Polymorphic Alu-Sine insert-SIze GEnotyper (PASSIGE): An efficient tool to simultaneously discover and genotype non-reference polymorphic Alu-SINE deletions in short read sequencing data.

INTRODUCTION

Genome-wide association studies have enjoyed monumental success over the past few decades in helping to detect disease associated loci. The NGRI catalog has recorded well over 4,000 significant associations to date (1). However, a now intensely studied source genetic variation that has been previously under-appreciated on a large scale is the contribution of repetitive, transposable elements, which collectively, comprise about half of the human genome (2).

Alu-SINES are a class of such repetitive elements that are primate specific, about 300 base pairs in size and collectively represent one of the most evolutionarily successful transposons, persisting in more than a million copies in the human genome and accounting for about 11% of human genome sequence (3). Alu elements have shown to have a significant contribution to human evolution and genetic diversity (4-6), influence gene expression (3), and have been considered for their utility in forensic analysis (7, 8).

Next-generation sequencing methods have accelerated the discovery and characterization of both rare single nucleotide variants (9), but a comprehensive catalog of mobile element polymorphisms, including Alus, has only recently been published (10). The identification of novel *insertions* of transposable elements has been achieved by several groups, using a variety of approaches, and the available methods for transposable element insertion detection already developed demonstrate robustness and reproducibility among their varied implementations (11-13). However, one aspect that is noticeably missing from recently developed methods is a means of also identifying transposable elements that are *absent* with respect to the reference genome.

Here, we present Polymorphic Alu-Sine insert-SIze GEnotyper (PASSIGE), an efficient program to accurately and rapidly detect and genotype non-reference polymorphic Alu deletions with respect to the reference genome.

METHODS

Description of algorithm for polymorphic Alu detection

We consider the case where we have the aligned sequence of a single human genome, generated from Illumina paired-end, short-read sequencing data, using a commonplace mapping algorithm, the

Burrows-Wheeler Aligner (BWA) (14). Paired-end sequencing consists of read pairs of a fixed length generated from opposing ends of a fragmented piece of DNA. The length of the fragment (also commonly referred to as the insert size) is readily estimated from aligned sequence data. The insert size of an Illumina library preparation is an experimentally defined parameter determined by factors such as the shearing technology used and library preparation method. Current recommended protocols for whole-genome sequencing specify an insert size distribution with a mean of 200-300 base pairs and tight standard deviation of less than 10%.

We explicitly make use of this defined distribution of insert size to evaluate regions where there is a polymorphic Alu-sine sequence that is present in the reference annotation but absent in our sequenced sample, known as deletion in structural variation terminology.

For each Alu site in the annotated reference genome, we define *A* as the region spanned by the Alu annotation. For each region *A*, we define a window *L* that falls to the left-hand side (with respect to the reference) that is *S* base pairs wide, where *S* is the size of the annotated Alu. We shift the start and end coordinates of this window by the mean insert size of the library, as determined computationally by considering the distribution of insert sizes across all read pairs in the genome. We also define a window, *S* base-pairs wide, that begins on the region immediately adjacent to the right of the annotated Alu sequence, and denote this region *R*.

Polymorphic Alu deletions in a human individual will either be homozygous or heterozygous in nature, and each scenario creates a different alignment profile of read pairs over the defined regions *L*, *A* and *R*. At a particular annotated Alu locus, assuming a diploid genome, we consider all pairs of reads surrounding the locus where the left-aligned read of a respective pair is wholly contained within the defined region *L*, and the other is contained within *A* or *R*. We denote this as the set of Alu-informative read pairs, *I*. We ignore the case where both reads from a respective read pair are contained within a together within single region of *L*, *A*, or *R*. A schematic of the informative reads considered by our algorithm is depicted in Figure 1.

Assuming reads are sampled at random from a diploid set of chromosomes, we expect read pairs from the set *I* as defined above to have distinct characteristics depending on the Alu genotype of the sample being considered. Homozygous deletions with respect to the reference have a representative paired end signature depicted in Figure 1a. In this case, all reads from set *I* will span the interval *L-R* and have an insert size distribution that is larger than that of the sequencing library *S* that is approximately equal to *S+A*, since all reads will also span the Alu deleted with the respect to the reference. In the heterozygous case, we expect exactly half of the paired reads fall in the Alu sequence, and half of the paired reads to come from the chromosome containing the deletion. Therefore the set *I* should contain paired reads spanning the regions *L-A* and *L-R* in equal proportions, with mean insert sizes of *S* and *S+A*, respectively.

Our method effectively involves partitioning the informative reads from *I* into these two categories, each of which distinguishes the genetic state of a polymorphic Alu; a direct benefit of this 1:1 ratio is that it allows us to adopt conventions typically used for genotyping SNPs. We determine two simple parameters on which to base a genotype call, read depth, and allelic balance. We define the allelic balance (*AB*) as:

$$AB = \frac{number\ of\ L-\quad read\ pairs}{number\ of\ L-\quad read\ pairs\ + number\ of\ L-R\ read\ pairs}$$

Since the inherent character of the reads considered by our methodology will differ slightly from the short read data used in genotype SNPs, we determine reasonable cutoffs for emitting a genotype call heuristically.

Algorithm Implementation

The implementation of our algorithm is written in Perl, and depends on the commonly used bioinformatics software Samtools (15) and Bedtools (16). The implementation is separated into two phases, an initial preprocessing step, and the subsequent analysis step. The main steps of the implementation are outlined in Figure 2. The input requirement is a BAM format alignment file and a file containing a list of Alu element annotations in BED format, readily available as a track from the UCSC Genome Browser (17). The output is a tab-delimited text file with the chromosomal coordinates of an annotated Alu, genotype calls, and several metrics useful for prioritizing and/or filtering of calls.

Running time of PASSIGE

The running time of our implementation is expected to depend heavily on the depth of sequencing, as the vast majority of time is devoted to preprocessing the data into a format amenable for Alu genotyping analysis with PASSIGE. The analysis of a single chromosome sampled at a simulated sequencing depth of 60x and aligned using BWA using default settings took on average 45 minutes to preprocess on a desktop computer with an Athlon A10-5800 4.0 GHz APU. Analysis of the preprocessed data and subsequent Alu genotyping required negligible computation time, taking only several minutes for chromosome 1.

