BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 13 2014, pages 1830–1836
doi:10.1093/bioinformatics/btu141

Sequence analysis

Advance Access publication March 10, 2014

A gradient-boosting approach for filtering *de novo* mutations in parent–offspring trios

Yongzhuang Liu[1,2], Bingshan Li[3], Renjie Tan[1,2], Xiaolin Zhu[2] and Yadong Wang[1,*]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China, [2]Center for Human Genome Variation, Duke University, Durham, NC 27708 and [3]Center for Human Genetics Research, Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN 37235, USA

Associate Editor: Dr Michael Brudno

ABSTRACT

Motivation: Whole-genome and -exome sequencing on parent–offspring trios is a powerful approach to identifying disease-associated genes by detecting *de novo* mutations in patients. Accurate detection of *de novo* mutations from sequencing data is a critical step in trio-based genetic studies. Existing bioinformatic approaches usually yield high error rates due to sequencing artifacts and alignment issues, which may either miss true *de novo* mutations or call too many false ones, making downstream validation and analysis difficult. In particular, current approaches have much worse specificity than sensitivity, and developing effective filters to discriminate genuine from spurious *de novo* mutations remains an unsolved challenge.

Results: In this article, we curated 59 sequence features in whole genome and exome alignment context which are considered to be relevant to discriminating true *de novo* mutations from artifacts, and then employed a machine-learning approach to classify candidates as true or false *de novo* mutations. Specifically, we built a classifier, named *De Novo* Mutation Filter (DNMFilter), using gradient boosting as the classification algorithm. We built the training set using experimentally validated true and false *de novo* mutations as well as collected false *de novo* mutations from an in-house large-scale exome-sequencing project. We evaluated DNMFilter's theoretical performance and investigated relative importance of different sequence features on the classification accuracy. Finally, we applied DNMFilter on our in-house whole exome trios and one CEU trio from the 1000 Genomes Project and found that DNMFilter could be coupled with commonly used *de novo* mutation detection approaches as an effective filtering approach to significantly reduce false discovery rate without sacrificing sensitivity.

Availability: The software DNMFilter implemented using a combination of Java and R is freely available from the website at http://humangenome.duke.edu/software.

Contact: ydwang@hit.edu.cn

Received on November 20, 2013; revised on February 2, 2014; accepted on March 4, 2014

1 INTRODUCTION

De novo mutations (DNMs) represent the most extreme form of rare variants and play an important role in human diseases (Veltman and Brunner, 2012). With rapid development of high-throughput-sequencing technology, large-scale whole-genome or -exome sequencing of parent–offspring trios or multiplex families is becoming a powerful approach to investigating DNMs associated with human disease. Recent sequencing studies have revealed that DNMs can affect genes with diverse biological consequences in several neuropsychiatric diseases, such as autism spectrum disorder (Michaelson *et al.*, 2012; Neale *et al.*, 2012; O'Roak *et al.*, 2012; Sanders *et al.*, 2012), intellectual disability (de Ligt *et al.*, 2012; Rauch *et al.*, 2012), schizophrenia (Girard *et al.*, 2011; Xu *et al.*, 2012, 2011) and epileptic encephalopathies (Epi4K Consortium & Epilepsy Phenome/Genome Project, 2013).

Here, we focus on a critical step in such studies, the detection of DNMs from whole genome/exome sequencing data in parent–offspring trios. The standard approach used by most studies is to call variants in each sample of a trio independently and then identify putative DNMs by comparing offspring against parental genotypes with Mendelian inconsistency. Therefore, a false positive variant call in offspring or a false negative variant call in either parent will result in a false positive DNM call; conversely, a false negative variant call in offspring or a false positive variant call in either parent will result in a false negative DNM call. Although there have been great improvements in development of single- and multiple-sample variant-calling approaches (Nielsen *et al.*, 2011), a variety of factors, including sequencing artifacts and alignment issues, lead to high rates of both false positive and false negative variant calls. Although allele frequency and linkage disequilibrium (LD) have been successfully leveraged to improve variant calling accuracy (Le and Durbin, 2011), it cannot apply to DNM calling because no such information is available for new mutations.

Distinct from standard approaches, methods that jointly model parent–offspring relationships within a trio have been developed specifically for DNM calling by utilizing Mendelian inheritance information within a trio. For example, DeNovoGear (Ramu *et al.*, 2013) calculates a posterior probability of being a true DNM call for every candidate-variant site by taking into account all three samples' genotype likelihoods under a prior based on genome-wide DNM rate. Polymutt (Li *et al.*, 2012) calculates maximum likelihoods of genotype configurations without and with Mendelian constraint, respectively, and then takes the ratio of the two resulting likelihoods as the cutoff. The larger the ratio, the more confident the DNM call is. Due to the use of

*To whom correspondence should be addressed.

extra information in the model, joint modeling approaches achieve much improved accuracy compared to standard approaches (Li *et al*., 2012; Ramu *et al*., 2013).

