Genetics of Hirschsprung disease
Schriemer, Duco

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2016

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 17-02-2020
Removing the maximum assigned value of the genotype quality score allows specific detection of de novo mutations in next-generation sequencing data

Duco Schriemer¹, Hongsheng Gui²³, Rutger W.W. Brouwer⁴, Erwin Brosens⁵, Clara S.M. Tang⁶, Christian Gilissen⁶, Wilfred F.J. van Ijcken⁴, Salud Borrego⁷⁸, Isabella Ceccherini⁹, Aravinda Chakravarti¹⁰, Stanislas Lyonnet¹¹, Paul K. Tam², Maria-Mercè Garcia-Barceló², Bart J.L. Eggen¹, Robert M.W. Hofstra⁵¹³

¹ Department of Neuroscience, section Medical Physiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands
² Department of Surgery, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, SAR, China
³ Centre for Genomic Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, SAR, China
⁴ Biomics Erasmus Center for Biomics, Erasmus Medical Center, Rotterdam, The Netherlands
⁵ Department of Clinical Genetics, Erasmus Medical Center, Rotterdam, the Netherlands
⁶ Department of Human Genetics, Donders Centre for Neuroscience, Radboud University Medical Center, The Netherlands
⁷ Unidad de Gestión Clínica de Genética, Reproducción y Medicina Fetal Hospitales, Universitarios Virgen del Rocío, Seville, Spain
⁸ CIBER de Enfermedades Raras (CIBERER), ISCIII, Seville, Spain
⁹ UOC Genetica Medica, Istituto Gaslini, Genova, Italy
¹⁰ McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, USA
¹¹ Département de Génétique, Faculté de Médecine, Université Paris Descartes, Paris, France
¹² INSERM U-781, AP-HP Hôpital Necker-Enfants Malades, Paris, France
¹³ Stem Cells and Regenerative Medicine, Birth Defects Research Centre UCL Institute of Child Health, London, UK

Manuscript in preparation
ABSTRACT

De novo mutations (DNMs) play a major role in the development of rare genetic syndromes and more common neurodevelopmental disorders. Next-generation sequencing (NGS) allows for the unbiased analysis of DNMs in the human genome or exome. Detecting DNMs in NGS data is challenging, as the number of sequencing artefacts is far greater than the number of real DNMs. Thresholds for reference allele ratio, sequencing depth, and Genotype Quality have been shown to effectively reduce the number of sequencing artefacts. In this study, we used a training set of 22 confirmed DNMs and 59 false positives to test how thresholds settings for these sequencing parameters affect the sensitivity and specificity of detecting DNMs. We found that removing the maximum assigned value to the Genotype Quality score allows to distinguish confirmed DNMs from false positives. Sequencing depth and reference allele frequency also had good and moderate discriminative power, respectively. By combining thresholds for all parameters, we were able to detect DNMs with 100% sensitivity and 98.3% specificity in our training dataset. Validation of the filtering method successfully detected all 33 confirmed and four novel DNMs in an independent dataset and produced 1 false positive. The findings of this study provide a guideline for the efficient detection of DNMs in NGS data.
INTRODUCTION

A de novo mutation (DNM) is an alteration in the genome that is present for the first time in a person as a result of a mutation in a germ cell (egg or sperm) of one of the parents or in the fertilized egg itself. Next-generation sequencing (NGS) has allowed the genome- and exome-wide detection of DNMs. Exome sequencing has revealed a role for DNMs in rare genetic syndromes and more common neurodevelopmental disorders\textsuperscript{1–13}. These studies found an average of 1 to 2 exonic DNMs per proband. However, technical artefacts in NGS data poses a challenge for the identification of DNMs. Due to the large number of genotyped base pairs in exome sequencing, the number of incorrectly called genotypes is several orders of magnitude larger than the actual number of DNMs. Currently, DNM detection methods are not fully accurate and all methods still require confirmation by validation techniques such as Sanger sequencing to discriminate between real DNMs and false positives. As cohort sizes in sequencing studies are expanding and exome sequencing is being replaced by whole-genome sequencing, the number of candidate DNMs that require experimental validation can reach thousands\textsuperscript{14–16}. It is therefore important to detect DNMs effectively and efficiently.

