The distribution of allele frequencies at heterozygous genomic loci in next generation sequencing data sets

Verena Heinrich [*], **Jens Stange** [†], **Thorsten Dickhaus** [†] , **Peter Imkeller** [†] , **Ulrike Krüger** [*] , **Sebastian Bauer** [*] , **Stefan Mundlos** [*] , **Peter N. Robinson** [*] , **Jochen Hecht** [*] , and **Peter M. Krawitz** [*]

[*]Institute for Medical and Human Genetics, Charite Universitaetsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, and [†]Department of Mathematics, Humboldt-University Berlin, Unter den Linden 6, 10099 Berlin

Submitted to Proceedings of the National Academy of Sciences of the United States of America

A deeper understanding of the distribution of the variant call frequencies at heterozygous loci innext-generation sequencing (NGS) data sets is a prerequisite for sensitive variant detection. We model the crucial steps in a second generation sequencing protocol as a stochastic branching process and derive the expected distribution of alleles at heterozygous loci after measurement. We confirm our theoretical results by analyzing technical replicates of human exome data and demonstrate that the variance of allele frequencies at heterozygous loci is higher than expected by a simple binomial distribution. Due to this high variance, mutation callers relying on binomial distributed priors are less sensitive for heterozygous variants that deviate strongly from the expected mean frequency. Our results also indicate that error rates can be reduced to a greater degree by technical replicates than by increasing sequencing depth.

variant calling | next-generation sequencing | allele distribution | branching process

Second generation DNA sequencing has revolutionized many biomedical areas, it especially accelerated the discovery of disease genes in Medical Genetics [1, 2] and is now about to enter diagnostics [3]. In order to translate this technology into a reliably tool for molecular diagnostics for human genetics and other fields, it will be necessary to further reduce error rates of sequence variant detection. Understanding the process of how the high-throughput sequencing data arises is crucial for the development of sensitive genotype calling algorithms. It is well known in the field that especially the error rates in detecting heterozygous mutations in diploid genomes are still considerably higher than the comparable error rates of homozygous variants - even at high levels of sequence coverage [4, 5].

It is currently widely assumed that the frequencies of calls at heterozygous sites in NGS data is binomially distributed, an assumption that has been incorporated into many variant calling programs for NGS data [6, 7, 8]. We were motivated to question this assumption by observations of more extreme probability distributions in whole-exome sequencing (WES) data sets, as well will demonstrate in this paper. We therefore analyzed the key steps in NGS data generation from a stochastic point of view and identified the amplification of sequence fragments during library preparation before measurement as crucial for the distribution of allele frequencies at heterozygous genomic loci.

We reasoned that the generation of fragments can be described as a Bienayme-Galton-Watson branching process with discrete time steps, which is a model that has been widely used by physicists and mathematicians in population genetics [9, 10, 11]. In this work we provide a detailed description of the fragment amplification process. We then show that our model accurately reflects allele frequencies in real WES datasets. One prediction of our model is that technical replication is more effective in reducing error rates than merely sequencing more reads from the same library, which we confirmed on a data set with nine technical WES replicates. Our results have important implications for understanding the causes of false-negative errors in NGS diagnostics.

Results

Fragment Amplification as a stochastic branching process. Suppose that we have a tube that initially contains a set of different alleles such as illustrated in Fig. 1A. We now perform K cycles of a polymerase chain reaction (PCR) on these alleles, which basically means adding a certain number of copies of these alleles to the tube at discrete time steps. This is an essential part of current NGS library preparation protocols that are used to enrich adapter-ligated DNA fragments [12].

