LFMD: a new likelihood-based method to detect low-frequency mutations without molecular tags

Rui Ye[1,2,7*], Jie Ruan[2,7*], Xuehan Zhuang[3*], Yanwei Qi[2,7], Yitai An[2,7], Jiaming Xu[2,7], Timothy Mak[4], Xiao Liu[2,7], Huanming Yang[2,6,7], Xun Xu[2,7], Larry Baum[1,4,5], Chao Nie[2,7#] & Pak Chung Sham[1,4,5#]

[1]Department of Psychiatry, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China;

[2]BGI-Shenzhen, Shenzhen 518083, China;

[3]Department of Surgery, Li Ka Shing Faculty of Medicine, The University of Hong Kong; Hong Kong, China;

[4]Center for Genomic Sciences, The University of Hong Kong, Hong Kong, China;

[5]State Key Laboratory of Brain and Cognitive Sciences, The University of Hong Kong, Hong Kong, China;

[6]James D. Watson Institute of Genome Sciences, Hangzhou, China

[7]China National GeneBank, BGI-Shenzhen, Shenzhen 518120, China

[*]These authors contributed equally to this work.

[#]Correspondence should be addressed to C.N. (niechao@genomics.cn) or P.C.S (pcsham@hku.hk).

V1.9 on 2019.06.04

Abstract

As next generation sequencing (NGS) and liquid biopsy become more prevalent in clinical and research area, especially cancer diagnosis, targeted therapy guidance and disease surveillance, there is an increasing need for better methods to reduce cost and to improve sensitivity and specificity. Since the error rate of NGS is around 1%, it is difficult to identify mutations with frequency lower than 1% accurately and efficiently because of low Signal-to-Noise Ratio (SNR). Here we propose a likelihood-based approach, low-frequency mutation detector (LFMD), combining the advantages of duplex sequencing (DS) and bottleneck sequencing system (BotSeqS) to maximize utilization of duplicate sequenced reads.

Compared with DS, the new method achieves higher sensitivity (improved ~16%), higher specificity (improved ~1%) and lower cost (reduced ~70%) without involving additional experimental steps, customized adapters and molecular tags. In addition, this method can also be used to improve sensitivity and specificity of other variant calling algorithms by replacing a step in traditional NGS analysis: removing polymerase chain reaction (PCR) duplication. Thus, LFMD can be a promising method used in genomic research and clinical fields.

Introduction

At the individual level, low-frequency mutations (LFMs) are defined as mutations with allele frequency lower than 5% or 1%. LFMs increase power to predict early stage of cancer and Alzheimer's Disease (AD)[1], distinguish samples with different age[2], identify disease-causing variants[3], diagnose before tri-parental in vitro fertilization[4], and track the mutational spectrum in viral genomes, malignant lesions, and somatic tissues[5,6]. To effectively improve signal-to-noise ratio (SNR) and detect LFMs, stringent thresholds, complex experimental skills[1,7], single cell sequencing[8-11], circle sequencing[12], and more precise models[13,14] were developed. The bottleneck sequencing system[15] (BotSeqS) and duplex sequencing[16] (DS) utilize duplicate reads generated by polymerase chain reaction (PCR), which are discarded by other methods, to achieve much higher accuracy. However, current methods still have some limitations in detecting LFMs.

Disadvantages of single cell sequencing and circle sequencing

For single cell sequencing, DNA extraction is laborious and exacting, with point mutations and copy number biases introduced during amplification of small amounts of fragile DNA. To increase specificity, only variants shared by at least two cells are accepted as true variants[11]. This method is not cost efficient and cannot be used in large-scale clinical applications because a large number of single cells need to be sequenced to identify rare mutations.

Circle sequencing only utilizes a single strand of DNA, so its specificity is limited by the error rate of PCR. It obtains errors at a rate as low as 7.6×10^{-6} per base sequenced[12] while DS can achieve 4×10^{-10} errors per base sequenced[16].

Disadvantages of BotSeqS

In contrast, BotSeqS uses endogenous molecular tags, the positions of the aligned read pair, to group reads from the same DNA template and construct double strand consensus reads. As a result, it can detect very rare mutations ($<10^{-6}$) while it is cheap enough to sequence the whole human genome[15]. But it introduces highly diluted DNA templates before PCR amplification to reduce endogenous tag conflicts and ensure sufficient sequencing of each DNA template. Thus, it has high specificity with poor sensitivity. In addition, it discards clonal variants and small insertions/deletions (InDels) in order to limit false positives.

Disadvantages of DS

Another compromising method to eliminate tag conflicts is Duplex sequencing (DS). It ligates exogenous random molecular tags (also known as unique molecular identifier, UID or UMI) to both ends of each DNA template before PCR amplification. Although sensitive and accurate, it wastes many data to sequence tags, fixed sequences and a large proportion of read families that contain only one read pair because of a sequencing error on a tag. Since random molecular tags are synthesized with customized adapters, batch effects might occur during DNA library construction. Additionally, DS only works on targeted small genome regions[6,13,17] rather than on the whole genome.

A new approach

In order to avoid the aforementioned problems, we present here a new, efficient approach that combines the advantages of BotSeqS and DS. It uses a likelihood-based model[13,14] to dramatically reduce endogenous tag conflicts. Then it groups reads into read families and constructs double strand consensus reads to detect ultra-rare mutations accurately while maximizing utilization of non-duplicate read pairs. Without exogenous molecular tags, our method can also work with the 50 bp short reads of BGISEQ as well as the longer reads of HiSeq. In summary, it simplifies the DNA sequencing procedure, saves data and cost, achieves higher sensitivity and specificity, and can be used in whole genome sequencing.

Using digital PCR to validate thousands of low-frequency sites is prohibitively expensive and laborious[18]. A new method which works on an independent platform can be used as a method to validate HiSeq results. Additionally, our new method is a statistical solution of the problem of PCR duplication in the basic analysis pipeline of next generation sequencing (NGS) data and can improve sensitivity and specificity of other variant calling algorithms without requiring specific experimental designs. As the price of sequencing is falling, the depth and the rate of PCR duplication are rising. The method we present here might help deal with such high depth data more accurately and efficiently.

Methodology

Intuitively, to distinguish LFMs (signal) from background PCR and sequencing errors (noise), we need to increase the SNR. To increase SNR, we need to either increase the frequency of mutations or inhibit sequencing errors. Single cell sequencing increases the frequency of mutations by isolating single cells from the bulk population, while BotSeqS and DS inhibit sequencing errors by identifying the major allele at each site of multiple reads from the same DNA template. In this paper, we only focus on the latter strategy.

To group reads from the same DNA template, the simplest idea is to group properly mapped reads with the same coordinates (i.e., chromosome, start position, and end position) because random shearing of DNA molecular can provide natural differences, called endogenous tags, between templates. A group of reads is called a read family. However, as the length of DNA template is approximately determined, random shearing cannot provide enough differences to distinguish each DNA template. Thus, it is common that two original DNA templates share the same coordinates. If two or more DNA templates shared the same coordinates, and their reads were grouped into a single read family, it is difficult to determine, using only their frequencies as a guide, whether an allele is a potential error or a mutation. Thus, BotSeqS introduced a strategy of dilution before PCR amplification to dramatically reduce the number of DNA templates in order to reduce the probability of endogenous tag conflicts. And DS introduced exogenous molecular tags before PCR amplification to dramatically increase the differences between templates. Thus, BotSeqS sacrifices sensitivity and DS sequences extra data: the tags.

Here we introduce a third strategy to eliminate tag conflicts. It is a likelihood-based approach based on an intuitive hypothesis: that if reads of two or more DNA templates group together, a true allele's frequency in this read family is high enough to distinguish the allele from background sequencing errors. The pipeline of LFMD is shown in Figure 1, and a comparison of DS and LFMD is shown in Figure 2.

Likelihood-based model

We aim to identify alleles at each potential heterozygous position in a read family (grouped according to endogenous tags). Then based on those heterozygous sites, we split the mixed read family into smaller ones, and compress each one into a consensus read. Finally, we detect mutations based on all consensus reads, which have much lower error rates than 0.1%.

