Supplementary Data

Consider an amplification mechanism with \( n_1(k) \) fragments of allele type \( A_1 \) and \( n_2(k) \) fragments of allele type \( A_2 \) after cycle number \( k \geq 0 \), where \( n_1(0) \) and \( n_2(0) \) denote the initial numbers of fragments. At each cycle, the fragments are copied with allele specific amplification probabilities \( p_1 \) and \( p_2 \). We use the single parameter \( p = p_1 = p_2 \) if the amplification is allele independent. If the number of fragments of allele type \( A_1 \) and allele type \( A_2 \) at cycle \( k + 1 \) is understood as the next generation of one parent generation \( k \), the distribution of descendants can be described as a Galton-Watson process, cf., e. g., [1].

To formalize the process mathematically, let for allele types \( i \in \{1, 2\} \) \( (\xi_{n,k}^i)_{n,k \in \mathbb{N}} \) denote two independent triangular arrays of stochastically independent random variables, each with identical Bernoulli-type distribution with parameter \( p_i \) respectively, such that

\[
P(\xi_{n,k}^i = 2) = p_i = 1 - P(\xi_{n,k}^i = 1).
\]

The parameter \( p_i, i = 1, 2 \), therefore reflects the probability that amplification of one fragment is successful. The index \( n \) ranges from 1 to \( n_i(k-1) \) in cycle \( k \).

This results in a two-type Bienaymé-Galton-Watson branching process \((Z_1^k, Z_2^k)_{k \geq 0}\) for the number of alleles of types \( A_1 \) and \( A_2 \) after \( k \) cycles, with initial values \( n_i(0) \) and recursive definition

\[
Z_i^k = \xi_{1,k}^i + \cdots + \xi_{Z_{k-1}^i,k}^i = \sum_{j=1}^{Z_{k-1}^i} \xi_{j,k}^i.
\]

Our statistical quantity of interest, the proportion of alleles of type \( A_i \) after \( k \) PCR cycles, can consequently be written as

\[
Q_i^k = \frac{Z_i^k}{Z_1^k + Z_2^k}.
\]

In the remainder, we aim at approximating the variance of \( Q_i^k \) for a fixed given total number \( k \) of PCR cycles, assuming \( n_1(0) \approx n_2(0) \approx N \), where \( N \) is assumed to be large. To this end, we make use of a central limit theorem for \( Q_i^k \) proven in [2] of the form

\[
(m_1(k) + m_2(k))\sqrt{N}(Q_i^k - q^i(k)) \xrightarrow{D} Y^i,
\]

where

\[
m_i(k) := E[Z_i^k|Z_0^i = 1] \quad \text{and} \quad q^i(k) = \frac{m_i(k)}{m_1(k) + m_2(k)} , \quad i = 1, 2.
\]
The limiting variable $Y^*$ is normally distributed with mean 0 and variance

$$a_i(k)^2 = \sigma_i^2(k)(1 - q_i(k))^2 + \sigma_i^2(k)(q_i(k))^2,$$

where $\sigma_i^2(k) = \text{Var}[Z_i^k|Z_0^i = 1]$, $i = 1, 2$.

It remains to calculate $m_1(k), m_2(k), \sigma_1^2(k),$ and $\sigma_2^2(k)$. For this, it is convenient to consider the probability generating function (pgf) of the branching process $(Z_k, Z_k^2)_{k \geq 1}$, given by

$$F_k(s_1, s_2) = \mathbb{E}[s_1^{Z_1^k} s_2^{Z_2^k} | Z_0^1 = n_1(0), Z_0^2 = n_2(0)],$$

where $n_i(0), i = 1, 2$, are the given fixed initial values. Due to independence of the two processes $Z_1^k$ and $Z_2^k$, it suffices to consider the marginal pgf for one allele type, say $A_1$, which we denote by $\tilde{F}_k$ in the remainder. Making use of (2), we obtain

$$\tilde{F}_k(s) = \mathbb{E}[s^{Z_1^k}|Z_0^1 = n_1(0)] = \mathbb{E}[\mathbb{E}[s^{Z_1^k}|Z_{k-1}^1]|Z_0^1 = n_1(0)]$$

$$= \mathbb{E}[\mathbb{E}[s^{Z_1^k}|Z_{k-1}^1]|Z_0^1 = n_1(0)].$$

By (1) we get for any $|s| \leq 1$ that $\mathbb{E}(s^{|Z_1^1}|) = s(1 - p_1) + s^2p_1$ and stochastic independence of the $\xi$’s leads to $\tilde{F}_k(s) = \mathbb{E}[(s(1 - p_1) + s^2p_1)Z_{k-1}^1|Z_0^1 = n_1(0)] = \mathbb{E}[\varphi(s)Z_{k-1}^1|Z_0^1 = n_1(0)]$ by setting $\varphi(s) = s(1 - p_1) + s^2p_1$. With the composition $\varphi^1 = \varphi$ and $\varphi^k = \varphi \circ \varphi^{k-1}, k \geq 2$, it follows

$$\tilde{F}_k(s) = \mathbb{E}[\varphi \circ \varphi^{k-1}(s)Z_0^1 = n_1(0)] = (\varphi^k(s))^{n_1(0)}.$$

