deCODE genetics

August 12, 2015

Reply to #2015:06:5348:0:1:REVIEW

Dear Sven Rahmann,

Thank you and the two reviewers for your comments on our manuscript “PopAlu: population-scale discovery of Alu polymorphisms”. We have now answered all questions and clarified the paper accordingly, as you requested.

As for the comparison to Mobster, we have added corresponding columns to Table 3 showing the performance of Mobster on the trio data set. Mobster does not report genotypes, which is why we could not include it in our other evaluations.

Please find detailed point-by-point replies below.

Best regards,

Yu Qian, Birte Kehr, Bjarni Halldórsson

Sturlugata 8
IS - 101 Reykjavik
Iceland
Phone: +354 570 1900
Fax: +354 570 1981
info@decode.is
www.decode.com

Replies to reviewer 1 (Alexander Schönhuth)

Basic reporting

Overall, the paper is very well written, and a pleasure to read.

Thank you!

I believe that the reader would appreciate two or three sentences on these questions:

What distinguishes Alu polymorphism discovery from deletion and insertion discovery in general? What, consequently, would be the drawbacks of methods such as NovelSeq (Hajirasouliha et al.), MindTheGap (Rizk et al.) or PopIns for detecting Alu insertions?

Of course, Alu polymorphism discovery is special because of the existence of repeat databases, against which one can map the reads, but you do not really mention this in the Introduction.

We have added a comment in the introduction about the difference between novel sequence insertion and mobile element polymorphism discovery. The tools you mention for insertion discovery are indeed not capable of discovering Alu polymorphisms because of the repetitive nature of Alu elements. The mentioned tools focus on insertions of **novel** sequences, sequences that are not present anywhere in the reference genome, unlike Alu and other repetitive sequences.

A minor remark: in the first sentence of the abstract, what exactly do you want to communicate with 'combined'? Do you just mean 'together'?

You are correct, we just mean ‘together’. We have replaced ‘combined’ by ‘together’ to avoid further confusion.

Experimental design

In general, the experimental design is very neat. I appreciate in particular to transform Alu insertion discovery into an Alu deletion problem, by inserting potential insertions and to re-align, which looks novel to me (is it?)

Thank you for this comment. We are also not aware of any other tool using such Alu insertion genotyping. A reason for this may be that only few tools exist to date that solve both the Alu deletion (genotyping) problem and the Alu insertion problem.

A few minor things:

- What is the motivation/intuition behind the formula in line 117, page 4? [P(H_0|G_1)]

We have added the following sentence:
“The intuition behind this estimate is that the probability for H_0 is relative to the length ratio of H_0 versus the sum of H_0 and H_1 where the length of H_1 is estimated as 2*||r|| base pairs and H_0 as 2*||r|| + l_{Alu} base pairs.”

- page 5, Figure 2: it would be helpful if you could also illustrate la and ra reads, as described from line 136 on page 5.

We have added references in the text to example la and ra reads in Figure 2.

- page 5, lines 143/144: What's the reasoning behind choosing these intervals (why exactly 3 sigma)?

Three sigma (the 99.7 percentile) is commonly treated as near certainty in statistics (see for example Wikipedia -> ‘three-sigma rule’).

- page 5: the procedures described on lines 150 till 163 (region forming) are irreproducible (hence do not adhere to PeerJ standards). What about a description of these heuristics in a supplementary file?

We have added more details to the two paragraphs, now describing details of the procedure. We hesitated to describe all details initially as we think that this part of our method leaves room for improvement.

- page 5, lines 167/168: Figure 3 isn't really a helpful illustration of TSD's (and the corresponding issues)

Could this be a misunderstanding? Only subfigure A of Figure 3 is meant to illustrate a TSD. The remaining subfigures B-D illustrate other possible complex situations at Alu insertion sites. We clarified the figure caption but, nevertheless, decided to move the figure to the end of the main text as a supplementary figure.

- page 6, line 177: why should AL be equal to AR at all?

In cases where there occurs no target site duplication, no deletion and no novel sequence insertion along with an Alu **insertion**, the insertion position for the left end of the Alu equals the insertion position for the right end, i. e. AL equals AR. In the Alu **deletion** problem AL is never equal to AR.

Validity of the findings

Overall, the results look sound and convincing. In general, I think that PopAlu makes a convincing tool!

Thanks!

One (somewhat bigger) question though: a comparison with Mobster is missing. We (as members of the Genome of the Netherlands consortium) believe that this is a state-of-the-art tool, possibly the best one currently available. If there no real reasons why a comparison does not make sense, please compare also with Mobster -- I argue that this is a necessity to really understand the value of your tool in terms of practical benefits.

We agree that a comparison with the recently published program Mobster has practical benefits and added it to Table 3. The numbers were taken from an output file of Mobster that was made available along with the Mobster paper and computed on the exact same data set as we have used in our evaluation. The results are comparable to both PopAlu and RetroSeq.

In order to show the most fair comparison of predicted breakpoints to PCR validated breakpoints, we re-calculated the numbers in the PopAlu distance column. For each Alu polymorphism PopAlu reports an interval, RetroSeq reports a single position, and Mobster reports both an interval and an “Insert Point”. In the initial submission, we calculated the distance for PopAlu based on the interval, whereas we now use the mid-point of the interval. We think that this is the most fair comparison to both Mobster and RetroSeq for the following reasons: On the one hand, it would be unfair to use the interval for PopAlu when using the insert point reported by Mobster. On the other hand, the intervals reported by Mobster are significantly larger than the intervals reported by PopAlu (and than potential target site duplications), so that the distance from these intervals is not informative. The table below provides all average numbers of basepairs that we considered. Columns used for Table 3 in the manuscript are highlighted in bold.

