LFMD: detecting low-frequency mutations in genome sequencing data 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], Xiuqing Zhang[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).

V4.0 on 2019.09.12

Abstract

As next-generation sequencing (NGS) and liquid biopsy become more prevalent in clinic and research, for cancer diagnosis, molecular target identification, and disease monitoring, there is an increasing need for better methods to reduce cost and improve sensitivity and specificity of mutation detection. Since NGS has an error rate of around 1%, it is difficult to accurately and efficiently identify mutations with less than 1% frequency in a sample. Here we propose a likelihood-based approach, called Low-Frequency Mutation Detector (LFMD), which combines the advantages of duplex sequencing (DS) and bottleneck sequencing system (BotSeqS) to maximize the utilization of duplicate sequence reads. Compared with DS, the new method achieves higher sensitivity (improved by ~16%), higher specificity and lower cost (reduced by ~70%) without involving additional experimental steps, customized adapters or molecular tags. In addition, this method can also be used to improve sensitivity and specificity of other variant calling algorithms by making it unnecessary to remove polymerase chain reaction (PCR) duplicates.

Introduction

At the individual level, low-frequency mutations (LFMs) are defined as mutations with allele frequency lower than 5% or 1% in an individual's DNA. LFMs can indicate early stages of cancer and Alzheimer's Disease (AD)(1), distinguish samples from people of different ages (2), identify disease-causing variants(3), predict potential drug resistance(4), diagnose mitochondrial disease before tri-parental *in vitro* fertilization(5), and track the mutational spectrum of viral genomes, malignant lesions, and somatic tissues(4,6). To effectively improve signal-to-noise ratio (SNR) and detect LFMs, researchers have developed methods with stringent thresholds, complex experimental procedures(1,7), single cell sequencing(8-11), circle sequencing(12), and more precise analytic models(2,13). The bottleneck sequencing system(14) (BotSeqS) and duplex sequencing(15,16) (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 the amplification of small amounts of fragile DNA. To increase specificity, only variants shared by at least two cells are accepted as true variants(11). At present, 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 controls errors to a rate as low as 7.6×10^{-6} per base sequenced(12) while DS can achieve 4×10^{-10} errors per base sequenced(15).

Disadvantages of BotSeqS

In contrast, BotSeqS uses endogenous molecular tags 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 being cheap enough to sequence the whole human genome(14). However, it requires highly diluted DNA templates before PCR amplification to reduce endogenous tag conflicts and ensure sufficient sequencing of each DNA template. Thus, it has a high specificity with poor sensitivity. Also, it discards clonal variants and small insertions/deletions (InDels) to limit false positives.

Disadvantages of DS

Another promising method to eliminate tag conflicts is Duplex sequencing (DS), which 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, much sequencing data is wasted on sequence tags, fixed sequences and a large proportion of read families that contain only one read pair, which arose from 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(2,15) rather than on the whole genome.

Disadvantages of Tag clustering

To solve the problem induced by the errors on tags, multiple methods have been developed to cluster similar tags(17-19), where one or two mismatches are allowed in merging two tags. Although tag clustering does improve the sensitivity, it is still not a straightforward way to solve the problem. Inappropriate tag clustering might occur because unstable synthesis of random tags can result in distinct but similar tags. The performance of this approach has not been well studied and reported yet.

A new approach

To avoid the aforementioned problems, we present here a new, efficient approach that combines the advantages of BotSeqS and DS. The method uses a likelihood-based model(2,13) 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 the utilization of non-duplicate read pairs. This simplifies the DNA sequencing procedure, saves data and cost, achieves higher sensitivity and specificity, and can be used in whole genome sequencing. In addition, our new method offers a statistical solution to 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.

