Published in final edited form as:
*Nat Methods*. 2013 October ; 10(10): 985–987. doi:10.1038/nmeth.2611.

# DeNovoGear: *de novo* indel and point mutation discovery and phasing

**Avinash Ramu**[1,6], **Michiel J. Noordam**[1,6], **Rachel S. Schwartz**[2], **Arthur Wuster**[3], **Matthew E. Hurles**[3], **Reed A. Cartwright**[2,4], and **Donald F. Conrad**[1,5]

[1]Department of Genetics, Washington University School of Medicine, St. Louis, MO USA

[2]Center for Evolutionary Medicine and Informatics, The Biodesign Institute, Arizona State University, Tempe, AZ USA

[3]Genome Mutation and Genetic Disease Group, Wellcome Trust Sanger Institute, Cambridge, UK

[4]School of Life Sciences, Arizona State University, Tempe, AZ USA

[5]Department of Pathology & Immunology, Washington University School of Medicine, St. Louis, MO USA

## Abstract

We present the DeNovoGear software for analyzing *de novo* mutations from familial and somatic tissue sequencing data. DeNovoGear uses likelihood-based error modeling to reduce the false positive rate of mutation discovery in exome analysis, and fragment information to identify the parental origin of germline mutations. We used our program to create a whole-genome *de novo* indel callset with a 95% validation rate, producing a direct estimate of the human germline indel mutation rate.

*De novo* mutations (DNMs) are an important source of human morbidity and mortality, and their detection is fundamental to the study of genetics. Mapping the location of germline and somatic mutations is revolutionizing our ability to diagnose and understand numerous severe diseases. Today, entire genomes can be screened for DNM using short-read sequencing, but

Correspondence: dconrad@genetics.wustl.edu.
[6]These authors contributed equally to this work

**Author Contributions**
A.R. implemented methods, analyzed data and wrote the paper; M.J.N. performed validation experiments, analyzed data and wrote the paper; R. S. S. performed simulations; A. W. provided code and performed early analysis demonstrating the utility of beta-binomials; M.E.H. and R. A. C. gave conceptual advice, supervised the project and wrote the paper; D.F.C. designed and supervised the project, implemented methods, analyzed data and wrote the paper.

**Competing Financial Interests**
The authors declare no competing financial interests.

**Software Availability**
Source code for *DeNovoGear* is available for download from http://genetics.wustl.edu/dclab/software/. A package containing the program, documentation, and example data files will be hosted on the Nature Methods website along with this paper.

**Data Access**
The CEU-Trio BAM files are available by FTP ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20120117_ceu_trio_b37_decoy/ The sequencing data used to generate these BAM files are available as part of Sequencing Project SRP003680 from the Sequencing Read Archives (http://www.ncbi.nlm.nih.gov/sra)

per-base error rates in variant discovery are still orders of magnitude larger than the frequency of DNMs, making careful calling, filtering and validation of calls essential.

While we and others have shown that it is feasible to accurately detect *de novo* point mutations from whole genome sequencing (WGS) data, our ability to reliably detect *de novo* indels is much less certain [1–4]. In the recently published Phase I analyses of the 1000 genomes project, experimental validation of an initial indel call set yielded an estimated false discovery rate of 35% [5]. Even after more extensive filtering, the authors noted that 18% of indel sites yielded inconsistent or ambiguous results. These numbers signal the need for extensive experimental validation and do not engender enthusiasm for the prospects of *de novo* indel calling.

In what we call the "basic" approach to DNM detection, genotypes are called on one sample at a time, and DNMs are identified as incompatible genotype calls between samples (e.g. parent-offspring trios or matched tissues). Here, we describe an approach that greatly improves the accuracy and interpretation of *de novo* point mutations and indels compared to this "basic" approach, by jointly analyzing a set of samples in a unified model-based framework. The DeNovoGear model consists of individual genotype likelihoods, transmission probabilities, and priors on the probability of observing a polymorphism or a *de novo* mutation at any given site in the genome (Online **Methods**). The DeNovoGear framework allows the user to specify the prior probability of observing a DNM, which in principle can be used as a lever to increase or decrease calling sensitivity. We performed simulations to show that increasing the mutation rate prior increases detection sensitivity and that use of a prior helps control Type I error at low sequencing depth (Supplementary Note, Supplementary Figures 1 and 2, Supplementary Tables 1 and 2).

