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).

V2.3 on 2019.06.23

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 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)(Hoekstra et al. 2016), distinguish samples with different age(Ding et al. 2015b), identify disease-causing variants(Wallace and Chalkia 2013), diagnose before tri-parental in vitro fertilization(Dimond 2015), and track the mutational spectrum in viral genomes, malignant lesions, and somatic tissues(Jabara et al. 2011; Schmitt et al. 2015). To effectively improve signal-to-noise ratio (SNR) and detect LFMs, stringent thresholds, complex experimental skills(Hoekstra et al. 2016; Marquis et al. 2017), single cell sequencing(Blandini et al. 1996; Navin et al. 2011; Baslan and Hicks 2014; Kang et al. 2016), circle sequencing(Lou et al. 2013), and more precise models(Guo et al. 2013; Ding et al. 2015a) were developed. The bottleneck sequencing system(Hoang et al. 2016) (BotSeqS) and duplex sequencing(Schmitt et al. 2012b; Kennedy et al. 2014) (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(Baslan and Hicks 2014). 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(Lou et al. 2013) while DS can achieve 4×10^{-10} errors per base sequenced(Schmitt et al. 2012b).

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(Hoang et al. 2016). 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(Schmitt et al. 2012a; Ding et al. 2015a; Schmitt et al. 2015) 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(Peng et al. 2015; Kou et al. 2016; Smith et al. 2017). One or two mismatches are allowed to merge two tags. Although tag clustering does improve the sensitivity, it is still not a straight forwards way to solve the problem. False tag clustering might occur because unstable synthesis of random tags contains relatively higher proportion

of similar tags. In the meanwhile, the evaluation of these methods on real data might be difficult and no conclusion has been reached.

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(Guo et al. 2013; Ding et al. 2015a) 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 50bp 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(Belmonte et al. 2016). 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.

Results

Comparison between DS and LFMD

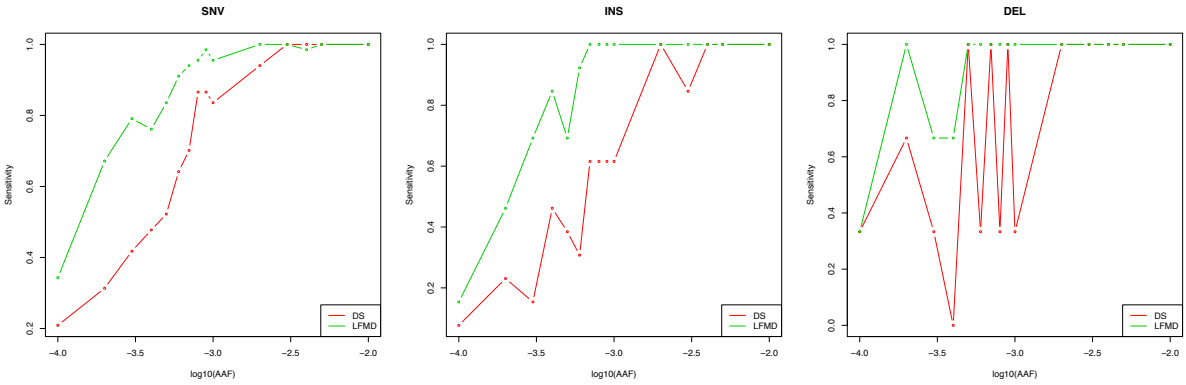
<u>Simulated data</u>

We used Python scripts developed by the Du novo(Stoler et al. 2016) 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 S1 and S2.

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. In the meanwhile, LFMD is much more sensitive than DS according to Figure 1 and Table S1.

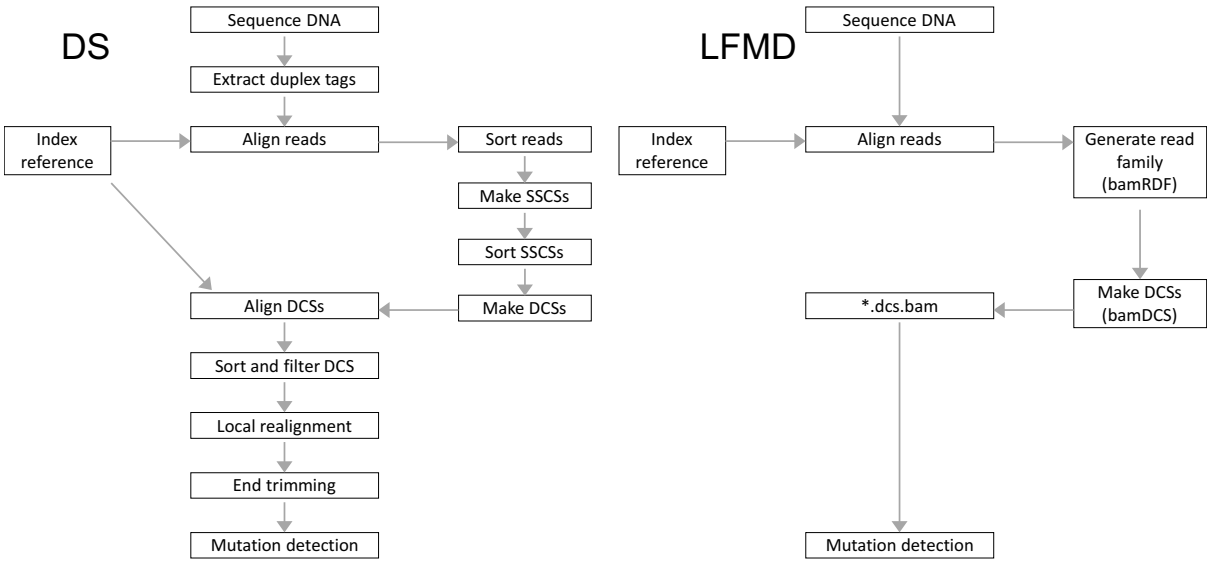
Figure 1. Sensitivity of DS and LFMD.