Since $m_1(k) = \lim_{s \uparrow 1} \frac{\partial}{\partial s} \tilde{F}_k(s)$ and $\sigma_i^2(k) = \lim_{s \uparrow 1} \frac{\partial^2}{\partial s^2} \tilde{F}_k(s) + m_1(k) - (m_1(k))^2$ for the special case $n_1(0) \equiv 1$ and $\varphi^k$ is a smooth function, we have to calculate the first two derivatives of $\varphi^k(s)$ with respect to $s$ in 1.

First, we notice that $\varphi(1) = 1 - p_1 + p_1 = 1$ and, consequently, $\varphi^k(1) = 1$ for all $k \geq 1$. Moreover, the first derivative of $\varphi$ with respect to $s$ is given by $\varphi'(s) = (1 - p_1) + 2sp_1$ leading to $\varphi'(1) = 1 + p_1$, and the second derivative is given by $\varphi''(s) = 2p_1$. Now, application of the chain rule yields

$$\frac{\partial}{\partial s} \varphi^k(s) = \frac{\partial}{\partial s} (\varphi(\varphi^{k-1}(s))) = \varphi'(\varphi^{k-1}(s)) \cdot (\varphi(s))^{k-1}$$

and iterating (4) leads to

$$\frac{\partial}{\partial s} \varphi^k(s)|_{s=1} = (\varphi'(1))^k = (1 + p_1)^k = m_1(k).$$

For calculating $\frac{\partial^2}{\partial s^2} \tilde{F}_k(s)|_{s=1}$, we utilize the generalized chain rule by Faà di Bruno (see [4]) for
second derivatives, given by
\[
\frac{\partial^2}{\partial t^2} (f \circ g)(t) = f''(g(t))(g'(t))^2 + f'(g(t))g''(t).
\]

For the composition \( \varphi^k = \varphi \circ \varphi^{k-1} \), the latter leads to
\[
\frac{\partial^2}{\partial s^2} (\varphi^k(s)) = \varphi''(\varphi^{k-1}(s)) \left( \frac{\partial}{\partial s} \varphi^{k-1}(s) \right)^2 + \varphi'(\varphi^{k-1}(s)) \left( \frac{\partial^2}{\partial s^2} (\varphi^{k-1}(s)) \right).
\] (6)

Iterative application of (6) and plugging in of (5) yields
\[
\frac{\partial^2}{\partial s^2} (\varphi^k(s)) \bigg|_{s=1} = \sum_{\ell=0}^{k-1} \varphi''(1)(\varphi'(1))^\ell(\varphi'(1))^{2(k-\ell-1)}
\]
\[
= \varphi''(1)(\varphi'(1))^{2k-2} \sum_{\ell=0}^{k-1} (\varphi'(1))^{-\ell}
\]
\[
= 2p_1(1 + p_1)^{2(k-1)} \sum_{\ell=0}^{k-1} (1 + p_1)^{-\ell}
\]
\[
= 2 \left[ (1 + p_1)^{2k-1} - (1 + p_1)^{k-1} \right]
\]
and we get
\[
\sigma_1^2(k) = \frac{\partial^2}{\partial s^2} (\varphi^k(s)) \bigg|_{s=1} + m_1(k) - m_1(k)^2
\]
\[
= (1 + p_1)^{2k} \left[ 2(1 + p_1)^{-1} - 2(1 + p_1)^{-k-1} + (1 + p_1)^{k-1} - 1 \right].
\]

The variance of the asymptotic normal distribution of \( Q_k^1 \) is finally given by
\[
\text{Var}(Q_k^1) = \frac{a_1(k)^2}{N(m_1(k) + m_2(k))^2}
\] (7)
with (see (3))
\[
a_1(k)^2 = \sigma_1^2(k) \left( 1 - q^1(k) \right)^2 + \sigma_2^2(k) \left( q^1(k) \right)^2
\]
\[
= \sigma_1^2(k) \left( \frac{(1 + p_2)^k}{(1 + p_1)^k + (1 + p_2)^k} \right)^2 + \sigma_2^2(k) \left( \frac{(1 + p_1)^k}{(1 + p_1)^k + (1 + p_2)^k} \right)^2.
\]