It has been proposed that the accuracy of genotype calling can be improved by jointly calling the genotypes of multiple individuals\textsuperscript{17,18}. Multi-sample genotype calling uses the observed alleles in other sequenced individuals to calculate the prior probability of finding a given genotype. For DNMs, which are typically found in only a single individual, the prior probability of finding a variant is low. This means that strong evidence is required to reliable call a variant. Joint genotype calling of multiple individuals has the additional benefit of annotating reference allele homozygous genotypes in the variant file at positions where a variant has been found in another individual. Since DNMs are absent in parents, it is important to be able to discriminate between reference allele homozygous genotypes and missing genotypes to efficiently detect DNMs.

When applied to parents-proband trios, joint genotype calling can improve the detection of DNMs\textsuperscript{19}. Several software packages were developed to detect DNMs in NGS data that all make use of trio-based joint genotype calling\textsuperscript{20–24}. Each software package has its additional methodology to specifically detect DNMs. PolyMutt and DeNovoGear use genomic mutation rates to calculate the likelihood of finding a DNM at a specific position\textsuperscript{20,21}. However, mutation rates vary between genomic regions and TrioDeNovo was developed to take region-specific mutation
races into account\textsuperscript{22}. DNMFilter uses machine learning to discriminate between DNMs and false positives based on several sequencing parameters\textsuperscript{23} and mirTrios uses sequencing parameters in a probabilistic model to detect DNMs\textsuperscript{24}. Three sequencing parameters were shown to efficiently eliminate incorrect genotype calls, while retaining most of the true variants: the reference allele ratio, sequencing depth and Genotype Quality score\textsuperscript{25,26}. In the present study, we aimed to assess the relative contribution of these parameters to DNM detection and analyzed how filtering thresholds for these parameters, in probands and parents, affects the sensitivity and specificity of detecting DNMs.

**METHODS**

**Exome sequencing**
DNA from 20 parents-proband trios was subjected to exome sequencing in three different centers (Table 1). Each center performed exome capture and NGS according to in house protocols. In short, the exomes of 5 trios were captured using the Illumina Truseq Exome Enrichment kit and sequenced on an Illumina GAIIx machine. The other two centers (5 and 10 trios, respectively) both used the Agilent SureSelect Exome Capture kit V4 and an Illumina HiSeq2000 sequencer. Sequencing reads from all 20 trios were aligned to the human reference genome hg19 (build 37) using standard BWA alignment\textsuperscript{27}. Multi-sample genotype calling was performed on all individuals simultaneously using the Genome Analysis Toolkit (GATK) Unified Genotyper 2.0\textsuperscript{28}.

<table>
<thead>
<tr>
<th>Center</th>
<th>Trios</th>
<th>Capture array</th>
<th>Sequencer</th>
<th>Mean depth</th>
<th>10X coverage</th>
<th>Validated candidates</th>
<th>Confirmed DNMs</th>
<th>Validation rate</th>
<th>DNMs/trio</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>Illumina Truseq</td>
<td>Illumina GAII</td>
<td>26.8 X</td>
<td>74%</td>
<td>27</td>
<td>1</td>
<td>3.7%</td>
<td>0.2</td>
</tr>
<tr>
<td>B</td>
<td>5</td>
<td>Agilent SureSelect V4</td>
<td>Illumina HiSeq 2000</td>
<td>41.9 X</td>
<td>89%</td>
<td>24</td>
<td>8</td>
<td>33.3%</td>
<td>1.6</td>
</tr>
<tr>
<td>C</td>
<td>10</td>
<td>Agilent SureSelect V4</td>
<td>Illumina HiSeq 2000</td>
<td>55.7 X</td>
<td>95%</td>
<td>32</td>
<td>15</td>
<td>46.9%</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Parents-proband trios were sequenced in three different centers. Sequencing depth differed per center and affected the number of confirmed DNMs and validation success accordingly.
Threshold setting to identify candidate DNMs
DNMs were identified as heterozygous variants in probands that were absent in both parents. The Mendelian Disease Arguments function of the KGGSeq software\(^{29}\) was used to extract DNMs from the multi-sample VCF file. Variants were selected to have: 1) call quality \(\geq 50\); 2) mapping quality \(\geq 30\); 3) read depth \(\geq 5\) in probands and parents; 4) Genotype Quality score \(\geq 10\) in probands and parents; 5) reference allele ratio \(\geq 0.10\) in heterozygous probands; 6) reference allele ratio \(\leq 0.10\) in reference allele homozygous parents; and 7) are in coding regions of the genome (Figure 1).

