A Comparison of Filters for Reducing False Positive Indels in Low Complexity Regions

Mike Huang

1 Summary

False Positive indels are known to occur in Low Complexity Regions (LCR) and/or in tandem repeats (TR). This report will explore the efficacy of various filters in reducing the number of false positive indels for SUREKids variant calls. The LCR filter reduced the most indels (54%), while the other filters reduced the indels by less than 5%. The validation for LCR showed that 69/4460(1.5%) of indels in LCR regions are found in ExAC vs 1140/3792(30%) of indels in non-LCR regions are found in ExAC. This supports the LCR filter as a suitable filter that reduces false positives of indels while losing few true positives.

2 Methods

2.1 Filters Compared (BED files)

1. *Low Complexity Region filter* generated by the original mDUST algorithm used in BLAST. A BED file of the LCRs for GChR37 can be found at https://github.com/lh3/varcmp/raw/master/scripts/LCR-hs37d5.bed.gz
2. *Tandem Repeat filter for repeats =>5* generated by Jeanie Lim. The BED file is named Duke800_PatternsRepeat5 and can be found in the supplementary files.
3. *Tandem Repeat filter for repeats =>10* generated by Jeanie Lim. The BED file is named Duke800_PatternsRepeat10 and can be found in the supplementary files.
4. *Tandem Repeat filter for repeats =<12* generated by Repeat Masker. The BED file of the Repeat Masker regions for hg19 is contained within chromTrf.tar.gz and can be found at http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/

2.2 Dataset

Trio Samples: CHH1676, CHH1677, CHH1678
Filter: pass_filtered

2.3 Reference Indel Databases:

1. dbsnp_135.b37.vcf.gz
2. ExAC.r0.3.1.sites.vep.vcf.gz

2.4 Filtering BED regions from VCF file

VCF files were filtered with the respective BED regions using the following command:

```
vcftools --gzvcf trio.vcf.gz --out trio_outputname --keep-only-indels --recode --exclude-bed
filter.bed
```

2.5 Validation

Validation was performed by comparison with reference variant databases, dbSNP and ExAC, which is considered as "ground truth." This was performed by intersecting the filtered VCF file with the reference VCF. The intersection was performed with the following command:

```
bcftools stats trio.vcf.gz reference.vcf.gz
```

2.6 Allele Frequency Analysis

Allele Frequency was analyzed from the ExAC population and compared between *LCR indels found in ExAC* vs *non LCR indels found in ExAC*. To find these two corresponding VCFs, the intersections were made with the following command:

```
bcftools isec -p outputfolder trio_indels.vcf.gz ExAC.r0.3.1.sites.vep.vcf.gz
```

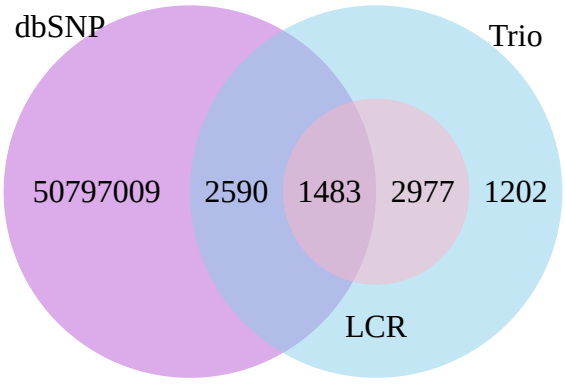
3 Results

3.1 Remaining variants after filtering

Filter	SNPs Remaining	Indels Remaining
LCR	47611/52521(90.7%)	3792/8252(46.0%)
TR =>5	50871/52521(96.9%)	7984/8252(97.8%)
TR =>10	52322/52521(99.6%)	8102/8252(98.2%)
TR <=12	51620/52521(98.3%)	7481/8252(90.7%)

The LCR filter that substantially reduced the number of indels at 46%. Furthermore, the SNPS reduced is at 9.3%.The TR filters were significantly less effective at reducing Indels and were excluded from further analysis.

3.2 Validation of LCR filtered indels with dbSNP



	Indels
$Trio \cap LCR$	4460
$Trio \setminus LCR$	3792
$dbSNP \cap Trio \cap LCR$	1483
$dbSNP \cap Trio \setminus LCR$	2590

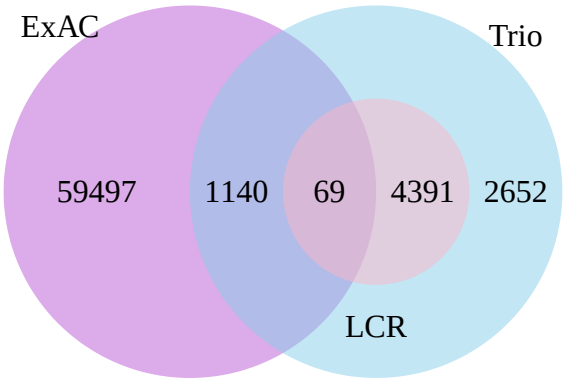
	TP	FP	Precision
Before LCR filter	4073	4179	4073/8252(49%)
After LCR filter	2590	1202	2590/3792(68%)