Synthetic dataset benchmarking

To test our method, we first created a simulated dataset derived from chromosome 17 of the hg18 version of the reference genome, downloaded from the UCSC Genome Browser. Using the RepeatMasker annotation track as a source of Alu annotations, we randomly selected 101 members of the AluY family of repeats and generated sequences for two haploid chromosomes with these AluY regions absent, such that the resultant diploid set contained 50 homozygous and 51 heterozygous AluY deletions. We purposefully selected AluY elements because they are the most frequently mutated Alus. We also required that there where no other repetitive elements within 300 base pairs of the Alu selected for deletion. While this is not necessarily representative of the mechanism of Alu transposon insertion, we did so in order to avoid confounding our heuristic determination of proper thresholds for deletion detection.

For the set of diploid simulation chromosomes, we simulated reads using the ART next-generation read simulator (18), which accurately simulates synthetic reads using a platform-specific error and base-pair quality profiles that are empirically derived from a large collection of sequencing datasets. We generated paired-end sequencing reads with a length of 75 base-pairs and a mean fragment size of 260 base-pairs and a standard deviation of 10. Reads were simulated at an average depth of 30x for each haploid chromosome, resulting in a total sequencing depth of 60x for the diploid set of chromosome 17. The parameters for read simulation were chosen to accurately reflect current sequencing data commonly being generated.

Simulated real data benchmarking

The HuRef reference genome is, to our knowledge, the only fully sequenced, whole-genome, shotgun assembly of a single individual (19). Therefore, it serves as useful reference upon which to benchmark the performance of PASSIGE as it is the only reference for which there exists a genome-wide validated "truth set" generated by parallel sequencing technology apart from short-read, next generation sequencing platforms. As before, we simulated reads on the combined HuRef and HuRef Prime reference genome assemblies for chromosome 1 to as before to a total diploid sequencing depth of 60x. Additionally, we simulated reads from chromosome X and Y, to evaluate any differences in the typing of haploid polymorphisms.

Evaluation of Alu genotyping using real data

We also tested our algorithm on chromosome 1 data publically available from the 1000 Genomes Project Consortium, Pilot 1 high coverage data for the European (CEU) and Yoruban (YRI) trios aligned to NCBI36, available at: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/data/

Evaluation of algorithm sensitivity

Finally, to assess how the sensitivity of our algorithm to detect polymorphic Alu elements is dependent on sequencing depth, we simulated reads on our synthetic and simulated real datasets at varying folds of coverage (5x, 10x, 20x, 30x, 60x, and 80x). In this context, sensitivity is assessed on the known polymorphic sites only, rather than all annotated sites.

RESULTS

Determination of heuristic cutoffs polymorphic site detection using PASSIGE

One expected consequence of short-read next generation sequencing on a whole genome scale that complicates delineation of polymorphic Alus is the mapping ambiguity of reads in repetitive regions. We note that our algorithm depends on the assumption that aligned reads mapped to the genome uniquely, which is generally reasonable given the current standard read length of 75-150 base pairs. However, this assumption is less valid when one considers the subset of reads that map to Alu repeats. As multimapping reads below a certain threshold are placed at random, we expect the coverage of pairs within the set within *I* as defined before will be slightly lower within these repetitive regions where a unique read and its repetitive pair maps to multiple possible locations. We noticed in our simulation studies that, even when reads are generated directly from the reference genome, many regions, especially those that are repeat rich, fail to achieve sufficient coverage of Alu-informative read pairs. Out of the 54,119 Alu-sine elements annotated on chromosome 17 in hg18, 3,740 failed to achieve any coverage of Alu-informative read pairs, leaving 50,379 Alus with non-zero coverage.

Since reads are independently sampled from each chromosome, accurate genotyping is ultimately a function of read depth over each allele. However, we found that a total read depth is not appropriate in the case of Alu polymorphism detection since the unique mappability of reads is expected to be different for *L-A* and *L-R* spanning reads and is highly dependent of genomic context. We chose to conservatively filter our method against polymorphic site detection by first applying a minimal coverage threshold of Alu-informative read pairs indicative of the non-reference allele at the particular locus under consideration (specifically, the coverage of *L-R* pairs in our terminology). Figure 3a shows the overall distribution of Alu-informative pairs for all annotated Alus. The frequency distribution of false-positives as a function of read depth of Alu-informative pairs is shown in Figure 3b. From this, a threshold of 15, or 1/4 of the total sequencing read depth, was chosen heuristically to produce no false positives in determination of polymorphic Alu sites. This resulted in a total of 48,086 annotated Alu sites on chromosome 17 attaining coverage sufficient for genotype consideration using our algorithm.

Empirical determination of allele-balance thresholds for genotyping calls

After determining this minimum threshold for reliable detection of polymorphic Alu sites, we next determined reasonable cutoffs for allelic balance and calling of zygosity. The range allele-balance values over our simulated heterozygous deletions ranged from 0.35 to 0.83 with an approximately normal distribution (data not shown). Setting these as the approximate limits for designation of heterozygosity led to proper classification of all homozygous reference sites on chromosome 17 that remained after the initial depth filtering. This heuristic range for allele balance is slightly higher than thresholds typically used for SNP calling in next generation sequencing data, which is expected given the increased ambiguity of mapping and higher variability in coverage around Alu elements. We that one out of the 50 homozygous non-reference sites we synthetically constructed, one was mistyped as heterozygous, with metrics that were not amenable to any filtering thresholds. Manual inspection of this site reveals that it has high homology to another region on chromosome 17 (Figure 4); we discuss methods to mediate this source of miss-calls further in the discussion section.

Based on our benchmark synthetic tests, we heuristically determined the following reasonable cutoffs for Alu-informative read-pair depth and allele balance. We require (1) that the coverage of Alu-informative pairs over the non-reference allele be equal to 1/4 of the average library sequencing depth to emit a polymorphic Alu call, and (2) that 0.35 < allele balance < 0.85 to emit a heterozygous variant call. While our thresholds only allow for just over 88% of the 54,119 annotated Alus on chromosome 17 to be effectively genotyped, we are able to achieve no false positives in terms of polymorphic site detection. The overall results of our synthetic benchmark tests are summarized in Table 1a and an example of output of the calls generated by PASSIGE alongside their concordance on all simulated polymorphic Alu deletions are available in Supplemental Table S1.

Overall, we achieved high accuracy on our synthetic dataset. PASSIGE exhibits very high sensitivity to detect isolated polymorphic Alu sites with an overall sensitivity of 98%. Interestingly, out method exhibits a slightly lower accuracy and sensitivity (97.9% and 96.0%, respectively) when assessing homozygous calls compared to heterozygous calls.

