Genetics and population analysis

novoCaller: a Bayesian network approach for *de novo* variant calling from pedigree and population sequence data

Anwoy Kumar Mohanty^{1,†}, Dana Vuzman¹, Laurent Francioli^{2,3}, Christopher Cassa¹, Brigham Genomic Medicine¹, Undiagnosed Diseases Network¹, Brigham and Women's Hospital FaceBase Project¹, Agnes Toth-Petroczy^{1,*} and Shamil Sunyaev^{1,4,*}

¹Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA, ²Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA, ³Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA and ⁴Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA

*To whom correspondence should be addressed.

[†]Present address: School of Electrical Sciences, Indian Institute of Technology Bhubaneswar, Odisha 752050, India
Associate Editor: Oliver Stegle

Received on January 22, 2018; revised on June 19, 2018; editorial decision on August 18, 2018; accepted on August 29, 2018

Abstract

Motivation: *De novo* mutations (i.e. newly occurring mutations) are a pre-dominant cause of sporadic dominant monogenic diseases and play a significant role in the genetics of complex disorders. *De novo* mutation studies also inform population genetics models and shed light on the biology of DNA replication and repair. Despite the broad interest, there is room for improvement with regard to the accuracy of *de novo* mutation calling.

Results: We designed novoCaller, a Bayesian variant calling algorithm that uses information from read-level data both in the pedigree and in unrelated samples. The method was extensively tested using large trio-sequencing studies, and it consistently achieved over 97% sensitivity. We applied the algorithm to 48 trio cases of suspected rare Mendelian disorders as part of the Brigham Genomic Medicine gene discovery initiative. Its application resulted in a significant reduction in the resources required for manual inspection and experimental validation of the calls. Three *de novo* variants were found in known genes associated with rare disorders, leading to rapid genetic diagnosis of the probands. Another 14 variants were found in genes that are likely to explain the phenotype, and could lead to novel disease-gene discovery.

Availability and implementation: Source code implemented in C++ and Python can be downloaded from <https://github.com/bgm-cwg/novoCaller>.

Contact: atoth-petroczy@bwh.harvard.edu or ssunyaev@bwh.harvard.edu

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

De novo (i.e. not present in either parent) mutations have an important role in medical genetics for both complex and monogenic disorders. Several recent studies have identified a substantial causal role

for *de novo* mutations in the etiologies of congenital heart disease (Zaidi *et al.*, 2013), schizophrenia (Fromer *et al.*, 2014), autism spectrum disorder (Iossifov *et al.*, 2014; O'Roak *et al.*, 2012; Sanders *et al.*, 2012) and epilepsy (Appenzeller *et al.*, 2014).

The effects of *de novo* mutations associated with monogenic Mendelian disorders are even more debilitating. The genomes of individuals with severe, undiagnosed, rare developmental disorders are enriched in damaging *de novo* mutations in developmentally important genes (DDD, 2017).

There are more than 4000 currently described Mendelian disorders for which a causal gene has not yet been identified. These disorders are often caused by a single, rare, large-effect variant in the coding sequence that is often hard to detect. Therefore, highly sensitive variant calling algorithms are required. Many variant calling methods have been developed to identify variation responsible for common traits in large-scale studies, and they are designed to generate calls of high quality. They are also designed to reduce false positives; however, this results in an increase in false negatives as well. When aiming for the detection of a single causal variant, it is crucial to increase sensitivity and eliminate false negatives.

Although affordable and widely used, currently available whole genome sequencing (WGS) and whole exome sequencing (WES) technologies are still hampered by uneven sequencing coverage and variability in variant call quality. Detecting *de novo* variation was recently improved by using trio-aware calling methods (Cleary et al., 2014; Francioli et al., 2014; 2017; Kong et al., 2012; Poplin et al., 2016) with the main focus on reducing false calls. The method used in the deCODE trio-sequencing project used empirical filters such as population frequency, read depth and allele balance. The random forest method applied in the GoNL project used 12 features of sequencing quality of the trio samples. PhaseByTransmission (PBT), the latest state-of-the-art caller of the GATK pipeline, calculates the joint probability of the data given the genotype (GT) likelihoods in the individual family members, the known familial relationships and a prior probability for the mutation rate (Francioli et al., 2017). DeepVariant, recently developed by Google, incorporates recent developments in image recognition by training a deep convolutional neural network to detect genetic variation by learning statistical relationships between images of the raw reads aligned around putative variant sites and ground-truth GT calls (Poplin et al., 2016). SNVMix, is a Bayesian method for predicting single nucleotide variants specifically in tumor samples (Goya et al., 2010). FamSeq developed by Peng et al. (Peng et al., 2013, 2014) uses pedigree information of large families to make high-quality GT calls independent of inheritance mode of the disease and takes as an input previously computed likelihoods.