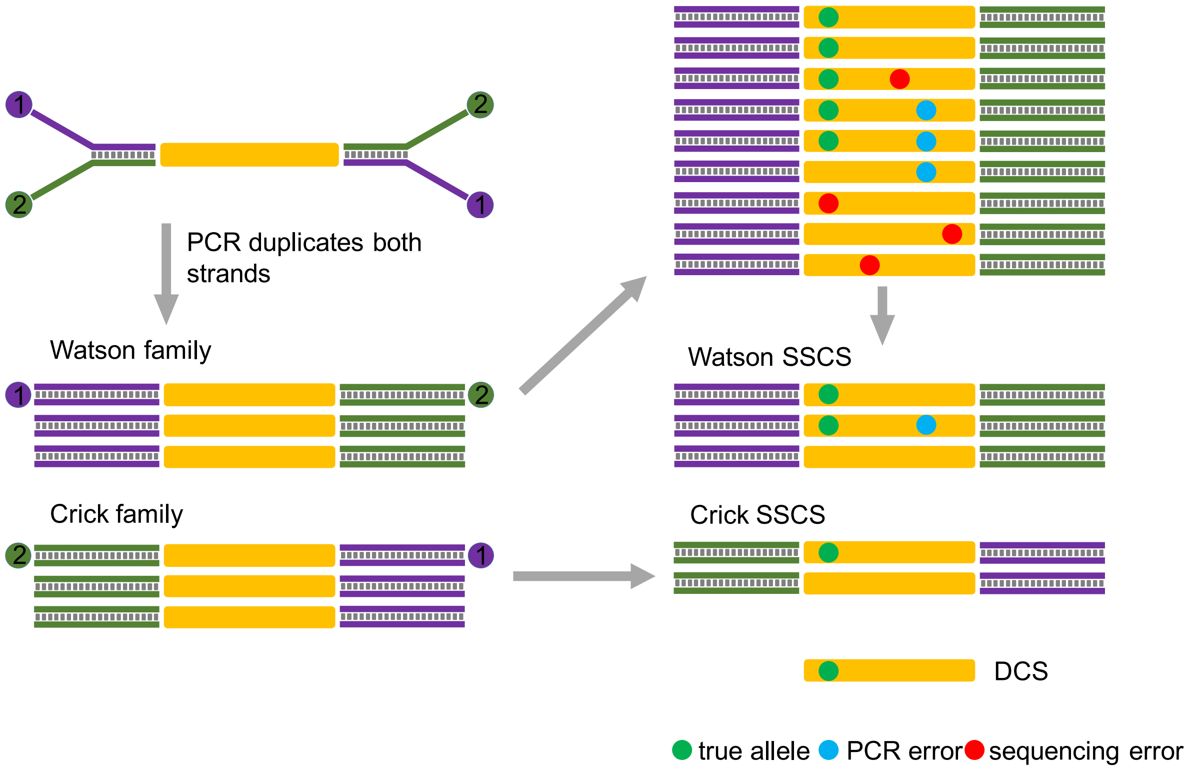
Methodology

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 reduce sequencing errors. Single cell sequencing increases the frequency of mutations by isolating single cells from the bulk population, while BotSeqS and DS reduce 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 molecules 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 are 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 overview of the LFMD pipeline is shown in Figure 1.

Figure 1. Overview of the 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 the likelihood-base model and given a combined error rate. Sequencing errors (red dots) are eliminated. Combining single-strand consensus sequences (SSCSs) of paired read families, high-quality double-strand consensus sequence (DCSs) with estimated error rates are generated and used in the downstream analysis.


1
2
2
1
PCR duplicates both strands
Watson family
Crick family
Watson SSCS
Crick SSCS
DCS
● true allele ● PCR error ● sequencing error

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 the 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 denote 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 represent 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

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

So the log-likelihood function can be written as

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

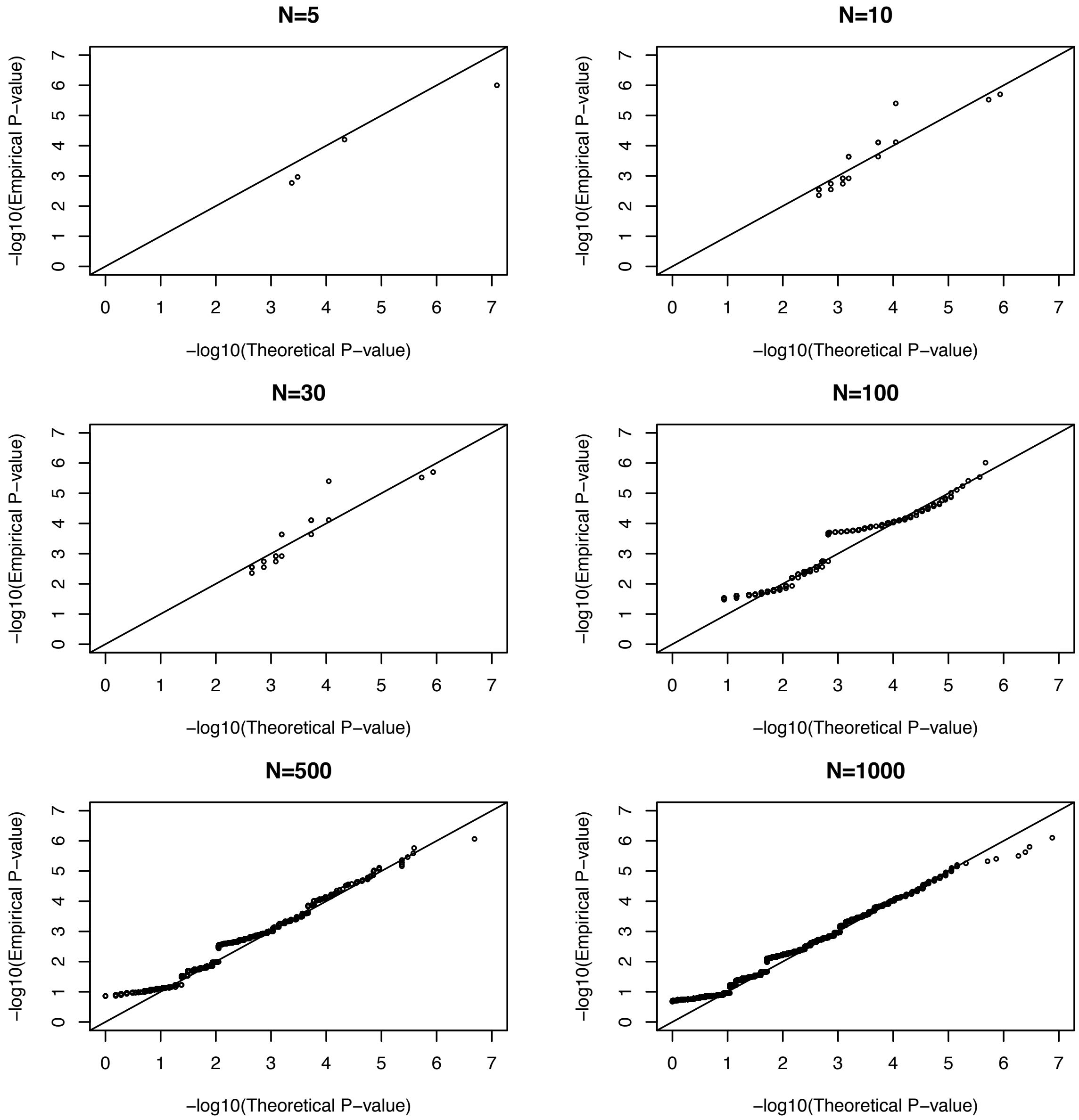
Thus, for each candidate allele g, under the null hypothesis $H_0\colon \theta_g = 0, \theta \in \Omega$, and the alternative hypothesis $H_1\colon \theta_g \neq 0, \theta \in \Omega$, the likelihood ratio test is

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

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(20) in 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(21) presented by Chen et al in 2017. (Details in Supplemental materials)

Interestingly, 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). Alternative approaches might be derived in the future. We also compared theoretical P-values with empirical P-values from Monte Carlo procedures (Figure 2), explored the power of our model under truly and uniformly distributed sequencing errors (Supplemental Material, Figure S1), and evaluated the accuracy of allele fraction (Supplemental Material, Figure S2). The simulation results support the theoretical conclusion sufficiently.