These low thresholds settings preserved most real variants in the data, yet removed many low quality variants\(^{25}\). To further reduce the number of DNM candidates, variants with a minor allele frequency above 1% in dbSNP137, 1000Genomes 201204, ESP6500AA and ESP6500EA were excluded.

Visual inspection and further selection of DNMs
Candidate DNMs after low threshold filtering were visually inspected in the Integrative Genomics Viewer (IGV)\(^{30}\) to select the most likely DNMs. Candidate DNMs were discarded if 1) all alternative allele reads were from the same strand, 2) more than two alleles were observed, 3) the proband showed the alternative allele in <20% of the sequencing reads and a parent in >5% of the sequencing reads, 4) only a single sequencing read supported the alternative allele, 5) \(\geq 50\)% of the sequencing reads had a base PHRED quality score <10.

Validation of DNMs
Candidate DNMs were Sanger sequenced in the proband and the parents by two individual sequencing runs in opposite directions.

Comparison between confirmed DNMs and false positives
The reference allele ratio, sequencing depth and (unmaximized) Genotype Quality score of confirmed DNMs and false positives were entered and plotted in GraphPad Prism 5. Parameters values were compared between confirmed and false positive DNMs and between probands and parents by Mann-Whitney test. GraphPad Prism 5 was subsequently used to generate receiver operating characteristic (ROC) curves and calculate the area under the curve (AUC).
Figure 1. DNM filtering pipeline. DNMs were identified as heterozygous variants in probands that were absent in both parents. Low thresholds for sequencing parameters removed low quality DNM candidates. The most likely DNMs were selected for Sanger sequencing validation by visual inspection in IGV. Based on the characteristics of confirmed DNMs, candidates after initial filtering were re-assessed to test if DNMs had been missed.
DNM validation cohort

A second, independent exome sequencing dataset of 25 parents-proband trios was generated by the Beijing Genomics Institute (BGI). Exomes were captured using Agilent SureSelect Exome Capture kit V4, NGS was performed on an Illumina HiSeq2000 and variants were called for each sample individually by the GATK Unified Genotyper 2.0. DNMs were selected as describe for the training cohort, with slightly more stringent thresholds for three parameters; 1) read depth ≥15 in probands, 2) reference allele ratio ≥0.35 in probands, 3) Genotype Quality score = 99 in probands.

RESULTS

Identification of DNMs

Exome sequencing data from 20 parents-proband trios, generated by three sequencing centers using two sequencing technologies, were used to screen for DNMs (Table 1). Candidate DNMs were identified by filtering at low thresholds for quality parameters and the 91 most likely DNMs (4.55 ± 3.35 per trio) were selected for Sanger sequencing validation (Figure 1). Eight DNM candidates could not be amplified successfully by PCR. Of the remaining 83 candidate DNMs, 24 were confirmed as DNMs and 59 candidates proved to be false positives.

The set of 91 candidate DNMs was obtained by variant filtering and selection of the best candidates in IGV. The selection of candidates in IGV was performed manually and was partially subjective. Therefore, all DNM candidates after initial filtering were filtered again, using more stringent threshold settings that were set after evaluation of the confirmed and false positive DNMs (Figure 1). This last, more stringent analysis identified three additional candidate DNMs that had been missed in the first selection in IGV. Sanger sequencing validation showed that all three additional candidate DNMs were false positives.