Mouse mtDNA

In order to evaluate the performance of LFMD, we compared LFMD with DS on a DS data from mouse mtDNA: SRR1613972. A comparison of DS and LFMD pipelines is shown in Figure 2. 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 S3) and unique proper mapped reads (Table S4). We found that mapping quality influenced the performance of both methods.

Figure 2. Pipelines of DS and LFMD.



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 67 LFMD-only point mutations) can be identified if we consider 1-2bp sequencing errors and PCR errors in the 24bp 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 23 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 2bp mismatches in tags, most of the last 23 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(Hoekstra et al. 2016)

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 S5.

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 S6 and Figure S1, 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(Tate et al. 2018). The annotation results of 5 LFMs are shown in Table S7.

Methodology

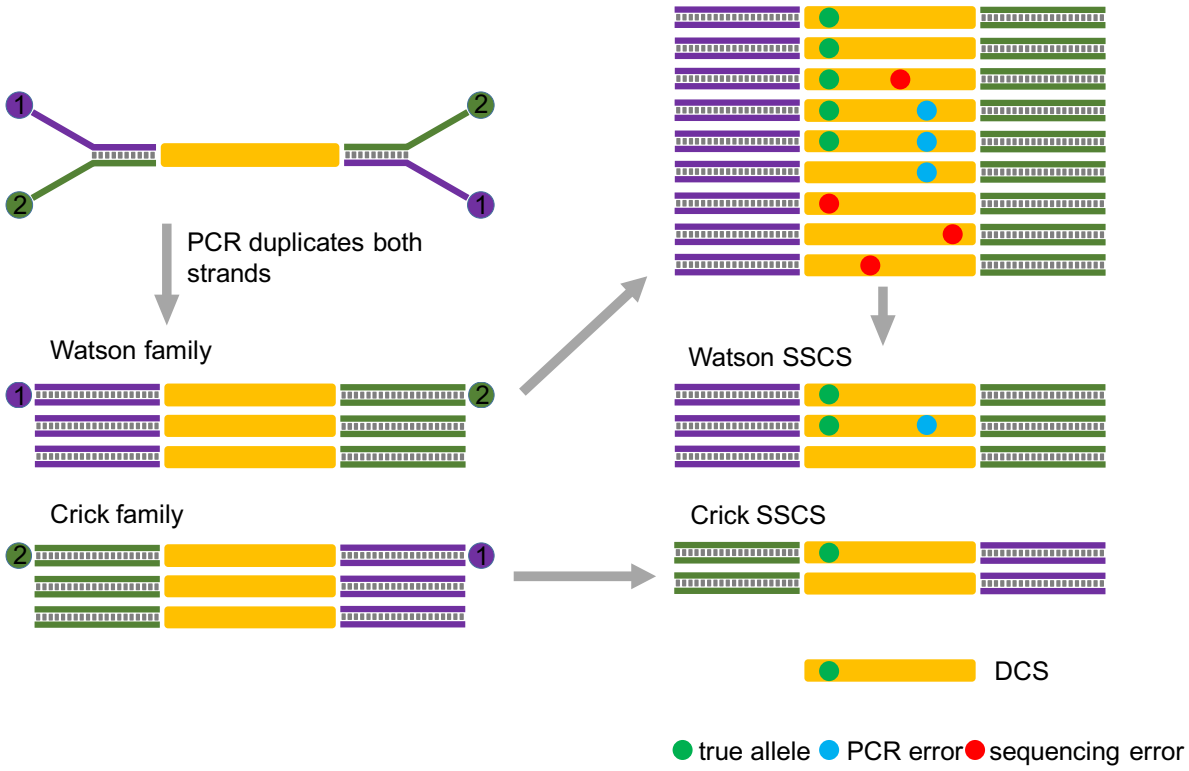
Intuitively, to distinguish LFMs (signal) from background PCR and sequencing errors (noise), we need to increase the SNR. To increase SNR, we need to either increase the frequency of mutations or inhibit sequencing errors. Single cell sequencing increases the frequency of mutations by isolating single 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 overview of LFMD pipeline is shown in Figure 3.

Figure 3. 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 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.



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

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

So the log-likelihood function can be written as

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

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(Drton 2009) 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(Chen et al. 2017) 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), and alternative approaches might be derived in the future. We also compared theoretical P-values with empirical P-values from Monte Carlo procedures (Supplemental Materials, Figures S2) and explored the power of our model under truly and uniformly distributed sequencing errors (Supplemental Materials, Figures S3). 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. The P-value of the allele g can then be given

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

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(Weiss and Von Haeseler 1997). 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} denote the coalescent PCR error rate and P_{pcr} the united PCR error rate of the double strand consensus allele. Then we get

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

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 Ding et al(Ding et al. 2015a) and Guo et al(Guo et al. 2013). 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.

Materials

Subject recruitment and sampling

A lymphoblastoid cell line (YH cell line) established from the first Asian genome donor(Wang et al. 2008) 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 S8).

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.

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 100bp and let the standard deviation (SD) of insert size equal 20bp. Let N represent the number of position families across one point. Then, N = (2 * 100) * (20 * 6) = 24000 if only considering $\pm$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 length of the tag, the fixed sequence 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 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 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 almost the same specificity with lower cost. It is a general method which can be used in several cutting-edge areas.

