RESEARCH ARTICLE

Misannotation of multiple-nucleotide variants risks misdiagnosis

[version 1; peer review: awaiting peer review]

Matthew N. Wakeling 1,2, Thomas W. Laver 1,2, Kevin Colclough 2, Andrew Parish 2, Sian Ellard 1,2, Emma L. Baple 1,3

1Institute of Biomedical and Clinical Science, University of Exeter Medical School, Exeter, Devon, EX2 5DW, UK
2Molecular Genetics Department, Royal Devon and Exeter NHS Foundation Trust, Exeter, Devon, EX2 5DW, UK
3Clinical Genetics Department, Royal Devon and Exeter NHS Foundation Trust, Exeter, Devon, EX2 5DW, UK

* Equal contributors

Abstract
Multiple Nucleotide Variants (MNVs) are miscalled by the most widely utilised next generation sequencing analysis (NGS) pipelines, presenting the potential for missing diagnoses that would previously have been made by standard Sanger (dideoxy) sequencing. These variants, which should be treated as a single insertion-deletion mutation event, are commonly called as separate single nucleotide variants. This can result in misannotation, incorrect amino acid predictions and potentially false positive and false negative diagnostic results. This risk will be increased as confirmatory Sanger sequencing of Single Nucleotide variants (SNVs) ceases to be standard practice. Using simulated data and re-analysis of sequencing data from a diagnostic targeted gene panel, we demonstrate that the widely adopted pipeline, GATK best practices, results in miscalling of MNVs and that alternative tools can call these variants correctly. The adoption of calling methods that annotate MNVs correctly would present a solution for individual laboratories, however GATK best practices are the basis for important public resources such as the gnomAD database. We suggest integrating a solution into these guidelines would be the optimal approach.

Keywords
multi nucleotide variants, GnomAD, GATK, variant calling, next generation sequencing, genetic testing
Corresponding authors: Matthew N. Wakeling (M.Wakeling@exeter.ac.uk), Emma L. Baple (e.baple@exeter.ac.uk)

Author roles: Wakeling MN: Formal Analysis, Writing – Review & Editing; Laver TW: Visualization, Writing – Original Draft Preparation; Colclough K: Investigation; Parish A: Data Curation; Ellard S: Funding Acquisition, Supervision; Baple EL: Conceptualization, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome [098395 to SE]; and the Newlife Foundation for Disabled Children [SG/16-17/02 to EB].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2019 Wakeling MN et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Wakeling MN, Laver TW, Colclough K et al. Misannotation of multiple-nucleotide variants risks misdiagnosis [version 1; peer review: awaiting peer review] Wellcome Open Research 2019, 4:145 (https://doi.org/10.12688/wellcomeopenres.15420.1)

First published: 01 Oct 2019, 4:145 (https://doi.org/10.12688/wellcomeopenres.15420.1)
Introduction

The rapid progress and reduced cost of Next Generation Sequencing (NGS) has transformed approaches to genomic research and clinical diagnostic testing. While single-gene tests, for instance using Sanger (dideoxy) sequencing, will produce a short list of variants which can be manually evaluated, this is not feasible for next generation analysis. Sequencing at this scale requires highly automated analysis pipelines. High throughput sequencing services are dependent on automated tools to annotate and classify variants by potential consequence. For this reason, it is particularly important that any tools used to call and annotate variants do so accurately without the need for any manual assessment to avoid potential misdiagnosis.

Multiple Nucleotide Variants (MNVs) present a particular challenge for automated NGS analysis pipelines. These variants consist of multiple Single Nucleotide Variants (SNVs) located very close together on the same strand of DNA. The Human Genome Variation Society (HGVS) guidelines state that in most circumstances, two adjacent substitutions should be classified as a single deletion-insertion mutation event, rather than two or more separate SNVs.

MNVs that contain multiple SNVs within the same codon may have a significantly different protein consequence than if the separate SNVs are annotated independently. For instance, a CTG codon (Leu) can be changed to TTG or CTC (two separate SNVs) without any protein coding consequence, but when changed to TTC (an MNV) the consequence is a missense (see Figure 1). Importantly, some MNVs would meet the evidence criteria for pathogenicity when called as a single mutational event, but would not when each SNV is treated separately. NGS pipelines that annotate these MNVs as two independent SNVs could fail to correctly identify a pathogenic variant, potentially negatively impacting on clinical care.