First, we define a Watson strand as a read pair for which read 1 is the plus strand while read 2 is the minus strand. A Crick strand is defined as a read pair for which read 1 is the minus strand while read 2 is the plus strand. The plus and minus strands are also known as the forward and reverse strands according to the reference genome. Read 1 and 2 are derived from raw pair-end fastq files. Thus a read family which contains Watson and Crick strand reads simultaneously is an ideal read family because it is supported by both strands of the original DNA template. Second, we select potential heterozygous sites which meet the following criteria: 1) the minor allele is supported by both Watson and Crick reads; 2) minor allele frequencies in both Watson and Crick read family are greater than approximately the average sequencing error rate, often 1% or 0.1%; 3) low quality bases (<Q20) and low quality alignments (<Q30) are excluded. Finally, we calculate genotype likelihood in the Watson and Crick family independently in order to eliminate PCR errors during the first PCR cycle.

At each position of a Watson or Crick read family, let X denotes the sequenced base and θ the allele frequencies. Let $P(x|\theta)$ be the probability mass function of the random variable X, indexed by the parameter $\theta = (\theta_A, \theta_C, \theta_G, \theta_T)^T$, where θ belongs to a parameter space Ω. Let $g \in \{A, C, G, T\}$, and θ_g represents the frequency of allele g at this position. Obviously, we have boundary constraints for θ: $\theta_g \in [0, 1]$ and $\sum \theta_g = 1$.

Assuming N reads cover this position, x_i represents the base on read $i \in \{1, 2, \dots, N\}$, and e_i denotes sequencing error of the base, we get

$$\begin{aligned} P(x_i|\theta) = {} & P(no\ sequencing\ error \mid the\ base\ is\ g) \cdot P(the\ base\ is\ g) \\ & + P(sequencing\ error\ with\ specific\ direction \mid the\ base\ is\ not\ g) \\ & \cdot P(the\ base\ is\ not\ g) \\ & = (1 - e_i)\,\theta_g + \frac{e_i}{3}(1 - \theta_g), \qquad g = x_i \end{aligned}$$

So the log-likelihood function can be written as

$$\ell(\theta) = \sum_{i=1}^{N} \log P(x_i|\theta) = \sum_{i=1}^{N} \log\left((1 - e_i)\,\theta_g + \frac{e_i}{3}(1 - \theta_g)\right), \qquad g = x_i$$

Thus, for each candidate allele g, under the null hypothesis H_0: $\theta_g = 0, \theta \in \Omega$, and the alternative hypothesis H_1: $\theta_g \neq 0, \theta \in \Omega$, the likelihood ratio test is

$$t_g = -2\{\ell_0(\theta) - \ell_1(\theta)\} \sim \chi_1^2$$

However, as $\theta_g = 0$ lies on the boundary of the parameter space, the general likelihood ratio test needs an adjustment to fit χ_1^2. Because the adjustment is related to calculation of a tangent cone[19] in a constrained 3-dimensional parameter space, and the computation is too complicated and time consuming for large scale NGS data, here we use a simplified, straightforward adjustment[20] presented by Yong et al in 2017.

In order to utilize Yong et al's method, we need to introduce conditional events. Let $\{\mathcal{A}_1, \dots, \mathcal{A}_K\}$ denotes the set of conditional events which are mapped to four alleles at the position. Let $k = 1, \dots, K, and\ K = 4$ be the total number of events, we get the log likelihood component

$$\ell_k\{\theta;\ \mathcal{A}_k(x_i)\} = \sum\nolimits_{x_i \in \mathcal{A}_k} \log P(x_i|\theta)$$

Then the composite conditional log likelihood can be constructed as

$$\ell_c(\theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} \omega_{ik} \ell_k\{\theta;\ \mathcal{A}(x_i)\}$$

in which we set

$$\omega_{ik} = 1$$

Let $\hat{\theta}_c = \arg\max_{\theta\in\Omega} \ell_c(\theta)$ be the maximum composite likelihood estimator, and define the composite score function, sensitivity matrix and variability matrix respectively as

$$U_c(\theta) = \frac{\partial \ell_c(\theta)}{\partial \theta}$$
$$H = \lim_{N\to\infty} -\frac{1}{N} E\left\{\frac{\partial^2 \ell_c(\theta)}{\partial \theta^T \partial \theta}\right\}$$
$$V = \lim_{N\to\infty} \frac{1}{N} E\left[\left\{\frac{\partial \ell_c(\theta)}{\partial \theta}\right\}\left\{\frac{\partial \ell_c(\theta)}{\partial \theta}\right\}^T\right]$$

The corresponding estimators of H and V are denoted by $\hat{H}$ and $\hat{V}$ evaluated at $\hat{\theta}_c$. The modified composite likelihood under boundary constraints was given by Yong et al[20] as

$$\ell_M(\theta) = \ell_c(\hat{\theta}_c) - \{T(\theta)^T \hat{H}_A T(\theta)\}\phi(\theta)$$

where

$$T(\theta) = N^{-1/2}\hat{H}^{-1}U_c(\hat{\theta}_c) - N^{1/2}(\theta - \hat{\theta}_c)$$
$$\hat{H}_A = \hat{H}\hat{V}^{-1}\hat{H}$$
$$\phi(\theta) = \frac{\ell_c(\theta) - \ell_c(\hat{\theta}_c)}{-T(\theta)^T\hat{H}T(\theta) + N^{-1}U_c(\hat{\theta}_c)^T\hat{H}^{-1}U_c(\hat{\theta}_c)}$$

Thus, we derive the adjusted likelihood ratio test

$$t_g = -2\{\ell_M(\theta_0) - \ell_M(\hat{\theta}_M)\} \sim \chi_1^2$$

where $\hat{\theta}_M = \arg\max_{\theta\in\Omega} \ell_M(\theta)$ and θ_0 is the parameter θ under null hypothesis H_0.

To facilitate the calculation of H and V, we let $pmf(e)$ denote the probability mass function of sequencing error rate e, the expected number of bases with e is represented as

$$\lim_{N\to\infty} N \cdot pmf(e)$$

Then, the expected number of bases g with e is

$$\lim_{N\to\infty} N \cdot pmf(e) \cdot \left\{(1-e)\,\theta_g + \frac{e}{3}(1-\theta_g)\right\}$$

Thus,

$$E\left[\frac{\partial \ell_c(\theta)}{\partial \theta_g}\right] = \lim_{N\to\infty} \sum\nolimits_e \left\{ N \cdot pmf(e) \cdot \left\{(1-e)\,\theta_g + \frac{e}{3}(1-\theta_g)\right\} \cdot \frac{1-\dfrac{4e}{3}}{(1-e)\,\theta_g + \frac{e}{3}(1-\theta_g)} \right\}$$

$$= \lim_{N\to\infty} N \cdot \sum\nolimits_e \left\{ pmf(e)\left(1-\frac{4e}{3}\right)\right\} = \lim_{N\to\infty} N \cdot C$$

where C is a finite constant. Then we derive

$$V = \lim_{N\to\infty} \frac{1}{N} E\left[\left\{\frac{\partial \ell_c(\theta)}{\partial \theta}\right\}\left\{\frac{\partial \ell_c(\theta)}{\partial \theta}\right\}^T\right] = \lim_{N\to\infty} NC^2 \begin{Bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{Bmatrix}$$

As a result, $\hat{V}^{-1}$ tends to $\boldsymbol{0}$ in the model, which means the adjustment is not necessary. To be clear, the special form of matrix V with all equal elements is due to the infinite N which insures all possible e and g occur in the function $\ell_c(\theta)$. The V does not have the special form when N is a finite number.

Thus, we finally arrive at a general conclusion that the further adjustment of χ_1^2 is not helpful in similar cases, although the asymptotic distribution we use is not perfect when N is small (e.g., N<5), and alternative approaches might be derived in the future. We also compared theoretical P-values with empirical P-values from Monte Carlo procedures (Supplementary Material, Figures S1) and explored the power of our model under uniformly and truly distributed sequencing errors (Supplementary Material, Figures S2). The simulation results support the theoretical conclusion sufficiently.

Because the null and alternative hypotheses have two and three free variables respectively, the Chi-square distribution has 1 degree of freedom. Type I error of the allele g can then be given

$$P_g = 1 - \mathrm{cdf}(t_g)$$

where $\mathrm{cdf}(x)$ is the cumulative density function of the χ_1^2 distribution. If $P_g$is less than a given threshold α, the null hypothesis is rejected and the allele g is treated as a candidate allele of the read family.