References

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

Belmonte FR, Martin JL, Frescura K, Damas J, Pereira F, Tarnopolsky MA, Kaufman BA. 2016. Digital PCR methods improve detection sensitivity and measurement precision of low abundance mtDNA deletions. *Sci Rep* **6**: 25186.

Blandini F, Greenamyre JT, Nappi G. 1996. The role of glutamate in the pathophysiology of Parkinson's disease. *Functional neurology* **11**: 3-15.

Chen Y, Huang J, Ning Y, Liang K-Y, Lindsay BG. 2017. A conditional composite likelihood ratio test with boundary constraints. *Biometrika* **105**: 225-232.

Dimond R. 2015. Social and ethical issues in mitochondrial donation. *British medical bulletin* **115**: 173.

Ding J, Sidore C, Butler TJ, Wing MK, Qian Y, Meirelles O, Busonero F, Tsoi LC, Maschio A, Angius A. 2015a. Assessing mitochondrial DNA variation and copy number in lymphocytes of~ 2,000 Sardinians using tailored sequencing analysis tools. *PLoS genetics* **11**: e1005306.

Ding J, Sidore C, Butler TJ, Wing MK, Qian Y, Meirelles O, Busonero F, Tsoi LC, Maschio A, Angius A et al. 2015b. Assessing Mitochondrial DNA Variation and Copy Number in Lymphocytes of ~2,000 Sardinians Using Tailored Sequencing Analysis Tools. *PLoS Genet* **11**: e1005306.

Drton M. 2009. Likelihood ratio tests and singularities. *The Annals of Statistics* **37**: 979-1012.

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

Hoang ML, Kinde I, Tomasetti C, McMahon KW, Rosenquist TA, Grollman AP, Kinzler KW, Vogelstein B, Papadopoulos N. 2016. Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc Natl Acad Sci U S A* **113**: 9846-9851.

Hoekstra JG, Hipp MJ, Montine TJ, Kennedy SR. 2016. Mitochondrial DNA mutations increase in early stage Alzheimer disease and are inconsistent with oxidative damage. *Annals of neurology* **80**: 301-306.

Jabara CB, Jones CD, Roach J, Anderson JA, 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.

Kang E, Wang X, Tippner-Hedges R, Ma H, Folmes CD, Gutierrez NM, Lee Y, Van Dyken C, Ahmed R, Li Y. 2016. Age-related accumulation of somatic mitochondrial DNA mutations in adult-derived human iPSCs. *Cell Stem Cell* **18**: 625-636.

Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, Prindle MJ, Kuong KJ, Shen JC, Risques RA et al. 2014. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc* **9**: 2586-2606.

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

Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, Press WH, Sawyer SL. 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.

Marquis J, Lefebvre G, Kourmpetis YA, Kassam M, Ronga F, De Marchi U, Wiederkehr A, Descombes P. 2017. MitoRS, a method for high throughput, sensitive, and accurate detection of mitochondrial DNA heteroplasmy. *BMC genomics* **18**: 326.

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

Peng Q, Satya RV, Lewis M, Randad P, Wang Y. 2015. Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes. *BMC genomics* **16**: 589.

Schmitt MW, Fox EJ, Prindle MJ, Reid-Bayliss KS, True LD, Radich JP, Loeb LA. 2015. Sequencing small genomic targets with high efficiency and extreme accuracy. *Nat Methods* **12**: 423-425.

Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. 2012a. Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A* **109**: 14508-14513.

Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. 2012b. Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences* **109**: 14508-14513.

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

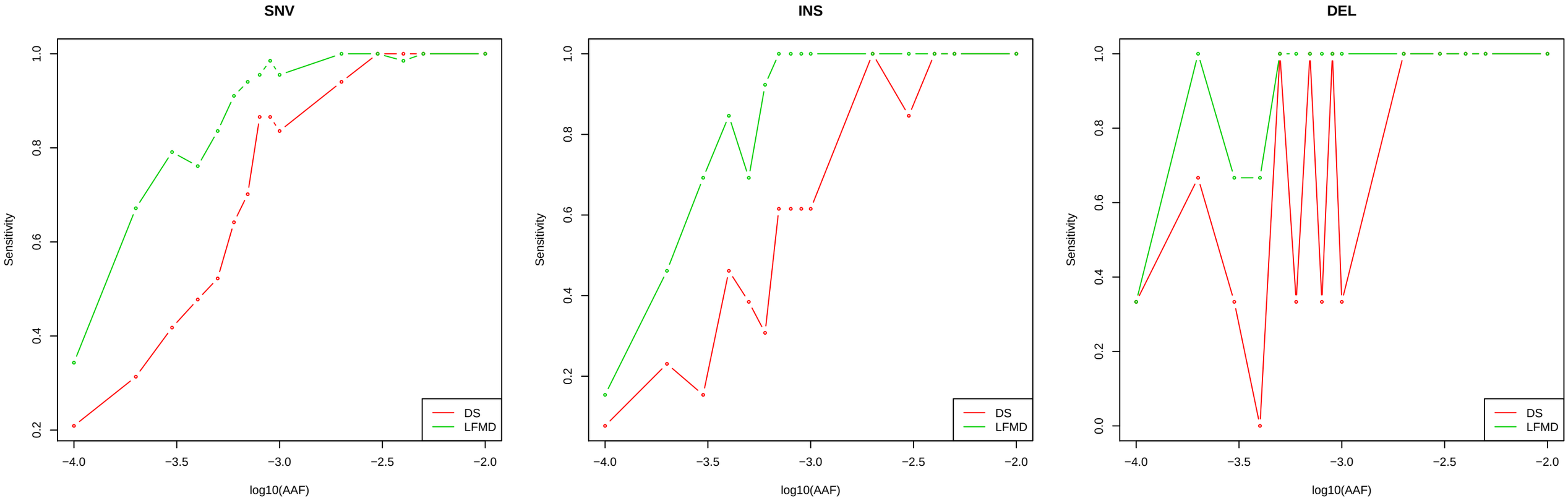
Stoler N, Arbeithuber B, Guiblet W, Makova KD, Nekrutenko A. 2016. Streamlined analysis of duplex sequencing data with Du Novo. *Genome biology* **17**: 180.