Most standard NGS variant calling pipelines, including the widely adopted GATK best practices, do not deal with MNVs correctly - calling them as separate SNVs. Consequently, most laboratories using NGS technologies are at risk of miscalling these variants. Some NGS variant callers incorporate haplotype information to correctly call MNVs. Another approach to correctly call MNVs is to re-process variant calls, for example using the Multi-Nucleotide Variant Annotation Corrector (MAC). There is also a GATK tool, ReadBackedPhasing, which performs phasing of SNVs based on the overlap between reads and uses this information to call variants. However, this tool is not part of the current versions of the widely followed GATK best practice guidelines.

Currently most of the accredited international genomic diagnostic laboratories confirm potentially pathogenic variants by Sanger sequencing, which will correctly identify MNVs that have been incorrectly annotated as separate missense changes. However, many diagnostic facilities are discontinuing this confirmatory analysis now that a low false positive rate for SNV calls has been established. Even if Sanger confirmation is performed, this can only correct false positives and not false negatives, as variants predicted (incorrectly or correctly) to be synonymous are typically not subjected to any further scrutiny.

The scale of the potential problem with MNVs was highlighted by the ExAC database. The variants within this data set were called using a GATK best practices pipeline which does not recognise MNVs as single mutation events. Lek et al. identified an average of 23 MNVs that were incorrectly annotated by the
Generating simulated MNVs
To determine whether the pipeline correctly annotates MNVs, we generated a BAM file containing five simulated MNVs in the HNF4A gene. These MNVs are detailed in Table 1. Each variant was generated as a homozygous call (GT 1/1 with no reads supporting the reference allele). Four variants were two base pair MNV events, and one was a three base pair MNV event. We processed these variants with the standard GATK best practices pipeline described above.

This dataset is publicly available at https://github.com/rdemolgen/MNV-test-data to provide a simple method for laboratories to test if their current analysis pipeline will annotate MNVs correctly.

Re-processing with alternative tools
To investigate whether using alternative tools results in correct annotation of MNVs, we re-processed the VCF file of simulated MNVs using GATK ReadBackedPhasing10, VarDict7, Platypus14 or MAC7. Secondly, we re-analysed a cohort of 1447 samples previously tested using a targeted panel of genes for diagnosis of monogenic diabetes and congenital hyperinsulinism15 to determine if any potential diagnoses were missed.

By simulating MNVs in NGS sequencing data and testing for them using a typical NGS pipeline employed by an NHS diagnostic laboratory, we demonstrate that MNVs are incorrectly annotated by standard diagnostic NGS pipelines, potentially generating false positive and false negative results and negatively impacting on patient care.

**Methods**

GATK best practices pipeline

The Molecular Genetics Laboratory at the Royal Devon & Exeter NHS Foundation Trust routinely uses a targeted NGS testing pipeline to interrogate an extended panel of genes associated with monogenic diabetes and congenital hyperinsulinism15. This pipeline aligns reads to the hg19/GRCh37 human reference genome with BWA mem16, applies Picard for duplicates removal17 and GATK IndelRealigner for local re-alignment.18 GATK haplotypecaller is then used to identify variants and these are annotated using Alamut batch version 1.5.2 (Interactive Biosoftware, Rouen, France). This analysis approach is based on the GATK best practice guidelines5.

<table>
<thead>
<tr>
<th>Variant Number</th>
<th>Genome position (GRCh37)</th>
<th>Nucleotide position</th>
<th>Codon position</th>
<th>Wild-type codon</th>
<th>Variant codon</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20:43052669_43052671</td>
<td>NM_175914:c.838_840</td>
<td>p.Leu280</td>
<td>CTG</td>
<td>TTC</td>
</tr>
<tr>
<td>2</td>
<td>20:43053017_43053019</td>
<td>NM_001030004:c.1186_1188</td>
<td>p.*396</td>
<td>TAA</td>
<td>TGG</td>
</tr>
<tr>
<td>3</td>
<td>20:43056977_43056979</td>
<td>NM_175914:c.1066_1068</td>
<td>p.Ser356</td>
<td>TCC</td>
<td>AGC</td>
</tr>
<tr>
<td>4</td>
<td>20:43058207_43058209</td>
<td>NM_175914:c.1261_1263</td>
<td>p.Ser421</td>
<td>TCT</td>
<td>TGA</td>
</tr>
<tr>
<td>5</td>
<td>20:43058219_43058221</td>
<td>NM_175914:c.1273_1275</td>
<td>p.Lys425</td>
<td>AAG</td>
<td>AGT</td>
</tr>
</tbody>
</table>