The method presented here, novoCaller, uses population frequency and pedigree data and a rigorous statistical approach in order to maximize accuracy in the *de novo* variant calling process and eliminate false negatives. Inspired by the errors that manual inspection revealed about low-quality variants, we aimed at utilizing read-level quality metrics and pedigree and control population data in order to reduce technical sequencing artifacts. novoCaller focuses solely on *de novo* variants and uses samples from parents rather than other relatives. Our method models GT frequency and quality of the data at any given location and learns model parameters from unrelated samples which in turn are then used to compute GT likelihoods. Thereby we combined overall statistical knowledge about sequencing quality and inspected the raw reads to evaluate GT quality.

We tested novoCaller on the GoNL dataset (Francioli et al., 2014) and on an autism cohort (De Rubeis et al., 2014). On an ROC curve, when novoCaller correctly calls 98% of 605 verified *de novo* variants, it generates only 20% false positives in the GoNL dataset whereas for a similar level of sensitivity, PBT gives a significantly larger false positive rate of 46% (Fig. 2). novoCaller outperforms both the PBT and the deCODE methods, and it reaches

sensitivity similar to that of the supervised random forest method without having to depend on a training dataset (AUC=0.969). Furthermore, it correctly identified 96% of the *de novo* calls in the autism dataset of 2023 variants. On average, it calls 1.6 ± 1.6 *de novo* variants in 48 clinical trios compared to a range of 10–50 raw calls, meaning a significant reduction in the need for manual inspection and experimental testing of the calls. It correctly identified 76/79 *de novo* variants, and called only two false positives. novoCaller was successfully used in clinical application on Mendelian disorders to produce genetic diagnostics of three cases and candidate variant identification in 14 cases in the (Haghighi et al., 2018) Brigham Genomic Medicine (BGM) gene discovery initiative.

2 Materials and methods

2.1 novoCaller

We designed a Bayesian network model that determines the likelihood of a variant being *de novo* based on observations in samples unrelated to the trio. These observations are used to account for technical sequencing artifacts observed at any given location as well as population parameters such as allele frequency. Specifically, we used allele depths (ADs) and the allele balance of the reference (R) and alternative (A) alleles. GT_m denotes the GT of the mother, GT_f denotes the GT of the father and GT_c denotes the GT of the child (proband). AD_m , AD_f and AD_c denote the AD vectors of the mother, father and child, respectively. Equation (1) describes the model and the parameters and Figure 1a shows the Bayesian network of this model.

$$\begin{aligned} &P(AD_c, AD_m, AD_f, GT_c, GT_m, GT_f | \rho, F, \mu) \\ &= P(AD_c | GT_c, \rho)P(AD_m | GT_m, \rho)P(AD_f | GT_f, \rho) \\ &\quad \times P(GT_c | GT_m, GT_f, \mu)P(GT_m | F)P(GT_f | F) \end{aligned} \quad (1)$$

The parameters in bold (ρ and F) represent sequencing quality and GT frequency at any given variant call location. ρ is a scalar value between 0 and 1, while F is a three-dimensional vector with entries ≥ 0 that sum to 1. ρ represents the strength of the relationship between the observed AD and the GT. The closer the value of ρ is to one, the stronger this relationship. F represents the parameter vector for a three-dimensional generalized Bernoulli distribution from which the GTs are sampled. These parameters are determined from the unrelated samples using a version of the expectation maximization (EM) algorithm. $P(GT_c | GT_m, GT_f, \mu)$ is the probability of the GT of the child GT_c being homozygous reference (RR), heterozygous (RA) or homozygous alternative (AA) depending on the nine different combinations of the GTs of the two parents. This probability also depends on the mutation rate μ . The larger the mutation rate, the more prone the algorithm will be to call a variant *de novo*. In other words, the μ parameter is the probability that the child is RA given that both the parents are RR without the effect of the data, i.e. $\mu = P(GT_c = RA | GT_m = RR, GT_f = RR)$. A small value of μ means that enough evidence in the data is needed to generate significantly high value of posterior probability that a *de-novo* event has occurred. However as Equation (1) shows us, $P(GT_c | GT_m, GT_f, \mu)$ is simply a scaling factor. Parameter μ 's primary function is to ensure that we do not get a zero value for the posterior probability for a *de novo* event. We tested the effect of changing mutation rates based on the mutation rate map from Francioli et al. (Francioli et al., 2015), and observed that the shape of the ROC is not sensitive to parameter μ (Supplementary Fig. S4). Therefore all analysis presented on this manuscript uses fixed $\mu = 1e-8$.