$Precision = \frac{TP}{TP+FP}$

An ideal filter should exclude false positives while keeping true positives. In this case, 1483/4460(33%) of indels in LCR regions are found in dbSNP vs 2590/3792(68%) of indels in non-LCR regions are found in dbSNP. Thus if dbSNP is a gold standard to determine true positives, 33% of the indels filtered by LCR are true positives. This is a non-trivial amount. However, it raised the precision from 49% to 68%.

However, dbSNP is not without inaccuracies. dbSNP is estimated to have a 15-17% false positive rate[1].

3.3 Validation of LCR filtered indels with ExAC



Intersection	Indels
$Trio \cap LCR$	4460
$Trio \setminus LCR$	3792
$ExAC \cap Trio \cap LCR$	69
$ExAC \cap Trio \setminus LCR$	1140

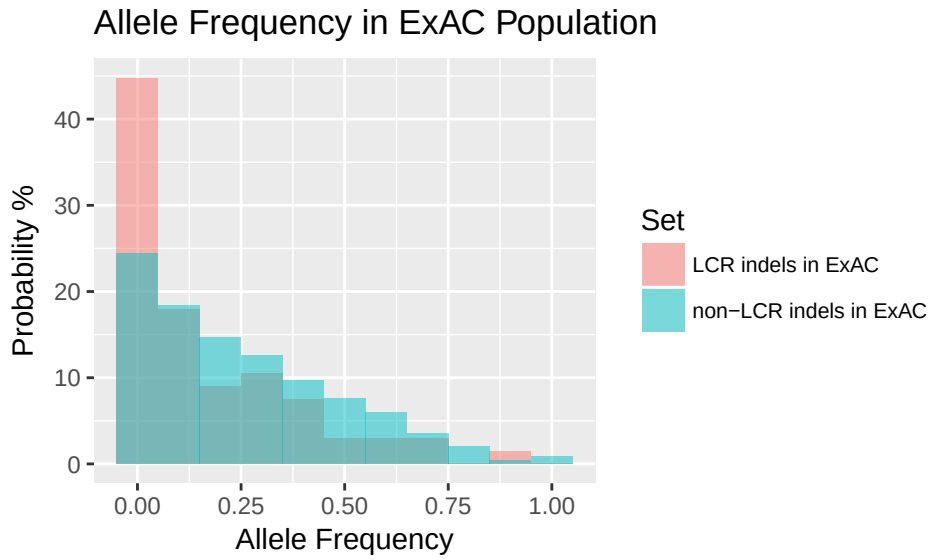
	TP	FP	Precision
Before LCR filter	1209	7043	1209/8252(15%)

	TP	FP	Precision
After LCR filter	1140	2652	1140/3792(30%)

69/4460(1.5%) of indels in LCR regions are found in ExAC vs 1140/3792(30%) of indels in non-LCR regions are found in ExAC. Thus, if ExAC was considered the gold standard, LCR filters very few TP at 1.5% while doubling the precision from 15% to 30%.

3.4 Allele Frequency Analysis

Perhaps the 1.5% of indels in LCR regions that are found in ExAC may be due to errors in ExAC itself. Low allele frequencies have been shown to lead to lowered confirmation of SNPs in public databases[1]. To determine the likelihood of errors in the indels found in ExAC that intersects with LCR, the distribtion of allele frequencies was determined between LCR indels in ExAC and non LCR indels in ExAC.



The LCR indels are 45% likely to have an allele frequency <10% while non-LCR indels are 24% likely to have an allele frequency <10%. This lower distribution in allele frequencies for LCR indels intersecting ExAC supports the hypothesis that those indels are errors in ExAC, and thus the percentage of true positives filtered by LCR likely less than 1.5%.

4 Conclusion

1. The LCR filter reduced the most indels (54%), while the other filters reduced the indels by less than 5%.
2. The LCR filter showed very strong selectivity for reducing false positives over true positives in its comparison with ExAC as a reference.
3. The analysis of the distribution of allele frequencies further supports the LCR filter's strong selectivity for false positives over true positives.

5 References

1. Mitchell, A. A., Zwick, M. E., Chakravarti, A., & Cutler, D. J. (2004). Discrepancies in dbSNP confirmation rates and allele frequency distributions from varying genotyping error rates and patterns. Bioinformatics, 20(7), 1022-1032. doi:10.1093/bioinformatics/bth034