It has recently been shown that there is a striking overdispersion in the distribution of alternate read frequency in whole exome sequencing data, compared to expectations of the binomial model typically used for genotype calling [6]. Specifically, at a heterozygous SNP position, the observed variance in the proportion of non-reference reads, var($p$), may be much larger than the theoretical binomial variance of $p(1\text{-}p)/n$, where $p = 0.5$ for a germline heterozygote and $n$ is the depth of coverage for the site.

We fit beta-binomial distributions to alternate read frequencies for SNPs and indels called from trio exome sequencing data generated by the 1000 genomes project[7] (Online **Methods**). The beta-binomial distribution fits the variance of the read depth distribution better than both the Poisson and the binomial distributions, and the improvement in fit is more dramatic for indels compared to SNPs (Figure 1). Model parameters estimated from all three CEU trio exomes were highly similar, but these estimates are significantly different from those made from experiments performed at a different center on different DNA ($p<0.0001$, Likelihood Ratio Test), meaning that beta-binomial models are not necessarily portable between labs or protocols (Supplementary Table 3). We have implemented a beta-binomial-based ('BB') caller within DeNovoGear that allows beta-binomials to be fit for arbitrary input datasets (**Methods**). The BB caller reduces the number of false positive *de novo* SNV calls by over 50% while maintaining the same power to detect true DNMs as

when using binomial likelihoods (Supplementary Table 4). We discuss the application of the BB caller to indels below.

We are aware of three tools for model-based discovery of germline DNMs from next-generation sequencing [7–9]. We compared the performance of SamTools, PolyMutt, GATK and DeNovoGear relative to a naive strategy based on single-sample calling, using simulated data and two datasets generated from a single parent-offspring trio by from the 1000 genomes project (WES and WGS datasets, Online **Methods**).

Sensitivity and specificity were determined using previously generated validation results for these samples (Supplementary Figures 3 and 4, Supplementary Tables 4 and 5). In this era of rapid technological development, it is a conservative assumption that our historical validated callset from 2011 captures all true positives, and the optimal performance comparison would attempt validation on all predictions from all programs.

All callers exhibit close to 50% sensitivity for detecting validated germline DNMs on the WGS dataset. For validated somatic mutations, DeNovoGear (n=930 calls, 99% sensitivity) and GATK (n=921 calls) outperform Samtools (n=890) and Polymutt (n=878). The WGS false discovery rates vary widely across callers. The naive approach is clearly impractical, producing 144,424 DNMs. Samtools produces 235,134 DNM calls with probability > 0.5, and after converting to posterior probabilities, 111,142 DNMs with probability > 0.9 (but note that Samtools is more similar to PolyMutt and DenovoGear when considering ROC curves, Supplementary Figures 3 and 4). GATK calls an order of magnitude fewer events with 15,141 DNMs at probability > 0.9. At this threshold, PolyMutt is even more conservative, with 6,215 DNMs called, and DeNovoGear is the most conservative with 4,474 DNMs called.

On the WES dataset, using the beta-binomial model clearly separates DeNovoGear-BB from the other callers, including the original DeNovoGear using binomial likelihoods (Supplementary Table 4). While DeNovoGear makes 153 total calls with a posterior probability > 0.9, DeNovoGear-BB makes only 70 calls at the same threshold with no loss of power for germline DNMs and only one less somatic DNM call.