Fig. 1. Bayesian network model. (a) Schematic representation of the framework of the Bayesian network model. We use allele depths (AD) and the allele balance of the reference (R) and alternative (A) alleles. GT_m denotes the genotype of the mother, GT_f denotes the genotype of the father and GT_c denotes the genotype of the child (proband). AD_m , AD_f and AD_c denote the allele depth vectors of the mother, father and child, respectively. (b) Bayesian network of the model for estimating F and ρ . Model parameters in bold (ρ and F) are determined from the unrelated samples using a version of the expectation maximization (EM) algorithm and the variational Bayes framework (Supplementary Methods). (c) Incorporating read-specific information into the Bayesian network. The Bayesian network that was applied on clinical data contained two independent parameters of forward and reverse allele depths conditional on genotype

We use a binomial model for the relationship between the GT and the AD as shown in Equation (2).

$$P(AD | GT, \rho) = \begin{cases} \rho^{AD^{(0)}} (1 - \rho)^{AD^{(1)}}, & \text{if } GT = RR \\ 0.5^{(AD^{(0)} + AD^{(1)})}, & \text{if } GT = RA \\ \rho^{AD^{(1)}} (1 - \rho)^{AD^{(0)}}, & \text{if } GT = AA \end{cases} \quad (2)$$

where $AD^{(0)}$ and $AD^{(1)}$ are the two elements of the AD vector. The method learns the model parameters for each variant site separately so that variations in the binomial parameter ρ are learned across all of the variant sites. $P(GT_m | F)$ and $P(GT_f | F)$ are discrete probability distributions of the three GTs RR, RA and AA based on the three-dimensional vector F , whose elements are ≥ 0 and sum to 1.

After the model parameters are determined, the probability of a *de novo* event is computed by calculating

$P(GT_c = RA, GT_m = RR, GT_f = RR | \dots)$, which is the conditional probability that the child is RA and the parents are RR conditional on the rest of the data. This is done by simply normalizing with respect to GT_c , GT_m and GT_f from Equation (1). If the posterior probability is above a certain pre-set threshold (i.e. 0.7 in our study), then a *de novo* variant is called.

2.2 Incorporating read-level information from BAM files in the *de novo* calling process

novoCaller was designed to generate calls rapidly from VCF files that contain AD information. This is a fast and easily applicable solution for large datasets. We implemented a python script which would open BAM files at the locations shortlisted by novoCaller and use the raw reads in the calling process. The strategy was that we would set the posterior probability threshold for making a call comparatively low on novoCaller thereby generating a long list of 50–100 calls per trio, followed by opening BAM files at those locations of the trio and unrelated samples. Opening BAM files allows access to both forward as well as reverse reads. We designed a model that assumes the forward and reverse reads are independent conditional on the unobserved GT of the samples as shown in Figure 1c. Hence, the model now uses two different ρ values, one for the forward read and one for the reverse read ($\rho_{forward}$ and $\rho_{reverse}$). Intuitively, discriminating between forward and reverse reads should result in greater performance. We did not see any improvement in

the overall calling process when tested on the GoNL dataset (Supplementary Fig. S5). However, we observed that this approach performed better on individual clinical cases. Unfortunately, the small number of such cases do not allow proper statistical comparison. We still decided to include this feature in the final release of the software.