Both standard and joint modeling approaches can achieve high sensitivity. However, in terms of specificity, despite the better performance of joint modeling over standard approaches, both approaches rely on information of single sites assuming all reads having been correctly mapped, so they cannot eliminate false positive DNM calls originating from alignment mistakes. Heuristic filtering strategies and visual alignment inspection via genome browsers (Robinson *et al*., 2011) are usually used to filter out such false positive DNM calls. However, it is inherently difficult to select appropriate filtering parameter combinations to accommodate sensitivity and specificity simultaneously; it is also impractical to manually inspect a large number of candidates. These challenges necessitate the development of effective and automated DNM filtering algorithms.

Machine learning is a powerful approach to modeling complex multidimensional data and has been successfully applied to next-generation sequencing (NGS) data to identify genetic variants. For example, Variant Quality Score Recalibration in Genome Analysis Toolkit (GATK) uses a semi-supervised machine-learning algorithm, Gaussian mixture model, to estimate the probability that each variant is a true polymorphism rather than a sequencer, alignment or data processing artifact, by evaluating sequence features extracted from true variants (typically HapMap 3 sites and polymorphic sites on the Omni 2.5M SNP chip array) (DePristo *et al*., 2011). Supervised machine-learning algorithms, which usually train a model with known true and false positive variants, are also widely used to classify candidates as real variants versus artifacts. For example, SNPSVM (O'Fallon *et al*., 2013) utilizes support vector machine (SVM) to detect single nucleotide variants (SNVs); the Atlas2 Suite (Challis *et al*., 2012) builds a logistic regression model to call SNVs, insertions and deletions (INDELs); forestSV (Michaelson and Sebat, 2012) and SVM^2 (Chiara *et al*., 2012) employs random forest (RF) (Breiman, 2001) and SVM to detect large structural variants (SVs). In addition, mutationSeq (Ding *et al*., 2012) makes use of four algorithms including RF, SVM, Bayesian additive regression tree (Chipman *et al*., 2010) and logistic regression to identify somatic mutations from tumor-normal paired-sequencing data. Because these machine-learning approaches can incorporate multidimensional sequence features into a model, they usually yield better results than approaches that are based on single or very few sequence features.

Since DNMs are extremely rare, oftentimes real mutations are buried in a mass of false calls. In this article, we develop a supervised machine-learning-based approach, namely DNMFilter, to effectively sift out false DNM calls from a large number of putative candidates. We choose gradient boosting as the classification algorithm for DNMFilter based on recent reports showing that it can achieve better performance than other supervised machine-learning algorithms in many conditions (Hastie *et al*., 2009) and our own preliminary comparative analysis (data not shown). DNMFilter is designed to train a model based on experimentally validated and collected DNMs and then classify each novel candidate as a true or false DNM probabilistically.

In the following sections, we describe DNMFilter, a gradient boosting approach for classifying and filtering DNM candidates identified from any computational or manual approaches. We investigate multidimensional sequence features in whole genome and exome alignment context that have been shown to be relevant to DNM calls. Then we illustrate how to employ gradient boosting to design DNMFilter based on these features. We evaluate DNMFilter's theoretical performance and evaluate the contribution of different sequence features. Finally, we apply DNMFilter on in-house whole-exome trios and one whole-genome CEU trio from the 1000 Genomes Project (1000GP) to investigate its general performance in practice.

2 METHODS

The basic assumption of our approach is that all true DNMs share similar sequence features in whole genome and exome alignment context, and so do non-DNMs. We formalize DNM filtering as a binary classification problem and use gradient boosting as the classification algorithm. For classification, all true DNMs are deemed positive examples, while non-DNMs including inherited variants and wild-types (no variants found in any of three samples in a trio) are deemed negative examples. Here, we demonstrate how to filter *de novo* SNVs, but our approach can be easily extended to filtering *de novo* INDELs as well as *de novo* SVs.

2.1 Dataset

For the development and assessment of our proposed approach, we use two real sequencing datasets in this article.

The first dataset is Illumina Hiseq whole-genome-sequencing data of one CEU trio (father NA12891, mother NA12892 and the female offspring NA12878) from the 1000GP, which was sequenced to >30X coverage and preprocessed at the Broad Institute. The alignment (.bam) files were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120117_ceu_trio_b37_decoy/. The DNMs of this trio were previously called and subjected to experimental validation (Conrad *et al*., 2011), including 49 germline DNMs, 952 cell line somatic DNMs, 129 inherited variants and 1304 false positive DNMs in autosomes and X chromosome.