DeNovoGear called 369 candidate *de novo* indels with probability > 0.5 on the WGS data using Samtools genotype likelihoods and indel-specific priors (Online **Methods**, Supplementary Figures 5 and 6, Supplementary Table 6). After excluding 241 sites by filtering and 71 sites by visual inspection, we attempted to validate 9 insertions and 48 deletions by Sanger sequencing (Supplementary Tables 7–9, Supplementary Figures 7–11, Online Methods). Remarkably, we validated 53 of the 56 sites for which we could design assays (95%, 6 insertions and 47 deletions). *De novo* indel calling with the beta-binomial greatly reduces the number of false positive calls compared to the standard binomial; at a probability threshold of 0.5, DeNovoGear calls 34 indels and DeNovoGear-BB only calls one, a reduction of 97% (Supplementary Table 6). We also compared the performance of Samtools genotype likelihoods for indel DNM detection to likelihoods from DINDEL, another well known indel calling algorithm based on a profile HMM[10] (Supplementary Note). Our results suggest that DINDEL genotype likelihoods are conservative, in that they

underestimate the evidence for an indel when one is present, but this is balanced by a major increase in specificity. Forty-four (79%) of the 56 candidate DNMs from our Samtools analysis were also called as DNMs with DINDEL likelihoods when considering the WGS dataset; in contrast two-thirds of the false positives were no longer supported as DNMs.

In our analyses of validated point mutations in these cell lines we observed a ratio of 49 germline DNMs to 952 somatic DNMs[1]. Assuming that these proportions hold for indels, this would provide a direct estimate of the sex-averaged *de novo* indel rate of $1.06 \times 10^{-9}$ (95% CI: $2.35 \times 10^{-10}$ $2.75 \times 10^{-9}$) per-base per generation or 9.06 indels per 100 point-mutations (given that the point mutation rate in the CEU trio offspring has been estimated to be $1.17 \times 10^{-8}$) (Online **Methods**). This estimate of the indel mutation rate is consistent with prior estimates from phylogenetic comparisons ($1.42 \times 10^{-9}$) and from the sequencing of Mendelian disease genes ($0.78 \times 10^{-9}$)[11,12]. We explored the influence of our assumptions on the rate that we obtained, and concluded that it is unlikely that the true indel DNM rate for this trio differs by more than a factor of five from our estimate (Supplementary Note).

Homopolymers and short tandem repeats are highly unstable in eukaryotic genomes and are known to mutate at rates orders of magnitude higher than point mutations in repeat-free sequence[13,14]. We applied filters that removed homopolymers and tandem repeat regions using standard annotations[15] and as a result we are underestimating the true small indel *de novo* rate (Online **Methods**). Curiously, 31 (58%) of our validated mutations fell within tandem repeats (8) or homopolymers (23) that were unannotated by Tandem Repeats Finder. We term these repeats “microrepeats” to reflect their extremely small size, 2–6 bp in the case of homopolymers and 2–5 copies of 3–4 bp repeats in the case of tandem repeats. No 2 bp repeat mutations were observed. This result suggests that replication slippage, well-known to cause repeat polymorphism at larger tandem repeat loci, is operating at even the smallest possible tandem repeats, blurring the boundary between simple diallelic indels and tandem repeat polymorphism.

DeNovoGear implements a fragment-based phasing algorithm that can determine the parent of origin for some DNMs (Online **Methods**). In the WGS dataset we were able to phase 24% (12/49) of validated germline *de novo* point mutations, 21% (205/952) of validated somatic DNMs and 28.5% (16/56) of validated indel calls. We are actively developing DeNovoGear to improve calling performance by implementing new genotype likelihood models and extending the inheritance model to cover arbitrary pedigree structures. New genotype likelihood models will be useful for different mutation types (e.g. VNTRs and CNVs), frequencies (e.g. in mosaic situations), and sample preparations (e.g. single cells).

## Online Methods

### Datasets

BAM files were generated from whole-exome and whole genome sequencing data of the CEU trio (NA12878, NA12891, NA12892), freely available from the 1000 genomes project FTP server and described previously[7]. Reads were aligned to an augmented reference sequence based on GRCh37, which is being used for the Phase Two of the 1000 genomes

project and available from the 1000 genomes website. BAMs were processed with best practices including PCR duplicate removal, local indel realignment and base quality recalibration. The package BCFtools was used to create BCF-format genotype likelihoods from each BAM file[8].