2.3 Clinical application

In order to apply novoCaller on clinical trios, we considered only coding variants with maximum allele frequencies of $<0.1\%$ in the ExAc population database. The allele frequencies in unrelated samples in our dataset were computed within the probabilistic framework, and we included only those calls that had allele frequencies in unrelated samples of $<1\%$. Equation (3) shows how the allele frequency in unrelated samples, $AF_{unrelated}$ is computed.

$$AF_{unrelated} = \sum_{i=1}^N \frac{P(GT_i = RA) + 2P(GT_i = AA)}{2N} \quad (3)$$

where GT_i is the GT of the i th sample.

For each case, a priority list of potential *de novo* calls was generated in decreasing order of posterior probability of a *de novo* event having taken place at any given variant location. First, an initial call list was generated using the same code used to generate Figure 2. The threshold on the posterior probability was kept low in order to generate a large number of calls, i.e. ~ 100 – 200 calls per trio. Then, for these ~ 100 calls, we parsed the BAM files of the trio (as well as the unrelated control samples) to obtain both the forward reads and the reverse reads and applied the read-level information-aware Bayesian network (Fig. 1c).

2.4 Trio datasets

- *Genome of the Netherlands (GoNL) dataset.* novoCaller was tested on the GoNL dataset, which consisted of 250 trios with whole genome sequences (Francioli et al., 2014). The authors used the Genome Analysis Toolkit (GATK) PBT caller to generate an initial set of *de novo* calls, from which 2279 calls were randomly selected for Sanger validation and 605 were Sanger confirmed as true *de novo* calls. The trained random forest algorithm predicted an additional 11 020 *de novo* mutations.

Fig. 2. novoCaller's performance on validated *de novo* calls. Comparison of novoCaller (blue) with GATK PhaseByTransmission (red, AUC=0.898), random forest trained on the GoNL data (Francioli *et al.*, 2014) (green circle) and the calling strategy used in Kong *et al.* (2012), labeled deCODE strategy (orange circle). Area under curve (AUC) value =0.969

- **ASD dataset.** A large-scale WES study on autism spectrum disorder comprised of 3871 autism cases had 2023 *de novo* variants and an additional 527 Sanger confirmed variants.
- **Clinical dataset of patients with rare diseases.** novoCaller was applied to 48 clinical trio samples that were collected via the BGM program, the Brigham and Women's Hospital FaceBase Project and the Undiagnosed Diseases Network. The patients presented rare forms of known or undiagnosed diseases for which routine sequencing panels could not identify a genetic cause. The BGM gene discovery initiative performed WGS/WES analysis as well as ordered Sanger sequencing of 17 interesting variants.

3 Results

3.1 Comparison of novoCaller to other trio-aware approaches

We applied novoCaller on the high-confidence *de novo* variant list of the GoNL dataset, which consists of 250 trio samples, and contains 605 validated true *de novo* calls and 1674 false *de novo* calls. We considered the parents of the trios (478 individuals: <http://www.nlgenome.nl/>) to be unrelated to each other and estimated model parameters from these samples. We also ran GATK PBT, which was initially used to generate the calls from this dataset. The comparison of the corresponding ROC curves is shown in Figure 2. We compared the performance of our method against two other *de novo* calling strategies: (i) the *de novo* calling strategy used in the Icelandic deCODE project (Kong *et al.*, 2012) (Fig. 2, red circle) and (ii) supervised random forest trained on 70% of the GoNL dataset of Francioli *et al.* (2014) (Fig. 2, green circle). novoCaller's true positive versus false positive rate is similar to that of the random forest even though it is unsupervised and was not trained on the Sanger sequenced list.

3.2 novoCaller identifies novel *de novo* mutations in the GoNL dataset

The GoNL consortium published a list of 11 020 high-quality calls that were prioritized by training a random forest model on the list of

Sanger sequenced calls as described in the previous sections (Francioli *et al.*, 2014). novoCaller not only correctly identified 100% of these calls but also found an additional 154 *de novo* calls with posterior probability >0.99 that were not in the previously published list. We visually inspected these calls in IGV and found 138 calls (90%) to be likely true (Supplementary Table S1). Thus, novoCaller was able to identify potentially true *de novo* variants that were missed by existing methods, including supervised methods.