The second dataset is from the published large-scale exome-sequencing project investigating DNMs in epileptic encephalopathies (Epi4K Consortium & Epilepsy Phenome/Genome Project, 2013). The DNA of a total of 264 trios was derived from either primary cells or lymphoblastoid cell lines (LCLs). All samples were captured using Illumina's TruSeq Exome Enrichment Kit. Raw sequencing reads were produced at Center for Human Genome Variation's Genomic Analysis Facility (Duke University). The alignment (.bam) files were generated as the following steps: all reads were aligned to 1000 Genomes Phase II reference genome using Burrows–Wheeler Alignment (Li and Durbin, 2009); PCR duplicates were removed using Picard (http://picard.sourceforge.net); recalibration of base quality scores and local realignment around INDELs were performed using GATK. In this dataset, 329 putative DNMs (309 *de novo* SNVs and 20 *de novo* INDELs) were confirmed by Sanger sequencing.

2.2 Model

Boosting is a powerful technique for combining multiple weak base classifiers to produce a form of committee whose performance can be significantly better than that of any of the base classifiers.

Given a training set of $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, boosting aims to find n approximation $\hat{f}(x)$ to a function $f^*(x)$ that minimizes the expected value of some specified loss function $L(y, f(x))$, as follows

$$f^* = \arg\min_f E_{y,x} L(y, f(x))$$

Boosting iteratively fits an additive expansion of the form

$$f(x; P) = \sum_{m=1}^{M} \beta_m h(x; \alpha_m)$$

Where β_m is the expansion coefficient, $h(x; \alpha_m)$ is the base classifier parameterized by α_m.

Gradient boosting is one kind of boosting algorithms that applies steepest descent to minimize the loss function on the training data. At iteration step m, the gradient is calculated by

$$g_{im} = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x_i)=f_{m-1}(x_i)}$$

Then f_m, $\mathrm{f} = \{f(x_1), f(x_2), \ldots, f(x_n)\}^T$ is updated as follows

$$\mathrm{f}_m = f_{m-1} - \rho_m g_m$$

Where ρ_m is the step length, which is calculated by

$$\rho_m = \arg\min_\rho L(\mathrm{f}_{m-1} - \rho g_m).$$

Gradient boosting machine (Friedman, 2001) makes use of decision trees as the base classifiers and implements the above generic gradient-boosting algorithm. In addition, stochastic gradient boosting (Friedman, 2002) incorporates the idea of bagging to gradient-boosting machine, which can improve the performance by fitting every base classifier with bootstrapped samples of the whole dataset at each iteration step.

In this article, we use gradient-boosting machine as well as stochastic gradient boosting implemented in the R gbm package (http://cran.r-project.org/web/packages/gbm/index.html). As to the parameter settings, Bernoulli distribution is chosen as the loss function, shrinkage is set to 0.001, tree construction depth is set to 1 and bag fraction is set to 0.5. Moreover, 10-fold cross-validation is used for tuning the number of iterations. The remaining parameters are all with gbm package's default settings. In addition, a score between 0 and 1 will be produced for each prediction, representing the probability of the classification as the true DNM.

2.3 Feature selection

In this article, we selected 59 sequence features which we believe are able to discriminate DNMs from non-DNMs. The description is shown in Table 1. All sequence features are directly extracted from three individuals' BAM files in a trio. The selected features can be generally divided into three categories: pileup features, alignment features and cross sample features. The pileup features include allele balance, mean base quality and read depth, which are usually employed by other DNM detection and heuristic filtering approaches. To characterize the DNMs which may be mistakenly detected by the effect of alignment mistakes, we incorporate alignment features into the model, including mean mapping quality, strand direction, strand bias, mean number of nearby mismatches, mean number of nearby INDELs, fraction of soft clipped reads and fraction of MQ0 (mapping quality is equal to 0) reads. Alignment errors usually show position-dependence and appear with greater frequency at some positions than others (Meacham *et al.*, 2011). Based on this character, Fisher exact test can be used to test reference and alternative allele counts of two samples at the same position, which can avoid the interference of sequencing errors. If the resulting *P*-value is significant, then the genotypes of these two samples are different. VarScan 2 (Koboldt *et al.*, 2012) applies the similar idea to ascertain somatic mutations in tumor-normal paired-sequencing data. For true DNMs, the genotypes of the parents should be different from that of offspring at the same position, so we borrow this idea to generate two cross sample features.