## The basic DeNovoGear model

DeNovoGear uses a genealogical modeling framework that can be used to evaluate the joint likelihood of all genotypes in a pedigree. Currently the pedigree structure is limited to parent-offspring trio or matched sample pairs (e.g. tumor-normal or monozygotic twins).

In the case of a trio we use subscripts to indicate genotype from mother, father and child: $G_M$, $G_F$, and $G_C$. Then we write the joint likelihood for the trio as

$$L(G_M, G_F, G_C|D)\,\alpha\,P(D_M|G_M)\,P(D_F|G_F)\,P(D_C|G_C)\,x\,P(G_C|G_M, G_F)\,P(G_M, G_F, R)$$

where $P(G_M, G_F, R)$ is the prior of drawing two genotypes $G_M$ and $G_R$ from the population and observing the base present in the public reference genome sequence, $R$. This prior is loosely derived from the standard neutral coalescent, and is modulated by a user-defined value for the population mutation rate θ. $P(G_C \mid G_M, G_F)$ is the 'transmission' likelihood; the likelihood that the child's genotype is $G_C$ given the parents genotypes are $G_M$ and $G_F$. Within this function also lies another 'pseudo prior', the assumed probability of observing a *de novo* mutation, which is used for evaluating $P(G_C \mid G_M, G_F)$ for Mendelian incompatible trio configurations. For Mendelian incompatible configurations we assume the minimum number of mutations required. The terms $P(G_C \mid G_M, G_F)$ and $P(G_M, G_F, R)$ are pre-calculated for all possible trio configurations and contained in a lookup table that is used by DeNovoGear to greatly reduce run time. The individual genotype likelihoods, or $P(D|G)$'s, are provided as input to DeNovoGear, a feature that allows users to benefit from extremely specialized sequencing error models implemented by other packages. We note that the DeNovoGear beta-binomial caller described below also models sequencing error and could be used as a stand-alone DNM caller without input from other packages.

## Prior sensitivity analysis

In order to assess the performance of DeNovoGear for different prior values we ran DeNovoGear by setting the mutation rate prior from $10^{-4}$ to $10^{-12}$ mutations/bp in geometric increments of $10^{-2}$. Our results show that varying the mutation rate prior does have a dramatic effect on the sensitivity and specificity of DNM calling when using a standard whole-genome sequencing study design such as the one generating the WGS dataset (Supplementary Tables 1 and 2, Supplementary Figs. 1 and 2). The total number of false positive calls increases over 5-fold when moving from $10^{-12}$ to $10^{-4}$, while 879/939 (94%) of validated DNMs are detected at the smallest rate prior, and 100% sensitivity for germline DNMs is achieved at $10^{-8}$.

## Controlling type I errors at low sequencing depth

**Depth of coverage analysis—**We generated low coverage datasets (1–20×) from the validated dataset of Conrad et al., which consists of 3038 candidate *de novo* sites [1]. The requisite number of reads was randomly subsampled from the BAM files for each of the individuals in the CEU trio. We determined the number of the 48 validated autosomal germline and the 888 autosomal somatic *de novo* mutations that were found using each coverage level. We calculated sensitivity and specificity for each coverage level.

**Simulated data for performance comparison—**We simulated three datasets of 100 million sites each. One dataset consisted of parents and a child that were entirely monomorphic. In the second, one parent and the child were heterozygous for an inherited mutation, and in the third, the parents were monomorphic and the child was heterozygous for a *de novo* mutation. Ten-fold (10X) coverage reads were randomly generated for each dataset with an error rate of 0.005 and equal probability of each allele for heterozygotes. Data were analyzed using SamTools trio caller and DeNovoGear.