3.3 Expectation maximization and variational Bayesian approach

Since the model parameters are learned from unrelated samples, if there are not enough unrelated samples to characterize technical artifacts at the location, it is possible that the model parameters are overfit. In order to test whether our EM estimation leads to overfitting, we applied a fully Bayesian solution. We treated model parameters as unobserved variables and computed the posterior probability of a *de novo event* conditional on the observed AD. We used the variational Bayesian method as described in Supplementary Methods and found no difference in the performance of the method. We tested our variational Bayes algorithm on the GoNL dataset (Supplementary Fig. S3) and compared the resulting ROC (variational Bayes, blue) with the ROC generated by the previous method (EM, green). Both methods have equally high performance with area under curve (AUC) values of 0.969. Since the performance rates of the two methods are almost identical, we concluded that we have enough samples in the test dataset and that the parameters are not overfit.

3.4 Testing novoCaller on autism cohort whole exome sequencing dataset

We tested the caller on an autism cohort WES dataset described in De Rubeis *et al.* (2014). In our previous analysis with the GoNL dataset, we fortunately had a ground-truth set of true positives and true negatives. This allowed us to generate an ROC curve, which is a traditional measure of performance for a binary classifier. For the dataset described in De Rubeis *et al.* (2014), however we had a list of Sanger confirmed true positive calls only. We made calls on 527 of those Sanger confirmed *de novo* calls and found only 7 false negatives, thus achieving 98% sensitivity overall. We also tested novoCaller on a list of 2023 *de novo* calls that were not further tested by Sanger sequencing (De Rubeis *et al.*, 2014). novoCaller called 96% of the mutations and identified only 78 calls as false positives (Supplementary Table S2), which are yet to be confirmed by experimental methods.

3.5 Application of novoCaller on clinical cases with rare disease conditions

After validating our method on high-confidence *de novo* mutation data, we wanted to predict novel *de novo* mutations within the scope of the BGM gene discovery initiative (Haghighi *et al.*, 2018). We applied novoCaller to 48 cases with Mendelian genetic disorders in which the suspected mode of inheritance was a *de novo* event.

In this clinical application, novoCaller creates a prioritized list of *de novo* variants and provides the posterior probability of the calls. On average, novoCaller identifies 1.6 ± 1.6 *de novo* events per trio, while raw calling of mutations results in an order of magnitude more variants that are tedious to verify (Table 1 and Supplementary Table S3). The calls that are found to be true in IGV inspection or are confirmed by Sanger sequencing are usually within the top calls

Table 1. Clinical application: *de novo* coding and splice site variants in trio WES/WGS samples of rare disease patients