Figure 2. The comparison between theoretical and empirical P-values from Monte Carlo procedures under truly distributed sequencing error rates. With the null hypothesis, one million simulations were conducted.



Because the null and alternative hypotheses have two and three free variables respectively, the Chi-square distribution has 1 degree of freedom. The P-value 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 $P_w(g)$ and $P_c(g)$ represent the relative error rates of the given allele in the Watson and Crick family respectively, and let $P_{wc}(g)$ denote the united error rate of the allele. Thus,

$$P_{wc}(g) = P_w(g) + P_c(g) - P_w(g)P_c(g)$$

For a read family which proliferated from $\boldsymbol{n}$ original templates, a coalescent model can be used to model the PCR procedure(22). The exact coalescent PCR error rate is too complicated to be calculated quickly, so we tried to give a rough estimate. According to the model, a PCR error proliferates and its fraction decreases exponentially with the number of PCR cycles, $\boldsymbol{m}$. 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 or less than

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

Let $e_{pcr}(g)$ denote the coalescent PCR error rate and $P_{pcr}(g)$ the united PCR error rate of the double strand consensus allele. Then we get

$$P_{pcr}(g) \approx \boldsymbol{n} * e_{pcr}(g)^2$$

Because the PCR fidelity ranges from 10^{-5} to 10^{-7}, we get $P_{wc}(g)P_{pcr}(g) \approx 0$, then the combined base quality of the allele on the DCS is

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

Then $Q(g)$ 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 Ding et al(2) and Guo et al(13). Ding 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.

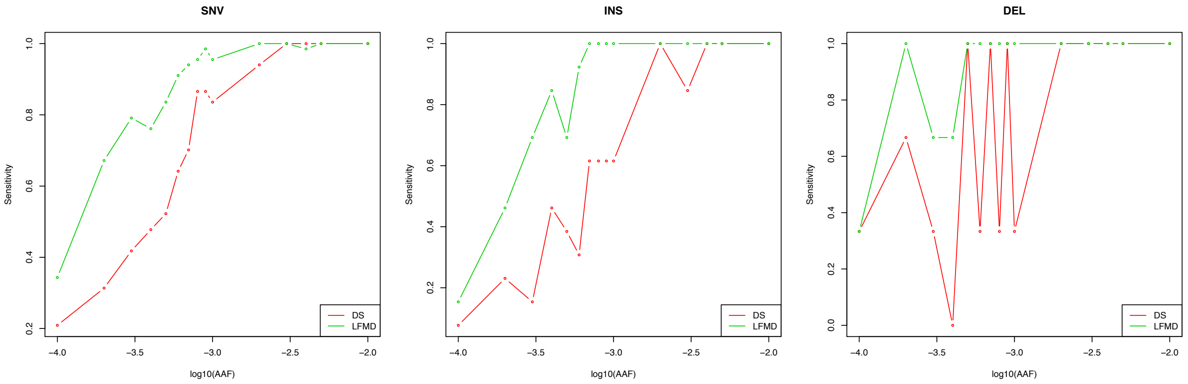
For those reads containing InDels, the CIGAR strings in BAM files contain I or D. It is obvious that reads with different CIGAR strings cannot fit into one read family. Thus, CIGAR strings can also be used as part of endogenous tags. In contrast, the soft-clipped part of CIGAR strings cannot be ignored 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

Simulated data

We used Python scripts developed by the Du Novo(23) team to simulate mixed double-strand sequencing data, which were analyzed using LFMD and DS. Although the simulation may not be entirely realistic, the results are still useful to evaluate the performance and the potential drawbacks of LFMD and DS. Because the underlying true mutations are known, the true positives and false positives can be calculated. We found that DS made two false positive rare variant calls due to mapping errors, which were not made by LFMD because it avoids mapping errors by skipping the realignment step. In the meanwhile, LFMD is much more sensitive than DS according to Figure 3 and Table S1. The specificity of LFMD is slightly higher than that of DS shown in Table S2.

Figure 3. The sensitivity of DS and LFMD. SNV, single nucleotide variant; INS, small insertion; DEL, small deletion; AAF, alternative allele fraction.



Mouse mtDNA

To evaluate the performance of LFMD in real data, we compared LFMD with DS on a DS dataset of mouse mtDNA: SRR1613972. A comparison of DS and LFMD pipelines is shown in Figure S3. We controlled almost all parameters to be the same in DS and LFMD and then compared the results although LFMD can achieve much higher sensitivity if we set the minimum number of supporting reads in each read family as 2 (DS suggests the parameter as 3 to balance sensitivity and specificity). We found that mapping quality influenced the performance of both methods. To reduce the influence of mapping quality, we only use all unique proper mapped read pairs to call mutations. The results are shown in Table S3.

We investigated the discordant mutations one by one. The DS_only mutations are all false positives due to read mapping errors (MT:7G>A and MT:3418T>del) and low sequencing quality (MT:5462T>G). The mapping errors are from the following: 1) DS assigns highest base quality for all bases of double-strand consensus sequence (DCS)(15) even though the base is ’N’ and 2) DS introduces more ‘N’ in DCS reads because the minimum proportion of “true” bases in a read family is set to 0.7 by default (it can be set to 0.5 to improve sensitivity).

Among the 53 LFMD_only mutations, 2 insertions and 3 deletions are missed by DS due to mapping errors of DCS, and 48 SNVs are not detected by DS due to three technical reasons: 1) sequencing and PCR errors on tags lead to smaller read families in DS, decreasing its sensitivity; 2) DS discards some low complexity tags; 3) DS assumes "true" mutations should occupy most reads (proportion >= 0.7) in the reads family, although this assumption is unreasonable considering PCR errors in the first PCR cycle.