Mosaic DNMs

Within the confirmed group of DNMs there were two outliers. In these two the reference allele ratios observed were 0.80 and 0.76, respectively. In line with the skewed distribution of reference and alternative alleles in the NGS data, Sanger sequencing validation of these two mutations showed a higher peak for the reference allele than for the alternative allele. This finding suggests that these
DNMs were present in only part of the DNA sample and are somatic, rather than germline DNMs. Since our focus was on the identification of germline DNMs, the two somatic DNMs were excluded from the comparison between confirmed DNMs and false positive DNM candidates.

**Characteristics of confirmed DNMs are different from false positives**

To assess what the discriminating features of true DNMs are in NGS data, reference allele ratio, sequencing depth and Genotype Quality score were compared between the 22 confirmed germline DNMs and 59 false positives.

**Reference allele ratio**

In probands, the average reference allele ratio did not differ between confirmed DNMs and false positives (p=0.0585), but confirmed DNMs fell within a smaller range (0.44-0.70) than false positives (0.20-0.86) (Figure 2A). In the homozygous reference parents, base calls for the alternative allele are not expected, but were nevertheless identified in a total of 5 sequencing reads in three different parents. All 5 alternative allele calls had base call quality scores below 10 and were therefore likely to be technical artefacts. In contrast, most of the reads underlying the DNM had a high quality (PHRED>20). If the low-quality base calls do not represent artefacts, then there likely is a low-level mosaicism at that position in the parent. Low-level mosaicisms of a few percent cannot be detected by Sanger sequencing, but can be identified by deep-sequencing NGS. Sequencing reads supporting the alternative allele were observed in 26.3% of the false positive DNM candidates and with higher base call quality scores. The presence of (low-quality) alternate reads in the parents therefore suggests that a DNM candidate is likely, but not necessarily, a false positive.

**Sequencing Depth**

The average sequencing depth of confirmed DNMs was higher than that of false positives, in both probands and their parents (p=1.87E-7 and p=1.30E-10, respectively) (Figure 2B). In probands, the lowest observed sequencing depth of confirmed DNMs was 17, while 53% of false positives had a lower sequencing depth (Figure 2B). With a minimal of 9 reads, parents showed a lower minimal sequencing depth for confirmed DNMs than probands, although the average sequencing depth did not differ between probands and parents (p=0.300).
Removing the max of the GQ score allows specific detection of DNMs in NGS data.

Figure 2. Characteristics of confirmed DNMs and false positives in probands and parents. Reference allele ratio, sequencing depth and second smallest genotype likelihood were plotted for all individual data points. Lines represent mean values. A) Reference allele ratio in probands ranges from 0.44 to 0.70, whereas this range was wider for false positives. Although many false positives showed reads with the alternative allele in parents, these were found at low frequency in confirmed DNMs as well. B) Sequencing depth of confirmed DNMs was higher than that of false positives, in both probands and parents. The minimal depth of confirmed DNMs was 17 in probands and 9 in parents. C) The second smallest PHRED-scaled genotype likelihood (ssPL) represents the unmaximized Genotype Quality score. With all values over 200, ssPL values were remarkably high for confirmed DNMs in probands.
Genotype Quality score

A genotype call is produced for each covered position in NGS data, which can be either AA, AB or BB. For each of these three genotypes, a likelihood score is calculated that indicates how likely the genotype is to be true, considering the number of reads and their qualities supporting the A and B alleles. The most likely one of the three genotypes is selected as the genotype at that position. The likelihood scores for the three possible genotypes are defined as a PHRED score ($-10\log_{10}[p\text{-value}]$), and is referred to as the PHRED-scaled Genotype Likelihood (PL). The PHRED-scaled Genotype likelihoods (PL) scores are normalized such that the most likely genotype receives a score of 0. The Genotype Quality (GQ) score (not to be confused with the Genotype Likelihood (PL) score) is defined as the second smallest of the three Genotype Likelihood (PL) scores (the smallest always being 0). So if the two alternative genotypes are very unlikely, they will have high Genotype Likelihood (PL) scores and hence the Genotype Quality (GQ) score will also be high. The maximum value that is assigned to the Genotype Quality (GQ) score is 99, meaning that all values above 99 will be set to this maximum.