Case id and phenotype	novoCaller	Genes	Chr	Position	Ref	Alt	PP	Validation
0001: Distal arthrogryposis	2(2)	PIEZO2*	chr18	10671599	ATCT	A	1.00	TP(S)
		ACSM4	chr12	7475034	G	A	1.00	TP(S)
0047: Craniofacial abnormalities	3(3)	CUL1	chr7	148494946	C	G	1.00	TP(S)
		FBN1*	chr15	48707969	A	C	1.00	TP(S)
		HNRNPC	chr14	21679608	TC	T	1.00	TP(S)
0063: GDD and syndactyly	1(1)	MED12**	chrX	70348505	C	T	1.00	TP(G)
0030: Infantile IBD and hearing loss	3(3)	PRR4	chr12	11083515	C	T	1.00	TP
		STXBP3*	chr1	109336268	AG	A	1.00	TP(S)
		RGS12	chr4	3430407	AT	A	1.00	TP
		PARD6A	chr16	67692501	A	AG	0.05	FN
0066: Acrofrontofacionasal dysostosis	3(3)	HEATR5B	chr2	37227894	T	C	1.00	TP
		POLR1A**	chr2	86271270	T	A	1.00	TP
		FRG1B	chr20	29624081	A	T	1.00	TP
0111: Congenital contractural arachnodactyly	2(2)	CHRNA1	chr2	175619031	G	A	1.00	TP
		FBN2**	chr5	127671679	C	T	0.99	TP
0108: Marshall-Smith syndrome	4(4)	SPRY2	chr13	80911319	G	T	1.00	TP
		NFIX**	chr19	13186487	T	C	1.00	TP
		MYO7B	chr2	128331589	C	T	1.00	TP
		HOXA13	chr7	27239274	C	T	1.00	TP
0008: Thymic hyperplasia, colonic adenomas	6(4)	FAM72B	chr1	120846059	G	A	1.00	FP(S)
		CASR	chr3	122000923	A	G	1.00	TP(S)
		ACSM4	chr12	7475034	G	A	1.00	TP
		KIAA1755*	chr20	36848054	C	A	1.00	TP(S)
		ANKRD36C	chr2	96517885	G	A	0.89	FP
		CTD-2031P19.4	chr5	55243420	T	TA	0.71	TP
0175: DD and oculomotor abnormalities, epilepsy	2(2)	EXT1	chr8	119122375	C	G	1.00	TP(G)
		MRPS12*	chr19	39423147	C	T	1.00	TP(G)
0202: ID, seizures, brain abnormalities, dysmorphic facies	1(1)	GDF11*	chr12	56143450	C	G	1.00	TP
0040: Hypertelorism	2(2)	CPNE1	chr20	34215309	G	A	1.00	TP
		CARD8*	chr19	48733704	C	A	0.99	TP
0069: Hearing loss and appearance of Apert syndrome	3(3)	EMLIN1	chr2	27306022	G	A	1.00	TP
		ZDBF2	chr2	207170282	A	G	1.00	TP
		FGFR2*	chr10	123276913	T	A	1.00	TP
0189: Cleft palate, aniridia, bilateral club feet, plagiocephaly	2(2)	SYNE2	chr14	64679559	A	C	0.99	TP
		PAX6*	chr11	31815351	C	A	0.95	TP
0227: Epilepsy, chronic hypertension	1(1)	KLHL32*	chr6	97562260	G	T	1.00	TP
		IL2RG	chrX	70330149	G	T	0.46	FN
0077: Pancreatic insufficiency, GI problems and DD	1(1)	PRKACA*	chr2	122125400	C	T	1.00	TP(S)
0117: Mild DD and cranial polyneuropathy	0(0)	ANKLE1	chr19	17397498	TG	T	0.00	TN(S)
		SDC3	chr1	31381369	G	19del	0.00	TN(S)
0052: Dysmorphic features	6(6)	ZBTB2	chr6	151687330	C	T	1.00	TP
		CENPJ	chr13	25486917	T	A	1.00	TP(S)
		CENPJ	chr13	25486918	T	A	1.00	TP(S)
		TRIM29	chr11	120007945	G	C	1.00	TP
		WNT6	chr2	219735845	G	A	1.00	TP
		POGLUT1	chr3	119209436	G	T	1.00	TP(S)
0074: GDD, congenital heart disease	4(4)	SMYD3	chr1	246021911	G	A	1.00	TP
		ITPR1PL1	chr2	96993674	ACC	A	1.00	TP
		ZNF451	chr6	56966637	T	C	1.00	TP(S)
		CLASP1	chr19	14208643	G	A	1.00	TP
0183: Retinal vasculitis, joint pain, recurrent hepatitis	2(2)	IVL	chr1	152882770	61nt	del	1.00	TP
		GDF11*	chr12	56143358	G	A	1.00	TP
0118: Metabolic disorder and MS-like brain lesions	1(1)	BMP3	chr4	81952521	C	A	1.00	TP(S)
0121: Craniosynostosis, bilateral cleft lip and palate	3(3)	OTOG	chr11	17632908	G	A	1.00	TP
		NUP93	chr16	56857716	G	A	1.00	TP
		FREM2*	chr13	39271841	T	T	1.00	TP
0135: Pierre Robin syndrome, cleft palate, cardiac defect	1(1)	MED13L*	chr12	116424184	AC	A	1.00	TP

Note: DD: developmental delay. GDD: global developmental delay. ID: intellectual disability. IBD: Inflammatory bowel disease. *: candidate gene for phenotype. **: genes that are known to be associated with the disease phenotype. Ref: reference allele. Alt: alternative allele. PP: posterior probability of the call. TP: true positive based on IGV inspection. FP: false positive based on IGV inspection. FN: false negative. TN: true negative. S: variant has been Sanger confirmed. G: variant has been observed both in WES and WGS data. The number of TP variants is shown in parenthesis.

of the priority list, having a high (>0.70) posterior probability. Thus, novoCaller reduces the amount of manual work needed to inspect visually the reads on IGV as well as the number of experimental Sanger confirmations needed. For these 48 trio cases with rare Mendelian disorders, the caller made 78 calls, of which 76 were visually verified on IGV to be true positives, 3 were false negatives and 2 were false positives (Table 1 and Supplementary Table S3). Sanger confirmation of 17 variants further strengthens the caller's high accuracy.