We will start by describing this process for a heterozygous, single nucleotide position in the genome as an inhomogeneous Markov chain (although we deal in this paper only with biallelic single nucleotide polymorphisms (SNPs) the process generalizes to all sequence variants). The preparation of a genomic DNA sample starts by shearing the chromosomal DNA into sequence fragments of a few hundred base pairs. We will discuss in the following only fragments that contain a variable base of a SNP, which means we can distinguish between two possible classes of fragments, those containing the base of allele a_1 and those that contain the base of allele a_2. We consider the fragmentation as random and unbiased. This means the extensions into both directions from the variable position is uniform and only limited by fragment size. We also assume that the numbers n_1 and n_2 of the fragments containing allele a_1 and a_2 are equal after fragmentation, as the DNA originates from many cells of a single diploid genome. Before sequencing (at time step $k = 0$), adaptor oligomers are ligated to the fragments and a PCR is run for K cycles. For successful amplification, adaptors must be attached to both ends of the fragment. The initial number of amplifiable fragments, $n_1(0)$ and $n_2(0)$, is in the order of dozens. For each such fragment the attachment of the polymerase to the adaptor is a prerequisite for amplification. We assume that the probability of this event depends only on the total number of polymerase molecules, which remains the same in every PCR cycle k, and

Reserved for Publication Footnotes

the sum of amplifiable fragments, $n_1(k) + n_2(k)$, but is independent of the variant itself. For not too large K we may assume that polymerase is always in excess of $n_1(k)+n_2(k)$, and thus a constant fraction of fragments will be bound by polymerase. We will use the parameter p in the main manuscript to describe the cycle and allele-independent probability that a fragment is copied (in the supporting information we perform the calculations for allele specific amplification probabilities, p_1 and p_2). We now describe the probabilities of the three possible transitions of a random allele in PCR cycle k:

$$
\begin{aligned}
Pr((n_1(k), n_2(k)) \rightarrow (n_1(k)+1, n_2(k)) &= \frac{n_1(k)}{n_1(k)+n_2(k)}p \\
Pr((n_1(k), n_2(k)) \rightarrow (n_1(k), n_2(k)+1) &= \frac{n_2(k)}{n_1(k)+n_2(k)}p \\
Pr((n_1(k), n_2(k)) \rightarrow (n_1(k), n_2(k)) &= 1-p
\end{aligned}
\qquad [1]
$$

The whole system thus transitions to:

$$(n_1(k+1), n_2(k+1)) = (n_1(k) + b_1(k), n_2(k) + b_2(k)) \qquad [2]$$

where $(b_1(k), b_2(k))$ are binomially distributed random variables $B(n_1(k), p)$ and $B(n_2(k), p)$ (see Fig 1 A).

The quotient $\frac{n_1(k)}{n_1(k)+n_2(k)}$ describes the proportion of allele a_1 after the k^{th} amplification cycle and this is the allele frequency that we *expect* to measure by sequencing multiple read fragments of this pool (note, that sequencing itself will contribute to the totally measured variance. Sequencing itself may be understood as a random sample of finite size - which is the sequencing depth - on the allele pool after amplification). We are thus primarily interested in the distribution of the random variable $Q(k)$ describing the ratio of alleles after amplification. All transitions meet the Markov condition, stating that the distribution of alleles after step k solely depends on the distribution of alleles in step $k-1$:

$$
\begin{aligned}
&P((n_1(k), n_2(k))|(n_1(k-1), n_2(k-1)) \\
&(n_1(k-2)n_2(k-2)), ..., (n_1(0), n_2(0))) = \\
&= P((n_1(k), n_2(k)))|(n_1(k-1), n_2(k-1))),
\end{aligned}
\qquad [3]
$$

The entire process is determined by the offspring of a probability generating function h and $Q(k)$ approaches a normal distribution [10]. We derived the first and second moments of the offspring distribution (see Supporting Information for a detailed calculus) to compute the variance of $Q(k)$:

$$Var\,(Q(k)) = \frac{2(1+p)^{-1} - (1+p_i)^{-k-1} + (1+p)^{-k} - 1}{8N} \qquad [4]$$

with $N = n_1(0) = n_2(0)$

According to a standard NGS protocol, we simulated the amplification process of our model depicted in Fig.1A for $K = [1, 30]$, $N = [1, 25]$, for p ranging from 0 to 1 and a sequencing depth of 20x. We computed the variance of the resulting allele frequency quotient for 10,000 SNPs (Fig.1B) which is the expected order of magnitude for heterozygous variant calls in a human exome. The behavior of the variances sampled from our simulations is well described by function [4] adapted by the additional contribution of variance introduced by sequencing. For fixed p and N the variance increases with a growing number of PCR cycles K and approaches a constant level for $K > 15$. For fixed K and N the variance has its maximum around $p = 0.2$ and decreases for p towards 1. This is clear as with perfect amplification, we expect the initial ratio of $\frac{n_1(0)}{n_1(0)+n_2(0)} \approx 0.5$ to remain constant. For fixed K and p the variance decreases with an increasing number of alleles before amplification. Intuitively speaking, its easier for one allele to gain predominance in the pool that is sequenced if the initial allele set is small, the amplification efficiency is low and enough PCR cycles are run.