Polymorphism discovery and genotyping accuracy on simulated real data using HuRef

We next aimed to assess the effectiveness of both polymorphic site discovery of PASSIGE on real data using simulated reads generated from chromosome 1 of the HuRef reference genome. To examine the nature of polymorphic Alus present in HuRef, we stratified the polymorphic calls made by Alu-family subtype. As expected, the overwhelming majority of polymorphic Alu calls are of the AluY family. Phylogenetic studies have demonstrated that the AluY family is the youngest family of the Alu-SINE transposons, and the only one considered to still be functionally active in the human genome (20). Results of the calls made by PASSIGE on chromosome 1 of the HuRef assembly are shown in Table 2 and Figure 5. The breakdown of polymorphic Alus by family is wholly consistent with previously published datasets (11, 19). We noticed that in the simulate HuRef data, the ratio of heterozygous to homozygous calls is considerably less than 2:1, which would expected for a diploid human genome. We note that this lower ratio is consistent with previously published results (19), and is most likely due to the lower coverage (9.1x, on average) used to generate the HuRef reference, as well as artifacts introduced on the whole-genome shotgun assembly process.

Since the sequence HuRef genome assembly has been extensively studied elsewhere, this allows for direct evaluation of the accuracy of polymorphic Alu calls generated by PASSIGE. We directly compared the results using our simulated read data to the polymorphic Alu calls determined by contig mapping to the reference published in Levy et al (Table 1b and Supplemental Table S2). In addition to the PASSIGE results for chromosome 1, we also compared calls for the haploid chromosomes X and Y, to evaluate performance on haploid chromosomes. This comparison of real data generated by whole-genome, shotgun assembly using gold-standard Sanger sequencing allows for the most accurate assessment of accuracy and sensitivity of PASSIGE apart from any inherit systematic bias introduced by next-generation, short-read sequencing technology. When applying the same heuristic cutoffs as previously determined, we see a markedly lower sensitivity on both diploid and haploid chromosomes as compared to synthetically generated data (91.5% and 83.3%, respectively vs. 98.0%). We also notice a marked decrease in determination of zygosity in homozygous calls (86.0% vs. 97.9%). We do note perfect concordance on haploid chromosomes, but the low number of polymorphic Alu sites annotated in Levy et al makes this increase in accuracy on haploid chromosomes only a suggestive one. Unfortunately, in their study, Levy et al only provide homozygous non-reference Alu deletions, precluding any direct assessment of the heterozygous Alus calls made by PASSIGE.

Genotyping performance and utility using 1000 Genomes data

The synthetic and simulated real datasets demonstrate that PASSIGE is useful for Alu polymorphism discovery genotyping by evaluating whole-genome sequencing data at all annotated Alus. One benefit of our implementation and an area in which we believe it will be of particular utility is the rapid genotyping when only a subset of known polymorphic Alu sites, rather than all annotated Alu sites, are considered. Only a small fraction of the annotated Alu sites within the human genome are actually polymorphic (20), and this will significantly reduce computational time as well as false-positive calls.

To demonstrate this, we ran PASSIGE on chromosome 1 for two complete high-coverage complete trios publically available from the 1000 Genomes Project Pilot 1 available for the CEU and YRI populations. Instead of scanning all Alu annotations, we focused on those Alu sites already determined to be polymorphic as determined in Stewart et al (10).

Out of the 106 known polymorphic sites we tested, we found evidence an average of 75 non-reference polymorphisms in all samples. Only 99 of the 106 sites tested showed evidence non-reference deletions. We utilized the fact that we have complete trios to estimate the false positive rate (FP%) in the child, by counting the number of polymorphisms detected in the child but absent in either parent in each respective trio. This provides a reasonable estimate for the upper bound of the FP% in the polymorphic Alu detection using PASSIGE at 10.38% and 3.90% for the CEU and YRI trio child sample, respectively. Results are summarized in Table 3.

Furthermore, to check for the possibility of any systematic, non-biological bias and evaluate the utility of Alu-genotyping in general, we performed PCA analysis on the Alu genotype calls generated by PASSIGE. We note that, even with a higher FP% rate and considerably lower number of genotyped sites than SNP genotype data, we are able to capture enough variation to appropriately stratify our trio samples by ethnicity on the first principal component (Figure 6). This reinforces the notion that the Alus, given their fixative nature in the genome and their ability to capture genetic diversity, are an ideal marker for genetic studies (6).

Evaluation of sensitivity based on read depth

In our synthetic, simulated, and real datasets, our samples are sequenced at extremely high depth (60x or greater). In our heuristic determination of thresholds, depth of coverage of Alu-informative pairs had the single largest effect on the sensitivity of PASSIGE and its ability to assess Alu polymorphisms. Therefore, a reasonable question is to ask how the sensitivity is affected by lower overall sequencing depth. We generated our synthetic and simulated real datasets at varying depths of coverage and assessed the sensitivity of our thresholds only at the known polymorphic loci. This data is summarized in Figure 7. We see that, at sequencing depths lower than 30x, there is a significant drop in sensitivity. Results are generally comparable between both synthetic data and data generated from HuRef, but lower overall for the simulated real dataset.

DISCUSSION

Comparison with other methods

While there have been several recent methods developed to discover novel Alu polymorphism insertions, few have focused on a systematic and rapid way to genotype non-reference deletions. We use an anchored read-pair algorithm similar to previous methods developed for detection of structural variation (21-23), but our method is tailored with specific heuristics developed for detection of polymorphic Alu non-reference deletions. Most of the methods are aimed specifically at indel discovery, including BreakDancer, Pindel, and Spanner, make use of all available read data, and do not explicitly consider the diploid nature of the genome. One exception to this is MoDil, but its implementation requires the use of mate-pair reads, which is much less prevalent than paired-end sequencing, due in part to the increased complexity of experimental library generation and associated costs.

We note that our method is algorithmically similar to one recently published method also tailored for Alu-polymorphism genotyping, PAIR (24). Similar to PASSIGE, they use the anchored read-pair specifically in the context of Alus. However, all other things being equal, our method for non-reference deletion detection theoretically requires 1/3 fewer reads, and therefore should computationally more efficient. We do report a slightly better sensitivity in synthetically generated data (98 vs 97%), but note this difference is likely representative of different synthetic benchmarks. We note that our method is also more directly comparable to SNP genotyping data, sampling from each haploid allele in equal proportions, and using established metrics like allele balance and depth as extended to genotyping of polymorphic Alus. Moreover, in their report, they to not develop any reasonable filtering thresholds and the lack of accessibility to their implementation prevents us from performing a direct performance comparison between the two methods.