The Genotype Quality (GQ) score is a commonly used quality parameter in NGS data and was therefore used to compare confirmed DNMs and false positives. All confirmed DNMs and 35 of 59 (59.3%) false positives had a Genotype Quality (GQ) score of 99 in probands, suggesting that confirmed and false positive DNMs had very similar Genotype Quality (GQ) scores. However, the Genotype Likelihoods (PL) of the other two possible genotypes are specified in a VCF file as well, and no maximum value is assigned to these scores. The second smallest PHRED-scaled Genotype Likelihood (ssPL) can therefore be regarded as the unmaximized Genotype Quality (GQ) score. These ssPL scores were significantly higher in confirmed DNMs than in false positive DNM candidates, in probands ($p=3.85\times10^{-14}$) and in parents ($p=2.61\times10^{-12}$) (Figure 2C). All ssPL scores in probands were above 200, whereas 84.7% of all false positives had ssPL values below 200. The ssPL score provides a good distinction between true and false positive DNMs. The high minimum ssPL values for confirmed DNMs were not observed in parents, where the lowest ssPL value was 27.

Discriminative power of different parameters

Having compared the characteristics of confirmed DNMs and false positives, we analyzed how the thresholds for the reference allele ratio, the sequencing depth and the ssPL influenced the sensitivity and specificity of detecting DNMs. DNMs
Removing the max of the GQ score allows specific detection of DNMs in NGS data

could not be accurately detected based on thresholds for reference allele ratio, neither in probands nor in parents (Figure 3A,B). A sequencing depth threshold around 30 allowed the detection of DNMs with both a sensitivity and specificity of ~75%, in both probands and parents (Figure 3C,D). An ssPL threshold of 200 in probands detected DNMs with 100% sensitivity and 84.7% specificity (Figure 3E). In comparison, selection of DNMs based on a maximum Genotype Quality score of 99 identified DNMs with a specificity of only 40.7%. In parents, 79.2% sensitivity and 88.0% specificity can be obtained by a ssPL threshold of 80 (Figure 3F).

Receiver Operating Characteristic (ROC) curves were generated to determine the discriminative power of the different sequencing quality parameters. The ssPL in probands was the best predictor for DNMs, with an area under the curve (AUC) of 0.9607 (Figure 4). ssPL in parents and sequencing depth in probands and parents were to a lesser extent able to discriminate DNMs from false positives (AUC = 0.8321, 0.8463 and 0.8077, respectively). The reference allele ratio had low predictive power. The AUC in probands was 0.6140, hardly better than a variable that has no predictive power (AUC of 0.5). The specificity of detecting DNMs on the basis of reference allele ratio in parents never reached above 26.3% and the corresponding AUC is 0.6052. By combining thresholds that allow maximum sensitivity (100%), only 1 false positive was observed (98.3% specificity). Six false positives were found (89.8% specificity) when respecting the maximum Genotype Quality score of 99.

**Validation on independent data set**
The characteristics of confirmed DNMs and false positives were based on NGS data from three different centers, using two different technologies. The parameters described thus reflect the characteristics of DNM across multiple datasets. The applicability on independent NGS datasets was further evaluated using a dataset with 33 confirmed DNMs. Using conservative filtering thresholds (see Methods), all confirmed DNMs were successfully detected with only 1 false positive. Moreover, four additional DNM candidates were identified that had been missed in the initial DNM screening of the confirmation set. These DNM candidates had been ignored because they showed the alternative allele in 1.2-3.0% of the sequencing reads in the parents. However, all four candidates were confirmed to be DNMs by Sanger sequencing validation. A total of 9 reads with the alternative allele were observed in these four DNMs. Three had a low base call quality (≤9), whereas the base call
Figure 3. Effect of filtering thresholds on sensitivity and specificity of DNM detection. The sensitivity and specificity with which DNMs were detected are plotted against filtering thresholds for the different parameters. The threshold for reference allele ratio in probands is a maximum value threshold, whereas other thresholds are minimum value thresholds.
Removing the max of the GQ score allows specific detection of DNMs in NGS data

quality for the other six was high (≥29). It is unclear whether these represent technical artefacts or parental mosaicism that was too low to detect by the Sanger sequencing.