High Variance of Heterozygous Allele Frequencies in real human exome data sets. After modeling the amplification step as stochastic process, we analyzed the distribution of allele frequencies at heterozygous genomic loci in real human exome data that were generated following a standard protocol with 18 PCR amplification cycles. In order to compare the empirically measured frequencies with our simulated data all heterozygous SNP positions that were covered by more than 20 reads, were downsampled to 20 reads per position. The allele frequencies were derived from these read sets. The variance of the measured reference allele distribution is 0.017 and thus markedly larger, than the variance of 0.012 that is expected for hypothetical sequencing before amplification (this is the variance of a Binomial distribution where n represents the sequencing depth and the success parameter is the ratio of the alleles in the starting solution, Fig. 2A). Thus, the sequence fragments in a short read alignment, on which the variant call is performed, are not properly represented by a random sample of the initial ratio of $\frac{n_1(0)}{n_1(0)+n_2(0)}$, but the effect of the amplification process on this distribution has to taken into account.

Our model assumes a constant amplification efficiency over all PCR cycles, which seems to be a reasonable simplification given the relatively low number of PCR cycles used in NGS library preparation protocols. A value of $p \in [0.3, 0.5]$ yielded a variance for the allele frequencies, that is close to the value determined on the real exome data (Fig. 1B and 2A). We measured the amount of fragmented DNA used as input in our WES experiments at $k = 0$ ($5ng$) and measured about $5 - 10\mu g$ after $k = 18$ cycles of amplification. This corresponds to an amplification by a factor of $1 - 2 \cdot 10^3$, and thus values of $p \in [0.3, 0.5]$ are realistic.

As already discussed, with fixed p and N the variation is approaching a limit for increasing K and for $K > 15$ it hardly changes. To check this experimentally we sequenced the exome of the same individual that was amplified with 36 PCR cycles instead of 18. As expected, no significant increase in the variance could be detected (Suppl. Fig. 2B).

Influence of allele frequency variance on error rates in heterozygous variant detection. Assuming comparable read qualities, the variant call is based on a random sample drawn from the set consisting of all alleles a_1 and a_2 after amplification which is of size $n_1(k) + n_2(k)$. The coverage or sequencing depth at a variant site is equivalent to the size of the random sample that the call is based on. We hypothesized that a certain rate of true heterozygous alleles will not be called due to the high variance in allele frequencies after amplification (i.e., false-negative call). To test this, we generated nine exome replicates of the same individual and classified genomic loci as heterozygous if they were called heterozygous in at least six out of nine replicates by two accepted calling algorithms (see Material and Methods). Fig. 3A shows the common polymorphism rs539412, that was called as heterozygous variant in the first four replicates, but failed to be called as heterozygous variant in the fifth replicate due to low frequency. Using this as a gold standard, we then measured the false-negative rate for calls based on each of the single WES datasets. Over the whole exome we measured a false negative rate between 1% and 3% depending on the coverage and the calling algorithm

(Fig 3B). In a usual exome one expects between 10,000-15,000 heterozygous variants. Our results indicate that one will miss around a hundred heterozygous variants by sequencing an exome just a single time simply due to the stochastic fluctuation of the allele frequencies after amplification. Furthermore a variant calling approach that is simply based on a heterozygous allele frequency interval f with $[14\% < f < 86\%]$, as suggested in [15], is more sensitive than a more sophisticated variant calling algorithm that uses the *wrong* prior distribution for the allele frequencies independent of the coverage (Fig 3B). Additionally for a sequencing depth above 30x the false negative rate does not decrease further. Thus, once a sufficient sequencing depth has been reached, only technical replication is able to further reduce the total error rates in a substantial fashion.