Extensions and future work

One shortcoming of PASSIGE that we noticed in our synthetic and simulated dataset is in the low accuracy of genotype differentiation between heterozygous and homozygous states. This is akin but opposite in direction to a similar effect in the genotyping SNPs, where the rare or alternate allele is often under-sampled in comparison to the reference. One possible way to improve upon this is the incorporation and development of a mappability metric for read-pair structural variation detection, similar to the Uniqueome (25). The mapping quality (MQ) metric commonly generated by alignment algorithms like BWA are assigned to a particular read, and not the pair as a whole. The utility of this conventional metric does not apply when one read is anchored in a repetitive element. Moreover, our method could be extended to incorporate split-reads, where the read is aligned in a way such that the alternate allele deletion is spanned by aligned parts of a single read. This is becoming more prevalent as read length continues increase, and will likely eclipse the utility of the anchored paired-read method once read length is able to span greater than the length of the typical 300 base pair Alu element.

Additionally, our method can easily be extended to incorporate other types of retrotransposons, provided different heuristic cutoffs are empirically derived. Moreover, in the context of Alus, we are evaluating the efficacy of PASSIGE in assessing Alu-mediated mobile element-associated deletions that incorporate non-homologous recombination of multiple Alu-elements, as these have been shown to be causal for many genetic disorders (26-31).

Conclusion

A goal of our implementation of PASSIGE is to produce a method simple and computationally efficient. Previous large scale studies that effectively genotype non-reference Alu deletions, such as those generated by studies of structural variation in 1000 Genomes Data, required extremely large numbers of samples, and involve consensus calls from dozens of methods and sequencing technologies implemented by multiple genome centers. The computational resources required often precludes such analysis in smaller re-sequencing studies performed by individual labs.

We developed PASSIGE specifically to circumvent this computational burden and enable a rapid assessment of non-reference Alu polymorphisms. We show that our method achieves reasonable sensitivity for the discovery of Alu polymorphisms, with the ability to assess of over 88% of all annotated Alus in our synthetically generated dataset. This sensitivity will likely improve with the generation of a background set of “un-type-able” Alu elements are cataloged, as described above.

Moreover, along this vein, we designed PASSIGE to explicitly make use of the large catalogs of structural variation generated in large-scale studies like those conducted by the 1000 Genomes Consortium. While we demonstrate that PASSIGE is useful for polymorphism discovery, as we demonstrate using high-coverage trio data that PASSIGE has a particular utility when polymorphic Alu-sites are known *a priori*, with an upper-bound on the FP% estimated to be as low as 3.90% in the YRI trio. The effective use of PASSIGE in genetic studies is expected to increase as large-scale population surveys of genetic variation, like those released by the 1000 Genomes Consortium, continue to improve both in their accuracy and comprehensiveness, facilitating more accurate and efficient genotyping of polymorphic Alu elements using PASSIGE.

Works Cited

1. Welter D*, et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research* 42(Database issue):D1001-1006.
2. Xing J*, et al.* (2009) Mobile elements create structural variation: analysis of a complete human genome. *Genome research* 19(9):1516-1526.
3. Deininger P (2011) Alu elements: know the SINEs. *Genome biology* 12(12):236.
4. Salem AH, Kilroy GE, Watkins WS, Jorde LB, & Batzer MA (2003) Recently integrated Alu elements and human genomic diversity. *Molecular biology and evolution* 20(8):1349-1361.
5. Ade C, Roy-Engel AM, & Deininger PL (2013) Alu elements: an intrinsic source of human genome instability. *Current opinion in virology* 3(6):639-645.
6. Antunez-de-Mayolo G*, et al.* (2002) Phylogenetics of worldwide human populations as determined by polymorphic Alu insertions. *Electrophoresis* 23(19):3346-3356.
7. Mamedov IZ*, et al.* (2010) A new set of markers for human identification based on 32 polymorphic Alu insertions. *European journal of human genetics : EJHG* 18(7):808-814.
8. Zhao GS, Chang L, & Mo YN (2010) [Applications of Alu family in forensic DNA analysis]. *Fa yi xue za zhi* 26(1):47-50.
9. Abecasis GR*, et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56-65.
10. Stewart C*, et al.* (2011) A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS genetics* 7(8):e1002236.
11. Hormozdiari F*, et al.* (2011) Alu repeat discovery and characterization within human genomes. *Genome research* 21(6):840-849.
12. Keane TM, Wong K, & Adams DJ (2013) RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics (Oxford, England)* 29(3):389-390.
13. Witherspoon DJ*, et al.* (2010) Mobile element scanning (ME-Scan) by targeted high-throughput sequencing. *BMC genomics* 11:410.
14. Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* 25(14):1754-1760.
15. Li H*, et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* 25(16):2078-2079.
16. Quinlan AR & Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* 26(6):841-842.
17. Karolchik D*, et al.* (2014) The UCSC Genome Browser database: 2014 update. *Nucleic acids research* 42(Database issue):D764-770.
18. Huang W, Li L, Myers JR, & Marth GT (2012) ART: a next-generation sequencing read simulator. *Bioinformatics (Oxford, England)* 28(4):593-594.
19. Levy S*, et al.* (2007) The diploid genome sequence of an individual human. *PLoS biology* 5(10):e254.
20. Bennett EA*, et al.* (2008) Active Alu retrotransposons in the human genome. *Genome research* 18(12):1875-1883.
21. Chen K*, et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods* 6(9):677-681.
22. Ye K, Schulz MH, Long Q, Apweiler R, & Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)* 25(21):2865-2871.

23. Lee S, Hormozdiari F, Alkan C, & Brudno M (2009) MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nature methods* 6(7):473-474.

24. Sveinbjornsson JI & Halldorsson BV (2012) PAIR: polymorphic Alu insertion recognition. *BMC bioinformatics* 13 Suppl 6:S7.

25. Koehler R, Issac H, Cloonan N, & Grimmond SM (2011) The uniqueome: a mappability resource for short-tag sequencing. *Bioinformatics (Oxford, England)* 27(2):272-274.

26. Aimi Y*, et al.* (2013) A novel break point of the BMPR2 gene exonic deletion in a patient with pulmonary arterial hypertension. *Journal of human genetics* 58(12):815-818.