	PopAlu			Mobster			
	Distance interval	**Distance mid-point**	Interval length	**Distance insert point**	Distance interval	Distance mid-point	Interval length
NA12878	2.1	**4.7**	7.6	**5.0**	1.2	7.0	17.2
NA12891	2.1	**4.6**	7.3	**6.4**	1.0	10.3	26.6
NA12892	2.2	**4.9**	7.7	**6.7**	0.9	7.9	20.6

We did not include Mobster in the remaining comparisons, the main reason being that it does not report genotypes for the discovered Alu polymorphisms. Nonetheless, we attempted to run Mobster on our simulated data set, which was generated using the read simulator Mason and aligned with BWA-mem. It turned out that Mobster does not support bam files generated with BWA-mem but only older versions of BWA (before Feb 2013 and earlier). We used BWA-mem v0.7.5a from May 2013. From our experience, the alignments of BWA-mem compared to those of older versions of BWA do differ, such that a fair comparison would require us to run PopAlu and RetroSeq likewise on the output of the old BWA version. However, we think that a comparison on an alignment generated with a more recent aligner is of more value for the general reader.

Replies to reviewer 2 (Tobias Marschall)

Experimental design

In general, the methods and experiments are described very clearly. However, there are three aspects where some additional details would aid understandability and reproducibility (without having to reverse engineer them from the provided software implementation, which is probably possible):

1) How does the heuristic partitioning mentioned on Page 5 (lines 155 and 163) work?

We added more details describing this (see answer to reviewer 1).

2) How is the consensus Alu sequence (mentioned on Page 6, lines 193 and 200) chosen? In my understanding, you use exactly ONE Alu sequence, right? Wouldn't it be helpful to use a collection, i.e. a complete "mobiome", as done by Mobster?

We agree that it would be a nice extension to use a collection of all known mobile elements as we already mention in the discussion. However, the current version of the code supports only Alu sequences. We have now clarified that we use a set of 51 different Alu sequences.

3) Page 6, Line 180: what exactly does "unambiguously determined" mean in this context?

We added “as described in the following” to the sentence as the following two paragraphs describe the two-level voting system which we use to choose breakpoints or discard a potential insertion site if there is too much ambiguity.

Validity of the findings

- Making the simulated benchmark data available to the community would be helpful.

We have uploaded the scripts we used for generating the simulated data to the PopAlu repository on GitHub and added a detailed description with all parameter settings for reproducing our benchmark data.

- Page 6: The description of SimIns is a bit ambiguous to me. You first say that the Alu elements where inserted back into the reduced reference (Line 208), presumably at the places they where deleted from, and then say that the Alu locations were chosen randomly (Line 212).

There are many more Alu elements on chromosome 21 than we used, so we randomly chose 100 of them. We deleted these randomly chosen Alu elements and then used different subsets of them to create our haplotypes by inserting them back at the places where they originated from. We reviewed the respective paragraph and hope that this is now clear.

Actually, wouldn't it be simpler to simulate just one data set and run it twice, once in "deletion/genotyping mode" using the full reference and once in "insertion/discovery mode" using a reduced reference? That would have the added benefit of showing whether performance correlates between the two experiments (e.g. some MEIs don't work in both settings because read mappability is poor in that region).

We agree that our set does not allow to compare the performance of PopIns in discovering deletions to its performance in discovering insertion. This would be possible when using the same set of Alu elements for both the deletion and insertion experiments. However, when using the exact same data sets, you get opposite frequencies for the deletion and insertion problem. For an unbiased comparison (low frequency events are generally more difficult to discover than high frequency events), you need to reverse the frequencies between the

two experiments, thus, simulate different sets of reads.

Comments for the author

This is a nice paper that complements the yet rather sparse literature on mobile element discovery/genotyping.

Thank you!

I've collected some minor comments/suggestions below:

- Discovery of deletions of known ALUs is the same thing as genotyping, right? I would point that out a bit more explicitly in the intro.

You are right that the main difficulty in the deletion problem is the genotyping since the Alu locations are known from the reference but all individuals have to be examined to discover a very rare Alu deletion. However, discovery of deletions distinguishes only between two states (fixed = no deletion vs polymorphic = deletion) while genotyping distinguishes three states (no deletion, heterozygous deletion, homozygous deletion). We have clarified this in the introduction and point it out explicitly at the beginning of the Alu deletion section.

- In my opinion, referring to your H_0 and H_1 as "alleles" rather than "haplotypes" would fit better, because you are only talking about a single locus.

We agree that this is more precise usage of terms and now replaced the term “haplotype” by “allele” in the places where we refer to H_0 and H_1. In places where we previously referred to “reads from haplotype H_x”, we now refer to “reads from a haplotype that carries allele H_x”.

- Section read classification: Could be clarified a bit more. From the description of category "I" for instances, it is not immediately clear to me why r_8 belongs to I while r_6 does not. Is "is mapped to" (Line 98/99) equivalent to "the alignment overlaps with"?

Thank you for pointing out this issue. We have corrected the figure such that r_6 does not overlap AR anymore.

- Page 4, Line 117: Motivating the choice of P(H_0|G_1) would be good.

We have added the motivation behind this choice (see answer to reviewer 1).

- Page 6, Line 192: The grammar of this sentence seems odd.

We have re-phrased this sentence.

- Page 6, paragraph on voting system: I think the manuscript could benefit from a figure explaining this voting system.

We added a figure showing an example instance of the voting system.