For the special case \( p_1 = p_2 = p \), we get that \( m_1(k) = m_2(k), q^1(k) = q^2(k) = 1/2 \) and \( \sigma_1^2(k) = \sigma_2^2(k) = \sigma^2(k) \) for all \( k \geq 1 \). Therefore, the asymptotic distributions of \( Q_k^1 \) and \( Q_k^2 \) coincide.
in this case if \( n_1(0) \approx n_2(0) \). The asymptotic variance of \( Q_1^k \) and \( Q_2^k \), respectively, in case of \( p_1 = p_2 = p \) and \( n_1(0) \approx n_2(0) \) is given by considering
\[
a(k)^2 = \frac{1}{2} \sigma^2(k) = \frac{1}{2} (1 + p) 2^k \left[ 2(1 + p)^{-1} - 2(1 + p)^{-k-1} + (1 + p)^{-k} - 1 \right]
\]
and utilization of (7). For \( i = 1, 2 \), we then get
\[
\text{Var}(Q_i^k) = \frac{(1 + p)^{2k} \left[ 2(1 + p)^{-1} - 2(1 + p)^{-k-1} + (1 + p)^{-k} - 1 \right]}{4N (1 + p)^{2k}} \frac{1}{8N}.
\]

Supplementary Figure 1: The variance of the allele frequency was sampled from simulations for \( p \) ranging from 0 to 1 (o) as well as analytically derived (eq. 8) (black line) for different initial values of \( N \).

Supplementary Table 1: False Negatives and Positive Rates (FNR, FPR) for genotype calling depending on different categories of read coverage. Default settings were used for samtools variant calling. The genotype calling approach proposed by Bell et al., classifies a genomic position as heterozygous if the measured allele frequency \( f \) is between 0.14 and 0.86. The last row reports error false negative error rates when a random additional technical replicate was considered, the detected genotype at a genomic position was counted as false negative if the measured allele frequency did not fall into the range of \([0.14, 0.86]\) in both samples.
Supplementary Table 2: Two different libraries (I, II) were generated for one individual. Library I was sequenced on a single lane of flowcell A (referred to as sample I1A). Library II was sequenced on two lanes (1,2) of flowcell B (II1B, II2B) and a third lane on yet another flowcell C (II1C). Sample II1C passed through a different cluster generation than II1B and II2B. The variance of the differences between allele frequencies was analyzed at common heterozygous positions that were covered by more than 20 reads in each sample of a compared pair. The variance of the allele frequency differences between (II1B, II2B), is of the same order of magnitude as the variance between (II1B, II1C) and (II2B, II1C). This indicates that the cluster generation step has hardly any effect on the allele frequency distribution. Note that the variance of the allele frequency differences between (I1A, II1B), (I1A, II2B), and (I1A, II2C) is about two orders of magnitude larger, which is due to different library preparations. Note also that the variances of (I1A, II1B), (I1A, II2B), and (I1A, II2C) vary to a greater extent than (II1B, II2B), (II1B, II1C), and (II2B, II1C) as sample I1A had a lower coverage resulting in a sixfold lower number of heterozygous loci to be comparable.

References

[1] Athreya KB, Ney PE (1972), Branching Processes, Springer

[2] Yakovlev AY, Yanev NM (2009), Relative Frequencies in Multitype Branching Processes, The Annals of Applied Probability Vol.19, No. 1, 1-14

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

[4] Faà di Bruno, F (1855), Sullo sviluppo delle Funzioni (in Italian), Annali di Scienze Matematiche e Fisiche Vol. 6, 479-480.
Supplementary Figure 2: Receiving operating characteristics for heterozygous genotype calling for frequency intervals and different categories of read coverage. A genomic position was called heterozygous, if the allele frequency $f$ of the alternating allele was smaller than $\text{abs}(0.5-f) < c$, with the cutoff $c$ ranging from $c = [0.02, 0.04, ..., 0.48]$. The area under the ROC curve (AUC) increases markedly for the first three categories of read coverage. For a read coverage above 35 there is no notable difference in AUC. For $\text{abs}(0.5-f) < 0.36$, the false positive error rates are comparable to the error rates of samtools with default parameter settings.
Supplementary Figure 3: Allele frequency distributions at heterozygous sites are position- and individual-independent. A) Frequency distribution of the alternating allele for all positions and all individuals pooled. B) Frequency distribution for a random set comprising 5% of all positions in a randomly chosen (representative) person. Comparison of the two distribution with Pearson’s Chi-squared test: $p = 0.234$. 
Supplementary Figure 4: An disequilibrium in the initial allele ratio $n_1:n_2$ shifts the expected mean away from 0.5 towards 0 or 1. Although the characteristic dependency on $p$ and $K$ is preserved, the variance of the allele ratio distribution decreases the more the initial ratio is skewed.