2.4 Construction of training set

The most common strategy for building a training set is to directly use experimentally validated DNMs, including true positive DNMs and false positive ones. However, this strategy is usually limited by the relatively small number of validated DNMs.

In this article, we use in-house sequenced whole-exome trios to build the training set. Specifically, we use the experimentally confirmed true DNMs as positives and the candidates failing validation as negatives. Since we have fewer negative candidates, we expand the negative class by including further candidates using the following criteria: (i) run commonly used DNM-detection approaches on trios to obtain a candidate list; (ii) exclude all confirmed true positive DNMs from the candidates; (iii) bootstrap samples from the results of step 2 and regard them as negative examples. This strategy for choosing false positive DNMs

Table 1. Description of all selected sequence features in whole-genome and -exome alignment context

Feature	Description
Allele balance	The fraction of alt alleles over ref + alt alleles (one value)
Mean base quality	The mean base quality of alt/ref alleles (two values)
Read depth	The number of reads in a position (one value)
Mean mapping quality	The mean mapping quality of reads with alt/ref alleles (two values)
Strand direction	If the strands of reads with alt/ref alleles are all in one direction, then the value is 0, otherwise the value is 1 (two values)
Strand bias	The Phred-scaled *P*-value of Fisher exact test for forward and reverse strand, alt alleles versus ref alleles
Mean distance to 3′	The mean distance from current position to 3′-end on reads with alt/ref alleles (two values)
Fraction of MQ0 reads	The fraction of MQ0 reads (mapping quality is 0) over reads with alt/ref alleles (two values)
Fraction of soft clipped reads	The fraction of soft clipped reads over reads with alt/ref alleles (two values)
Mean number of nearby mismatches	The mean number of nearby mismatches on reads with alt/ref alleles (two values)
Mean number of INDELs	The mean number of INDELs on reads with alt/ref alleles (two values)
Paired samples test	The Phred-scaled *P*-value of Fisher exact test for father/mother and offspring, alt alleles versus ref alleles (two values)

Each feature in this table will be calculated for father, mother and offspring except Paired Samples Test.

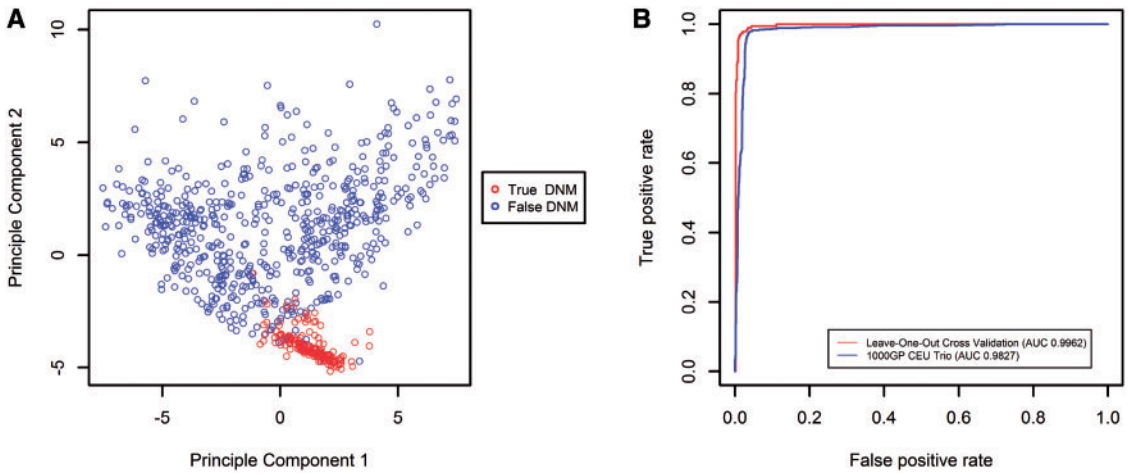

Fig. 1. (**A**) The first two principal components after PCA projecting for the training set built. (**B**) The ROC curves of leave-one-out cross validation and prediction on validated DNMs of a 1000GP CEU trio

works because the commonly used DNM detection approaches can generate hundreds of candidate DNM calls and majority of the candidates are false based on the mutation rate. Using this procedure we believe that such an augmented negative set not only contains preponderant false positive DNMs but also serves as a more representative sample of false positives that are likely to be generated by most popular DNM callers. With this strategy, our trained model has the power to filter out a wide range of false positives and can obtain an unbiased prediction of candidate DNM calls.

2.5 DNMFilter: DNMs filter

We develop DNMFilter based on the approach described above. DNMFilter consists of two core modules: (i) that extracts sequence features of known DNMs to build the training set; (ii) that selects sequence features to train gradient-boosting model and applies the trained model to filter out false positive DNMs. DNMFilter is designed to work on any candidate DNM call set obtained from any computational or manual approaches. DNMFilter is implemented using a combination of Java and R.