## Alternative genotype likelihoods: Beta-binomial caller

The beta-binomial distribution is parameterized by two variables, α and β, so the likelihood function for the homozygous reference class of sites could be written as

$$L_{RR}(D|RR)=\mathrm{Beta}(k+\alpha_{RR}, n-k+\beta_{RR})/\mathrm{Beta}(\alpha_{RR}, \beta_{RR})$$

where *n* is the total number of reads observed at a site, and *k* the number of those reads with the alternate allele. In our model we make the simplifying assumption that $L_{RR}$ and $L_{AA}$ are symmetric, that is $\alpha_{AA} = \beta_{RR}$ and $\alpha_{RR} = \beta_{AA}$. We consider these fitted BB distributions to be informative about sequencing error, but we only consider the possibility of two alleles at any given site for all genotype classes (RR, RA, AA). Therefore the data used for model fitting and likelihood calculation simply consists of the total read depth and count of the most common alternate allele. Training a beta-binomial ('BB') model is an iterative process performed one exome at a time. We conduct a first round of SNP genotype calling using a standard approach such as implemented by Samtools. We then fit beta-binomial distributions to a set of high confidence sites representing heterozygous or homozygous non-reference genotypes using maximum likelihood [16]. During our preliminary analyses we observed that the our fitted models for $L_{AA}$ and $L_{RR}$ provided a poor fit to sites where only one base was observed (e.g. all reads contained "A"), so for this class of sites we have hardcoded the likelihoods for these sites. For example, at a site with only reference reads the phred scaled likelihood function is: RR-0, RA-255, AA-255. It is this version of the BB model that we describe in the main text. Because the beta-binomial framework only considers two alleles, we implemented a simple filtering strategy to remove DNM calls at triallelic sites.

### Performance comparisons

We used the following packages for calling mutations: DeNovoGear 0.5, samtools version: 0.1.17, polymutt 0.0.4, DINDEL 1.01, and the trio-aware Bayesian caller of GATK 2.1–8. All packages were run with default settings. Samtools and Polymutt output likelihood ratio statistics of the form L(*de novo* mutation)/L(no *de novo* mutation) which we convert to posterior probabilities for comparison to DeNovoGear.

### Estimating Indel priors

It is well known that indel mutation rates are size dependent smaller indels are far more likely to form than large ones and that conditional on size, deletions occur more frequently than insertions [14]. We used the Watterson estimator to generate size-specific mutation rate estimates for insertions and deletions separately, using the indel callset from Phase I of the 1000 genomes project [17]. We next fit these mutation rate estimates to a log-linear model, to allow priors to be assigned to indels of arbitrary size (Supplementary Fig. 6). The prior function implemented by default for insertions is log(μ) = *c* * (−22.8689− (0.2994*insertion length)) and for deletions, log(μ) = *c* * (−21.9313− (0.2856*deletion length)), where in both cases *c* represents a scaling constant that can be altered by the user.

### Filtering and mutation rate calculation

With standard experimental design, it is thought that a large portion of false positive DNM calls is due to alignment error. Repeat-rich regions in particular are prone to both alignment and sequencing artifacts. We implement here a small number of filters to remove potentially artifactual indel DNM calls. We removed DNM calls that intersected the “Simple Repeats” and “Segmental Duplications” tracks downloaded from the UCSC genome browser (n=13 and n=109, respectively). We remove DNM calls that fall at sites of reported indel polymorphism found in dbSNP (n = 98), as such variants may be the result of undercalling indels in the parents, and we removed calls around CNVs known to exist in these cell lines (n=21)[18]. By visual inspection of sequencing read alignments, we identified three types of obvious artifacts that we removed manually (Supplementary Figures 7–9).

In order to produce an estimate of the per generation indel mutation rate for this trio, we used an equation that accounts for DNM discovery power (*p*; estimated at 95% from previous work), the proportion of validation assays we were able to attempt (*a*; 56/57), the proportion of validated DNM sites that segregate in the germline (*s*; estimated at 49/1001 in this cell line from previous work), the total number of validated DNMs (*d*; 53), and the number of bases we were able to effectively screen for DNM in this trio (*b*; 2,631,436,052). Then our rate estimate is

$$\widehat{\mu}=\frac{\left(s\times d\times\frac{1}{a}\right)}{b}\times\frac{1}{p}$$

