spanned a repeat of that sequence is governed by the percentage of the genome covered in that barcode. This can be estimated from the active inferred molecules. For our data, the chances that a read is not generated from a long molecule is roughly 1%. And the average barcode contains roughly 0.04% of the genome. After the appropriate probability corrections, the same normalization across all alignments for a given read is used. Thus an alignment in an active molecule can be much higher confidence given it's supporting molecular information.

Furthermore, we must cap the alignment confidence on our confidence that the other candidate molecules in contest are actually inactive. This is done via the difference in scores when making a move from candidate A to candidate B and vice versa.

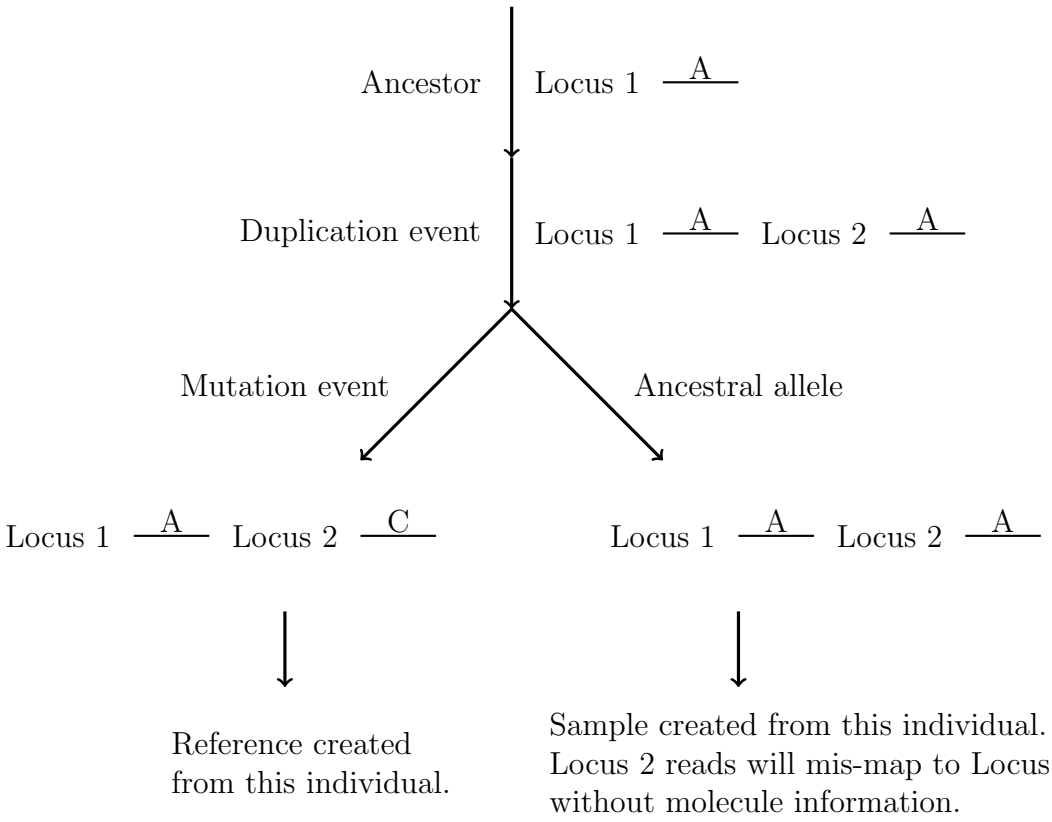
7 Implementation

Our implementation is written in the Go programming language using Cgo to build an API into the internal functions of the well known and widely used aligner, BWA-mem [citation]. Since generating all of the potential alignments is the most computationally intensive step, and BWA-mem does this internally already, Lariat is not substantially slower than running BWA-mem which is already a part of most genomics pipelines. The code is open source and available at https://github.com/10XGenomics/lariat [5].

8 Reference bias

There are several types of reference bias. For instance, if your sample contains sequence whose true locus is not represented in the reference you are using, but the reference does contain a distant paralogue, this is one type of reference bias which will cause mismappings and incorrect mapping qualities for those reads. The type of reference bias I want to address here is when a read from your sample originates from locus A but has a more similar sequence to the reference at locus B than to the reference at locus A. One way in which this can happen is gene conversion. But there is a more subtle mechanism that only occurs when you think about population genetics. Consider the following sequence of events outlined in figure 5.