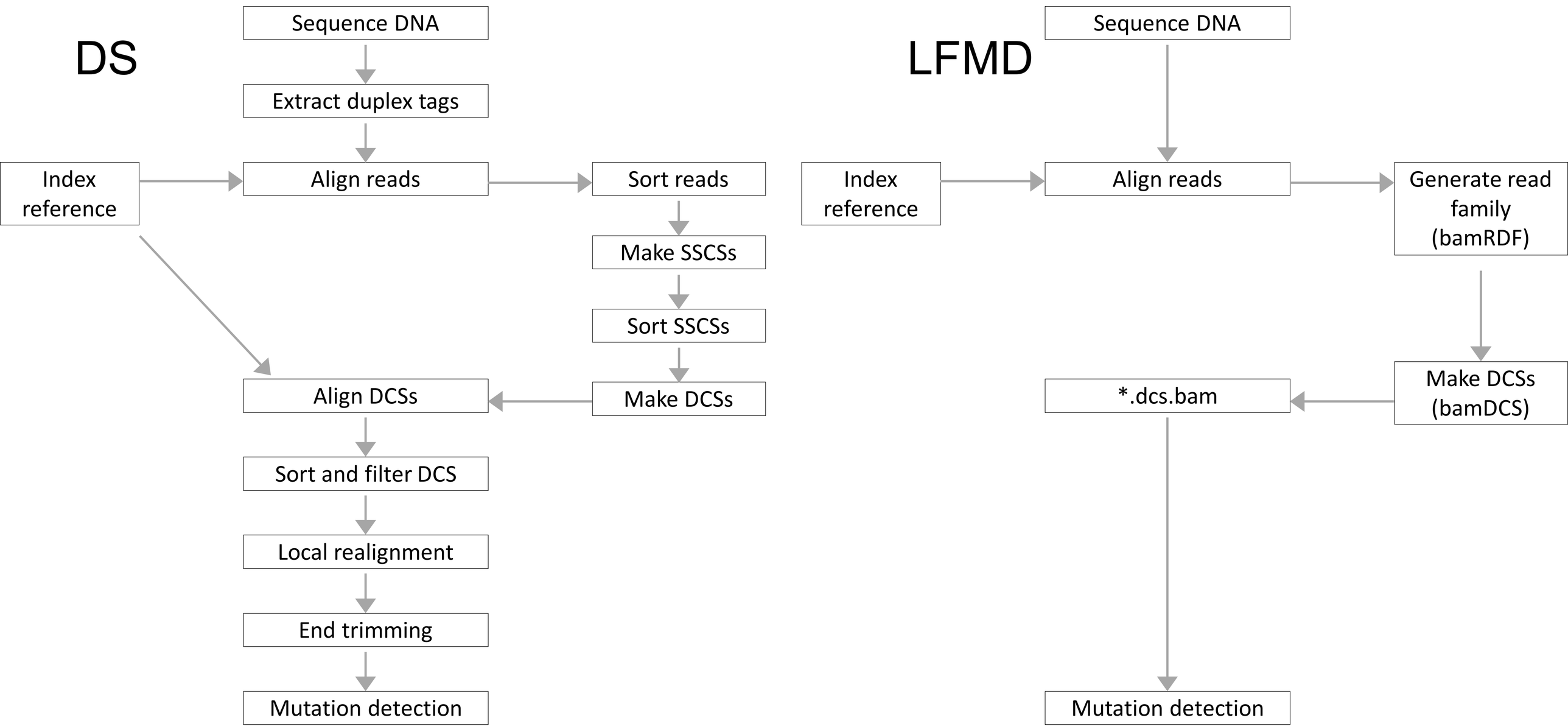
Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E. 2018. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic acids research* **47**: D941-D947.