Although P_g cannot be interpreted as the probability that $H_{0,g}$ is true and allele g is an error, it is a proper approximation of the error rate of allele g. We only reserve alleles with $P_g \leq \alpha$ in both Watson and Crick families and substitute others with "N". Then Watson and Crick families are compressed into several single strand consensus sequences (SSCSs). The SSCSs might contain haplotype information if more than one heterozygous site is detected. Finally, SSCSs which are consistent in both Watson and Crick families are claimed as double strand consensus sequences (DCSs).

For each allele on a DCS, let P_w and P_c represent the relative error rates of the given allele in the Watson and Crick family respectively, and let P_{wc} denote the united error rate of the allele. Thus,

$$P_{wc} = P_w + P_c - P_w P_c$$

For a read family which proliferated from $\boldsymbol{n}$ original templates, a coalescent model can be used to model the PCR procedure[21]. According to the model, a PCR error proliferates and its fraction decreases exponentially with the number, $\boldsymbol{m}$, of PCR cycles. For example, an error that occurs in the first PCR cycle would occupy half of the PCR products, an error that occurs in the second cycle occupies a quarter, the third only 1/8, and so on. As we only need to consider PCR errors which are detectable, the coalescent PCR error rate is defined as the probability to detect a PCR error whose frequency $\geq 2^{-\boldsymbol{m}}/\boldsymbol{n}$, and it is equal to

$$1 - (1 - error\ rate\ per\ cycle)^{2^m - 1}$$

Let e_{pcr} denote the coalescent PCR error rate and P_{pcr} the united PCR error rate of the double strand consensus allele. Empirically we get

$$P_{pcr} \approx 10 * e_{pcr}^2$$

Because $P_{wc}P_{pcr} \approx 0$, the combined base quality of the allele on the DCS is

$$Q = -10 \log_{10}(P_{wc} + P_{pcr})$$

Then Q is transferred to an ASCII character, and a series of characters make a base quality sequence for the DCS. Finally, we generate a BAM file with DCSs and their quality sequences.

With the BAM file which contains all the high quality DCS reads, the same approach is used to give each allele a P-value at each genomic position which is covered by DCS reads. Adjusted P-values (q-values) are given via the Benjamin-Hochberg procedure. The threshold of q-values is selected according to the total number of tests conducted and false discovery rate (FDR) which can be accepted.

A similar mathematical model was described in detail in previous papers by Jun et al[13] and Yan et al[14]. Jun et al. used this model to reliably call mutations with frequency > 4%. In contrast, we use this model to deal with read families rather than non-duplicate reads. In a mixed read family, most of the minor allele frequencies are larger than 4%, so the power of the model meets our expectation.

For those reads containing InDels, the CIGAR strings in BAM files contain I or D. It is obvious that reads with different CIGAR strings cannot fit into one read family. Thus, CIGAR strings can also be used as part of endogenous tags. In contrast, the soft-clipped part of CIGAR strings cannot be ignored when considering start and end positions because low-quality parts of reads tend to be clipped, and the coordinates after clipping are not a proper endogenous tag for the original DNA template.

Results

Comparison between DS and LFMD

Simulated data

We used Python scripts developed by the Du novo[22] team to simulate mixed double-strand sequencing data and then compared the results of LFMD and DS. Although the simulation was not perfect, the analysis was still useful to demonstrate the power and the potential drawbacks of LFMD and DS because we knew the true mutations explicitly, and true positive (TP) and false positive (FP) could be defined and calculated clearly. The numbers of TP and FP are shown in Tables 1 and 2.

We found that DS induces several false positives due to mapping errors. LFMD eliminates mapping errors of DCSs by outputting DCSs directly into BAM files. LFMD is much more sensitive than DS according to Figures 3, 4, and 5.

Mouse mtDNA

In order to evaluate the performance of LFMD, we compared LFMD with DS on a DS data from mouse mtDNA: SRR1613972. The analysis pipeline is shown in Figure 4. We controlled almost all parameters to be exactly the same in DS and LFMD and then compared the results. Because DS is the current gold standard, we treated the DS results as the true set and then calculated the true positive rate (TP), false positive rate (FP), and positive predictive value (PPV) of LFMD based on all proper mapped reads (Table 1) and unique proper mapped reads (Table 2). We found that mapping quality influenced the performance of both methods.

Although the majority of mutations are identified by both methods, some mutations are detected only by DS or only by LFMD. We investigated these discordant mutations one by one. It is interesting that most of them (42 out of 62 LFMD-only point mutations) can be identified if we consider 1-2 bp sequencing errors and PCR errors in the 24 bp tag sequences of DS. Two of them are potential true positive mutations because there is only one support read in one of the 2 families. The last 18 LFMD-only mutations did not have matched tags to make DCSs. They are potential FPs of LFMD or FNs of DS. But when we consider more than 2 bp mismatches in tags, most of the last 18 LFMD-only mutations had double strand support. This phenomenon implies contamination of DS tags or potential false positive hints of LFMD which should be validated in future research.

Twenty-six samples from Prof. Kennedy's laboratory[1]

We compared the performance of DS and LFMD on 26 samples from Prof. Scott R. Kennedy's laboratory. Only unique mapped reads were used to detect LFMs. The majority of LFMs were detected by both tools. Almost all LFMs only detected by DS were false positives due to alignment errors of DCS, while LFMD outputs BAM files directly and avoids alignment errors. LFMs only detected by LFMD are supported by raw reads if considering PCR and sequencing errors on molecular tags. As a result, LFMD is much more sensitive and accurate than DS. The improvement on sensitivity is about 16% according to Table 5.

YH cell line

We sequenced the YH cell line, passage 19, 8 times in order to validate the stability of the method. All results, shown in Table 6 and Figure 6, are highly consistent.

ABL1 data

Using the duplex sequencing method in 2015, Schmitt et al. analyzed an individual with chronic myeloid leukemia who relapsed after treatment with the targeted therapy imatinib (the Short Read Archive under accession SRR1799908). We analyzed this individual and found 5 extra LFMs. Two of them were in the coding region of the ABL1 gene. It was reported that E255G (E255VDK, Dasatinib, Imatinib, Nilotinib) and V256G (V256L, Imatinib) were associated with drug resistance[23]. The annotation results of 5 LFMs are shown in Table 7.

Materials

Subject recruitment and sampling

A lymphoblastoid cell line (YH cell line) established from the first Asian genome donor[24] was used. Total DNA was extracted with the MagPure Buffy Coat DNA Midi KF Kit (MAGEN). The DNA concentration was quantified by Qubit (Invitrogen). The DNA integrity was examined by agarose gel electrophoresis. The extracted DNA was kept frozen at -80°C until further processing.

Mitochondrial whole genome DNA isolation

Mitochondrial DNA (mtDNA) was isolated and enriched by double/single primer set amplifying the complete mitochondrial genome. The samples were isolated using a single primer set (LR-PCR4) by ultra-high-fidelity Q5 DNA polymerase following the protocol of the manufacturer (NEB) (Table 8).

Library construction and mitochondrial whole genome DNA sequencing

For the BGISeq-500 sequencing platform, mtDNA PCR products were fragmented directly by Covaris E220 (Covaris, Brighton, UK) without purification. Sheared DNA ranging from 150 bp to 500 bp without size selection was purified with an Axygen™ AxyPrep™ Mag PCR Clean-Up Kit. 100 ng of sheared mtDNA was used for library construction. End-repairing and A-tailing was carried out in a reaction containing 0.5 U Klenow Fragment (ENZYMATICS™ P706-500), 6 U T4 DNA polymerase (ENZYMATICS™ P708-1500), 10 U T4 polynucleotide kinase (ENZYMATICS™ Y904-1500), 1 U rTaq DNA polymerase (TAKARA™ R500Z), 5 pmol dNTPs (ENZYMATICS™ N205L), 40 pmol dATPs (ENZYMATICS™ N2010-A-L), 1 X PNK buffer (ENZYMATICS™ B904) and water with a total reaction volume of 50 µl. The reaction mixture was placed in a thermocycler running at 37°C for 30 minutes and heat denatured at 65°C for 15 minutes with the heated lid at 5°C above the running temperature. Adaptors with 10 bp tags (Ad153-2B) were ligated to the DNA fragments by T4 DNA ligase (ENZYMATICS™ L603-HC-1500) at 25°C. The ligation products were PCR amplified. Twenty to twenty-four purified PCR products were pooled together in equal amounts and then denatured at 95°C and ligated by T4 DNA ligase (ENZYMATICS™ L603-HC-1500) at 37°C to generate a single-strand circular DNA library. Pooled libraries were made into DNA Nanoballs (DNB). Each DNB was loaded into one lane for sequencing.