In summary, the discordant mutations in this dataset are mainly false positives and false negatives of DS. All LFMD_only mutations are manually checked and well supported by read families under the same criteria of DS considering 1 or 2 sequencing and PCR errors on tags. Therefore, LFMD achieves higher sensitivity and specificity than DS in this dataset.

Twenty-six human mtDNA 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 proper mapped reads were used to detect LFMs. The majority of LFMs were detected by both tools. Almost all LFMs detected only by DS were false positives due to alignment errors of DCS, while LFMD outputs BAM files directly and avoids alignment errors (Supplemental Material, Figure S3). 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 of sensitivity is about 16% according to Table S4.

YH cell line

We sequenced the YH cell line (passage 19) in 8 independent experiments to evaluate the stability of LFMD. The experimental details can be found in the Materials part. The results, shown in Supplemental Materials, Table S5 and Figure S4, from the 8 parallel samples are highly consistent in terms of numbers of mutations detected (range 61~68). Under the hypothesis that true mutations should be identified from at least two samples, we detected 68 "true" mutations and the mean true positive rate (TPR) and false discovery rate (FDR) are around 91.36% and 2.36% respectively.

ABL1 data

Using the duplex sequencing method in 2015, Schmitt et al. analyzed an individual with chronic myeloid leukemia who relapsed after targeted therapy with the drug, Imatinib (the Short Read Archive under accession SRR1799908). We analyzed this individual and found 5 extra LFMs. Two of them are in the coding region of the *ABL1* gene and change amino acids: E255G and V256G. In the drug resistance database of COSMIC(24), we found that E255VDK, change of the 255th amino acid, is associated with resistance to the drugs Dasatinib, Imatinib, and Nilotinib, and V256L is related to resistance to the drug Imatinib. Although the directions of amino acid changes, in this case, are not the same as those in the database, these two additional LFMs still inferred potential resistance to the drug Imatinib and provided an additional explanation for the clinical relapse of leukemia. The annotation results of 5 LFMs are shown in Supplemental Materials, Table S6.

Materials

Subject recruitment and sampling

A lymphoblastoid cell line (YH cell line) established from the first Asian genome donor(25) was used. Total DNA was extracted with the MagPure Buffy Coat DNA Midi KF Kit (MAGEN). The DNA concentration was quantified by Qubit (Invitrogen). 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 S7).

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

150bp to 500bp 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 10bp 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 PE100 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 100bp paired-end reads. The libraries were processed for high-throughput sequencing with a mean depth of ~60000x.

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. Analysis codes used in this paper can be accessed at https://github.com/RainyEricYe/LFMD.

Discussion

LFMD is still expensive for target regions >2 Mbp in size because of the need for high sequencing depth. However, as the cost of sequencing continues to fall, it will become increasingly practical. In order to sequence a larger region at reduced cost, the dilution step of

BotSeqS can be introduced into the LFMD pipeline. Because LFMD can deal with tag conflicts, the dilution level might be decreased several magnitudes to increase the sensitivity. Additional experiments will be done soon.

Only accepting random sheared 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 the alignment software. Although tags are excluded in the model of LFMD, LFMD still has the potential to utilize tags and deal with amplicon sequencing data. The basic assumption in our model, that error rates are the same in all three directions, is not close enough to reality according to experimental data at present. These remain issues may be solved in the next version of LFMD.

To estimate the theoretical limit of LFMD, let read length be 100bp and the standard deviation (SD) of insert size be 20bp. Furthermore, 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).

LFMD reduces the cost dramatically mainly because it discards tags. First, for a typical 100bp read, the lengths of the tags and the fixed sequences between the tag and the true sequence are 12bp and 5bp respectively. So (12+5) / 100 = 17% of data are saved if we discard tags directly. Second, the efficiency of target capture decreases by about 10% to 20% because of the tags, according to in-house experiments. Third, LFMD can work on short read data of BGISEQ and then 30% to 40% of the cost can be saved because of the cheaper sequencing platform. Totally, the cost can be reduced by about 70%.

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 specificity with lower cost comparing with by far the best method, DS. It is a general method that can be used in several cutting-edge areas and its mathematical methodology can be generalized to increase the power of other NGS tools.

Acknowledgments

We thank the Shenzhen Municipal Government of China for its support (No: JCYJ20170412153100794).

Funding

This work is supported by the Shenzhen Municipal Government of China (No: JCYJ20170412153100794). The funders had no role in study design, data collection, analysis, decision to publish, or preparation of the manuscript.

References

1. Hoekstra, J.G., Hipp, M.J., Montine, T.J. and Kennedy, S.R. (2016) Mitochondrial DNA mutations increase in early stage Alzheimer disease and are inconsistent with oxidative damage. *Annals of neurology*, **80**, 301-306.
2. Ding, J., Sidore, C., Butler, T.J., Wing, M.K., Qian, Y., Meirelles, O., Busonero, F., Tsoi, L.C., Maschio, A. and Angius, A. (2015) Assessing mitochondrial DNA variation and copy number in lymphocytes of~ 2,000 Sardinians using tailored sequencing analysis tools. *PLoS genetics*, **11**, e1005306.
3. Wallace, D.C. and Chalkia, D. (2013) Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. *Cold Spring Harbor perspectives in biology*, **5**, a021220.
4. Schmitt, M.W., Fox, E.J., Prindle, M.J., Reid-Bayliss, K.S., True, L.D., Radich, J.P. and Loeb, L.A. (2015) Sequencing small genomic targets with high efficiency and extreme accuracy. *Nature methods*, **12**, 423.
5. Dimond, R. (2015) Social and ethical issues in mitochondrial donation. *British medical bulletin*, **115**, 173.
6. Jabara, C.B., Jones, C.D., Roach, J., Anderson, J.A. and Swanstrom, R. (2011) 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.
7. Marquis, J., Lefebvre, G., Kourmpetis, Y.A., Kassam, M., Ronga, F., De Marchi, U., Wiederkehr, A. and Descombes, P. (2017) MitoRS, a method for high throughput, sensitive, and accurate detection of mitochondrial DNA heteroplasmy. *BMC genomics*, **18**, 326.
8. Kang, E., Wang, X., Tippner-Hedges, R., Ma, H., Folmes, C.D., Gutierrez, N.M., Lee, Y., Van Dyken, C., Ahmed, R. and Li, Y. (2016) Age-related accumulation of somatic mitochondrial DNA mutations in adult-derived human iPSCs. *Cell Stem Cell*, **18**, 625-636.