### PCR validation of de novo indel calls

As described in the main text, we selected 57 potential *de novo* indel sites for validation using PCR amplification followed by Sanger sequencing. We performed 57 PCR assays on DNA samples from Coriell cell lines GM12878, GM12891, and GM12892 using a Biorad T100 Thermocycler 2.0. The primers and specific PCR conditions used for each assay are as described in Supplementary Table 8. In brief, the PCR conditions were 3 minutes at 95 °C, followed by 35 cycles of 1 minute at 95 °C, 1 minute at 58 or 61 °C and 20 or 60 seconds at 72 °C, ending with 1 cycle of 1 minute at 72 °C. The PCR mix consisted of 12.5 μl 2× PCR Master Mix (Cat No. M750B; Promega), 1.0 μl primer set (10 μM), 1.0 μl genomic DNA (50 ng per μl) and 10.5 μl dH20. The final reaction volume was 25 μl. All PCR products were run on a 3%, 1× TBE agarose gel to be analysed for size and subsequently sent off for Sanger sequencing (Genewiz).

### Haplotype phasing

DeNovoGear implements a fragment-based phasing algorithm that can determine the parent of origin for some DNMs. The phaser looks at reads or read pairs that cover both the *de novo* site and a phase-informative site that is close to the *de novo* site. The phasing routine produces counts of maternal and paternal variants observed on the same fragment as the *de novo* mutation. These counts should be directly interpretable in a qualitative manner (e.g. an observation of one paternal variant and no maternal variants indicates a paternal origin of the DNM). However, counts could also be included in a testing framework for count data to control for possible index switching, although such experimental artifacts should be rare.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

We thank R. Hardwick for assistance with primer design, V. Plagnol and H. Li for helpful discussion, and members of the 1000 genomes community for generating software, data, and resources that were used as part of this project. This research was supported in part by Wellcome Trust grant WT098051.

## References

1. Conrad DF, et al. Nature Genetics. 2011; 43:712–714. [PubMed: 21666693]
2. Roach JC, et al. Science. 2010; 328:636–639. [PubMed: 20220176]
3. Kong A, et al. Nature. 2012; 488:471–475. [PubMed: 22914163]
4. Cartwright RA, Hussin J, Keebler JE, Stone EA, Awadalla P. Stat Appl Genet Mol Biol. 2012; 11
5. Abecasis GR, et al. Nature. 2012; 491:56–65. [PubMed: 23128226]
6. Heinrich V, et al. Nucleic Acids Res. 2012; 40:2426–2431. [PubMed: 22127862]
7. DePristo MA, et al. Nat Genet. 2011; 43:491–498. [PubMed: 21478889]
8. Li H. Bioinformatics. 2011; 27:2987–2993. [PubMed: 21903627]
9. Li B, et al. PLoS Genet. 2012; 8:e1002944. [PubMed: 23055937]
10. Albers CA, et al. Genome Res. 2011; 21:961–973. [PubMed: 20980555]
11. Lynch M. Proc Natl Acad Sci U S A. 2010; 107:961–968. [PubMed: 20080596]
12. Lunter G. Bioinformatics. 2007; 23:i289–296. [PubMed: 17646308]
13. Lynch M, et al. Proc Natl Acad Sci U S A. 2008; 105:9272–9277. [PubMed: 18583475]

14. Kvikstad EM, Tyekucheva S, Chiaromonte F, Makova KD. PLoS Comput Biol. 2007; 3:1772–1782. [PubMed: 17941704]
15. Benson G. Nucleic Acids Res. 1999; 27:573–580. [PubMed: 9862982]
16. Smith DM. Appl Stat. 1983; 32:196–204.
17. Watterson GA. Theor Popul Biol. 1975; 7:256–276. [PubMed: 1145509]
18. Conrad D, et al. Nature. 2010; 464:704–712. [PubMed: 19812545]