Discussion

In this work we studied the distribution of alleles at heterozygous genomic positions as measured in next generation sequencing data sets. A solid knowledge of distribution and variance of allele calls at heterozygous loci is important as it is essential prior information for many variant calling approaches. In this work, we have demonstrated that the amplification step contributes significantly to the total variance of this distribution. We demonstrated that the amplification step contributes significantly to the total variance of allele frequencies. We modeled the fragment generation process as a Bienayme-Galton-Watson branching process and showed that the variance is accurately described by equation (4). For typical values of the efficiency of the amplification process (p) and sequencing depth, this is substantially higher than the variance of the binomial distribution (Fig. 2A). Clearly, the higher the variance of allele calls at heterozygous loci, the higher the false negative error will be.

From our analytical results one may draw some conclusions about how to reduce the stochastic fluctuactions coming from the amplification step: Increasing the efficiency of the adaptor ligation (which is increasing N), increasing p and reducing the number of PCR cycles K in a second generation protocol will all help to reduce the variance of heterozygous alleles. Ultimately calling errors arising from stochastic events during library preparation and fragment amplification will be avoided in single molecule sequencing techniques of the future [19].

Next-generation technologies such as whole-exome and genome sequencing are beginning to be used for diagnostic purposes. In this setting, it is critical to provide an estimation of the false-negative rate of the methodology. Clearly, it is important to report the regions of the exome that are not sufficiently covered for reliable variant calling. Our results suggest that it is also important to evaluate the variance at heterozygous SNP positions as it might serve as an indicator for the overall false-negative error rate in an experiment.

Materials and Methods

Exome Sequencing and Variant Detection. Human blood or tissue samples of 17 anonymized donors were used for exome sequencing. For one of these individuals nine technical replicates were generated. This means nine independent samples of the same individual were collected and further processed independently. For each sample genomic DNA was enriched for the target region of all human CCDS exons [13] with Agilent's SureSelect Human All Exon Kit and subsequently sequenced on a Illumina Genome Analzyer II with 100bp single end reads. The enrichment of adapter-modified DNA fragments before sequencing includes an amplification step of 18 PCR cycles in the standard protocol. For one exome 36 cycles of PCR were run to analyze the effect of the cycle number onto the allele frequency distribution. For The raw data of $\approx$ 5 GB per exome was mapped to the haploid human reference sequence hg19 with novoalign [14] resulting in a mean coverage of the exome target region of 50x. In this study heterozygous sequence variant detection was restricted to positions of high human variability as defined by dbSNP132 positions, in order to decrease the probability of false positive calling errors. A genomic position was called as a heterozygous variant, if more than 20 sequence reads covered this position in the reference based sequence alignment and if the ratio of the alternating allele to the sum of the alternating allele and the reference allele was in between 0.14 and 0.86. This heterozygous detection algorithm was shown to be highly sensitive for a coverage above 20 [15]. For the replicates we classified a locus as truly heterozygous, if it was classified as heterozygous by the above described calling criterion and by samtools [18] in at least six out of nine replicates.

Heterozygous Allele Frequencies. The reference allele frequency at a genomic position that was classified as heterozygous as described above is defined as the number of fragments that map to this position, cover the variable base and show the reference allele, divided by all fragments covering this site. There are two well known biases that shift the detected mean reference allele frequency from the expected value of 0.5 to slightly higher values but do not influence the variance of the distribution: SureSelect baits that were used for exon enrichment are designed as 120 bp antisense oligonucleotides to the haploid reference sequence of the latest Human Genome Build. This means DNA hybridization between sample DNA fragments containing common variants, that differ from the reference sequence, may be weaker as compared to hybrids without mismatches. This may lead to a slightly more effective enrichment of sequence fragments containing the reference allele. After sequencing, all short sequence reads are mapped to the haploid reference sequence. Sequence fragments containing non-reference allele variants have a lower mapping quality. For short read lengths, reads with low base quality, and low sequence complexity, this may result in a slightly reduced mapping ratio of non-reference allele fragments [16, 17]. Due to these **in vitro** as well as **in silico** biases, the detected mean reference allele frequency was shifted from 0.5 to 0.54 in our analyzed exome data sets.