Sequencing was performed according to the BGISeq-500 protocol (SOP AO) employing the PE50 mode. For reproducibility analyses, YH cell line mtDNA was processed four times following the same protocol as described above to serve as library replicates, and one of the DNBs from the same cell line was sequenced twice as sequencing replicates. A total of 8 datasets were generated using the BGISEQ-500 platform. MtDNA sequencing was performed on the BGISeq-500 with 50 bp paired-end reads. The libraries were processed for high-throughput sequencing with a mean depth of ~20000x.

The data that support the findings of this study have been deposited in the CNSA (https://db.cngb.org/cnsa/) of CNGBdb with accession code CNP0000297.

Discussion

LFMD is still expensive for target regions >2 Mbp in size because of the high depth. As the cost of sequencing continues to fall, it will become increasingly practical. Only accepting random sheered DNA fragments, not working on short amplicon sequencing data, and only working on pair-end sequencing data are known limitations of LFMD. Moreover, LFMD's precision is limited by the accuracy of alignment software. Although tags were excluded in this paper, LFMD still has the potential to utilize tags and deal with amplicon sequencing data.

To estimate the theoretical limit of LFMD, let read length equal 100 bp and let the standard deviation (SD) of insert size equal 20 bp. Let N represent the number of position families across one point. Then, N = (2 * 100) * (20 * 6) = 24000 if only considering ±3 SD. As the sheering of DNA is not random in the real world, it is safe to set N as 20,000. Ideally, the likelihood ratio test can detect mutations whose frequency is greater than 0.2% in a read family with Q30 bases. Thus, the theoretical limit of minor allele frequency is around 1e-7 (= 0.002 / 20000).

Conclusion

To eliminate endogenous tag conflicts, we use a likelihood-based model to separate the read family of the minor allele from that of the major allele. Without additional experimental steps and the customized adapters of DS, LFMD achieves higher sensitivity and almost the same specificity with lower cost. It is a general method which can be used in several cutting-edge areas.

Figures and tables

Table 1. Number of true positives detected by DS and LFMD. There are 67 single nucleotide variants (SNVs), 13 insertions (INSs), and 3 deletions (DELs) in the simulated data at every level of alternative allele frequency (AAF).

AAF	**SNV**		**INS**		**DEL**	
	DS	**LFMD**	**DS**	**LFMD**	**DS**	**LFMD**
1.0E-04	14	23	1	2	1	1
2.0E-04	21	45	3	6	2	3
3.0E-04	28	53	2	9	1	2
4.0E-04	32	51	6	11	0	2
5.0E-04	35	56	5	9	3	3
6.0E-04	43	61	4	12	1	3
7.0E-04	47	63	8	13	3	3
8.0E-04	58	64	8	13	1	3
9.0E-04	58	66	8	13	3	3
1.0E-03	56	64	8	13	1	3
2.0E-03	63	67	13	13	3	3
3.0E-03	67	67	11	13	3	3
4.0E-03	67	66	13	13	3	3
5.0E-03	67	67	13	13	3	3
1.0E-02	67	67	13	13	3	3

Table 2. Number of false positives detected by DS and LFMD.

AAF	**SNV**		**INS**		**DEL**	
	DS	**LFMD**	**DS**	**LFMD**	**DS**	**LFMD**
1.0E-04	0	0	0	0	0	0
2.0E-04	0	0	0	0	0	0
3.0E-04	0	0	0	0	0	0

4.0E-04	0	0	0	0	0	0
5.0E-04	1	0	0	0	0	0
6.0E-04	0	0	0	0	0	0
7.0E-04	0	0	0	0	0	0
8.0E-04	0	0	0	0	0	0
9.0E-04	0	0	0	0	0	0
1.0E-03	0	1	0	0	0	0
2.0E-03	0	0	0	0	0	0
3.0E-03	2	0	0	0	0	0
4.0E-03	0	0	0	0	0	0
5.0E-03	0	0	0	0	0	0
1.0E-02	0	0	0	0	0	0

Table 3. Results of DS and LFMD based on all proper mapped reads. FNR, TPR, and PPV are calculated based on the assumption that results of DS are the complete and true mutation sets.

	DS_only	Overlap	LFMD_only	FNR	TPR	PPV
A>C	0	7	12	0.00%	100.00%	36.84%
A>G	3	118	19	2.48%	97.52%	86.13%
A>T	0	43	5	0.00%	100.00%	89.58%
A>del	0	3	1	0.00%	100.00%	75.00%
A>ins	0	4	1	0.00%	100.00%	80.00%
C>A	0	12	6	0.00%	100.00%	66.67%
C>G	0	3	1	0.00%	100.00%	75.00%
C>T	0	83	20	0.00%	100.00%	80.58%
C>del	0	4	3	0.00%	100.00%	57.14%
C>ins	0	4	0	0.00%	100.00%	100.00%
G>A	2	39	18	4.88%	95.12%	68.42%
G>C	0	5	0	0.00%	100.00%	100.00%
G>T	0	3	5	0.00%	100.00%	37.50%

G>del	0	4	1	0.00%	100.00%	80.00%
G>ins	0	2	1	0.00%	100.00%	66.67%
T>A	0	20	4	0.00%	100.00%	83.33%
T>C	0	137	19	0.00%	100.00%	87.82%
T>G	1	12	5	7.69%	92.31%	70.59%
T>del	1	11	1	8.33%	91.67%	91.67%
T>ins	0	1	0	0.00%	100.00%	100.00%
total	7	515	122	1.34%	98.66%	80.85%

Table 4. Results of DS vs LFMD based on all unique proper mapped reads. FNR, TPR, and PPV are calculated based on the assumption that results of DS are the complete and true mutation sets.

	DS_only	**Overlap**	**LFMD_only**	**FNR**	**TPR**	**PPV**
A>C	0	5	11	0.00%	100.00%	31.25%
A>G	2	70	9	2.78%	97.22%	88.61%
A>T	0	28	4	0.00%	100.00%	87.50%
A>del	0	2	0	0.00%	100.00%	100.00%
A>ins	0	3	1	0.00%	100.00%	75.00%
C>A	0	8	4	0.00%	100.00%	66.67%
C>G	0	2	1	0.00%	100.00%	66.67%
C>T	0	57	10	0.00%	100.00%	85.07%
C>del	0	2	2	0.00%	100.00%	50.00%
C>ins	0	4	0	0.00%	100.00%	100.00%
G>A	1	19	5	5.00%	95.00%	79.17%
G>C	0	4	0	0.00%	100.00%	100.00%
G>T	0	2	4	0.00%	100.00%	33.33%
G>del	0	1	1	0.00%	100.00%	50.00%
G>ins	0	1	1	0.00%	100.00%	50.00%
T>A	0	11	2	0.00%	100.00%	84.62%
T>C	0	82	11	0.00%	100.00%	88.17%
T>G	0	10	1	0.00%	100.00%	90.91%
T>del	1	7	0	12.50%	87.50%	100.00%

T>ins	0	1	0	0.00%	100.00%	100.00%
total	4	319	67	1.24%	98.76%	82.64%

Table 5. DS vs LFMD on 26 samples from Prof. Kennedy's laboratory.

Sample	DS-only	Overlap	LFMD-only	DS-only /Overlap	LFMD-only /Overlap
1440B	27	928	110	2.91%	11.85%
1440E	10	491	66	2.04%	13.44%
2384H	13	500	171	2.60%	34.20%
2384P	4	200	60	2.00%	30.00%
3080H	5	231	68	2.16%	29.44%
3080P	23	504	104	4.56%	20.63%
334B	14	592	100	2.36%	16.89%
334E	13	1332	142	0.98%	10.66%
409B	20	649	76	3.08%	11.71%
409E	10	994	134	1.01%	13.48%
511H	15	669	104	2.24%	15.55%
523B	2	494	57	0.40%	11.54%
523E	6	675	73	0.89%	10.81%
533B	1	216	52	0.46%	24.07%
533E	1	111	35	0.90%	31.53%
547H	4	411	104	0.97%	25.30%
547P	10	799	94	1.25%	11.76%
552B	14	467	87	3.00%	18.63%
552E	12	576	76	2.08%	13.19%
558P	7	82	40	8.54%	48.78%
626H	6	189	101	3.17%	53.44%
626P	5	165	76	3.03%	46.06%
652B	10	684	78	1.46%	11.40%
652E	3	595	54	0.50%	9.08%
670B	8	753	73	1.06%	9.69%
670E	1	116	41	0.86%	35.34%

Median	/	/	/	2.02%	16.22%

Table 6. Number of mutations found in mtDNA of 8 YH cell lines. Under the hypothesis that true mutations should be identified from at least two samples, we detected 68 "true" mutations and then calculated TP, FP, TPR, and FPR.