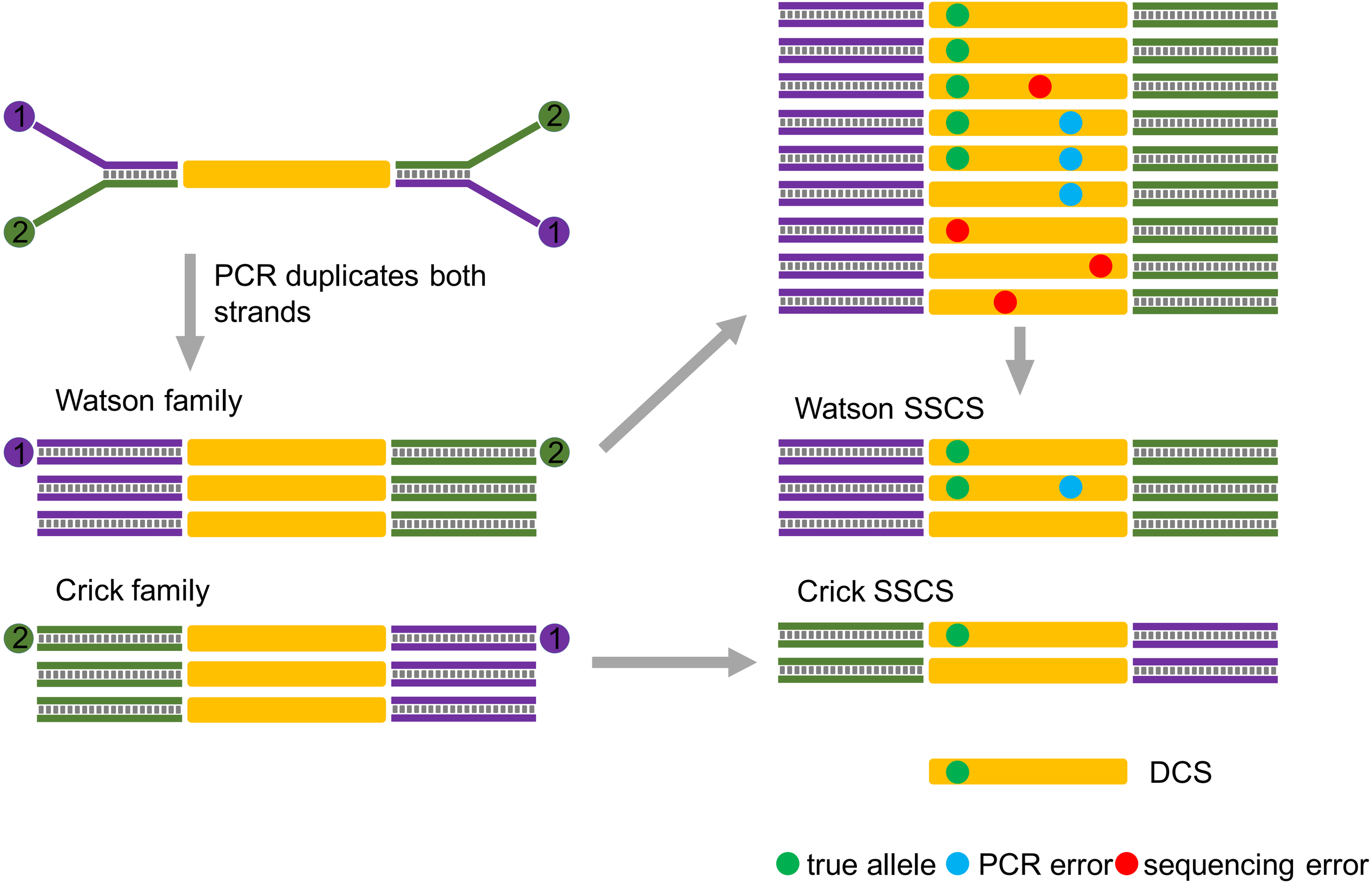
Wallace DC, Chalkia D. 2013. Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. *Cold Spring Harbor perspectives in biology* **5**: a021220.

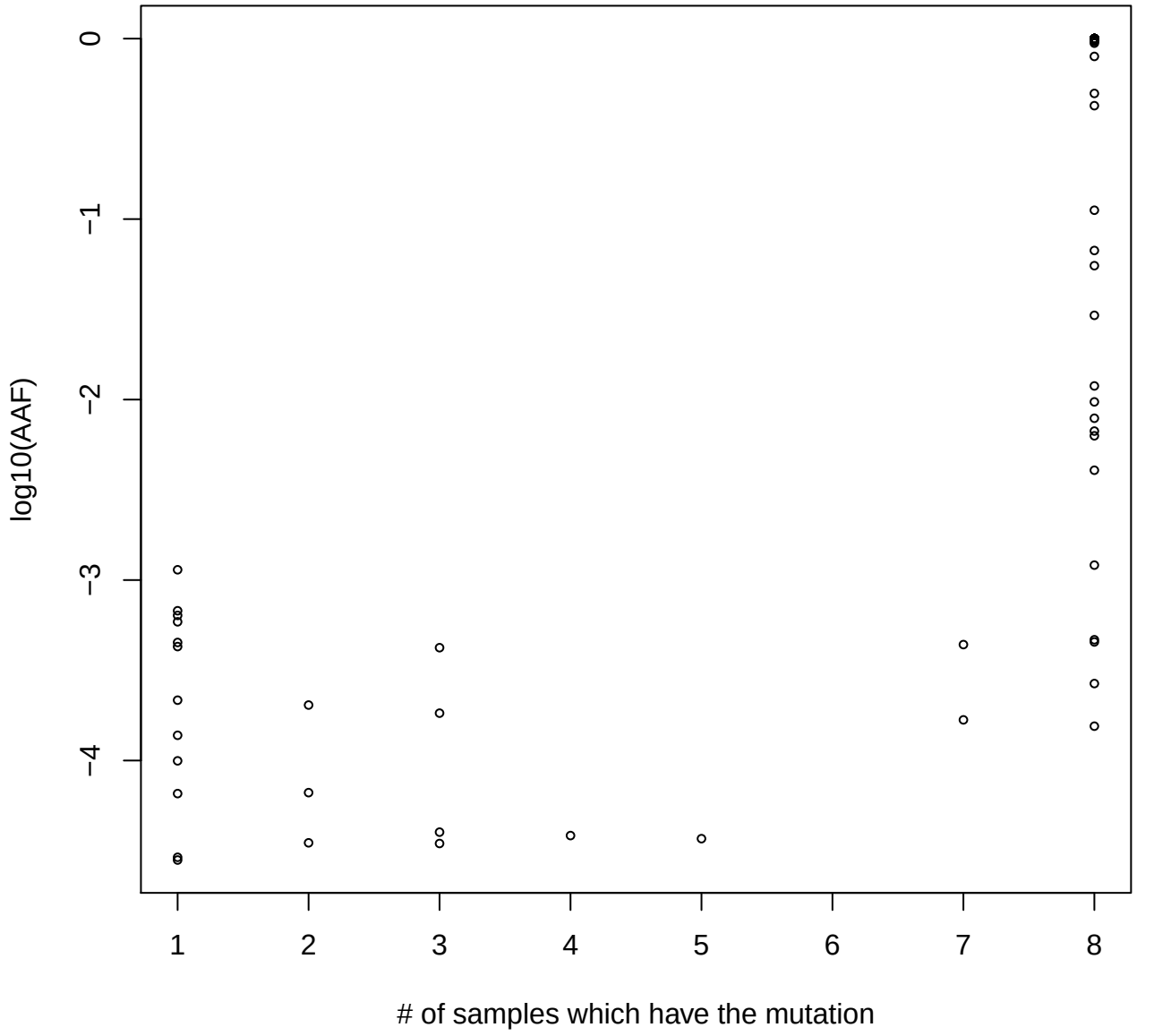
Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J et al. 2008. The diploid genome sequence of an Asian individual. *Nature* **456**: 60-65.