27. Harmel EM*, et al.* (2013) Alu-mediated recombination defect in IGF1R: haploinsufficiency in a patient with short stature. *Hormone research in paediatrics* 80(6):431-442.

28. Mizunuma M*, et al.* (2001) A recurrent large Alu-mediated deletion in the hypoxanthine phosphoribosyltransferase (HPRT1) gene associated with Lesch-Nyhan syndrome. *Human mutation* 18(5):435-443.

29. Schanze D*, et al.* (2014) Deletions in the 3' Part of the NFIX Gene Including a Recurrent Alu-Mediated Deletion of Exon 6 and 7 Account for Previously Unexplained Cases of Marshall-Smith Syndrome. *Human mutation*.

30. Seabra CM*, et al.* (2014) A novel Alu-mediated microdeletion at 11p13 removes WT1 in a patient with cryptorchidism and azoospermia. *Reproductive biomedicine online*.

31. Wada T*, et al.* (2013) Alu-mediated large deletion of the CDSN gene as a cause of peeling skin disease. *Clinical genetics*.

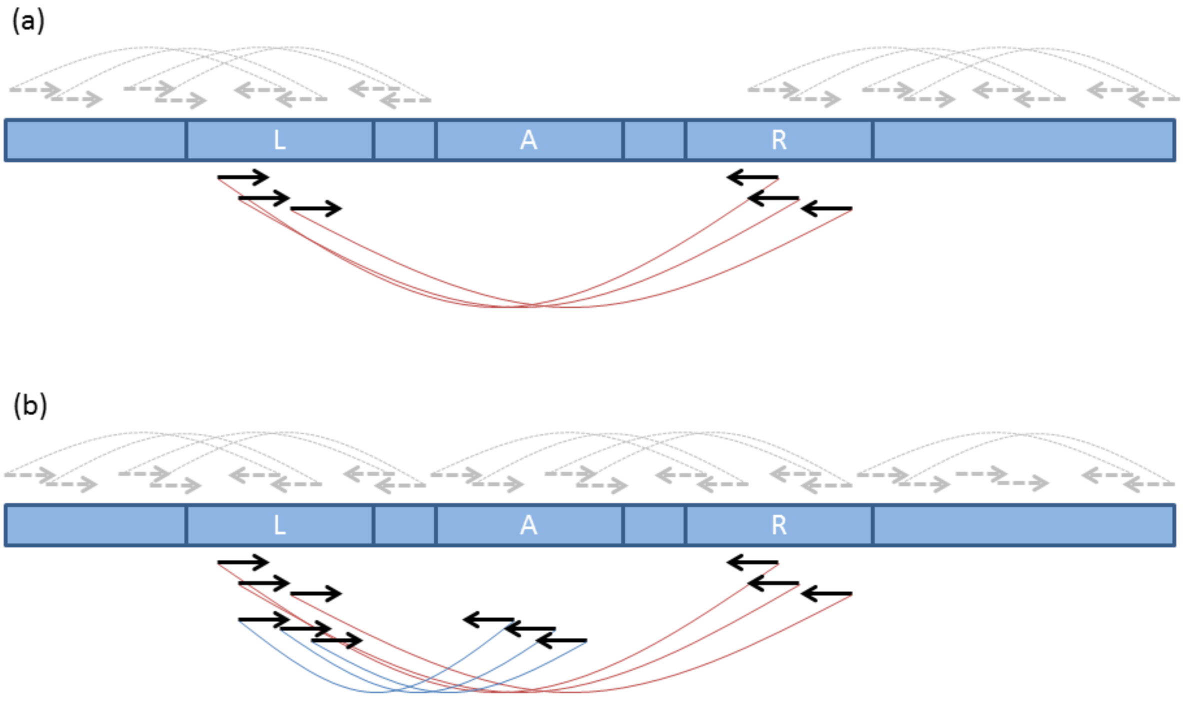

Figure 1. Schematic overview of reads considered for Alu deletion detection. Depicted is the representative paired-read signature of a presence of (a) a homozygous deletion and (b) a heterozygous deletion with respect to the reference. L and R denote equal-sized regions to the left and right of an annotated Alu region, A. The reads pairs (black arrows) considered by our algorithm are colored according to their respective insert sizes. The red curves represent read pairs spanning an Alu deletion that are larger than the expected insert size. The blue curves represent reads derived from a chromosome that does not contain a deletion and have insert sizes that fall within the expected size. Gray-dotted curves and arrows denote sequenced read pairs that are ignored by our algorithm.

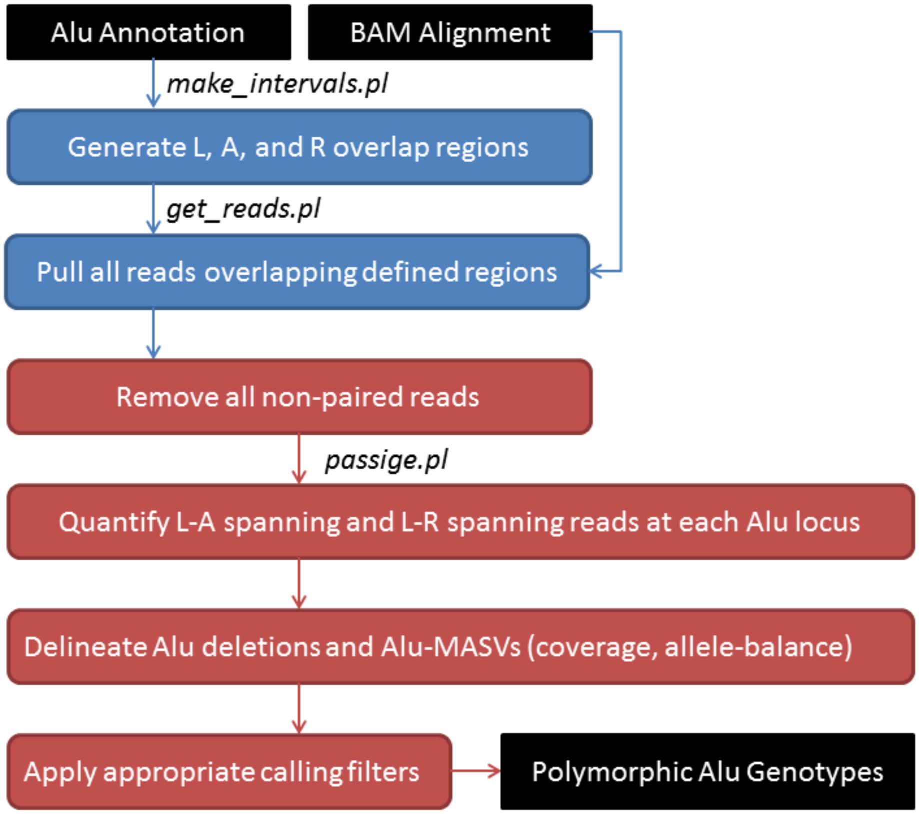
Alu Annotation
BAM Alignment
make_intervals.pl
Generate L, A, and R overlap regions
get_reads.pl
Pull all reads overlapping defined regions
Remove all non-paired reads
passige.pl
Quantify L-A spanning and L-R spanning reads at each Alu locus
Delineate Alu deletions and Alu-MASVs (coverage, allele-balance)
Apply appropriate calling filters
Polymorphic Alu Genotypes