Distributions of Heterozygous Allele Frequencies are position- and individual-independent. The dependency of the allele frequency distribution on genomic position as well as on the individual was tested on human exome data sets. Position dependency was tested by comparing the distribution of all heterozygous allele frequencies in an individual to a smaller random subset of these positions (see supporting Fig S2). The comparison between these distributions did not show significant differences by Chi-square testing. The dependency on the individual was tested by comparing the differences of heterozygous allele distributions between different individuals and technical replicates of the same individual. The difference in frequency distributions between different individuals is not significant and fluctuations in these distributions are comparable to those observed in technical replicates of the same individual.

ACKNOWLEDGMENTS. author contributions: P.M.K. and J.H. designed research; P.M.K., V.H., U.K., J.H., S.M., P.N.R., P.I., T.D., J.S performed research, analyzed the data, P.M.K and P.N.R. wrote the paper. We acknowledge that the conclusion of this paper about the value of technical replicates was obvious to our lab technicians even before starting this project.

1. Ng SB, et al. (2010) Exome sequencing identifies the cause of a mendelian disorder Nat Genet 42:30-5
2. Robinson PN, Krawitz P, Mundlos S (2011) Strategies for Exome and Genome Sequence Data Analysis in Desisease Gene Discovery Projects Clin Genet doi: 10.1111/j.1399-0004.2011.01713
3. Choi M, et al. (2009) Genetic diangosis by whole exome capture and massively parallel DNA sequencing Prox Natl Acad Scie USA 106:19096-101
4. Nothnagel M, et al. (2010) Technology specific error signatures in the 1000 Genomes Project data Hum Genet doi:10.1007/s00439-011-0971-3)
5. Harismendy O, et al. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies Genome Biol doi:10.1186/gb-2009-10-3-r32
6. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores Genomes Research 18:1851-8
7. Li R, et al. (2008) SOAP: short oligonucleotide alignment program Bioinformatics 24:713-4.
8. Goya R, et al. (2010) SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors Bioinformatics 26:730-6
9. Athreya KB, Ney PE (1972), Branching Processes, Springer
10. Yakovlev AY, Yanev NM (2009) Relative Frequencies in Multitype Branching Processes Tha Annals of Applied Probability 19:1-14
11. Polya G,Szegö G (1970) Problems and Theorems in Analysis I Springer

12. Bentley DR, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry Nature 456:53-59.

13. Pruitt R, et al. (2009) The consensus coding sequence(CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Research 19:1316-1323

14. Hercus C, et al. (2011) Novoalign V2.07, www.novocraft.com

15. Bell CJ, et al. (2011) Carrier testing for severe childhood recessive disease by next generation sequencing Science Translational Medicine 3:64-69

16. Degner JF, et al. (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data Bioinformatics 25:3207-3212

17. Krawitz, et al. (2010) Microindel detection in short-read sequence data Bioinformatics 26:722-729

18. Li H, et al. (2010) The Sequence Alignment/Map format and SAMtools Bioinformatics 25:2078-9

19. Timp W, et al. (2010) Nanopore Sequencing: Electrical Measurements of the Code of Life IEEE Trans Nanotechnol 9:281-294