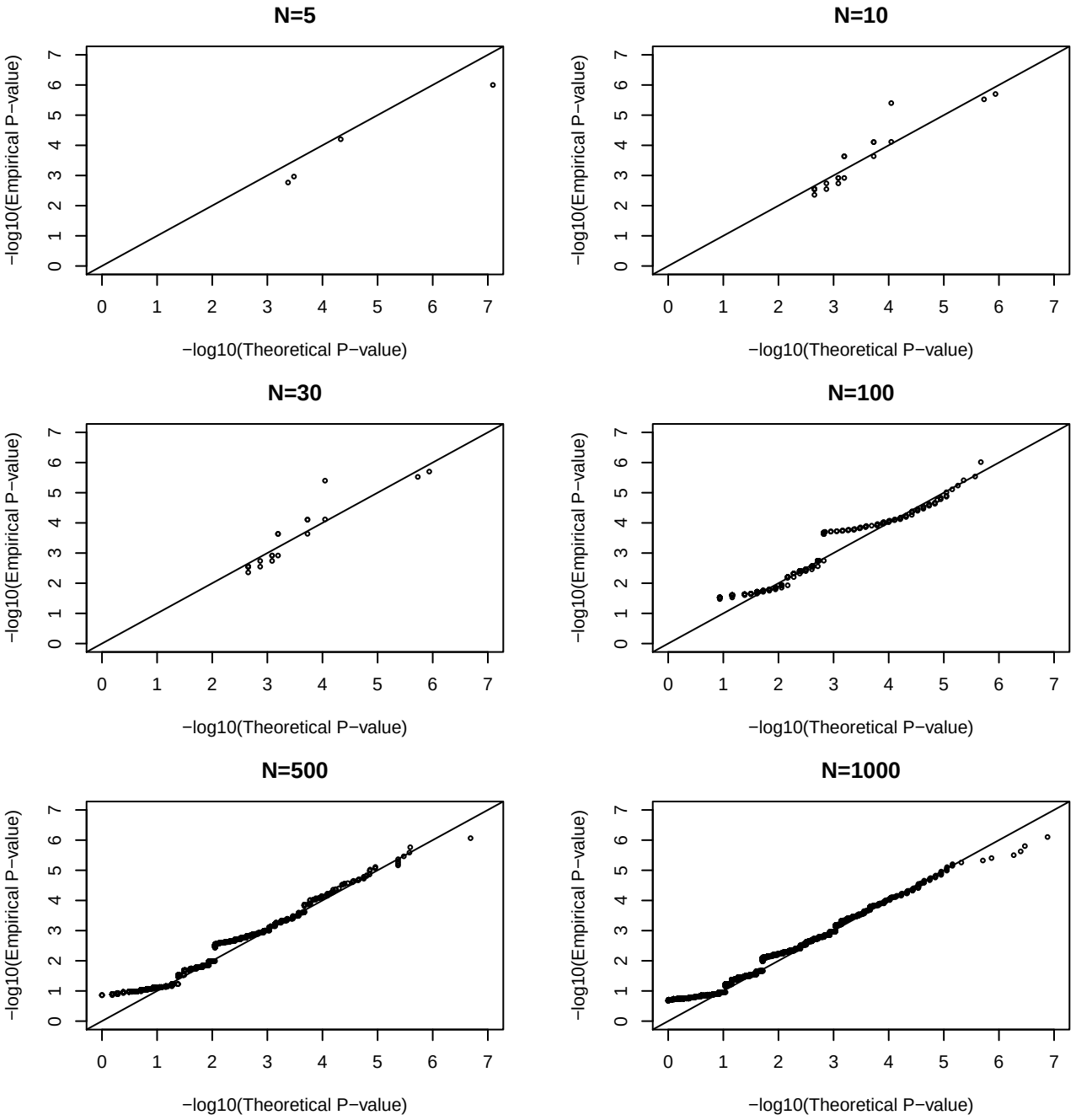
Weiss G, Von Haeseler A. 1997. A coalescent approach to the polymerase chain reaction. *Nucleic acids research* **25**: 3082-3087.