DISCUSSION

The specific detection of DNMs is difficult due to the relatively high number of incorrect genotype calls in NGS data. Joint genotype calling in trios reduces the number of potential DNM candidates and several software packages have been developed that use probabilistic modelling to further improve the specificity of DNM detection\textsuperscript{20–24}. Reference allele ratio, sequencing depth and Genotype Quality have been shown to efficiently filter Mendelian errors in NGS data\textsuperscript{25,26}. Here we assessed if and how these parameters can discriminate between confirmed DNMs and false positives. To this purpose, we used a set of 22 confirmed germline DNMs and 59 false positives that were initially deemed plausible DNMs.

Using threshold settings for reference allele ratio, sequencing depth and Genotype Quality, we were able to detect DNMs with 100% sensitivity and 98.3% specificity. The key to reaching high specificity was to remove the maximum assigned value to the Genotype Quality score. By using the second smallest PHRED-scaled genotype likelihood instead of the (maximized) Genotype Quality score, the specificity of detecting DNMs using this single parameter increased from 40.7% to 84.7% (at 100% sensitivity). This raises the question why the Genotype Quality score has a maximum value? First of all, it requires more space to save values of more than 2 digits, especially considering that thousands of Genotype Quality scores are saved in each VCF file. Secondly, a Genotype Quality score of 99 corresponds to a chance of 1.26E-10 that the genotype call is incorrect. Higher Genotype Quality scores are therefore not deemed more informative. Nonetheless, we found that all 22 confirmed DNMs had a second smallest genotype likelihood score of at least 200 in probands, which translates to a ≤1E-20 chance of making an incorrect genotype call. Since the majority of false positive DNMs had ssPL scores below 200, the discriminative power of the ssPL in probands was high, with an AUC of 0.9607. Also in parents, the ssPL proved a useful parameter to distinguish DNMs from false positives, just like sequencing depth in both probands and parents. The reference allele frequency in probands and parents had limited discriminative power.
In addition to the 22 germline DNMs, two DNMs were identified that had likely occurred somatically. These mutations showed an underrepresentation of the alternative allele in both the NGS data and the Sanger sequencing validation. This finding is in line with a study reporting that 6.5% of all DNMs are mosaic and have occurred somatically\(^\text{31}\). The accurate detection of somatic DNMs is important, as they are involved in cancer and other disorders\(^\text{32,33}\). However, mosaic DNMs are more difficult to detect than germline DNMs, especially if the frequency of the mutation is low.

Figure 4. Receiver Operating Characteristic (ROC) curves of the sequencing parameters. The Area Under the ROC Curve (AUC) was calculated as a measure of the discriminative power of a parameter. The ssPL value in probands was the most powerful parameter to distinguish between confirmed DNMs and false positives.
Removing the max of the GQ score allows specific detection of DNMs in NGS data

Existing software packages for the detection of DNMs in NGS data rely on joint genotype calling and probabilistic models to predict DNMs. Although different programs use different models, they reach similar sensitivity and specificity\(^2\). In the present study we show that filtering thresholds for the unmaximized Genotype Quality score, in combination with sequencing depth and reference allele frequency, provides an easy and straightforward filtering approach to detect DNMs with high sensitivity and specificity across exome sequencing data from different sequencing centers. We propose that incorporation of the ssPL score in probabilistic models may further enhance the specificity of detecting DNMs and urge the use of this parameter in future analyses. Whether removing the maximum value of the Genotype Quality score has benefit in the detection of other types of genetic variation is unclear, but deserves further investigation.
REFERENCES

Removing the max of the GQ score allows specific detection of DNMs in NGS data