9. Blandini, F., Greenamyre, J.T. and Nappi, G. (1996) The role of glutamate in the pathophysiology of Parkinson's disease. *Functional neurology*, **11**, 3-15.

10. Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D. and Esposito, D. (2011) Tumour evolution inferred by single-cell sequencing. *Nature*, **472**, 90.

11. Baslan, T. and Hicks, J. (2014) Single cell sequencing approaches for complex biological systems. *Current opinion in genetics & development*, **26**, 59-65.

12. Lou, D.I., Hussmann, J.A., McBee, R.M., Acevedo, A., Andino, R., Press, W.H. and Sawyer, S.L. (2013) High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proceedings of the National Academy of Sciences*, **110**, 19872-19877.

13. Guo, Y., Li, J., Li, C.-I., Shyr, Y. and Samuels, D.C. (2013) MitoSeek: extracting mitochondria information and performing high-throughput mitochondria sequencing analysis. *Bioinformatics*, **29**, 1210-1211.

14. Hoang, M.L., Kinde, I., Tomasetti, C., McMahon, K.W., Rosenquist, T.A., Grollman, A.P., Kinzler, K.W., Vogelstein, B. and Papadopoulos, N. (2016) Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proceedings of the National Academy of Sciences*, **113**, 9846-9851.

15. Schmitt, M.W., Kennedy, S.R., Salk, J.J., Fox, E.J., Hiatt, J.B. and Loeb, L.A. (2012) Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences*, **109**, 14508-14513.

16. Kennedy, S.R., Schmitt, M.W., Fox, E.J., Kohrn, B.F., Salk, J.J., Ahn, E.H., Prindle, M.J., Kuong, K.J., Shen, J.C., Risques, R.A. *et al.* (2014) Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc*, **9**, 2586-2606.

17. Peng, Q., Satya, R.V., Lewis, M., Randad, P. and Wang, Y. (2015) Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes. *BMC genomics*, **16**, 589.

18. Kou, R., Lam, H., Duan, H., Ye, L., Jongkam, N., Chen, W., Zhang, S. and Li, S. (2016) Benefits and challenges with applying unique molecular identifiers in next generation sequencing to detect low frequency mutations. *PloS one*, **11**, e0146638.

19. Smith, T., Heger, A. and Sudbery, I. (2017) UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res*, **27**, 491-499.

20. Drton, M. (2009) Likelihood ratio tests and singularities. *The Annals of Statistics*, **37**, 979-1012.

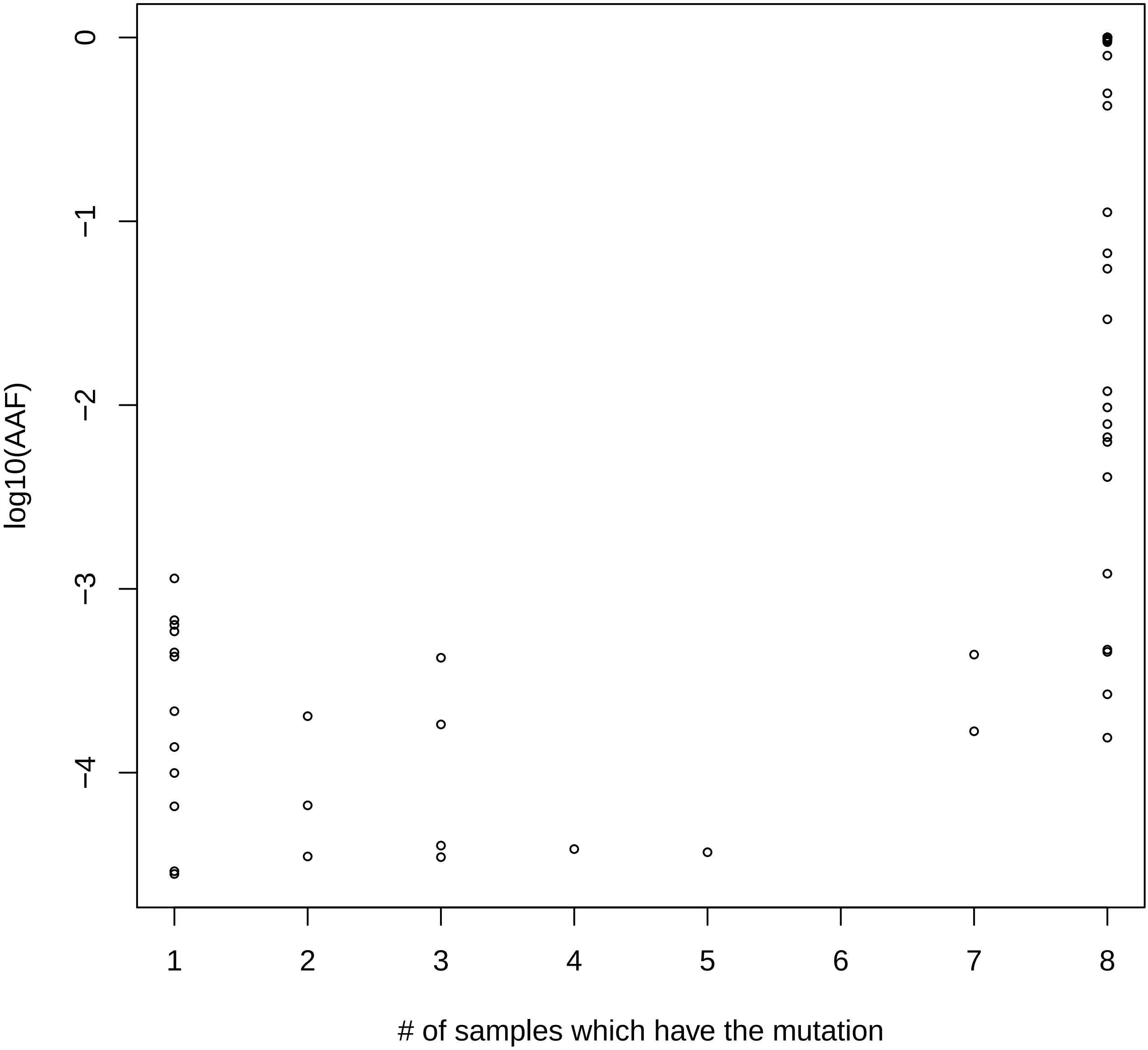
21. Chen, Y., Huang, J., Ning, Y., Liang, K.-Y. and Lindsay, B.G. (2017) A conditional composite likelihood ratio test with boundary constraints. *Biometrika*, **105**, 225-232.