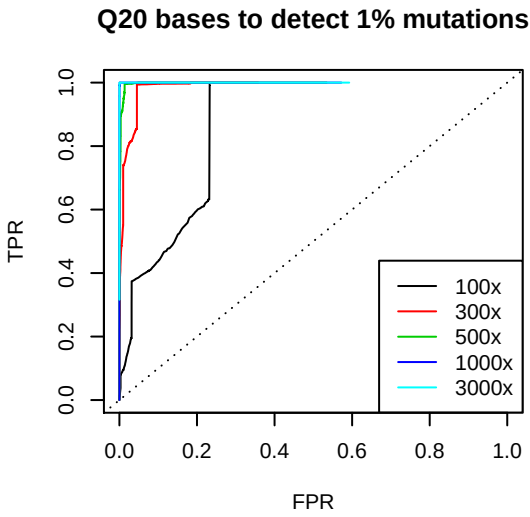

SNV
INS
DEL
Sensitivity
log10(AAF)
DS
LFMD
0.0
0.2
0.4
0.6
0.8
1.0
−4.0
−3.5
−3.0
−2.5
−2.0

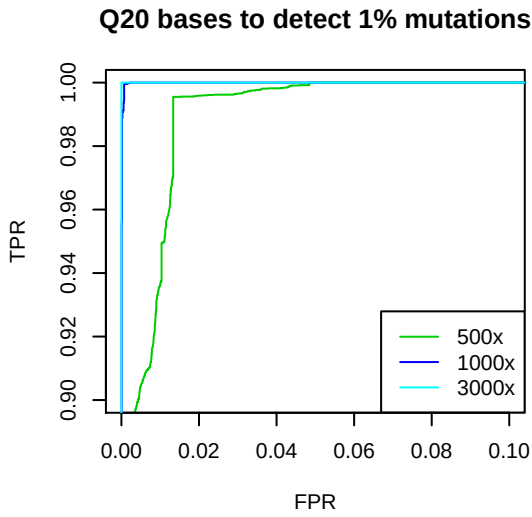

DS
Sequence DNA
Extract duplex tags
Index reference
Align reads
Sort reads
Make SSCSs
Sort SSCSs
Make DCSs
Align DCSs
Sort and filter DCS
Local realignment
End trimming
Mutation detection
LFMD
Sequence DNA
Index reference
Align reads
Generate read family (bamRDF)
Make DCSs (bamDCS)
*.dcs.bam
Mutation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

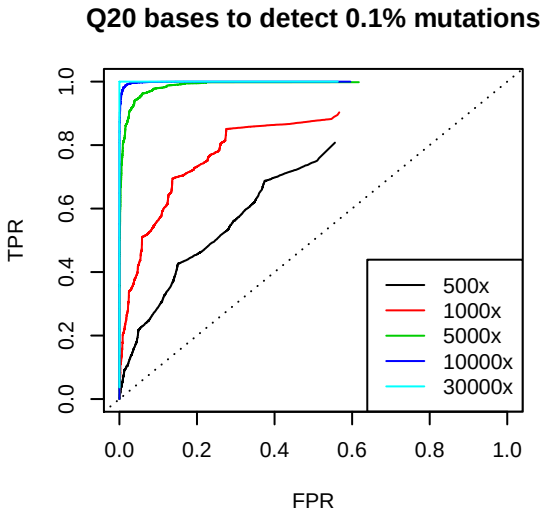

1
2
2
1
PCR duplicates both strands
Watson family
1
2
Crick family
2
1
Watson SSCS
Crick SSCS
DCS
true allele
PCR error
sequencing error

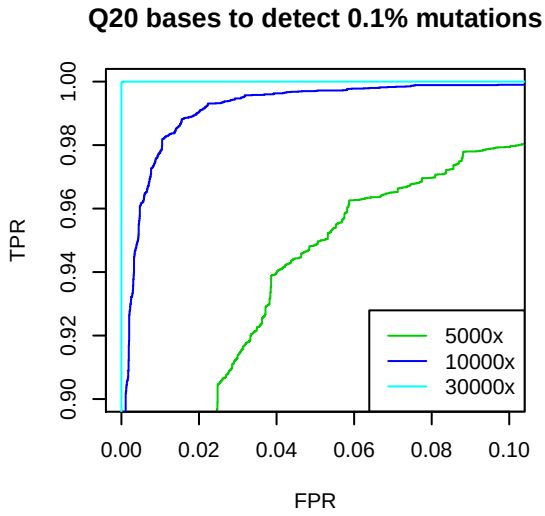

log10(AAF)
0
−1
−2
−3
−4
1
2
3
4
5
6
7
8
# of samples which have the mutation

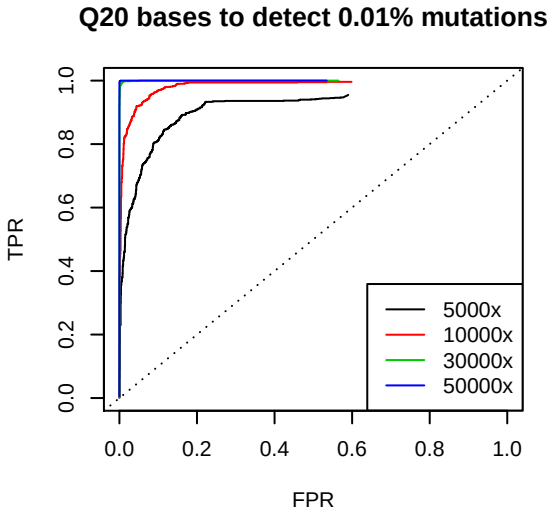

N=5
N=10
N=30
N=100
N=500
N=1000
−log10(Empirical P−value)
−log10(Theoretical P−value)
0
1
2
3
4
5
6
7