For 17 out of the 48 cases, the crowdsourced team of clinicians and researchers in the BGM program selected a set of *de novo* variants that could explain the phenotype of a given case (Table 1, possibly causal genes marked with an asterisk). In two cases, a novel disease–gene association was discovered and published, and for 12 cases the validation is still ongoing. For example, a *de novo* deletion in PIEZO2 (p.Glu2727del) was found to be associated with distal arthrogryposis [MIM: 108145] (Coste *et al.*, 2013). Further biochemical studies on the variant proved a causal relationship between a *de novo* deletion and the disease. A *de novo* deletion (p.Lys343fsX6) in the STXBP3 gene has been linked to infantile IBD and hearing loss (Kelsen *et al.*, 2018).

Further, novoCaller identified three novel variants in known disease genes in patients with severe craniofacial abnormalities. A *de novo* missense mutation in the POLR1A gene (p.N1043Y) was found in a patient with acrofrontofacionasal dysostosis [MIM: 616462]. A *de novo* splice site mutation was identified in the NFIX gene in a patient with Marshall–Smith syndrome [MIM: 602535], and a splice mutation in FBN2 was found in a patient with congenital contractural arachnodactyly [MIM: 121050]. All three of these diseases are congenital autosomal dominant disorders known to be caused by heterozygous mutations. Identifying these three new variants in known disease genes that led to genetic diagnosis of the probands, illustrate the power of our approach to identifying rare *de novo* variants.

4 Discussion

The advent of genomic sequencing has revolutionized genetic diagnosis, allowing inspection of much of the coding sequence. Clinical diagnostic sequencing is now widely available, and it is estimated that 1.6 million individuals will have some form of genomic sequencing completed by 2017 (Kurreck and Stein, 2016). Even in clinical cases with largely severe disorders, exome sequence analysis uncovers an initial molecular diagnosis around 30–35% of the time. There are many possible reasons for this incomplete diagnostic rate, including incomplete penetrance of disease mutations, too much focus on coding variation, and technical issues such as uneven sequencing coverage, alignment issues and low-quality variant calls. The detection of rare *de novo* mutations is especially important in rare monogenic Mendelian disorders.

Since most variant caller algorithms are geared toward common variations, the detection of rare variants remains a challenge. Several recent advancements in *de novo* mutation detection were made when trio data were utilized to rule out rare variants that segregate in a family (GoNL and deCODE). A random forest trained on orthogonally confirmed *de novo* mutations used 12 features, including sequencing quality measures of the trio samples. We found our method to be superior to the above-mentioned algorithms. novoCaller detects *de novo* events in the child of a trio sample by using unrelated samples to detect technical artifacts present in the sequencing and alignment process. It uses familial information to minimize the error of undercalled heterozygous parents. Our method extends the power of joint calling as described in Cleary

et al. (2014) by gauging the presence of technical artifacts and estimating the prevalence of an allele at any particular variant call location. The unified Bayesian framework allows our method to make the best decision about whether a variant is present or whether alternate reads align at the location as technical artifacts. We find that our method can better reduce the number of false positive variant calls and identify potential false negative calls compared with other commonly used methods. Our work is an example of how hand-crafted statistical models specifically designed for a given problem can perform just as well as machine learning methods without having to depend on high-quality training datasets. novoCaller could be used to sort through variants to generate high-confidence *de novo* calls initially without having to depend on a previously generated high-confidence dataset of a given population to take into account sequencing and/or technical artifacts. It is important to note that novoCaller cannot distinguish between mosaic parents who transmitted the variant to the offspring.

novoCaller generates a priority list of *de novo* calls based on their posterior probability. When applied on clinical Mendelian cases at the BGM program, the potential causal gene was found among the top 3 candidates in the priority list and led to diagnosis or candidate gene identification in 17/48 cases (Table 1 and Supplementary Table S2). By changing the posterior probability threshold, the behavior of the caller can be adjusted and the number of false negatives reduced by allowing for false positives. In rare disease cases, we may wish to use the caller with a lower threshold on the posterior probability to minimize the chance of missing potential causal *de novo* variants. For studying *de novo* variation in common disorders, the opposite mode of operation may be preferable in order to generate a set of high-confidence calls and minimize the number of false positives in experimental verification like Sanger sequencing. novoCaller is readily applicable to other species and in large sequencing studies in which pedigree information is known between samples.

