# Addressing Ambiguous Indels with the indelGeneralizer

Robert E. Denroche[1], Andrew M.K. Brown[1], Kathy Chun[2,3], B.F. Francis Ouellette[1,4], John D. McPherson[1,5]

1. Ontario Institute for Cancer Research, Toronto, Ontario
2. Genetics Program, North York General Hospital, Toronto, Ontario
3. Department of Laboratory Medicine and Pathobiology, University of Toronto, Ontario
4. Department of Cell and Systems Biology, University of Toronto, Ontario
5. Department of Medical Biophysics, University of Toronto, Ontario

## Abstract

The genomic position of an indel is ambiguous when the bases being inserted or deleted are in the context of a homopolymer or simple repeat (as can be seen in Figure 1). Many aligners (Novoalign, bwa) report ambiguous indels at the leftmost (or most 5' in the reference strand) genomic position, however, local realignment steps (GATK) can shift the position of the indel in the repeat. Furthermore, there exist conventions that handle ambiguous indels differently; HGVS guidelines dictate that the most 3' position in the direction of transcription should be arbitrarily assigned. Ambiguous indel positions become troublesome when comparing calls between datasets. Due to variation in alignment, local realignment or calling conventions, the same indel may be reported at different positions and considered to be different variants during downstream analysis. Comparing lists of variants is a common practice in genomic analysis; for example, contrasting a tumour-normal pair to identify somatic variants or searching a database such as dbSNP.

We developed the indelGeneralizer, an open-source software tool capable of identifying ambiguous indels and converting their genomic position to a consistent, user selected format. In a recent study evaluating the use of NGS for clinical BRCA genotyping, variants in 96 samples were called using a typical pipeline (Novoalign, GATK) and compared to the clinical, HGVS formatted calls derived from Sanger sequencing. Of the 61 pathogenic indels reported, we initially found that 27 (44%) of the NGS calls were inconsistent with the clinical data due to alignment ambiguity. By applying the indelGeneralizer we resolved 24 (89%) of the ambiguous indels. Alignment of the outstanding 3 variants involved multiple ambiguous events – a current limitation of the software which we hope to address in future versions. In its current state, the indelGeneralizer could be beneficial for many types of analyses where variants called from different pipelines must be compared.

## Implementation

The indelGenerlizer is a perl script which is available at https://github.com/robdenroche/indelGeneralizer. It uses bioperl (http://www.bioperl.org) to index and read reference fasta files.

indelGeneralizer is capable of identifying whether a given indel is ambiguous and returning a generalized alignment. The user can specify a generalization convention to use (e.g. output the leftmost position) and also control whether anchor bases are used and how the positions are reported. Currently there are two modes the script can be run in:

**Interactive Mode:**
- Attempts to generalize a single indel which is supplied on the command line;
- Can output the indel aligned to the *left* or *right*, output *all* alignments, or output a *full* display of all the ambiguous alignments (Figure 2).

**VCF Mode:**
- Efficiently runs over an entire sorted VCF file by storing chunks of the fasta reference in memory;
- Outputs indels aligned to the *left* or *right*, skipping SNPs or multi-allelic variants;
- Has options to preserve the input VCF's order, or (in most cases) produce a sorted VCF by ensuring that generalized indels are placed correctly in the file.


**Figure 1:** The deletion of one T from a homopolymer run of 3 Ts can be called in three separate positions. While each position is a valid way to report the deletion, differences between conventions or pipelines can result in calling the same indel different ways. If these discrepancies are not taken into account when comparing lists of variants, ambiguous indels that happen to be called at separate positions will be erroneously treated as separate variants.

```
-c chr17 -p 41244988 -i +GC -m full

                          |41244984
(aligned consensus)       TGAGCGCATC
chr17 41244986 . A AGC    TGA**GCATC
chr17 41244988 . C CGC    TGAGC**ATC <--
(aligned consensus)       TGAGCGCATC
                           41244991|
```

```
-c chr17 -p 41246043 -i -ATTTA -m full

                                |41246037
(reference sequence)            CGCTTTAATTTATTT
chr17 41246039 . CTTTAA C       CGC-----TTTATTT
chr17 41246040 . TTTAAT T       CGCT-----TTATTT
chr17 41246041 . TTAATT T       CGCTT-----TATTT
chr17 41246042 . TAATTT T       CGCTTT-----ATTT
chr17 41246043 . AATTTA A       CGCTTTA-----TTT <--
(reference sequence)            CGCTTTAATTTATTT
                                      41246051|
```

**Figure 2:** The *full*-mode output of two runs of the indelGenerlizer. *Left*: An insertion of GC can be reported as occurring either before or after the GC in the reference sequence. *Right*: A deletion of two As and three Ts can occur in many positions, and the order of the reference bases being deleted can change and still result in an equivalent deletion. Prior to generalization, both of these indels were reported as errors in a clinical BRCA sequencing experiment.