Samples	# of mutations	TP	FP	TPR	FPR
L01_501	64	63	1	92.65%	1.56%
L01_502	68	67	1	98.53%	1.47%
L01_503	62	62	0	91.18%	0.00%
L01_504	65	63	2	92.65%	3.08%
L01_505	62	60	2	88.24%	3.23%
L01_506	61	59	2	86.76%	3.28%
L01_507	65	62	3	91.18%	4.62%
L01_508	62	61	1	89.71%	1.61%
Mean	63.63	62.13	1.50	91.36%	2.36%
SD	2.33	2.42	0.93	3.55%	1.45%

Table 7. Five low-frequency SNVs found only by LFMD

SNP	Variant	Transcript	Function	cDNA Position	CDS Position	AA Position	AA Change
chr9:133738364	A>G	NM_005157	coding	767	764	255	E>G
chr9:133738364	A>G	NM_007313	coding	1260	821	274	E>G
chr9:133738367	T>G	NM_005157	coding	770	767	256	V>G
chr9:133738367	T>G	NM_007313	coding	1263	824	275	V>G
chr9:133748236	C>T	NM_005157	intronic				
chr9:133748236	C>T	NM_007313	intronic				
chr9:133748343	T>G	NM_005157	coding	1007	1004	335	V>G
chr9:133748343	T>G	NM_007313	coding	1500	1061	354	V>G
chr9:133756073	A>C	NM_005157	intronic				
chr9:133756073	A>C	NM_007313	intronic				

Table 8. Long range polymerase chain reaction (LR-PCR) primer sets

Name	Sequence (5'->3')	Start	Stop	Product Length

LR-PCR1	AACCAAACCCCAAAGACACC	550	569	9290
	GCCAATAATGACGTGAAGTCC	9839	9819	
LR-PCR2	TCCCACTCCTAAACACATCC	9592	9611	7626
	TTTATGGGGTGATGTGAGCC	645	626	
LR-PCR4	AAGAGTGCTACTCTCCTCGCTCCG	16432	16455	16569
	GTGCGGGATATTGATTTCACGGAGG	16431	16407	

Figure 1. Overview of LFMD pipeline. The Y-shaped adapters determine read 1 (purple bar) and 2 (green bar). The directions of reads determine +/- strands. So after the first cycle of the PCR amplification, the Watson and Crick families are well defined. Then within a read family, true alleles (green dots) and accumulated PCR errors (blue dots) are detected via likelihood-base model and given a combined error rate. Sequencing errors (red dots) are eliminated. Combining SSCSs of paired read families, high quality DCSs with estimated error rates are generated and used in the downstream analysis.

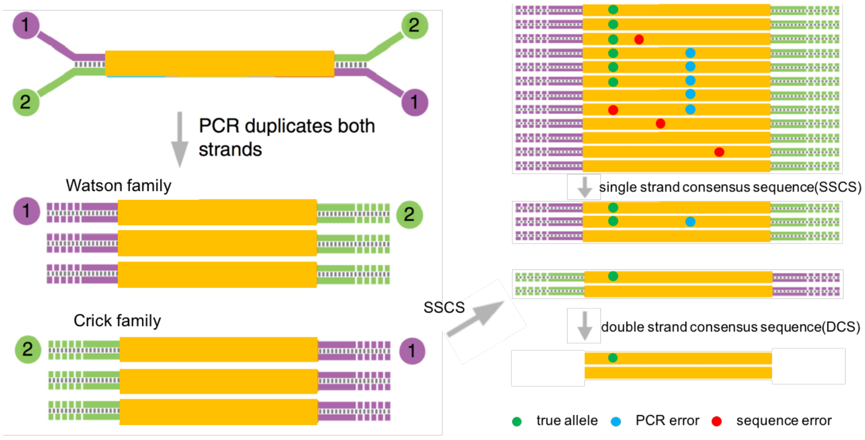

1
2
2
1
PCR duplicates both strands
Watson family
1
2
Crick family
2
1
SSCS
single strand consensus sequence(SSCS)
double strand consensus sequence(DCS)
● true allele
● PCR error
● sequence error

Figure 2. Pipelines of DS and LFMD

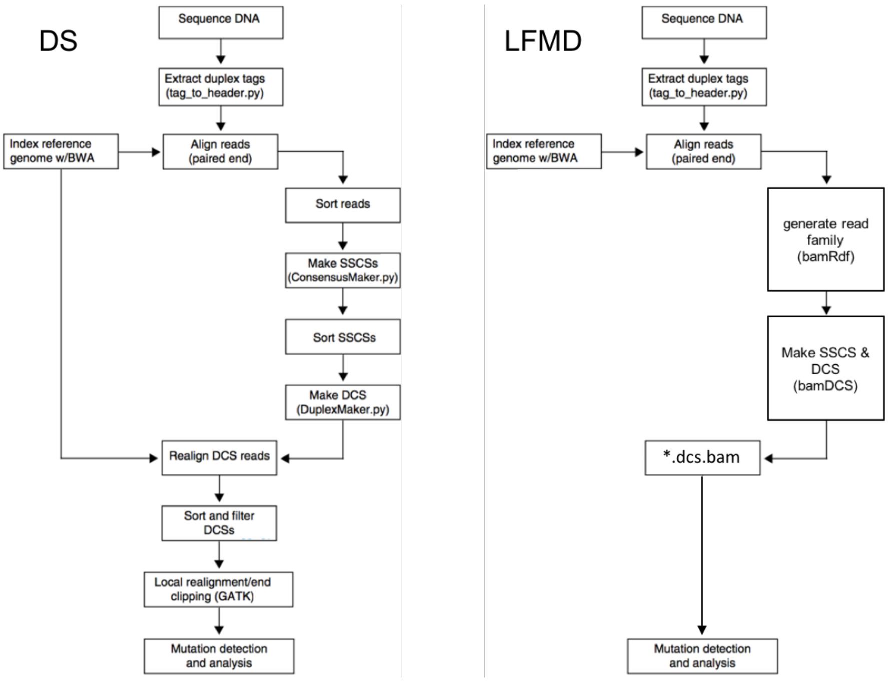

DS
Sequence DNA
Extract duplex tags (tag_to_header.py)
Index reference genome w/BWA
Align reads (paired end)
Sort reads
Make SSCSs (ConsensusMaker.py)
Sort SSCSs
Make DCS (DuplexMaker.py)
Realign DCS reads
Sort and filter DCSs
Local realignment/end clipping (GATK)
Mutation detection and analysis
LFMD
Sequence DNA
Extract duplex tags (tag_to_header.py)
Index reference genome w/BWA
Align reads (paired end)
generate read family (bamRdf)
Make SSCS & DCS (bamDCS)
*.dcs.bam
Mutation detection and analysis