Figure 2. Outline of software implementation of algorithm. Required input files and resultant output are shown in black boxes. Preprocessing steps are shown in blue. Analysis steps are shown in red. The PERL files that implement the different steps are italicized.

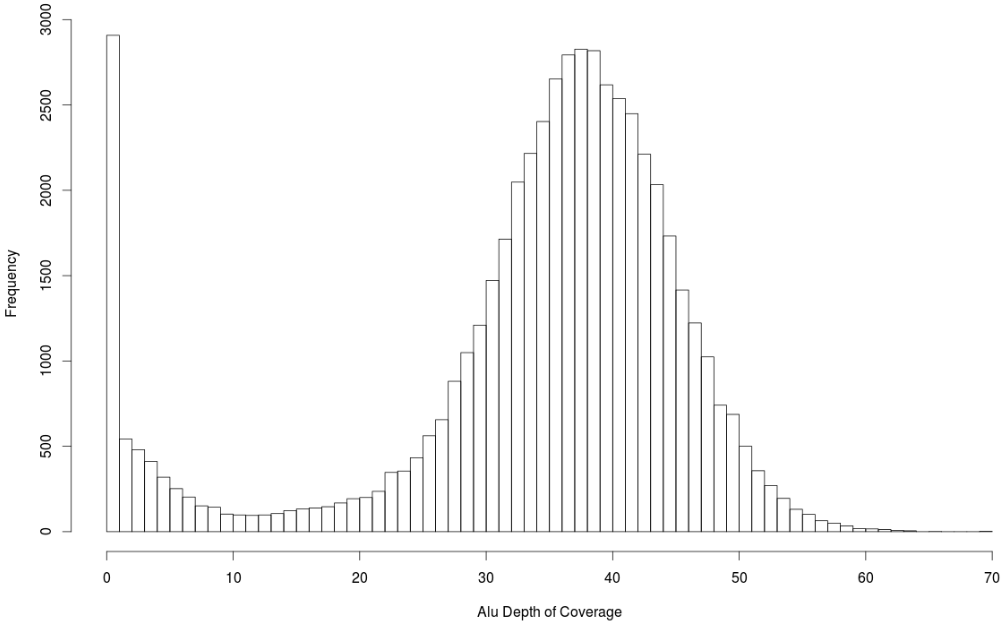

Figure 3a. Distribution of Alu informative pair counts at all annotated Alus on chromosome 17.

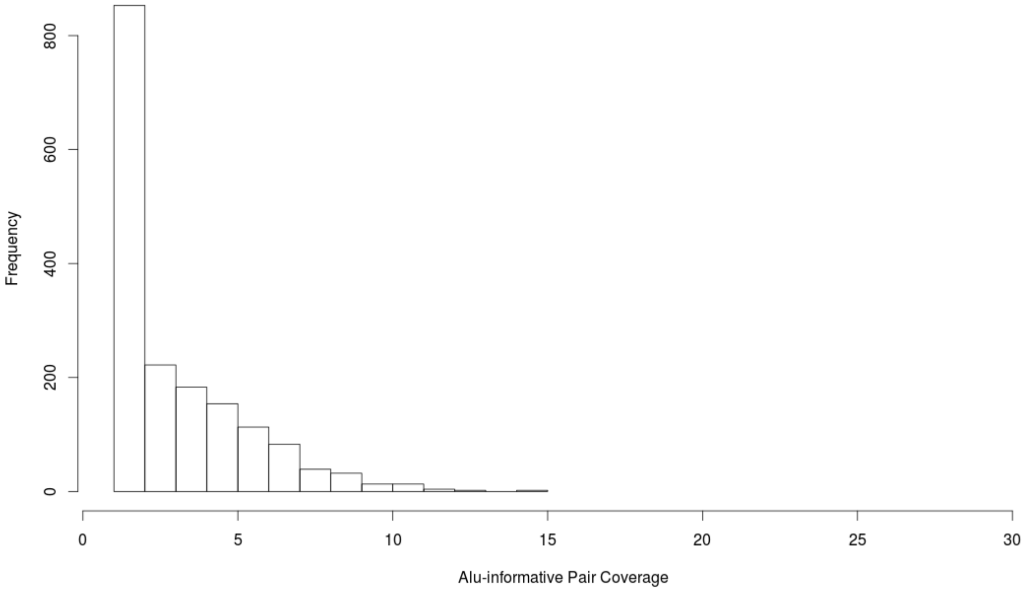

Figure 3b. Frequency of false positives at various depths of non-reference Alu informative pair coverage.