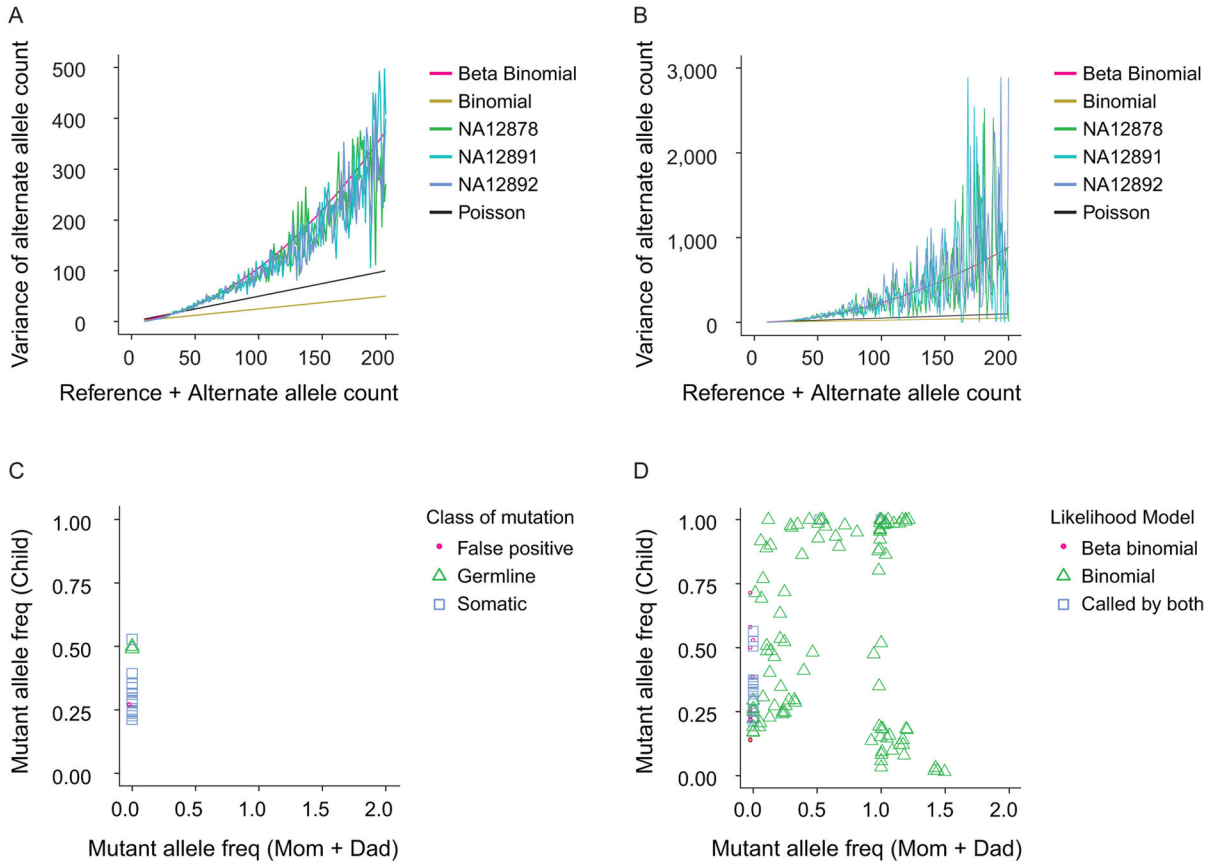


**Figure 1.**

Using beta-binomial ('BB') likelihoods to model exome data. A) We fit three functions to the read count data from the exomes of NA12878, NA12891, and NA12892, considering only high-confidence SNP sites. For each function, we plot the expected variance in the number of alternate alleles sampled as a function of read depth. One the same scale we show the observed variance for all three exome datasets. B) We performed the same analysis using indel sites and observed an even larger difference in fit between beta-binomial and binomial models than with SNPs. When we examine mutant allele frequencies in the CEU WES dataset at the sites called by using binomial (n=153), beta-binomial (n=66), or both models (n=40), we see that the BB model primarily reduces false positives by eliminating undercalling of heterozygotes in the parents. (C) Distribution of mutant allele frequencies for previously validated sites. On the x-axis, sites are positioned by the cumulative mutant read frequency in the parents. On the y-axis sites are positioned by the mutant read frequency in the trio offspring. Points are colored by validation status. (D) Distribution of mutant allele frequencies for sites called in this study and not validated previously. Sites are colored green if called only by binomial model, red if called only by BB, blue if called both models.