22. Weiss, G. and Von Haeseler, A. (1997) A coalescent approach to the polymerase chain reaction. *Nucleic acids research*, **25**, 3082-3087.

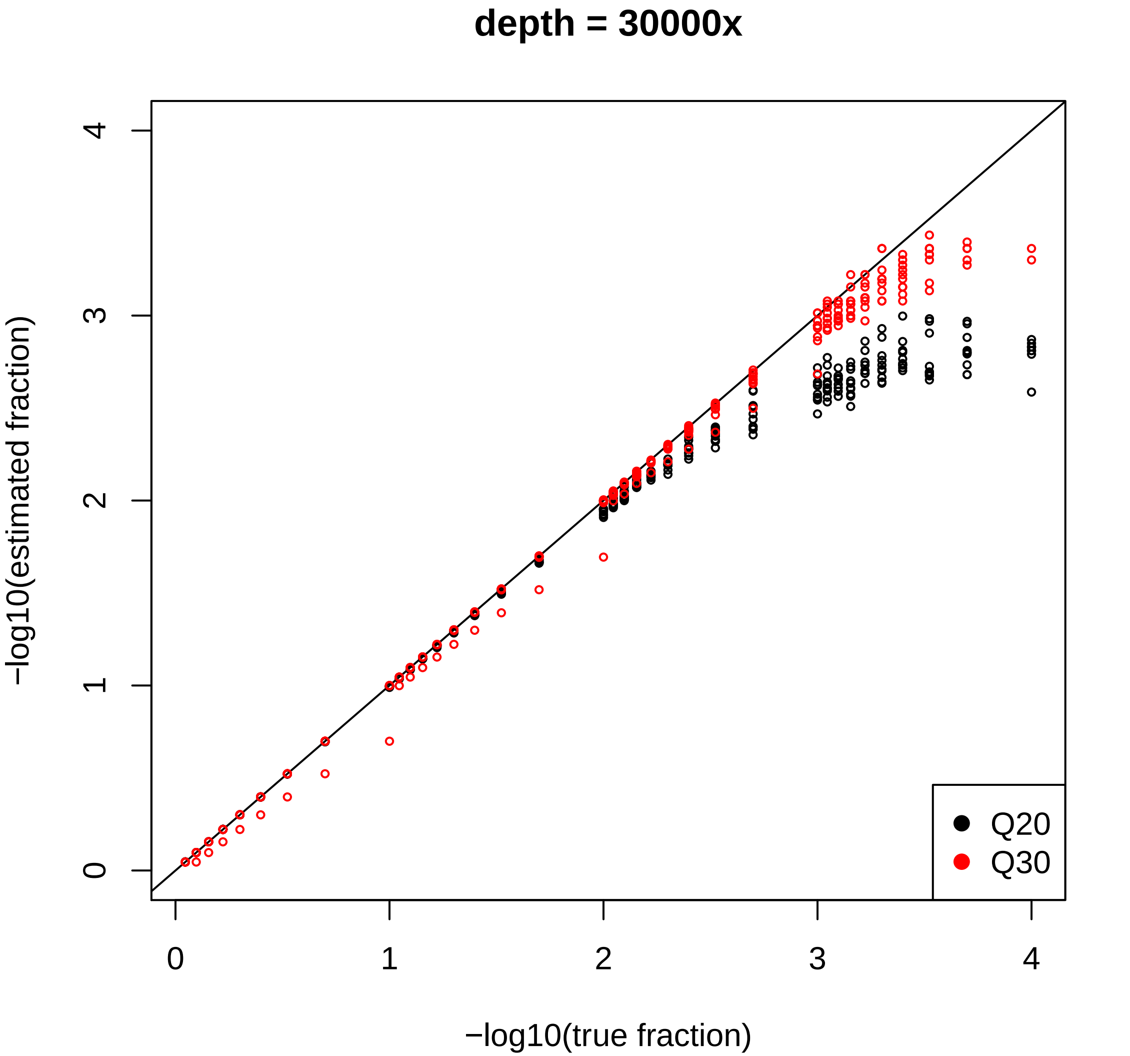
23. Stoler, N., Arbeithuber, B., Guiblet, W., Makova, K.D. and Nekrutenko, A. (2016) Streamlined analysis of duplex sequencing data with Du Novo. *Genome biology*, **17**, 180.