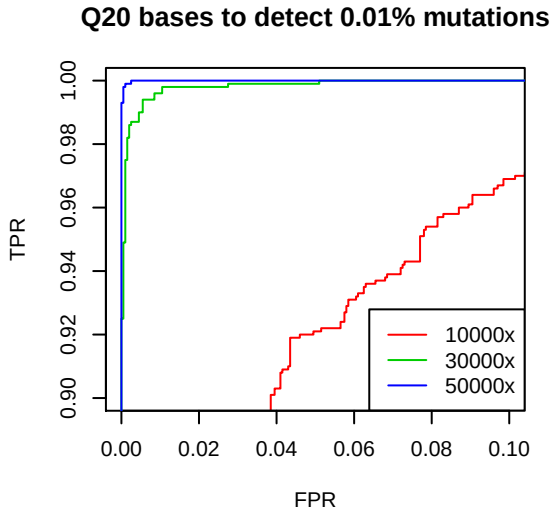

Q20 bases to detect 1% mutations
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0
100x
300x
500x
1000x
3000x

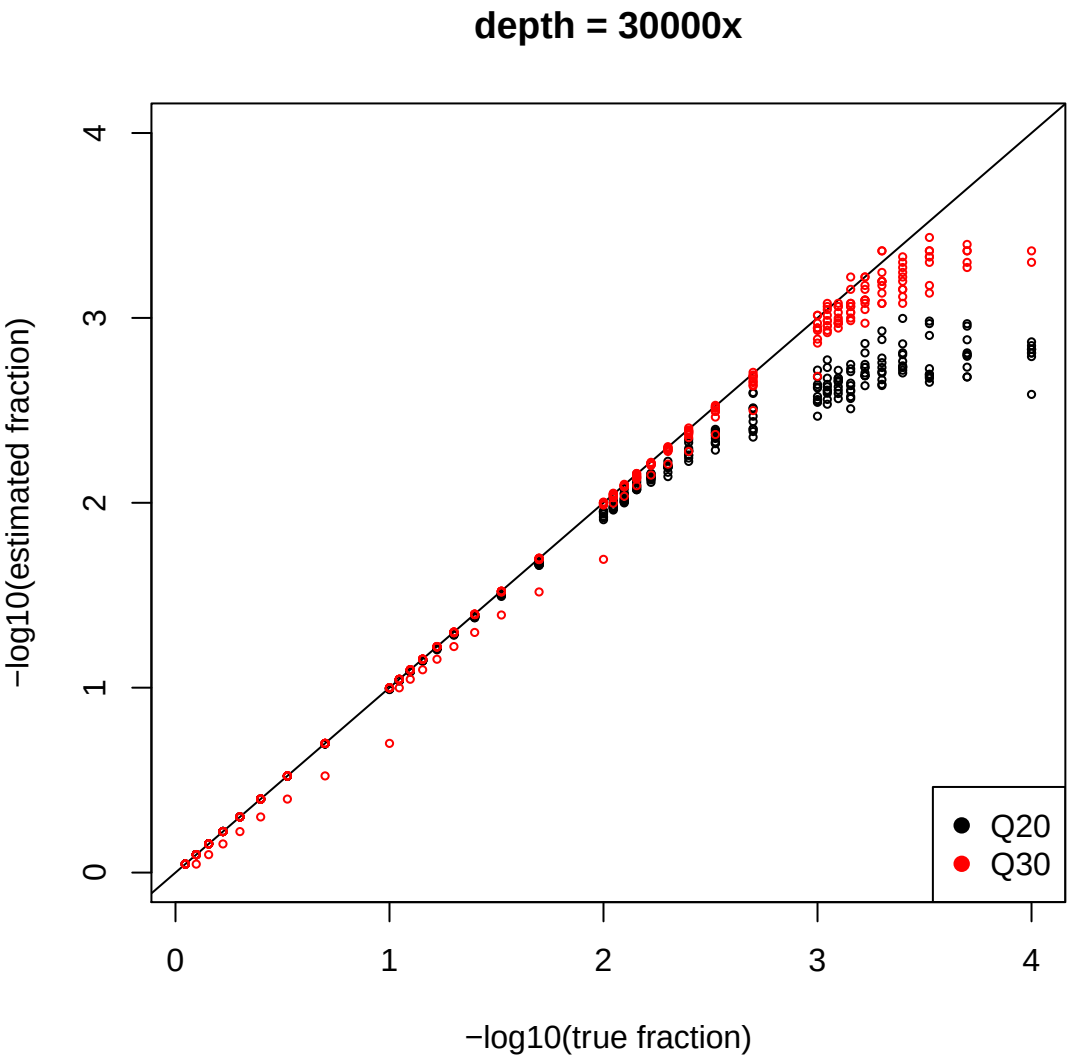

Q20 bases to detect 1% mutations
TPR
FPR
0.90
0.92
0.94
0.96
0.98
1.00
0.00
0.02
0.04
0.06
0.08
0.10
500x
1000x
3000x


Q20 bases to detect 0.1% mutations
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0
500x
1000x
5000x
10000x
30000x


Q20 bases to detect 0.1% mutations
TPR
FPR
0.90
0.92
0.94
0.96
0.98
1.00
0.00
0.02
0.04
0.06
0.08
0.10
5000x
10000x
30000x


Q20 bases to detect 0.01% mutations
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0
5000x
10000x
30000x
50000x


Q20 bases to detect 0.01% mutations
TPR
FPR
0.90
0.92
0.94
0.96
0.98
1.00
0.00
0.02
0.04
0.06
0.08
0.10
10000x
30000x
50000x


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