Figure 3. SNV sensitivity of DS and LFMD

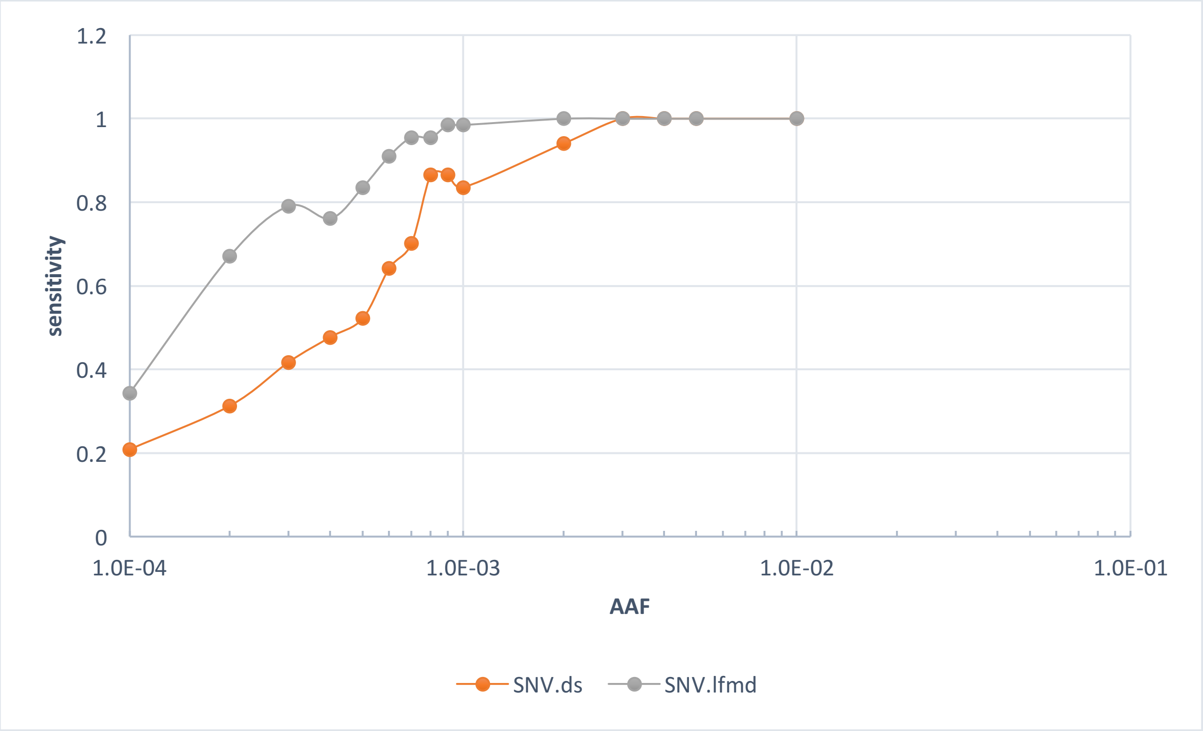


Figure 4. INS sensitivity of DS and LFMD

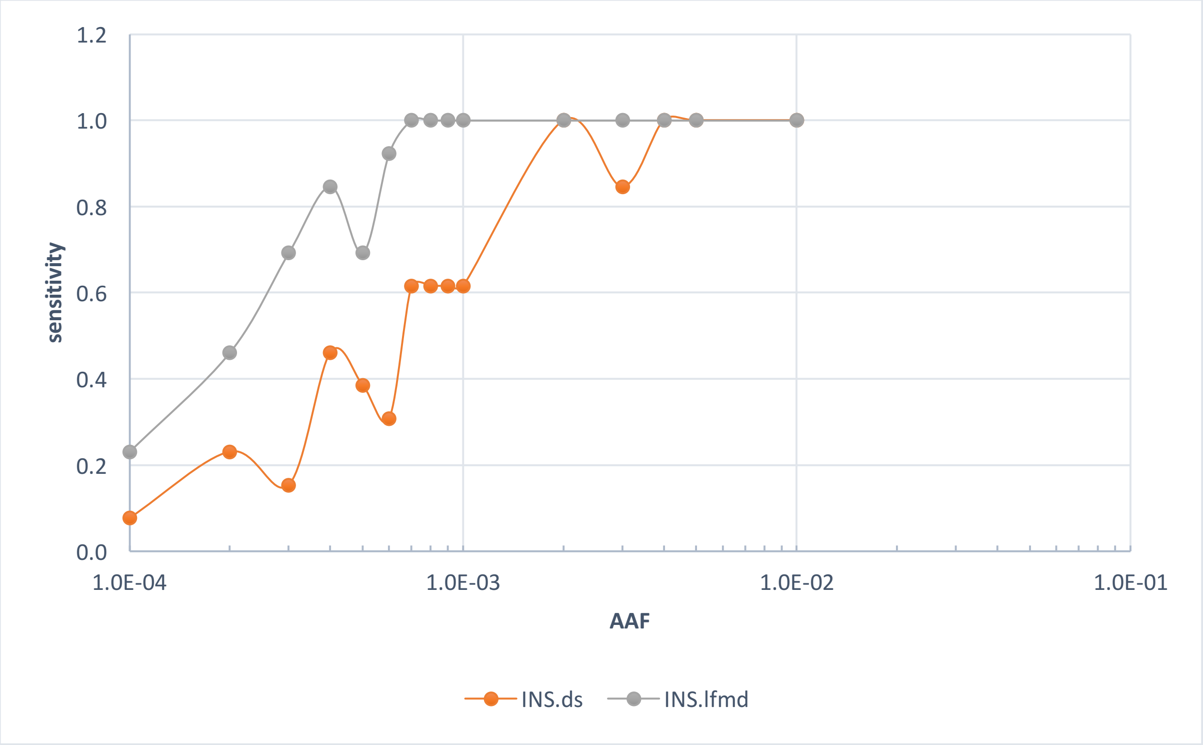


Figure 5. DEL sensitivity of DS and LFMD

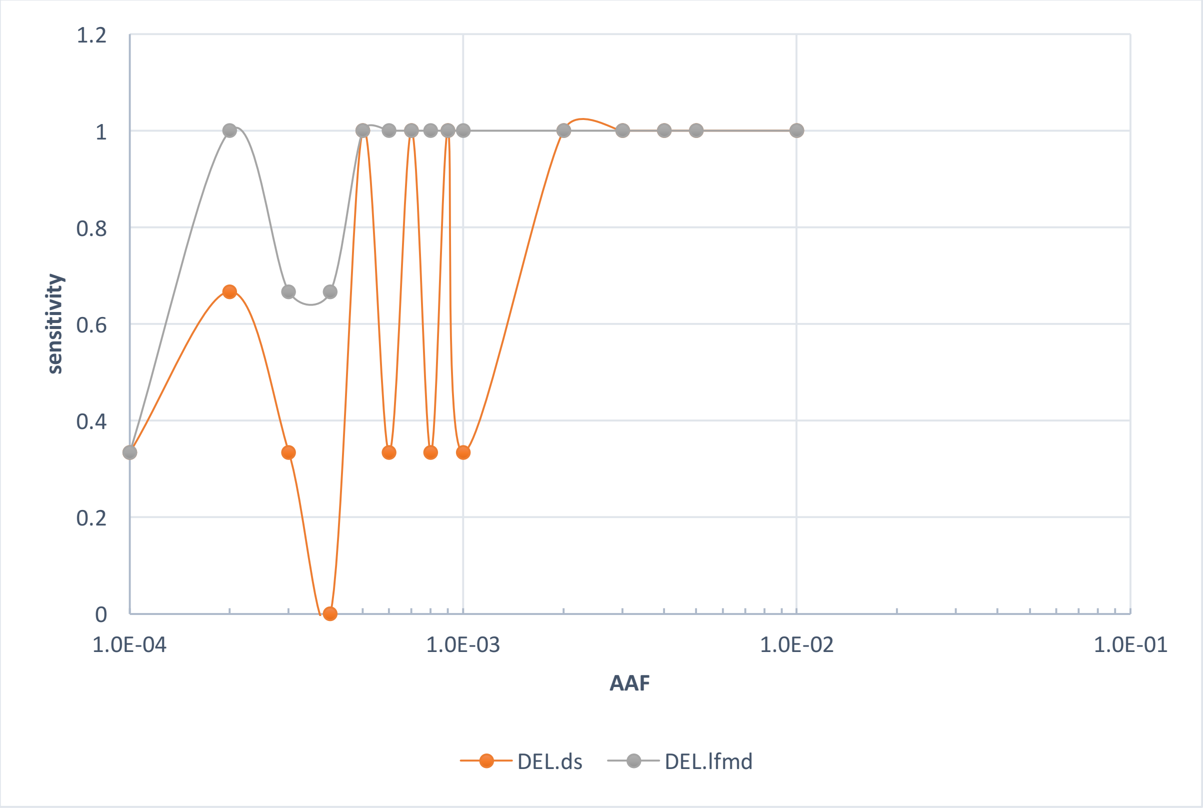


Fig. 6. Distribution of mutations found in mtDNA of YH cell lines compared with human Revised Cambridge Reference Sequence (rCRS).