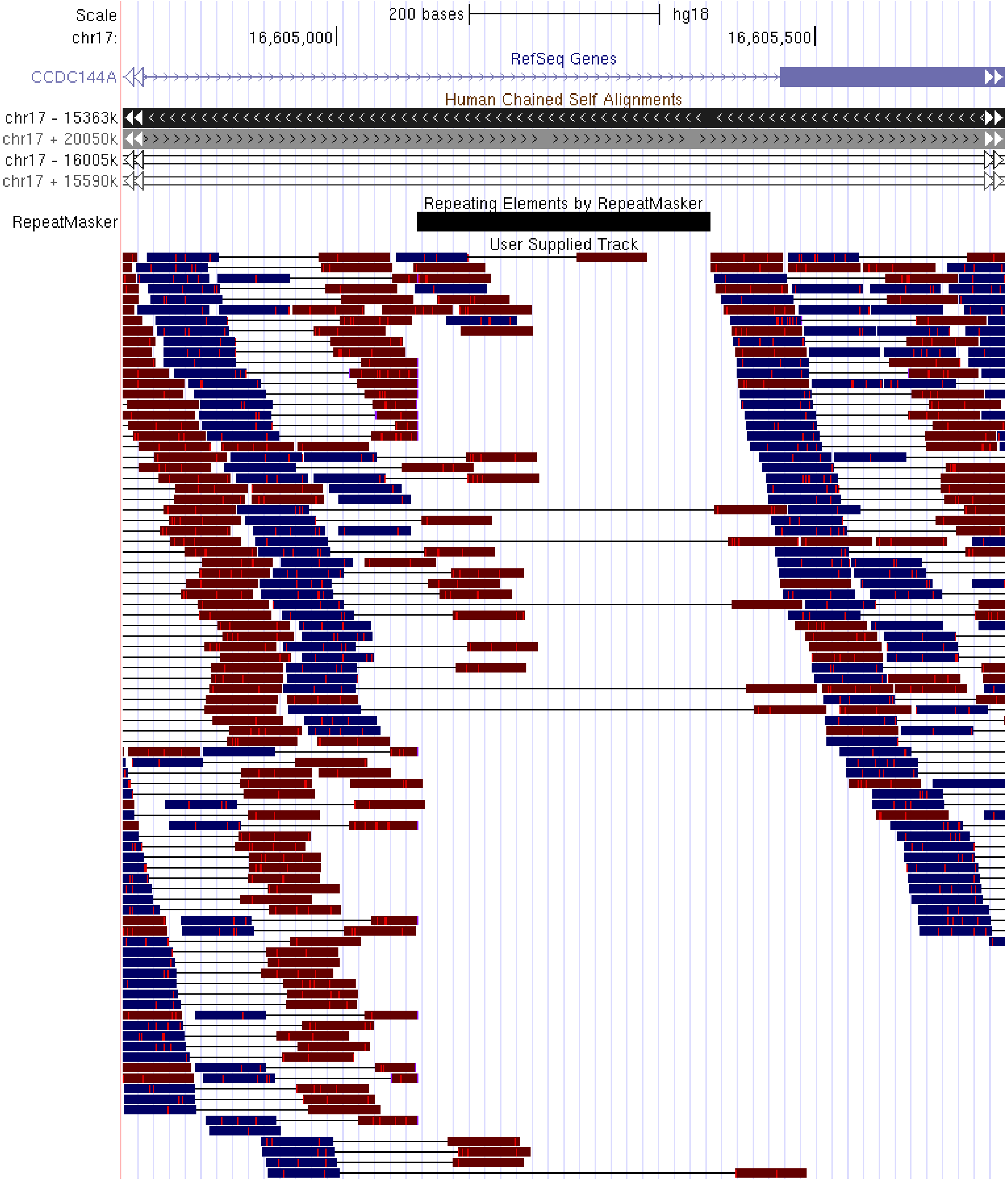

Figure 4. False-positive Alu call made by PASSIGE. The single homozygous Alu region misstyped as a heterozygote in our synthetic dataset on chromosome 17 is shown on the UCSC genome browser. The self-chain alignment track reveals that this region has high homology to another region on the same chromosome.

	Total	Considered	Correctly Called
Polymorphic Sites	101	99(98.0%)	98(98.9%)
Homozygous	50	48(96.0%)	47(97.9%)
Heterozygous	51	51(100%)	51(100%)

Table 1a. Summary of results from the synthetic dataset.

	Total	Considered	Correctly Called
chr1	47	43(91.5%)	37(86.0%)
chrX,Y	12	10(83.3%)	10(100%)

Table 1b. Summary of results from simulated HuRef dataset Note that all sites summarized here are homozygous (see text).

Alu Family	Homozygous	Heterozygous	Total
AluJ*	0	0	0
AluS*	0	2	2
AluY	17	8	25
AluYa5	30	4	34
AluYa8	2	3	5
AluYb8	11	15	26
AluYb9	4	1	5
AluYc	1	2	3
AluYc3	0	0	0
AluYd8	0	0	0
AluYf4	0	2	2
AluYg6	2	3	5
AluYh9	0	0	0
AluYk1	0	1	1
AluYk4	0	0	0
Totals	67	41	108

Table 2. Summary of PASSIGE polymorphic Alu calls on simultated reads from HuRef chromosome 1.

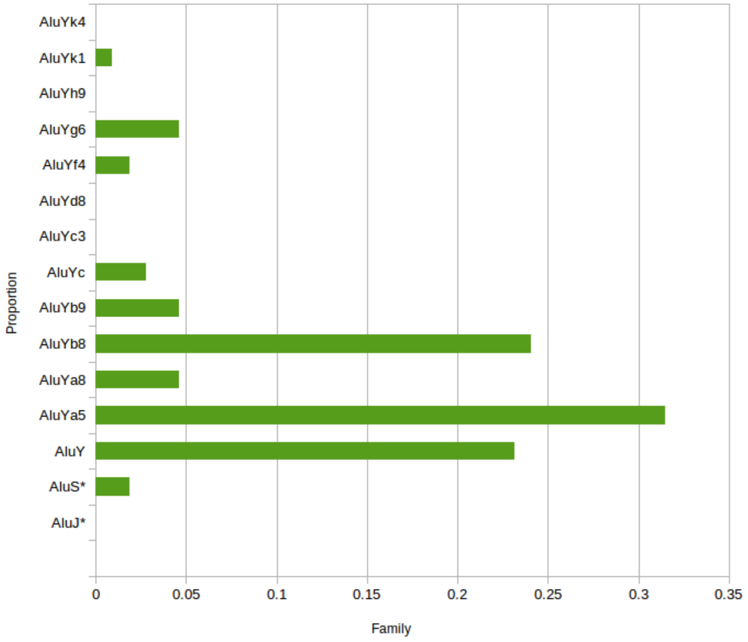

Figure 5. Polymorphic Alu elements classed by family detected by PASSIGE on chromosome 1 of the combined HuRef and HuRef Prime assemblies.

Sample	Relationship	Population	Polymorphic Calls	Mendelian Errors	FP%
NA12892	Daugher	CEU	77	8	10.38%
NA12891	Father	CEU	73	NA	NA
NA12878	Mother	CEU	67	NA	NA
NA19240	Daughter	YRI	76	3	3.90%
NA19239	Father	YRI	78	NA	NA
NA19238	Mother	YRI	79	NA	NA

Table 3. Summary of polymorphic Alu calls made on 106 polymorphic loci on chromosome 1. Discordant genotype counts (present in the child, but not the parents) as well as an estimated upper-bound of the false positive rate (FP) for the child are shown as well.