**Table 1. Simulated Multiple Nucleotide Variants within the HNF4A gene.** Variants are described according to Human Genome Variation Society sequence variation nomenclature guidelines15.
Table 2). Variant 1 was incorrectly called as two separate synonymous variants (p.Leu280Leu), whereas the correct annotation is an in-frame deletion-insertion that results in the missense variant p.Leu280Phe. If used diagnostically this would result in a false negative result. Variant 2 alters a stop codon – when the MNV is treated correctly this results in a stop loss, however when each SNV is called separately the original stop codon is maintained presenting the potential for a false negative result. Variant 3 should result in annotation of a synonymous variant when correctly called, however GATK best practices incorrectly recognises this as two separate missense variants (p.Ser356Thr and p.Ser356Cys), which could result in a false positive testing result. When treated correctly, variant 4 should create a stop codon resulting in a nonsense variant, however it is inaccurately annotated as two variants, a missense (p.Ser421Cys) and a synonymous variant (p.Ser421Ser). Variant 5 is called as p.Lys425Arg and p.Lys425Asn, whereas it should be called as a different missense variant, p.Lys425Ser. This could result in either a false positive or a false negative result depending on the clinical interpretation of the missense variants.

Simulated MNVs were correctly called using alternative software

As described above, when our simulated MNVs are called using GATK best practices they are incorrectly called as two separate variants. In contrast when re-analysed using GATK ReadBackedPhasing\textsuperscript{10}, MAC\textsuperscript{9} and Platypus\textsuperscript{14} the separate SNVs are correctly merged into a single MNV in all five cases and the MNVs were correctly annotated by Alamut batch 1.5.2 as in-frame insertion-deletions. VarDict\textsuperscript{7} correctly calls four variants but fails to call variant 1, which is a CTG to TTC non-consecutive change, as a single event.

Variants identified through an NGS diagnostic targeted panel are miscalled by GATK best practices

The Molecular Genetics Laboratory at the Royal Devon & Exeter NHS Foundation Trust utilises an NGS analysis pipeline based on GATK best practices. Having established, using simulated data that GATK ReadBackedPhasing\textsuperscript{10} correctly called MNVs, we re-analysed 1447 samples tested on a diagnostic panel for monogenic diabetes and congenital hyperinsulinism\textsuperscript{15} to examine if any MNVs had been incorrectly annotated using the GATK best practices pipeline.

On four occasions MNVs were found to have been miscalled as two separate single base substitution variants (Table 3). In three cases the correct annotation for the MNV was a missense variant, however GATK best practices resulted in two different missense variants being called. The fourth MNV should also have been called as a missense variant, but was called as a nonsense variant and a different missense variant. In all four cases the variants were confirmed by Sanger sequencing prior to reporting thus the correct diagnosis was made. In the absence of Sanger sequencing confirmation these incorrect annotations have the

### Table 2. Simulated Multiple Nucleotide Variants within the HNF4A gene as annotated by GATK best practices.

<table>
<thead>
<tr>
<th>Variant</th>
<th>Wild-type codon</th>
<th>Variant codon</th>
<th>GATK best practices annotation 1</th>
<th>GATK best practices annotation 2</th>
<th>Correct annotation</th>
<th>Likely implication for diagnostic testing\‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CTG</td>
<td>TTC</td>
<td>c.838C&gt;T p.Leu280Leu</td>
<td>c.840G&gt;C p.Leu280Leu</td>
<td>c.838_840delinsTTC p.Leu280Phe</td>
<td>False negative result</td>
</tr>
<tr>
<td>2</td>
<td>TAA</td>
<td>TGG</td>
<td>p.<em>396</em></td>
<td>p.<em>396</em></td>
<td>p.<em>396Trpext</em>26</td>
<td>False negative result</td>
</tr>
<tr>
<td>5</td>
<td>AAG</td>
<td>AGT</td>
<td>c.1274A&gt;G p.Lys425Arg</td>
<td>c.1275G&gt;T p.Lys425Asn</td>
<td>c.1274_1275delinsGT p.Lys425Ser</td>
<td>False positive or negative result</td>
</tr>
</tbody>
</table>

\‡Based on testing for dominant acting heterozygous, pathogenic loss of function variants.

### Table 3. Multiple Nucleotide Variants found in the re-analysed data from the diagnostic panel to be incorrectly annotated as separate variants.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Wild-type codon</th>
<th>Variant codon</th>
<th>GATK best practices annotation 1</th>
<th>GATK best practices annotation 2</th>
<th>Correct annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>INSR</td>
<td>GCC</td>
<td>TTC</td>
<td>p.Ala752Val</td>
<td>p.Ala752Ser</td>
<td>p.Ala752Phe</td>
</tr>
<tr>
<td>EIF2AK3</td>
<td>GAT</td>
<td>TCT</td>
<td>p.Asp615Ala</td>
<td>p.Asp615Tyr</td>
<td>p.Asp615Ser</td>
</tr>
<tr>
<td>GCK</td>
<td>TAC</td>
<td>CAA</td>
<td>p.Tyr61*</td>
<td>p.Tyr61His</td>
<td>p.Tyr61Gln</td>
</tr>
</tbody>
</table>
potential to result in false positive or false negative results depending on the clinical interpretation of the missense variants.