3 RESULTS

We build the training set using the approach described in Section 2.4 and evaluate the model's theoretical classification performance. In addition, we evaluate different sequence features' contribution to the model's performance. Furthermore, we combine DNMFilter with commonly used DNM detection approaches and apply them on in-house whole-exome trios and one 1000GP CEU trio to look into its performance in the general case.

3.1 Theoretical performance evaluation

According to the approach in Section 2.4, we build a training set with 185 experimentally confirmed true autosomal DNMs and 587 collected false autosomal DNMs identified in 2/3 (176) in-house exome trios. We use principal component analysis (PCA) to project 59 dimensional sequence features to two principal components (see Fig. 1A). The result shows the selected sequence features can confidently discriminate true positive from false positive DNMs, suggesting that the training set constructed as in Section 2.4 is able to capture a broad range of false positive patterns and expected to be effective in filtering out false positive candidates. We also explore higher dimensions and find that more principal components can further facilitate separating the two classes (data not shown). With the training set built, we train the model and evaluate its performance using leave-one-out cross-validation and testing its predictive power on 2434 experimentally confirmed (true and false) DNMs of one 1000GP CEU trio. Two receiver operating characteristic (ROC) curves are shown in Figure 1B, indicating that the model is robust and can achieve high sensitivity and specificity theoretically.

3.2 Feature importance

To evaluate the contribution of each selected sequence feature, we employ the feature relative importance measure approach available in R gbm package. The relative importance of all 59 features for the above training set in Section 3.1 is shown in Figure 2. Not surprisingly, three allele balance features and offspring's mean mapping quality for both reference and alternative alleles are among the top-ranked features. In addition, paired samples test introduced in this article contribute significantly to the performance. Alignment features except mean mapping quality also have non-zero relative importance, suggesting alignment mistake is an important cause of false DNMs.

3.3 Performance on in-house whole-exome trios

To investigate DNMFilter's performance in general, we apply it on the remaining 1/3 (88) in-house exome trios. The process is as follows: we first detect DNMs using common DNM-detection approaches, including Naïve Caller, polymutt (version 0.15) and DeNovoGear (version 0.5.2), and then use DNMFilter to filter candidate DNMs obtained by each approach, respectively. Here the standard approach (also named Naïve Caller) refers to using GATK UnifiedGenotyper to call variants jointly for all individuals within a trio and then comparing genotypes to identify candidate DNMs. To make a fair comparison, we design several