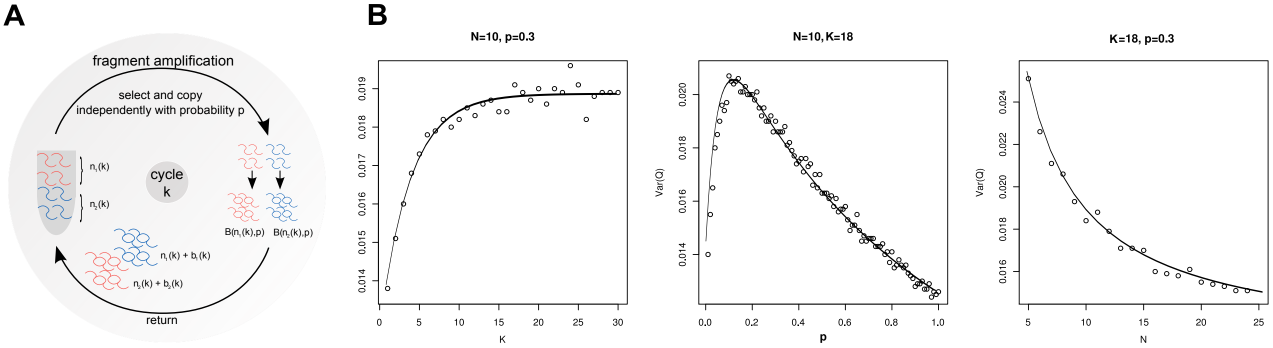

Fig. 1. A) **The amplification of heterozygous alleles before sequencing is similar to drawing balls from an urn and replacing these balls by adding some additional balls of the same color–diesen Satz wohl loeschen, er macht ohne Polya nicht mehr soviel Sinn!**. The distribution of the allele frequencies depends on parameters p that represents the efficiency of the PCR reaction and the probability that an allele is amplified, the cycle number K, and on the initial number of alleles N B) The variance of the allele frequency after amplification was sampled from simulations for p ranging from 0 (no amplification) to 1 (perfect duplication in each PCR cycle), for different cycle numbers K and numbers of starting alleles N. The measurement process of sequencing was simulated for a read coverage of 20x. The variance sampled from 10,000 simulated heterozygous SNPs and depicted as black circles (o), is well approximated by the analytical results of eq. 4 (black line). For a cycle number of $K > 20$ the variance does not change significantly. The variance reaches its maximum for an amplification probability around p=0.3. For an increasing number of alleles before amplification the variance approximates a fixed level, explained solely by the variance introduced by the measurement process of sequencing.