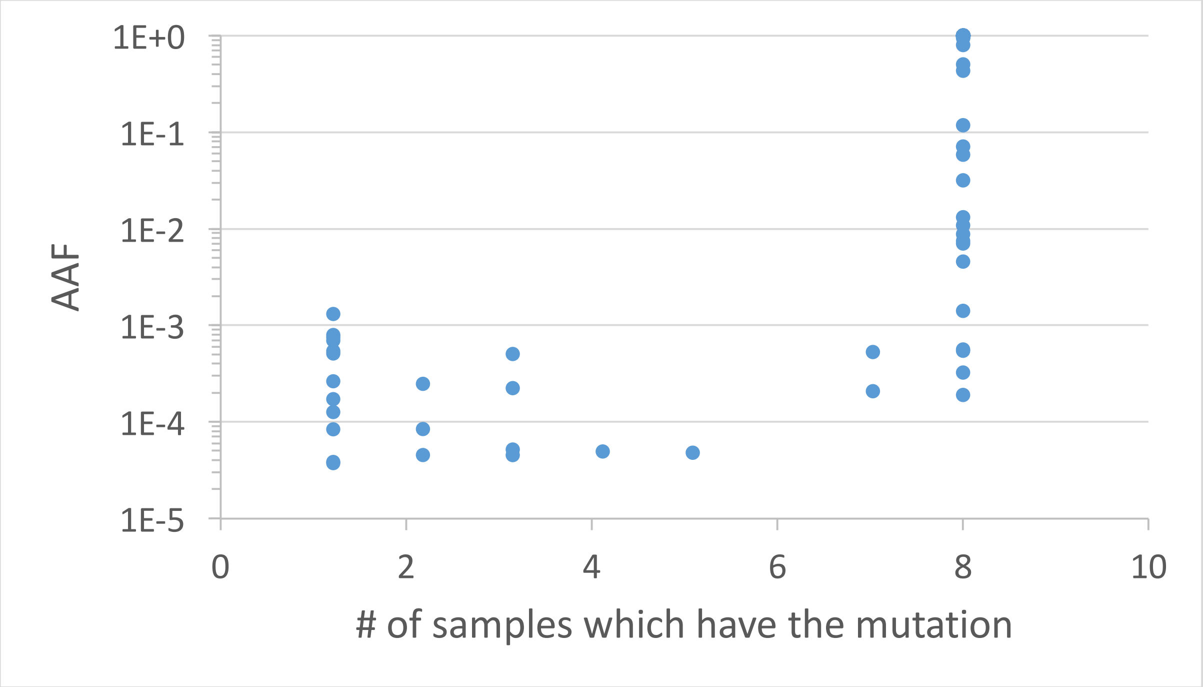


Figure S1. The comparison between theoretical and empirical P-values from Monto Carlo procedures under truly distributed sequencing error rates. With the null hypothesis, 1e6 times of simulations were conducted.

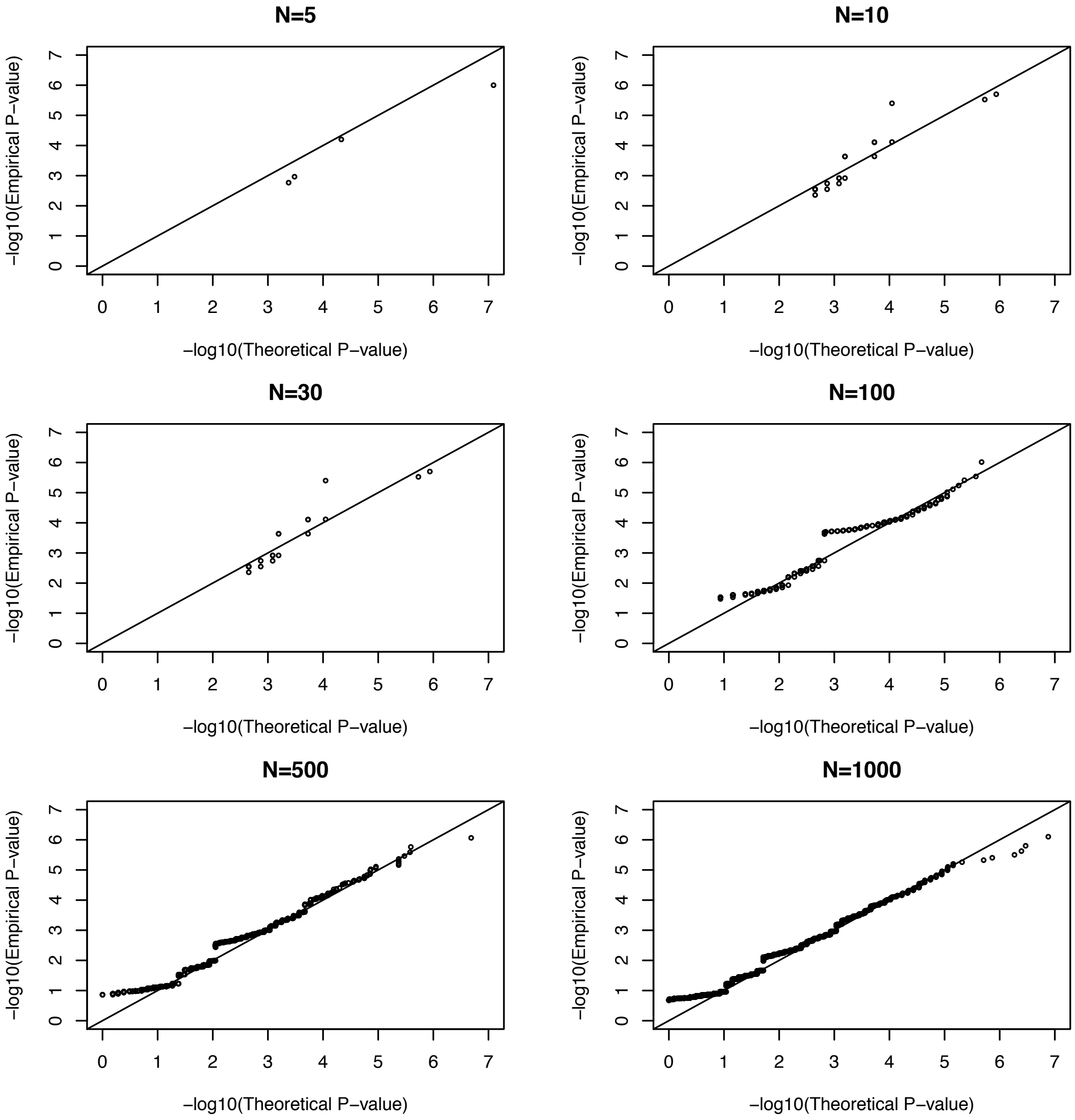
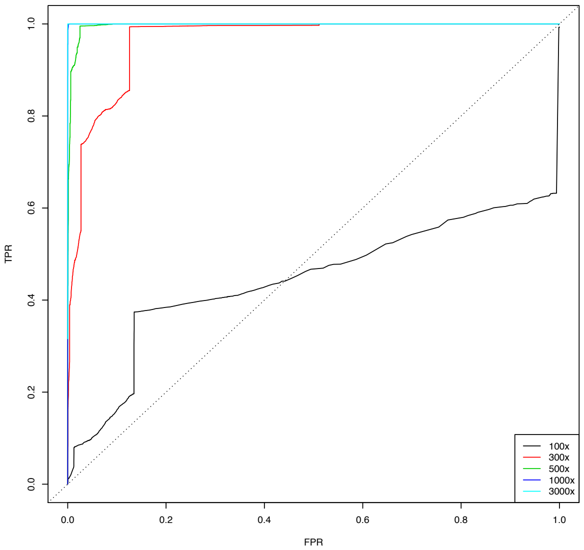
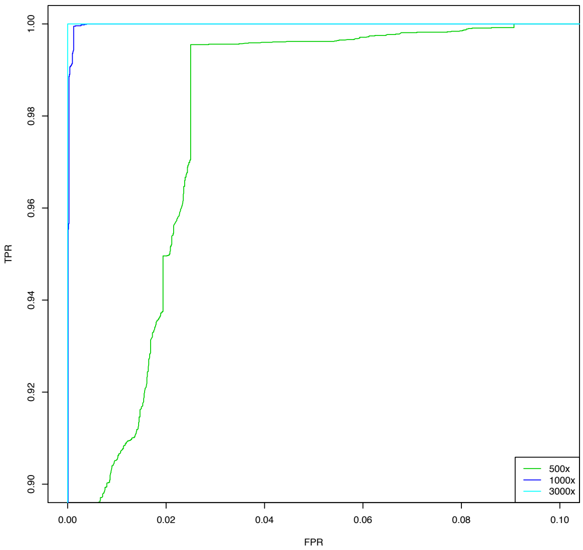


Figure S2. ROC curves of the likelihood-based model under uniformly distributed sequencing errors (Q20).