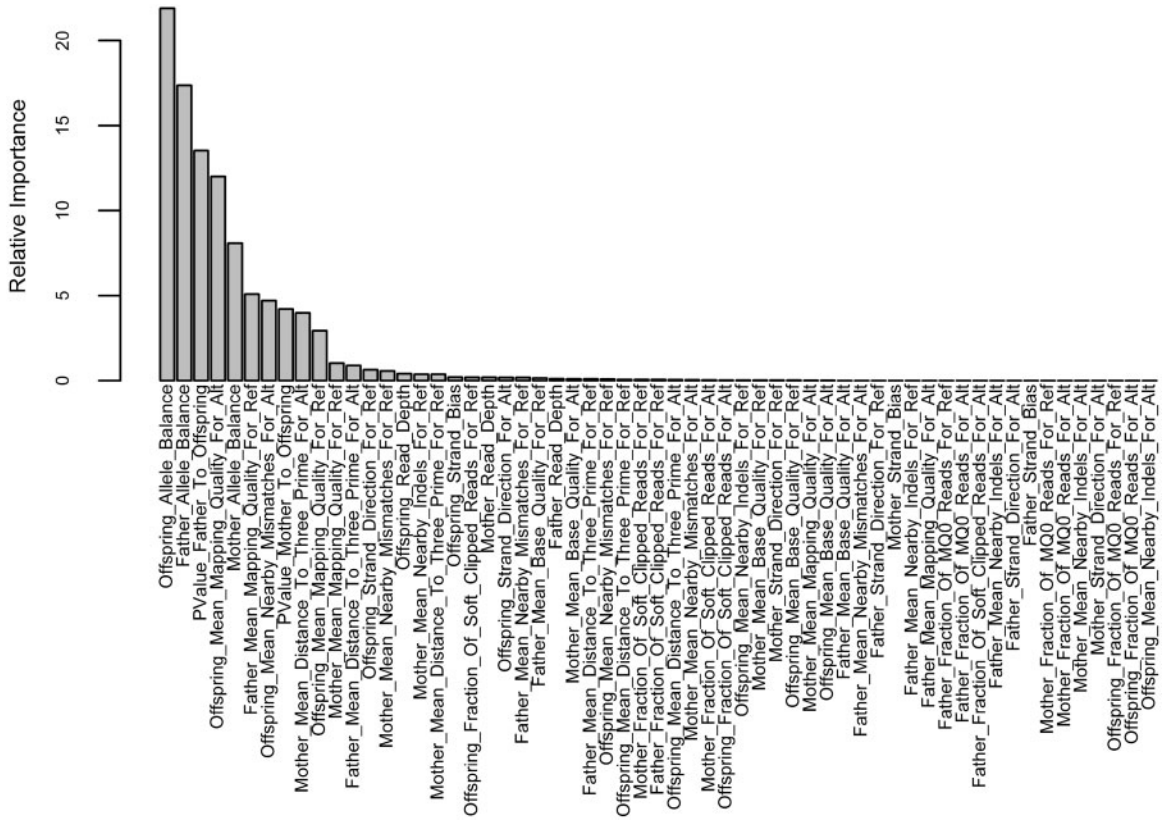

Fig. 2. The relative importance of 59 sequence features for the training set in Section 3.1

Table 2. The sensitivity and average DNM calls per trio of three commonly used DNMs-detection approaches combined with heuristic filtering as well as the sensitivity and average DNM calls per trio after DNMFilter (cutoff = 0.4) in 77 primary trios and 11 LCL trios

Programme	Raw DNM calls with heuristic filtering (DNM calls after DNMFilter)			
	PRIMARY		LCL	
	Sensitivity	Average DNM calls per trio	Sensitivity	Average DNM calls per trio
Naïve caller	98.9% (94.4%)	796.8 (8.2)	100% (94.7%)	702.7 (15.1)
DeNovoGear	96.6% (92.1%)	274.1 (8.9)	100% (94.7%)	296.2 (16.3)
Polymutt	96.6% (93.9%)	379.1 (7.5)	94.7% (89.5%)	341.5 (13.4)

heuristic filtering strategies as follows: (i) only keep the call where heterozygous genotype is present in offspring and homozygous reference genotypes are present in both parents; (ii) each of three samples in a trio is covered by at least 10 reads; (iii) minimum DQ for polymut and minimum pp_dnm for DeNovoGear are both with default settings to ensure highest sensitivity, and Naïve Caller's Phred-scaled likelihood (PL) for the genotypes AA, AB and BB, where A is the reference allele and B is the alternate allele, are set as >20, 0, >20 for the offspring and 0, >20, >20 for both parents. There are a total of 109 autosomal DNMs in the 88 trios that were confirmed in the previous study, but 1 DNM is of low coverage (<10) that does not meet the above heuristic filtering criteria, so this DNM is excluded in the following analysis. Table 2 shows that DNMFilter can significantly reduce the average number of DNM calls per trio compared with DNM-detection approaches with heuristic filtering strategies, and also maintain a high sensitivity, indicating that the vast majority of candidates removed by DNMFilter are false positives. Due to the presence of somatic mutations in cell lines, 11 LCL trios have more putative DNM calls than the 77 primary trios.

To evaluate DNMFilter's scoring performance, we check DNMFilter scores of all 108 confirmed DNM calls as well as