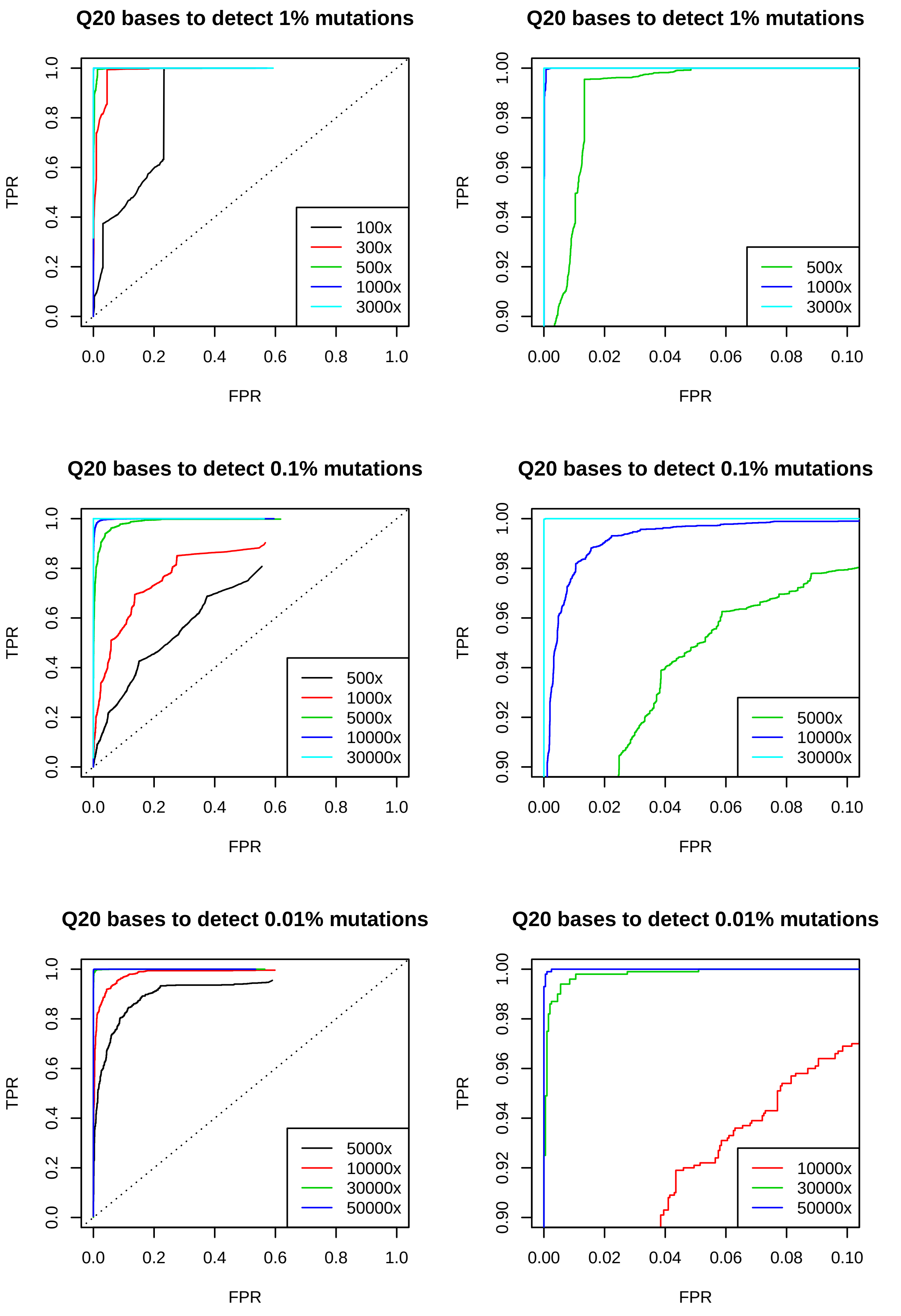
24. Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C. and Dawson, E. (2018) COSMIC: the catalogue of somatic mutations in cancer. *Nucleic acids research*, **47**, D941-D947.

25. Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60-65.


0
−1
−2
−3
−4
log10(AAF)
1
2
3
4
5
6
7
8
# of samples which have the mutation

DS
Sequence DNA
Extract duplex tags
Index reference
Align reads
Sort reads
Make SSCSs
Sort SSCSs
Make DCSs
Align DCSs
Sort and filter DCS
Local realignment
End trimming
Mutation detection
LFMD
Sequence DNA
Index reference
Align reads
Generate read family (bamRDF)
Make DCSs (bamDCS)
*.dcs.bam
Mutation detection


depth = 30000x
−log10(estimated fraction)
−log10(true fraction)
0
1
2
3
4
0
1
2
3
4
● Q20
● Q30