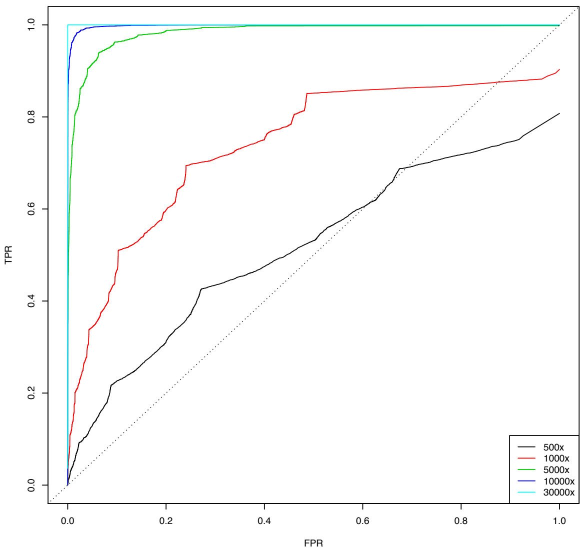
Q20 bases to detect 1% mutations



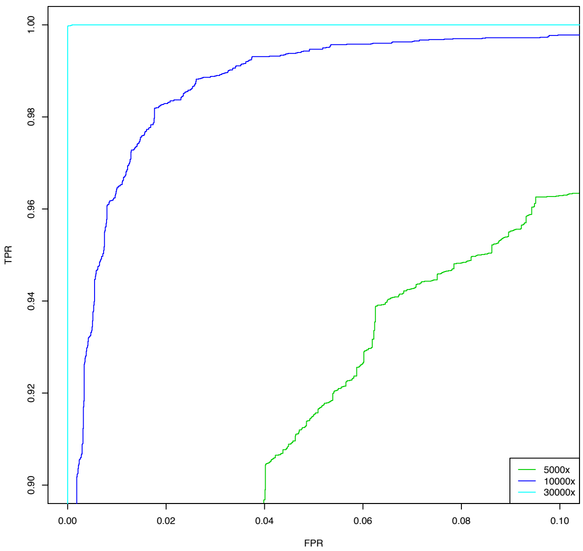
Q20 bases to detect 1% mutations



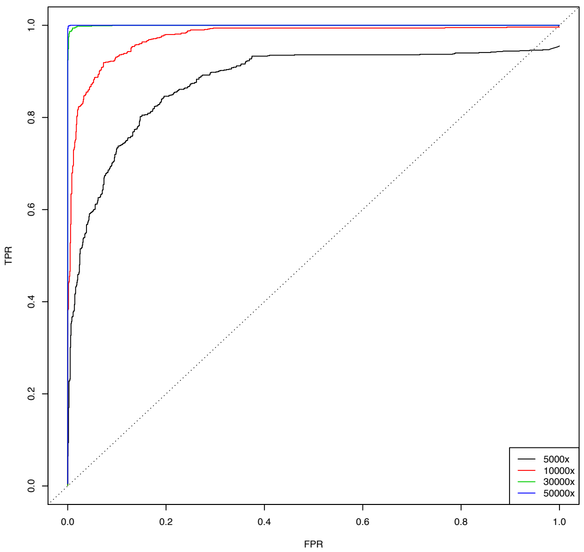
Q20 bases to detect 0.1% mutations



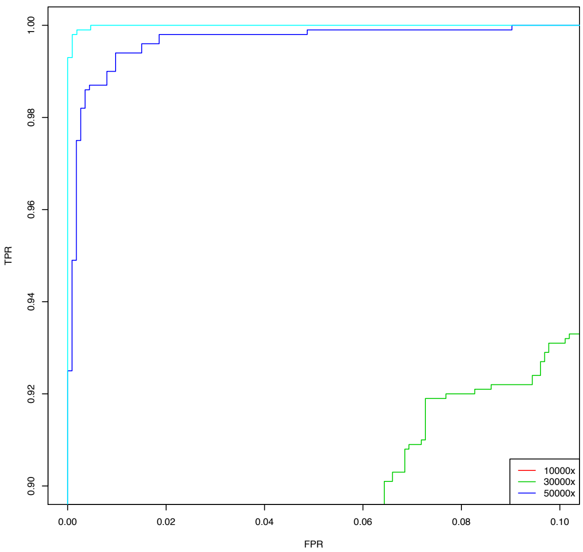
Q20 bases to detect 0.1% mutations



Q20 bases to detect 0.01% mutations



Q20 bases to detect 0.01% mutations



References

1. Hoekstra, J.G., Hipp, M.J., Montine, T.J. & Kennedy, S.R. Mitochondrial DNA mutations increase in early stage Alzheimer disease and are inconsistent with oxidative damage. *Annals of neurology* **80**, 301-306 (2016).
2. Ding, J. *et al.* Assessing Mitochondrial DNA Variation and Copy Number in Lymphocytes of ~2,000 Sardinians Using Tailored Sequencing Analysis Tools. *PLoS Genet* **11**, e1005306 (2015).
3. Wallace, D.C. & Chalkia, D. Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. *Cold Spring Harbor perspectives in biology* **5**, a021220 (2013).
4. Dimond, R. Social and ethical issues in mitochondrial donation. *British medical bulletin* **115**, 173 (2015).
5. Jabara, C.B., Jones, C.D., Roach, J., Anderson, J.A. & Swanstrom, R. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proceedings of the National Academy of Sciences* **108**, 20166-20171 (2011).
6. Schmitt, M.W. *et al.* Sequencing small genomic targets with high efficiency and extreme accuracy. *Nat Methods* **12**, 423-5 (2015).
7. Marquis, J. *et al.* MitoRS, a method for high throughput, sensitive, and accurate detection of mitochondrial DNA heteroplasmy. *BMC genomics* **18**, 326 (2017).
8. Kang, E. *et al.* Age-related accumulation of somatic mitochondrial DNA mutations in adult-derived human iPSCs. *Cell Stem Cell* **18**, 625-636 (2016).
9. Blandini, F., Greenamyre, J.T. & Nappi, G. The role of glutamate in the pathophysiology of Parkinson's disease. *Functional neurology* **11**, 3-15 (1996).
10. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90 (2011).
11. Baslan, T. & Hicks, J. Single cell sequencing approaches for complex biological systems. *Current opinion in genetics & development* **26**, 59-65 (2014).
12. Lou, D.I. *et al.* High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proceedings of the National Academy of Sciences* **110**, 19872-19877 (2013).
13. Ding, J. *et al.* Assessing mitochondrial DNA variation and copy number in lymphocytes of~ 2,000 Sardinians using tailored sequencing analysis tools. *PLoS genetics* **11**, e1005306 (2015).
14. Guo, Y., Li, J., Li, C.-I., Shyr, Y. & Samuels, D.C. MitoSeek: extracting mitochondria information and performing high-throughput mitochondria sequencing analysis. *Bioinformatics* **29**, 1210-1211 (2013).
15. Hoang, M.L. *et al.* Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc Natl Acad Sci U S A* **113**, 9846-51 (2016).
16. Schmitt, M.W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences* **109**, 14508-14513 (2012).
17. Schmitt, M.W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A* **109**, 14508-13 (2012).
18. Belmonte, F.R. *et al.* Digital PCR methods improve detection sensitivity and measurement precision of low abundance mtDNA deletions. *Sci Rep* **6**, 25186 (2016).

19. Drton, M. Likelihood ratio tests and singularities. *The Annals of Statistics* **37**, 979-1012 (2009).
20. Chen, Y., Huang, J., Ning, Y., Liang, K.-Y. & Lindsay, B.G. A conditional composite likelihood ratio test with boundary constraints. *Biometrika* **105**, 225-232 (2017).
21. Weiss, G. & Von Haeseler, A. A coalescent approach to the polymerase chain reaction. *Nucleic acids research* **25**, 3082-3087 (1997).
22. Stoler, N., Arbeithuber, B., Guiblet, W., Makova, K.D. & Nekrutenko, A. Streamlined analysis of duplex sequencing data with Du Novo. *Genome biology* **17**, 180 (2016).
23. Tate, J.G. *et al.* COSMIC: the catalogue of somatic mutations in cancer. *Nucleic acids research* **47**, D941-D947 (2018).
24. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60-5 (2008).