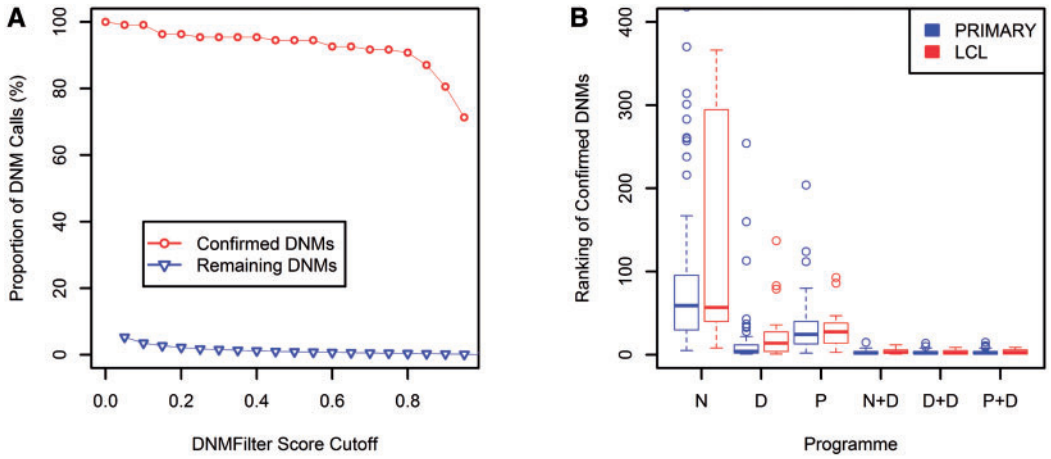

Fig. 3. (**A**) Illustrations of proportion of confirmed DNM calls and remaining DNM calls (excluding confirmed DNM calls from Naïve caller, DeNovoGear and polymutt' resulting calls) in terms of DNMFilter's different score cutoffs. (**B**) The ranking distribution of confirmed DNM calls for different DNM detection approaches in 77 primary trios and 11 LCL trios. Here N, D and P denote Naïve caller, DeNovoGear and polymutt respectively; N + D, D + D and P + D denote Naïve caller, DeNovoGear and polymutt combined with DNMFilter, respectively

Table 3. The results of three commonly used DNMs-detection approaches (with heuristic filtering) as well as the results after DNMFilter on 1000GP CEU trio

Programme	Raw DNM Calls with Heuristic Filtering (After DNMFilter)		
	Sensitivity on known DNMs		
	Germline	Somatic	Total DNMs
Naïve caller	100% (100%)	99.4% (98.6%)	47869 (2835)
DeNovoGear	100% (100%)	98.5% (98.1%)	16109 (2809)
Polymutt	100% (100%)	99.3% (98.6%)	29476 (2791)

scores of the remaining DNM calls (excluding confirmed DNM calls from results of Naïve Caller, DeNovoGear and polymutt). We rank all candidates by DNMFilter score. Figure 3A shows that DNMFilter score can clearly discriminate confirmed DNMs from the remaining DNM calls. Even at a low cutoff, DNMFilter can eliminate a large proportion of false positive DNM calls. To evaluate DNMFilter's ranking performance, minimum genotype quality (GQ) for Naïve Caller, DQ for polymut and pp_dnm for DeNovoGear, along with DNMFilter score are used to rank all putative DNM calls, respectively. Figure 3B shows that most validated true DNMs are ranked at the top by DNMFilter, demonstrating effectiveness of the algorithm in removing false positives that are mistakenly regarded as highly confident by other callers. This suggests that DNMFilter score as well as predictions from other callers can be combined to more reliably cull out true DNMs from a large number of candidate calls for experimental validation and further analysis.

3.4 Performance on 1000GP CEU trio

We also combine DNMFilter with three commonly used DNM-detection approaches as did in Section 3.3 and apply them on one 1000GP whole-genome-sequenced CEU trio that is independent of the sequencing data used for building the training set. Table 3 shows that DNMFilter significantly reduces the number of false positive DNMs, compared with other common DNM-detection approaches, while maintaining the high sensitivity of detecting true germline and somatic DNMs. It's worth noting that although the final number of DNMs obtained by DNMFilter is greater than the number of validated germline and somatic DNMs, it is assumed that there are a number of cell line somatic or even germline DNMs missed by Conrad *et al.* because of the limitation of early sequencing technology and data-preprocessing pipeline as well as the originally low sequencing coverage. In addition, despite that the training set is constructed with whole-exome-sequencing data, our result suggests that it can be effectively applied to whole-genome-sequencing data as well.

4 DISCUSSION

In summary, we developed DNMFilter, a novel gradient boosting-based approach for filtering DNMs identified in parent–offspring trios. We curated 59 sequence features in whole-genome and -exome alignment context and employed gradient boosting as the classification algorithm. We built the training set with confirmed true and false positive DNMs as well as collected

false positive DNMs. The evaluation of theoretical performance demonstrates that DNMFilter works confidently for its designed purpose. According to feature relative importance measure, we showed that alignment error is a significant cause of false DNM calls. We also applied DNMFilter on in-house whole-exome trios and one 1000GP CEU trio, and found that DNMFilter could maintain the high sensitivity and significantly reduce false positive DNMs when coupled with commonly used DNM detection approaches.

All results indicate that DNMFilter is a valuable complement for existing DNM detection approaches. By combining DNMFilter with any DNM detection approach(es) into a pipeline, users can first relax the confidence of detection step to ensure sensitivity, and then DNMFilter can be employed to filter out false positive DNM calls, which eventually leads to a reasonable size of highly confident DNM call set for experimental validation and further analysis. In particular, DNMFilter is expected to work best when it is applied to samples from the same sequencing and alignment pipeline as the ones used in the training set.