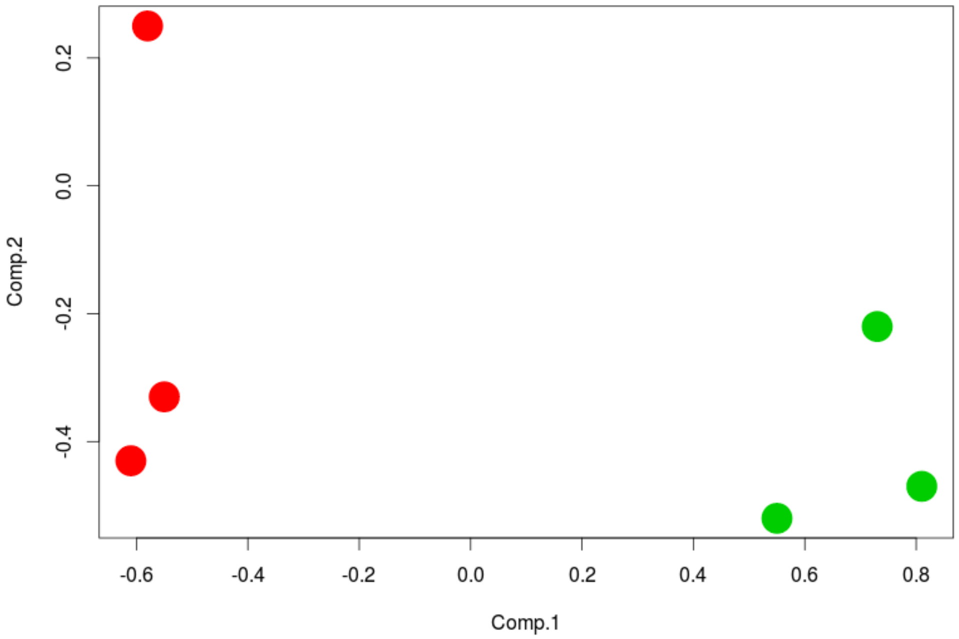

Figure 6. PCA plot of the first two principle components generated for the 1KG trios. Trios from CEU (red) and YRI (green) demonstrate good separation by ethnicity using genotype calls on polymorphic loci on chromosome 1 generated by PASSIGE.

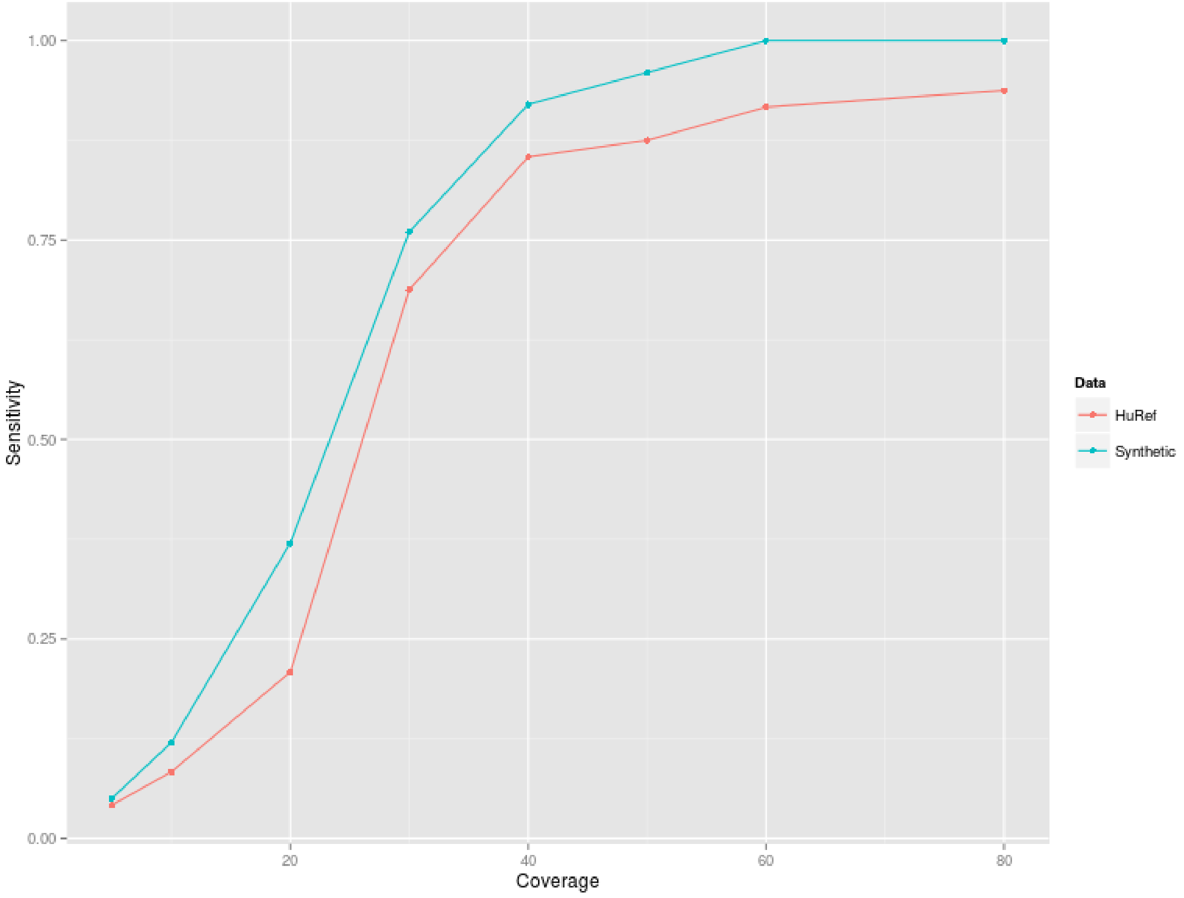

Figure 7. Sensitivity of PASSIGE at varying depths. The sensitivity of PASSIGE to assess known polymorphic loci is plotted at varying sequencing depths in both our synthetic data set as well as HuRef simulated dataset. The location of the polymorphic loci are known a priori. We see a significant drop of sensitivity at a sequencing depth lower than 30x.