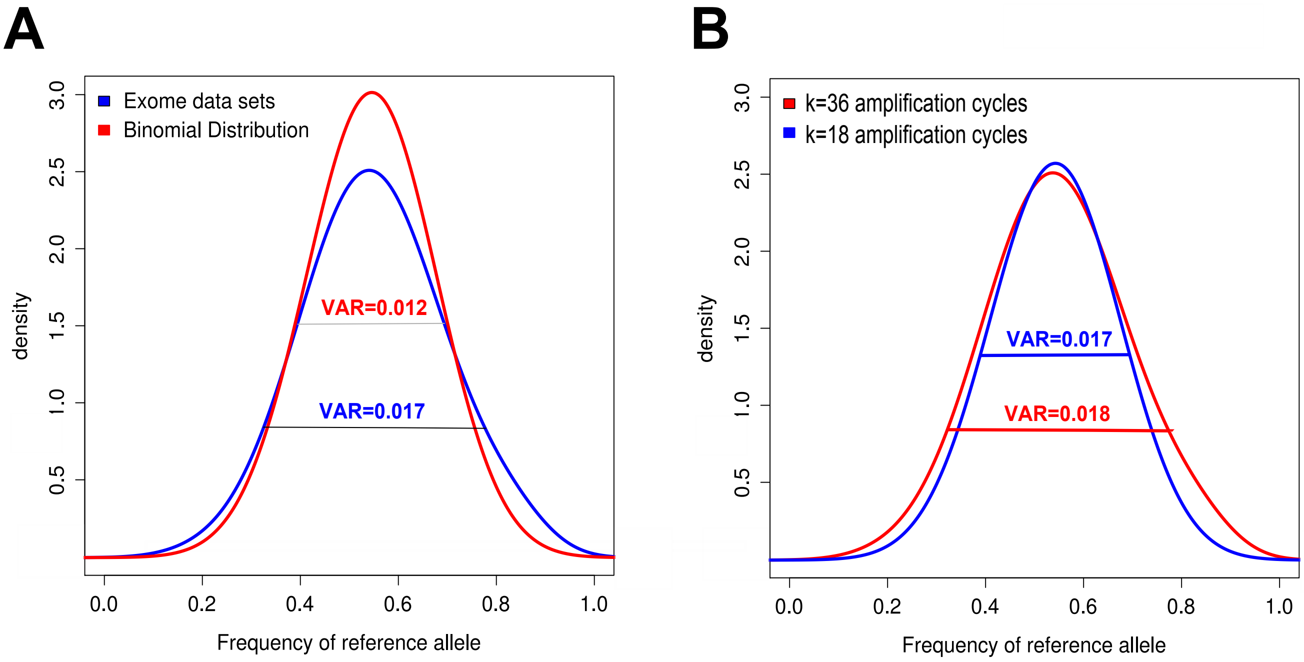

Fig. 2. Variance of the measured allele frequency at heterozygous genomic positions in NGS exome data sets A) The distribution of heterozygous allele frequencies measured in exome data sets at 20x coverage (blue) compared to the theoretical distribution expected before amplification (red). The variance of the real distribution after amplification is significantly larger. B) An exome of the same individual was sequenced following 18 and 36 cycles of amplification. As expected from theory the variance of the allele frequencies only sightly increases after the additional 18 cycles of amplification.

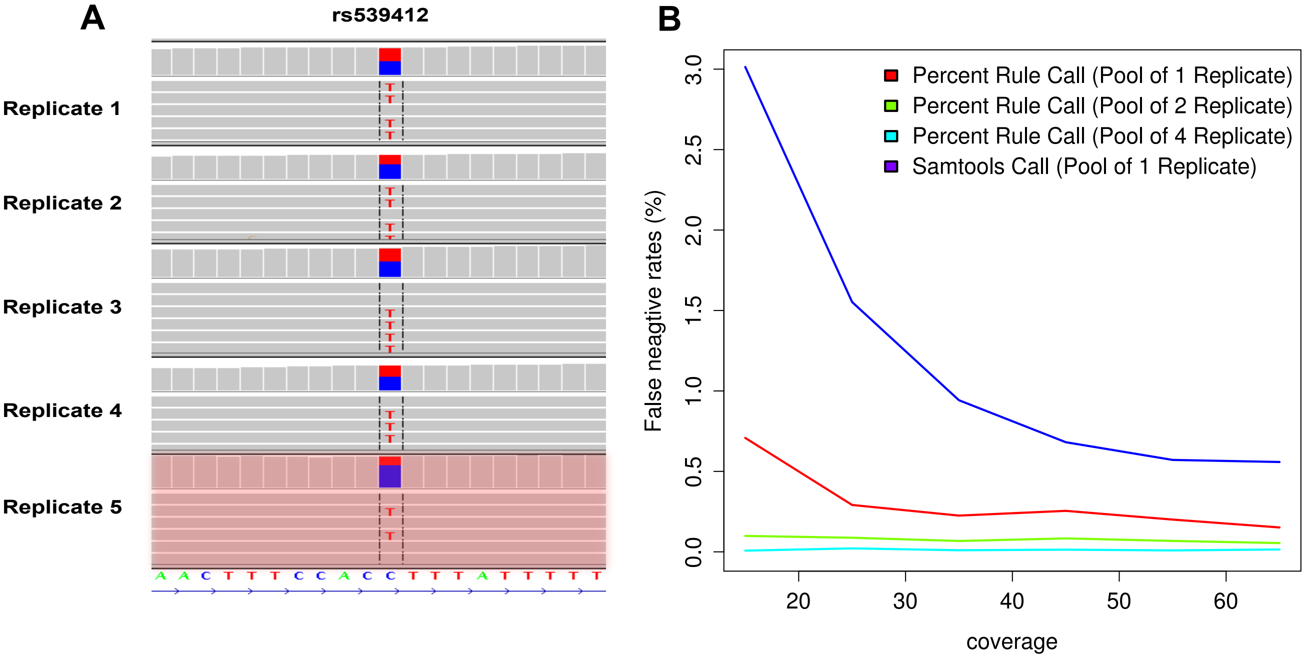

Fig. 3. **Influence of variance in measured allele frequency on variant calling** A) The genotype at the SNP position rs539412 has been called as heterozygous variant in the first four replicates, but was not detected in the fifth replicate due to low frequency B) The false negative error rate decreases with increasing sequencing depth. At low total sequencing depth the error rate is markedly reduced by considering pools of technical replicates. The classification of a genotype as heterozygous based on a simple frequency intervall (heterozygous if alternating allele frequency is inbetween 14% and 86%) is more sensitive than a calling algorithm that uses a binomial prior distribution as default setting for the allele distribution.