In future, we will consider extending DNMFilter to other kinds of DNMs, such as INDELs and SVs. Currently, DNMFilter's power is largely limited by the small number of known DNMs, especially for potential *de novo* INDEL and SV filtering. As more parent–offspring trios are sequenced in future, more DNMs within a more complete variant spectrum will be validated and incorporated into the training set. We plan to actively update DNMFilter with new whole-exome and -genome data to make it more effective and robust. We also consider incorporating additional relevant sequence features to capture a more comprehensive pattern discriminating true and false DNM calls that might not be represented by existing sequence features. We hope that DNMFilter is useful to the community as either a stand-alone tool for detecting DNMs or a filtering strategy combined with other DNM detection tools to boost both sensitivity and specificity.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Dr David Goldstein for his helpful comments and suggestions and Dr Qinghua Jiang for his help in manuscript editing.

Funding: The Epilepsy Phenome/Genome Project NIH grant U01-NS053998; Epi4K Project 1-Epileptic Encephalopathies NIH grant U01-NS077364; Epi4K-Sequencing, Biostatistics and Bioinformatics Core NIH grant U01-NS077303; Epi4K-Phenotyping and Clinical Informatics Core NIH grant U01-NS077276; Natural Science Foundation of China [grant numbers: 61173085, 61102149]; Governmental scholarship from China Scholarship Council (CSC) (to Y.L. and R.T.).

Conflict of Interest: none declared.

REFERENCES

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
Challis,D. *et al.* (2012) An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinform.*, **13**, 8.
Chiara,M. *et al.* (2012) SVM^2: an improved paired-end-based tool for the detection of small genomic structural variations using high-throughput single-genome resequencing data. *Nucleic Acids Res.*, **40**, e145.
Chipman,H.A. *et al.* (2010) Bart: bayesian additive regression trees. *Ann. Appl. Stat.*, **4**, 266–298.
Conrad,D.F. *et al.* (2011) Variation in genome-wide mutation rates within and between human families. *Nature genetics*, **43**, 712–714.
de Ligt,J. *et al.* (2012) Diagnostic exome sequencing in persons with severe intellectual disability. *New England J. Med.*, **367**, 1921–1929.
DePristo,M.A. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.
Ding,J. *et al.* (2012) Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics*, **28**, 167–175.
Epi4K Consortium & Epilepsy Phenome/Genome Project. (2013) De novo mutations in epileptic encephalopathies. *Nature*, **501**, 217–221.
Friedman,J.H. (2001) Greedy function approximation: a gradient boosting machine. *Ann. Stat.*, **29**, 1189–1232.
Friedman,J.H. (2002) Stochastic gradient boosting. *Comput. Stat. Data An.*, **38**, 367–378.
Girard,S.L. *et al.* (2011) Increased exonic de novo mutation rate in individuals with schizophrenia. *Nat. Genet.*, **43**, 860–863.
Hastie,T. *et al.* (2009) *The Elements of Statistical Learnin*. Springer, New York.
Koboldt,D.C. *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.
Le,S.Q. and Durbin,R. (2011) SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res.*, **21**, 952–960.
Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
Li,B. *et al.* (2012) A likelihood-based framework for variant calling and de novo mutation detection in families. *PLoS Genet.*, **8**, e1002944.
Meacham,F. *et al.* (2011) Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinform.*, **12**, 451.
Michaelson,J.J. and Sebat,J. (2012) forestSV: structural variant discovery through statistical learning. *Nat. Methods*, **9**, 819–821.
Michaelson,J.J. *et al.* (2012) Whole-genome sequencing in autism identifies hot spots for *de novo* germline mutation. *Cell*, **151**, 1431–1442.
Neale,B.M. *et al.* (2012) Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature*, **485**, 242–245.
Nielsen,R. *et al.* (2011) Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.*, **12**, 443–451.
O'Fallon,B.D. *et al.* (2013) A support vector machine for identification of single-nucleotide polymorphisms from next-generation sequencing data. *Bioinformatics*, **29**, 1361–1366.
O'Roak,B.J. *et al.* (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, **485**, 246–250.
Ramu,A. *et al.* (2013) DeNovoGear: *de novo* indel and point mutation discovery and phasing. *Nat. Methods*, **10**, 985–987.
Robinson,J.T. *et al.* (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.
Rauch,A. *et al.* (2012) Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet*, **380**, 1674–1682.
Sanders,S.J. *et al.* (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, **485**, 237–241.
Veltman,J.A. and Brunner,H.G. (2012) *De novo* mutations in human genetic disease. *Nat. Rev. Genet.*, **13**, 565–575.
Xu,B. *et al.* (2012) *De novo* gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat. Genet.*, **44**, 1365–1369.
Xu,B. *et al.* (2011) Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nat. Genet.*, **43**, 864–868.