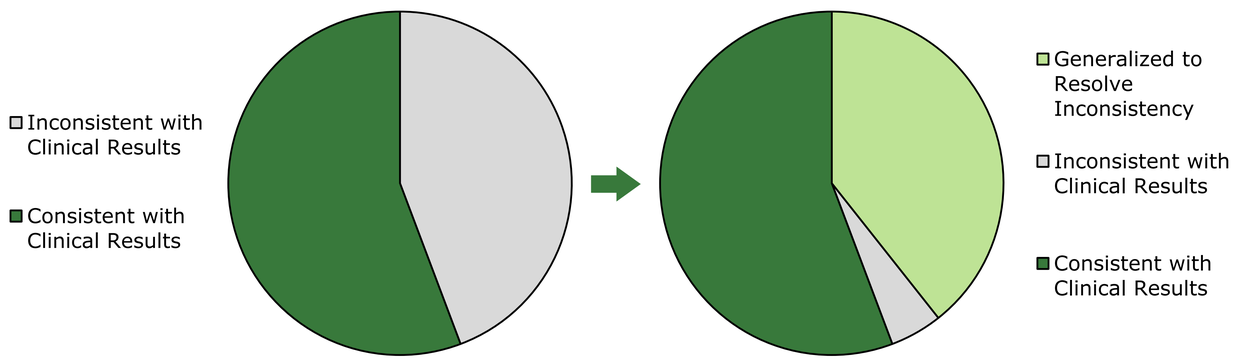


**Figure 3:** *Left:* Before being processed with the indelGeneralizer, 27 (44%) of the 61 pathogenic indels identified in a BRCA sequencing experiment were inconsistent with those reported in the clinical results due to ambiguity. *Right:* 24 (89%) of the 27 inconsistent indels were successfully resolved by generalization. The remaining 3 were reported as multiple edit events in the clinical results and required additional steps to resolve.

| Input VCF | # Indels | # Ambiguous Indels | # Duplicate Indels | Run Time |
|---|---|---|---|---|
| Whole Exome GATK (average of 90) | 47,081 | 34,853 (74.03%) | 1,215 (3.49%) | 62 m |
| Whole Genome GATK (average of 5) | 900,196 | 792,193 (88.06%) | 13,579 (1.71%) | 109 m |
| dbSNP 137 | 5,335,713 | 4,093,464 (76.72%) | 157,882 (3.86%) | 124 m |

**Table 1:** Metrics generated by running the indelGeneralizer over variant list produced by GATK on exome and whole genome sequencing as well as over the entirety of dbSNP 137. Ambiguous indels are variants that could potentially be aligned differently, and duplicate indels are variants where two different but equivalent alignments are present in the same file.

## Experiments and Results

Ambiguous indels created an interesting challenge in a recent experiment which compared variants called from deep, targeted NGS sequencing of BCRA1 and 2 with clinical Sanger genotyping results for 96 patients. It was necessary to compare variants between the following formats:

**NGS VCF calls:**
- Produced by typical Novoalign[3]-GATK[2] pipeline;
- Ambiguous indels aligned to the most 5' position in the reference.

**Clinical HGVS calls:**
- Called by hand from Sanger traces;
- Ambiguous indels aligned to the most 3' position in the direction of transcription.

Initially 27 (44%) of the 61 pathological indels called in both sets were inconsistent and were reported as false (Figure 3). However, by leveraging the indelGeneralizer and the HGVS nomenclature tool Mutalyzer[1], we were able to resolve 24 (89%) of the inconsistent indels. Further effort was needed to show that the 3 remaining mutations were also consistent - they involved multiple events reported at the same position (e.g. a deletion of 3 bases and an insertion of 1 base in HGVS format is a SNP and a deletion of 2 bases in VCF format).

In another experiment, we examined the prevalence of ambiguous indels and tested the indelGeneralizer's performance on three larger datasets: unfiltered GATK variants from 90 exomes, unfiltered GATK variants from 5 whole genomes, and the entirety of dbSNP 137. Performance over these datasets (with the multiple GATK runs averaged) can be seen in Table 1. The script was run on a single core with 2 GB of available RAM (though the memory actually used was much smaller).

## Conclusion and Future Direction

When working with a list of indel variants, and especially when comparing between lists that were produced by different pipelines or follow different conventions, it is important to consider that ambiguous indels may have been reported at different or even multiple positions. In order to address this issue, we have developed the indelGeneralizer, a perl script capable of identifying ambiguous indel calls and converting them to a user defined format. Examination of single variants or batch processing of VCF files are possible with the script, which is available as open source software.

A current limitation of the tool is that it only considers one variant at a time – ambiguous local realignments that involve multiple SNPs and indels are also possible, and it would be ideal if conventions and logic for generalizing these could be incorporated into the script. Additional code optimization and tuning could potentially increase performance over both large batch and single indel jobs. Extending the script into a perl module would be beneficial as it would allow developers to easily incorporate the functionality into their own projects.

1. Wildeman M et al. Improving sequence variant descriptions in mutation databases and literature using the MUTALYZER sequence variation nomenclature checker. Hum. Mutat. 29:6-13 (2008)
2. McKenna A et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297-303 (2010)
3. www.novocraft.com

Ontario

Funding for the Ontario Institute for Cancer Research is provided by the Government of Ontario.