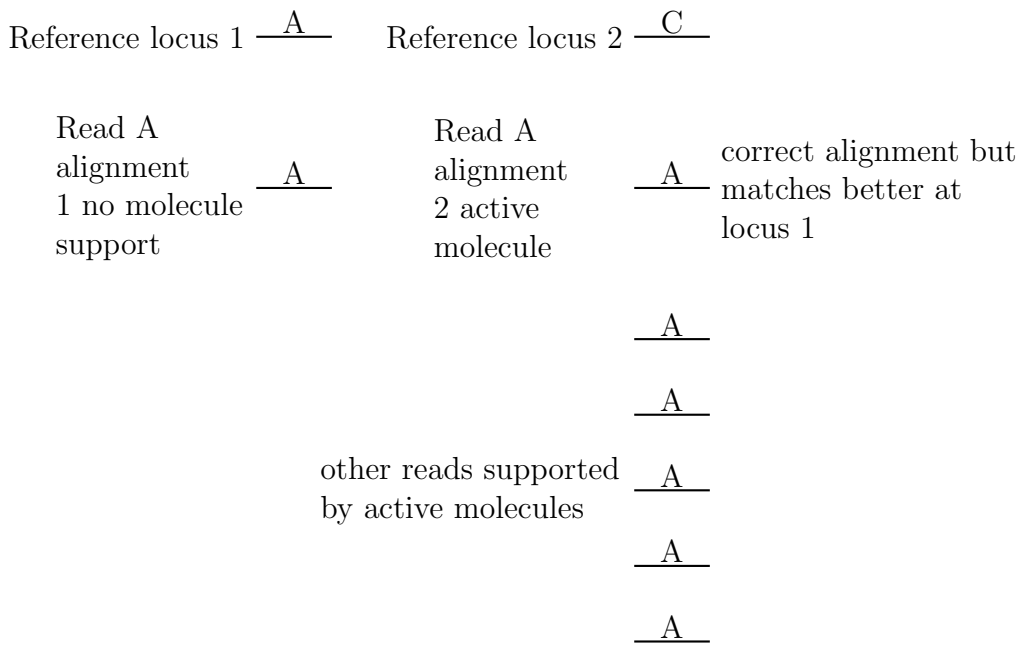
Figure 5: Population inheritance of segmental duplication mutations reference bias.



This explains why we see this phenomenon much more than expected from having random mutations on top of the reference sequence. Lariat tends to place these reads correctly, but cannot give them a very high mapping quality because without the molecule information, the alignment probability is better at the other locus.

9 Using cross barcode information to rescue reference biased mappings.

Figure 6: Using cross barcode correlative information to increase mapping confidence.



As you can see from figure 6, Read A's correct location is not the best alignment based on sequence similarity alone. The probabilistic model placed this read here due to it's supporting molecule information. But the mapping quality may be quite poor since it has a competing alignment with a higher probability before molecule information. The key thing here is the assumptions that we made in our mapping quality calculation. If the sample actually contains the A allele at locus 2, all reads will have this mismatch and be penalized for it (or may be mismapped to locus 1). The key assumption that we wish ti avoid here is the assumption that all mismatches are sequencing or biochemistry errors. To do this, we allow all reads at locus 2 to "share" the probability penalty of this "mismatch" because it is more likely that this represents a genetic variation in the sample than many many sequencing errors. When we do this, the probability before molecule information of alignment 2 goes up, and further up when considering molecule information thus attaining high confidence mapping.

10 Validation

We have several methodologies for validating mappings and novel variants uncovered by Lariat. One is via orthogonal long read technologies such as PacBio or Moleculo. We align the long reads to the reference and then align the reads to the reference with the alternate

allele in place of the reference allele at that locus and ask if the scores of these two alignments are different. If they are, we say that read supports the allele that got the higher alignment score. If they are the same, we say they do not support one over the other. For validation we require a certain number of supporting reads in addition to a certain percentage of the total reads to support the novel variant to say that it is validated. We also have a pedigree inheritance with phasing validation that requires all individuals to agree on the inheritance pattern. And finally we have finished BAC sequences from the human genome project not used in the final reference on 2 individuals for whom we also have samples from which to run 10X genomics library prep and sequencing. We then can compare these novel variants to the BAC sequences [4]. Via all of these methods we get roughly a 95% PPV on our novel variants via Lariat and cross barcode rescue methods.

References

[1] Andrew Adey, Jacob O Kitzman, Joshua N Burton, Riza Daza, Akash Kumar, Lena Christiansen, Mostafa Ronaghi, Sasan Amini, Kevin L Gunderson, Frank J Steemers, et al., *In vitro, long-range sequence information for de novo genome assembly via transposase contiguity*, Genome research **24** (2014), no. 12, 2041–2049.

[2] Alex Bishara, Yuling Liu, Ziming Weng, Dorna Kashef-Haghighi, Daniel E Newburger, Robert West, Arend Sidow, and Serafim Batzoglou, *Read clouds uncover variation in complex regions of the human genome*, Genome research **25** (2015), no. 10, 1570–1580.

[3] Eric S Lander and Michael S Waterman, *Genomic mapping by fingerprinting random clones: a mathematical analysis*, Genomics **2** (1988), no. 3, 231–239.

[4] Heng Li, *Aligning sequence reads, clone sequences and assembly contigs with bwa-mem*, arXiv preprint arXiv:1303.3997 (2013).

[5] Heng Li and Richard Durbin, *Fast and accurate short read alignment with burrows–wheeler transform*, Bioinformatics **25** (2009), no. 14, 1754–1760.

[6] Rajiv C McCoy, Ryan W Taylor, Timothy A Blauwkamp, Joanna L Kelley, Michael Kertesz, Dmitry Pushkarev, Dmitri A Petrov, and Anna-Sophie Fiston-Lavier, *Illumina truseq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements*, PloS one **9** (2014), no. 9, e106689.

[7] Brock A Peters, Bahram G Kermani, Andrew B Sparks, Oleg Alferov, Peter Hong, Andrei Alexeev, Yuan Jiang, Fredrik Dahl, Y Tom Tang, Juergen Haas, et al., *Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells*, Nature **487** (2012), no. 7406, 190–195.

[8] Grace XY Zheng, Billy T Lau, Michael Schnall-Levin, Mirna Jarosz, John M Bell, Christopher M Hindson, Sofia Kyriazopoulou-Panagiotopoulou, Donald A Masquelier, Landon Merrill, Jessica M Terry, et al., *Haplotyping germline and cancer genomes with high-throughput linked-read sequencing*, Nature biotechnology (2016).