Discussion

Using simulated MNVs and re-analysing data from a diagnostic NGS targeted gene panel test we have demonstrated that the current approach employed by most NGS variant pipelines, including GATK best practices, can result in MNVs being mis-called. There are important implications to this inaccuracy. In a diagnostic setting this could result in false positive or negative test results for patients with potentially serious consequences to an individual’s healthcare outcome. The significant number of previously published MNVs within known human disease genes that are listed within the HGMDPro database provides some context to the potential scale of this problem.

The GATK best practice guidelines4 have been widely adopted and are employed in the analysis pipelines for the majority of diagnostic and research NGS facilities worldwide. Our analysis pipeline, based on GATK best practices, which is currently in use at our diagnostic laboratory, failed to correctly call our simulated MNVs and four MNVs identified by reanalysis of targeted gene panel data.

Our analysis demonstrated that in contrast to GATK best practices, alternative tools are available which merge the nearby SNVs correctly into a single MNV, which is essential for correct annotation of variant consequence. Thus a potential solution for individual laboratories to resolve this issue would be the integration of other tools within their NGS pipelines that deal with MNVs correctly. However, this depends on laboratory awareness of this ongoing problem and the potential for patient harm that it presents.

In the current version of the GATK best practices, phasing is performed by GATK HaplotypeCaller, so the ReadBackedPhasing software, which previously performed this role, is no longer being actively maintained. However, while HaplotypeCaller builds haplotypes we have demonstrated that it does not correctly utilise the information to call MNVs. ReadBackedPhasing calls MNVs but does not provide the quality score information for them that is produced for variants by HaplotypeCaller, which prevents them from being filtered by quality. Thus we suggest that the ideal solution would be for the features of software which enable correct calling of MNVs, namely the appropriate use of haplotype information, to be incorporated into HaplotypeCaller.

Adoption of a solution into the GATK best practices is the optimal solution as it does not require individual laboratories to be aware of the problem and adopt bespoke solutions. GATK is widely adopted for its ease of use: it provides an integrated suite of tools with inputs and outputs in standard formats, it has excellent documentation and a large user community solving shared problems.

Another important consideration to note is that publicly available online variant frequency resources such as gnomAD and ExAC are currently based on GATK best practices pipelines. These resources are critical to variant interpretation in rare genetic disorders as a key criterion for pathogenicity assignment is allele frequency. Currently MNVs are flagged, but still represented as multiple separate SNVs within gnomAD and ExAC. This means that even where laboratories make changes to their local pipeline to correctly call MNVs, their local data for these variants will be incompatible with these public resources, with allele frequency information being unavailable for those MNVs.

In summary, the issue of MNVs being miscalled by the most commonly employed NGS analysis pipelines continues to be an important issue. We anticipate that this issue will be compounded by the change in practice as laboratories cease routine Sanger sequencing to confirm SNVs. Although there are a number of tools available that call MNVs correctly, these are not currently being widely adopted.

Addressing this issue by implementing changes within GATK best practices would have the greatest impact on prevention of misdiagnoses resulting from MNV calling inaccuracies and also importantly provide compatibility with the online public variant frequency databases that are central to current diagnostic variant classification.

Key points

- Multi-nucleotide variants (MNVs) are misannotated by the most widely used next generation sequencing analysis pipelines
- Misannotation of MNVs can result in genetic misdiagnosis
- We suggest that individual laboratories should consider implementing alternative software to avoid misannotation of these variants
- The test data described in this manuscript has been made publicly available at https://github.com/rdemolgen/MNV-test-data so that laboratories can verify if their analysis pipeline correctly annotates multi-nucleotide variants
- We suggest that GATK best practices pipeline should implement a solution for MNV misannotation to ensure widespread adoption

Data availability

Underlying data

Simulated MNV data is available at https://github.com/rdemolgen/MNV-test-data

Archived simulated MNV data at time of publication: http://doi.org/10.5281/zenodo.3375579

License: GNU General Public License v3.0

The dataset of 1447 samples previously sequenced cannot be shared due to patient confidentiality issues, as the genotype data
could be used to identify individuals and so cannot be made openly available. Requests for access to the anonymised data by researchers will be considered following an application to the Genetic Beta Cell Research Bank (https://www.diabetesgenes.org/current-research/genetic-beta-cell-research-bank/) with proposals reviewed by the Genetic Data Access Committee.

References