Acknowledgements

We thank all the families who participated in the Brigham Genomic Medicine program and the physicians who contributed to subject recruitment and clinical investigation. We thank Richard Maas, Joel Krier, Nikkola Carmichael, Elizabeth Fieg and Victoria Perroni for their work at BGM that facilitated the assembly and analysis of the clinical dataset. Undiagnosed Diseases Network members are listed in Supplementary Material.

Funding

This work was supported by the Brigham Research Institute (BRI) (DirectorFLS Transformative Award) and the National Institutes of Health grants R01-GM078598 and HG007229. Research reported in this manuscript was supported by the NIH Common Fund, through the Office of Strategic Coordination/Office of the NIH Director under Award Numbers U01HG007690 (Loscalzo) and by the NIH-NIDCR National Institute of Dental and Craniofacial Research under Award Numbers 5U01DE024443 (Maas). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest: none declared.

References

Appenzeller, S. *et al.* (2014) De novo mutations in synaptic transmission genes including DNMT1 cause epileptic encephalopathies. *Am. J. Hum. Genet.*, **95**, 360–370.

- Cleary, J.G. *et al.* (2014) Joint variant and de novo mutation identification on pedigrees from high-throughput sequencing data. *J. Comput. Biol.*, **21**, 405–419.
- Coste, B. *et al.* (2013) Gain-of-function mutations in the mechanically activated ion channel *piezo2* cause a subtype of distal arthrogryposis. *Proc. Natl. Acad. Sci. U S A*, **110**, 4667–4672.
- DDD (2017) Prevalence and architecture of de novo mutations in developmental disorders. *Nature*, **542**, 433–438.
- De Rubeis, S. *et al.* (2014) Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, **515**, 209–215.
- Francioli, L.C. *et al.* (2014) Whole-genome sequence variation, population structure and demographic history of the dutch population. *Nat. Genet.*, **46**, 818–825.
- Francioli, L.C. *et al.* (2015) Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.*, **47**, 822–826.
- Francioli, L.C. *et al.* (2017) A framework for the detection of de novo mutations in family-based sequencing data. *Eur. J. Hum. Genet.*, **25**, 227–233.
- Fromer, M. *et al.* (2014) De novo mutations in schizophrenia implicate synaptic networks. *Nature*, **506**, 179–184.
- Goya, R. *et al.* (2010) SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, **26**, 730–736.
- Haghighi, A. *et al.* (2018) An integrated clinical program and crowdsourcing strategy for genomic sequencing and Mendelian disease gene discovery. *NPJ Genom. Med.*, **3**, 21.
- Iossifov, I. *et al.* (2014) The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, **515**, 216–221.
- Kelsen, J.R. *et al.* (2018) *Sa2008* - Mutations in *Strxbp3* contribute to very early onset of IBD immunodeficiency and hearing loss. *Gastroenterology*, **154**, S–445.
- Kong, A. *et al.* (2012) Rate of de novo mutations and the importance of fathers age to disease risk. *Nature*, **488**, 471–475.
- Kurreck, J. and Stein, C.A. (2016) *Molecular Medicine: An Introduction*. Wiley-Blackwell, Hoboken, NJ, USA.
- O’Roak, B.J. *et al.* (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, **485**, 246–250.
- Peng, G. *et al.* (2013) Rare variant detection using family-based sequencing analysis. *Proc. Natl. Acad. Sci. USA*, **110**, 3985–3990.
- Peng, G. *et al.* (2014) FamSeq: a variant calling program for family-based sequencing data using graphics processing units. *PLoS Comput. Biol.*, **10**, e1003880.
- Poplin, R. *et al.* (2016) Creating a universal SNP and small indel variant caller with deep neural networks. *bioRxiv*. doi: 10.1101/092890.
- Sanders, S.J. *et al.* (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, **485**, 237–241.
- Zaidi, S. *et al.* (2013) De novo mutations in histone-modifying genes in congenital heart disease. *Nature*, **498**, 220–223.