Q20 bases to detect 1% mutations

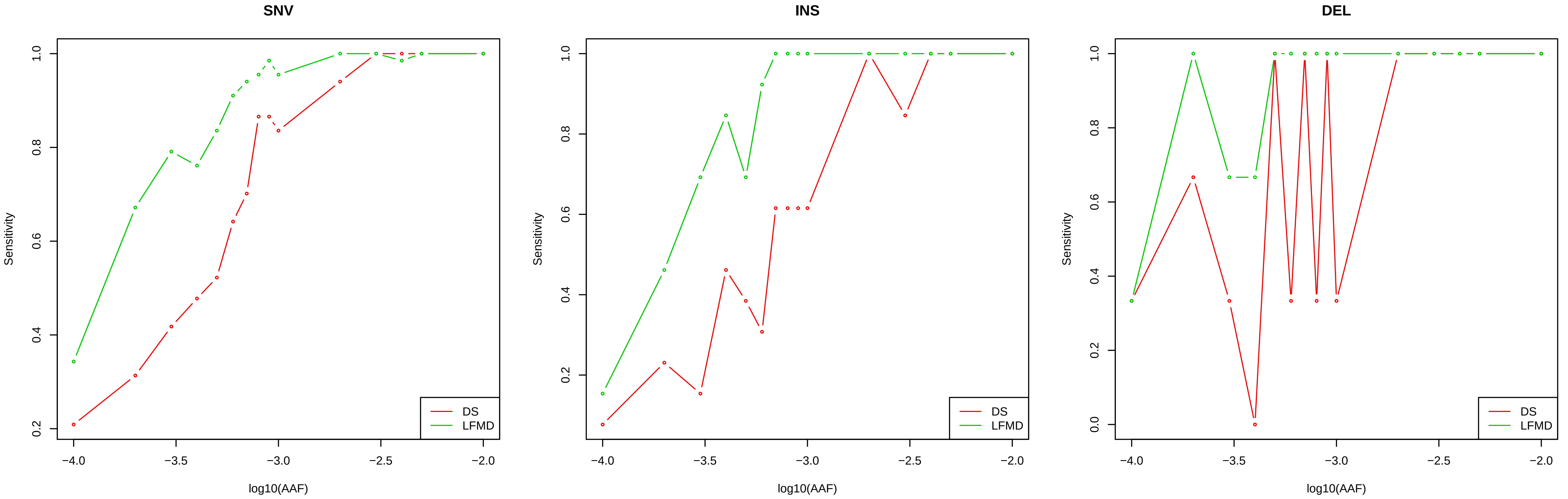
Q20 bases to detect 1% mutations

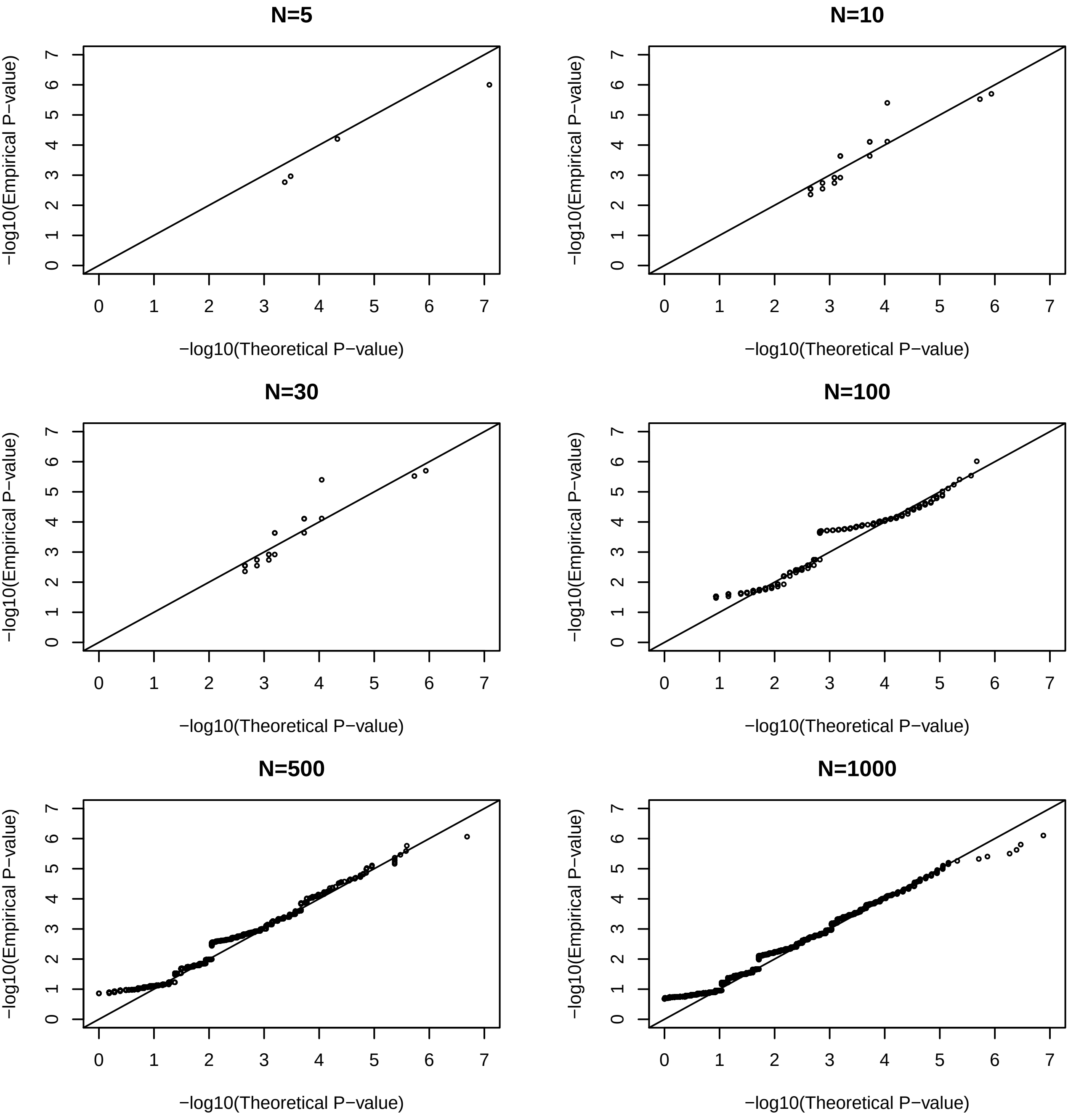
Q20 bases to detect 0.1% mutations

Q20 bases to detect 0.1% mutations

Q20 bases to detect 0.01% mutations

Q20 bases to detect 0.01% mutations


SNV
INS
DEL
Sensitivity
log10(AAF)
−4.0
−3.5
−3.0
−2.5
−2.0
0.0
0.2
0.4
0.6
0.8
1.0
DS
LFMD


N=5
N=10
N=30
N=100
N=500
N=1000
−log10(Empirical P−value)
−log10(Theoretical P−value)
0
1
2
3
4
5
6
7

1
2
2
1
PCR duplicates both strands
Watson family
1
2
Crick family
2
1
Watson SSCS
Crick SSCS
DCS
● true allele
● PCR error
● sequencing error